Eighty years of literary audio archives at the Library of Congress: Preserving collections from the physical to digital

Received (in revised form): 23rd April, 2021

Kristy Darby

Digital Collections Specialist, Library of Congress, USA

Kristy Darby is a digital collections specialist in the Digital Collections Management and Services Division of the Library of Congress. She works on various digital projects, including supporting a growing collection of open access e-books and developing workflows for processing digital content.

Digital Content Management Section, Library of Congress, 101 Independence Avenue S.E., Washington, DC 20540, USA
Tel: +1 202 707 9066; E-mail: kdar@loc.gov

Catalina Gómez

Reference Librarian and Curator, Library of Congress, USA

Catalina Gómez is a reference librarian in the Hispanic Division of the Library of Congress. She curates and manages the PALABRA Archive, develops collections on Latin American art and works on a variety of outreach and digital humanities projects.

Hispanic Reading Room, Library of Congress, 101 Independence Avenue S.E., Washington, DC 20540, USA
Tel: +1 202 707 6404; E-mail: cgom@loc.gov

Anne Holmes

Program Specialist, Library of Congress, USA

Anne Holmes is a programme specialist in the Literary Initiatives Division of the Library of Congress, where she curates and manages the Archive of Recorded Poetry and Literature and oversees a variety of digital literary programmes.

Literary Initiatives, Library of Congress, 101 Independence Avenue S.E., Washington, DC 20540, USA
Tel: +1 202 707 3498; E-mail: aholm@loc.gov

Marcus Nappier

Digital Collections Specialist, Library of Congress, USA

Marcus Nappier is a digital collections specialist in the Digital Collections Management and Services Division of the Library of Congress, where he supports staff in building a universal digital collection and ensuring enduring access to users.

Digital Content Management Section, Library of Congress, 101 Independence Avenue S.E., Washington, DC 20540, USA
Tel: +1 202 707 9087; E-mail: mnap@loc.gov

Camille Salas
Assistant Head of Digital Content Management, Library of Congress, USA

Camille Salas is an assistant head in the Digital Content Management Section of the Library of Congress. She oversees a variety of digital projects to document and sustain digital preservation processes and procedures to grow and maintain the Library's permanent digital collections.

Digital Content Management Section, Library of Congress, 101 Independence Avenue S.E., Washington, DC 20540, USA
Tel: +1 202 707 0930; E-mail: csala@loc.gov

Abstract This paper reviews the 80-year history of two Library of Congress literary audio archives — the Archive of Recorded Poetry and Literature and the PALABRA Archive — and details the challenges and opportunities that the dawn of the digital era posed for such collections. Curators and archival professionals who had been accustomed to analogue collection frameworks and workflows began to develop strategies for digitisation and digital access, paving the way for the establishment of the Library's Digital Content Management Section. This new section's Digital Collections Management Compendium outlines the institution's policy and guidance for its digital content managers. New complexities with handling digital files highlighted the need to develop innovative digital processing workflows, as well as the importance of documenting these workflows and techniques for future processing efforts. Continuous documentation and efforts to process digital files have led to increased confidence in utilising scripting for batch processing as well as an improved understanding of the requirements for making this content accessible. The collaboration between literary audio archives curators and digital content managers laid the foundation for similar digital preservation practices that the institution continues to build upon for other projects, and continues to ensure the successful transition of these historic literary collections into the digital era.

KEYWORDS: audio archives, literature, digital collections, workflows, digital preservation, collaboration, analogue to digital conversion

INTRODUCTION

In the early 1940s, the Library of Congress began recording prominent poets and authors reading from their work. Two culturally significant literary audio archives emerged during this time: the Archive of Recorded Poetry and Literature (ARPL) and the Archive of Hispanic Literature on Tape (AHLOT), collections that today comprise more than 2,000 audio recordings featuring some of the world's most important literary figures of both the 20th and 21st centuries, including Jorge Luis Borges, Czesław Miłosz, Gabriel García Márquez, Rita Dove and Elizabeth Bishop. These archives exist in large part due to the generosity of scholar and philanthropist Archer Huntington, who donated funds to the Library of Congress (Figure 1) in the mid-1930s for both the establishment of a Hispanic Division and the position of Chair in Poetry.

This paper will delve into the history of both ARPL and AHLOT (the latter was recently rebranded as the PALABRA Archive), but most importantly into the transition of both of these historic collections from the analogue to the digital era. This story will be presented as a case study meant to shine a light on the important shifts and processes that enabled the Library of Congress (hereafter referred to as 'the Library') — whose mission is

Figure 1: Library of Congress Thomas Jefferson Building, Washington, DC
Source: Prints and Photographs Division, Library of Congress

'to engage, inspire and inform Congress and the American people with a universal and enduring source of knowledge and creativity'[1] — to adapt to the technological changes of the 21st century by focusing on the crucial collaboration between archive curators and digital content experts. The release of digitised audio recordings for ARPL and AHLOT represents two of the Library's first formalised projects that helped establish formative institutional workflows to document digital processing, metadata best practices and publishing procedures. These laid the foundation for similar digital preservation practices that the institution continues to build upon for other projects, and they demonstrate the vital role that digitally accessible collections play in creating and sustaining cultural value and access.

ARCHIVAL BEGINNINGS AND EVOLUTION

The origins of ARPL are tied closely to the creation of the Poetry Office (now part of the Literary Initiatives Division) and the Chair in Poetry position, to which Joseph Auslander was first appointed in 1937. This position was known publicly as Consultant in Poetry until 1985, when by act of Congress it became Poet Laureate Consultant in Poetry.

In 1939, President Franklin D. Roosevelt appointed US poet Archibald MacLeish to serve as Librarian of Congress. MacLeish believed that libraries and librarians should play an essential role in preserving national and international culture, and put a spotlight on poetry as a cultural necessity. As such, MacLeish and Chair in Poetry Joseph Auslander inaugurated a poetry series at the Library in 1941, bringing poets to give public readings and to record them reading their work.

The idea to record poets began in part as a cultural response to the Second World War — specifically, as a way to expand access to contemporary US voices in poetry by creating purchasable record albums, with the view that US literature could spread more effectively than European fascist rhetoric.[2] The first recorded literary event at the Library was a reading and lecture by Robinson Jeffers in 1941 titled 'The

Poet in a Democracy'. During this event, Jeffers asserted that 'it may be the destiny of America to carry culture and freedom across the twilight of another dark age. … Our business is to … keep alive, through everything, our ideal values, freedom, courage, mercy and tolerance'. The public readings and recordings continued that year with American poets Robert Frost, Carl Sandburg and Stephen Vincent Benét.[3]

Allen Tate's appointment as Consultant in Poetry in 1943 brought about the official establishment of the recording project. Public poetry events paused during Tate's tenure, largely due to the war, but the consultant continued to invite writers to the Library to record their work in the newly equipped Recording Laboratory. Although Tate secured only two poets to record in private laboratory sessions during his term (Katherine Garrison Chapin and himself), this duty — inviting writers to the Library to record their work in private sessions and/or to participate in recorded public literary events — would become an integral part of the consultant's role for the next several decades. Robert Penn Warren, who took up the consultant role after Tate in 1944, brought 13 poets and writers to the Library to record their work, including William Carlos Williams and Henry Miller. In 1949, during Leonie Adams' term, Librarian of Congress Luther Evans stipulated that consultants must give a public reading early in their tenure; because these readings would be recorded, this edict also meant that the Archive of Recorded Poetry and Literature would capture the work of each consultant.

To date, the Archive of Recorded Poetry and Literature comprises nearly 2,000 recordings collected between 1941 and 2007. Many of these recordings feature consultants and poets laureate like Elizabeth Bishop, Robert Hayden, Gwendolyn Brooks, Rita Dove and Mark Strand giving their opening readings and closing lectures in the Library's Coolidge Auditorium. The archive also contains many other prominent 20th-century voices — including Marianne Moore, James Baldwin, Sandra Cisneros, Audre Lorde, Allen Ginsberg, Adrienne Rich, Kurt Vonnegut and Ralph Ellison — reading publicly and in private Recording Laboratory sessions. In the 1970s and 1980s, consultants and Library staff brought more international writers to the institution to read and record, often in collaboration with US and foreign agencies — prominent voices like Russian poet Andrei Voznesensky, Danish writer Klaus Rifbjerg and Barbadian poet Kamau Brathwaite. ARPL also contains many recordings that were donated by or acquired from other literary institutions, radio stations and private benefactors. All of these recordings were captured in analogue formats, mostly on magnetic tape reels. Curators in the Library's Literary Initiatives Division are committed to ensuring that all recordings in this historic and unparalleled collection are made available and accessible to anyone in the world with an internet connection.

Shortly after the creation of ARPL, the Hispanic Division at the Library (then called the Hispanic Foundation) began a literary audio archive of its own. In 1943, the Archive of Hispanic Literature on Tape (AHLOT) was founded by Francisco Aguilera — assistant chief of the division, specialist in Hispanic literature, and poet — who was closely acquainted with some of the most important literary figures in Spain and Latin America. In contrast to ARPL, which contained audio recordings of both public literary readings and private sessions in the Recording Lab, AHLOT comprised solely private sessions recorded in the studio, and would include literary figures from the Iberian peninsula, Latin America and the Caribbean (it would later include US poets and writers of Hispanic descent).

The first recordings for AHLOT included sessions with towering figures such as Spanish poet Pedro Salinas and Nobel laureates Juan Ramón Jiménez and Gabriela Mistral. But it was not until the late 1950s and early 1960s that the archive took form. Thanks to a grant from the Rockefeller Foundation, Aguilera obtained financial assistance to conduct travel and record writers in Latin America, permitting him to record 70 writers in Argentina, Chile, Peru and Uruguay in 1958; 49 writers in Mexico, Panama and Guatemala in 1960; and 45 writers in Colombia, Ecuador and Venezuela in 1961. With time, other offsite recordings were made possible through the cooperation of US public and cultural affairs officers at posts abroad and Library of Congress overseas centres such as the Rio de Janeiro Office.[4]

To date, the Archive of Hispanic Literature on Tape comprises nearly 800 recordings representing 34 countries and in multiple languages such as Spanish, Portuguese, English, Catalan, Galician, Basque, French and Dutch, as well as indigenous languages like Quechua, Maya, Nahuatl, Aymara and Zapotec. It contains recordings featuring some of the most prominent literary voices of the Luso-Hispanic world, such as Jorge Luis Borges, Elena Poniatowska, Carlos Drummond de Andrade and Julio Cortázar, and includes recordings with nine Nobel laureates including Pablo Neruda (Figure 2), Gabriel García Márquez, Octavio Paz and Mario Vargas Llosa. Today, the archive's curators in the Library's Hispanic Division continue recording literary figures and growing this important and unique archive, which has become a cultural treasure for Luso-Hispanic nations and the world. In 2020, the collection was rebranded as the PALABRA Archive (*palabra* being Spanish for 'word'). With the new brand, the Library marked this archive's transition from analogue to digital and celebrated a new era for the collection.

TRANSITIONING THE COLLECTIONS FROM ANALOGUE TO DIGITAL

The transition of both ARPL and PALABRA from analogue to digital has been a complex and gradual process, due to factors such as the size of the institution and the radically different infrastructures that are required to preserve and serve each format. For the past decade, collection curators have been working through technological

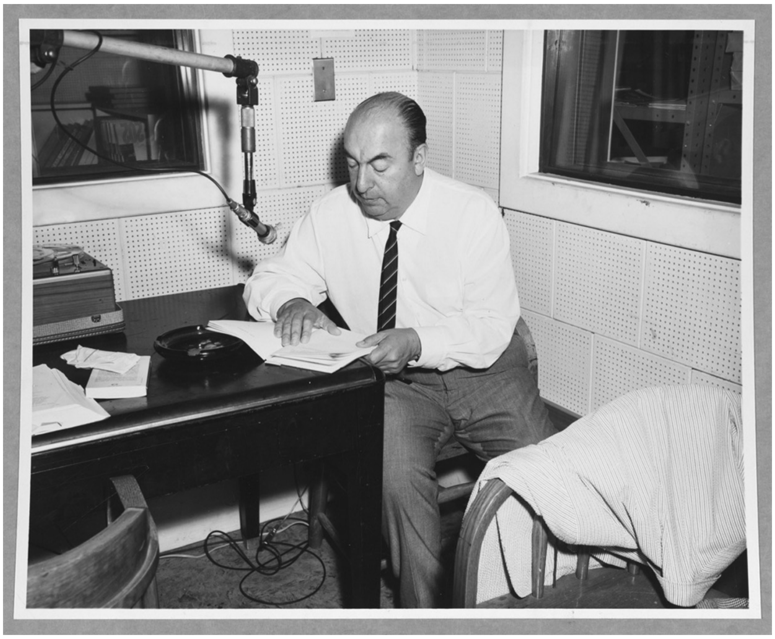


Figure 2: Pablo Neruda recording for the Library of Congress in its recording studio

Source: Prints and Photographs Division, Library of Congress transitions and reconfiguring workflows and interdepartmental relationships to support the continuation and preservation of the archives.

This transition began in the mid-2000s. Recording for ARPL halted, in part because the Library began to capture its public literary events on video, making them available to stream on the institution's website. The scope of the poet laureate position and Poetry Office had also changed in the past few decades: laureates were no longer expected to host the Library's literary event series, and the office was broadening its programming and outreach. As mentioned above, however, the PALABRA Archive did continue to record poets and writers in the Recording Lab and abroad.

Another important shift occurred in 2007, when the Library's Motion Picture, Broadcasting and Recorded Sound Division (MBRS) — the custodial division of the Library's film and audio collection materials — moved from the institution's main campus in Washington, DC, to the Packard Campus for Audio Visual Preservation in Culpeper, Virginia. A crucial entity in the management of these collections, this division not only stores and preserves the analogue objects for the archives, but it is responsible for digitising them and ensuring that the digitised content is inventoried and organised. When MBRS moved to Culpeper, this altered many of the ways in which material from both ARPL and PALABRA was processed, as well as the way in which recordings were captured. The move in 2007 marked the end of analogue and the beginning of born-digital recording in the Library's Recording Lab.

It was not until 2012–13 that the curatorial divisions of both archives began to prepare a framework for a transition to digital formats and to find new processes to adapt to the recent changes that had taken place. After meetings between the Hispanic Division, Poetry Office and MBRS, a plan was established to begin digitising the tapes from both collections, and curators began working with the Library's web developers to make recordings from both ARPL and PALABRA available for online streaming. MBRS began digitising the collections, and in 2015 the Library launched online collections for ARPL[5] and PALABRA.[6] Both sites launched with 50 streaming recordings, featuring some of the most prominent poets and writers in the archives. Since then, content from both collections has been released on a yearly basis: ARPL launches 50 newly digitised recordings every April during US National Poetry Month, and PALABRA does likewise during National Hispanic Heritage Month, which is celebrated from 15th September to 15th October. Today, five years into the initial release, more than 300 recordings from each collection are available for online streaming.

The above steps marked the important beginning of the digital era for these literary audio archives, but more work still needed to be done. The amount of digitised material from both ARPL and PALABRA continued to grow, as well as born-digital content from the latter, but a solid programme to store, preserve and process these digital files was still lacking at the Library. In 2018, the Library established the Digital Content Management (DCM) Section, which began a dedicated effort to manage and preserve these digital archives properly. While the institution remains fully committed to preserving its analogue collections, with all analogue tapes from the audio archives stored in some of the most advanced and protected environments suited for these formats, the consolidation of DCM marked a new era for the Library's literary audio archives as well as for the management of the institution's digital collections as a whole.

DIGITAL CONTENT MANAGEMENT AT THE LIBRARY

The DCM section facilitates born-digital acquisitions, maintains custodial responsibility for digital general collections materials (and digital collections not otherwise under curatorial control), conducts research and testing on digital content management methods and practices, and develops policy and procedures for the management of digital content for the Library.

DCM was created as the result of the approval of a New and Expanded Program Request in the 2017 Congressional budget.[7] DCM fills an essential role that was previously not fully formalised: managing and caring for born-digital content at the Library of Congress.[8]

DCM staff crafted a mission statement and vision statement, both of which reflect the mission of the Library as stated earlier, as well as the Library's vision, which is that all Americans are connected to the Library of Congress. DCM's mission is to 'enable and support staff across the Library in building a universal digital collection and ensuring enduring access for all users', and under its vision, DCM staff work to maintain the section as a trusted and essential partner, supporting all units building and enabling use of the universal collection. The Library's Digital Strategy[9] also guides DCM's goals and everyday work in multiple ways, including supporting the growth and maximising the use of content; reaching users in broader and more inclusive ways; and ensuring enduring access to content and cultivating a culture of innovation and creativity.

DCM aligning its vision and mission with that of the Library as a whole drives the section's approach to working across multiple divisions and directorates. DCM works in partnership with colleagues across the Library, supporting their work and enabling a shared goal of engagement and connection with collections and users. At a high level, DCM brings its processes, workflows and methods to the table, and works with curatorial and content experts to care for and manage these collections and make them available to the Library's user community.

BUILDING A CULTURE OF COLLABORATION AND CONTINUOUS IMPROVEMENT

As a new section, DCM also needed to establish a daily operations framework so that the new team could execute its primary responsibility of digital content management. Before digital collections staff joined the section in 2018, DCM's section head developed a work plan partly informed by meetings and communications with colleagues throughout the Library. There were assignments and projects to be assumed from staff who had been conducting work that would better fit within the responsibilities of the new section. New projects would entail experimentation and the creation and documentation of new workflows. In other words, DCM had a considerable backlog of work to address.

In collaboration with the senior digital collection specialists, the section head prioritised projects, including the annual releases for ARPL and PALABRA, while planning how the staff would organise itself in a way that made it clear who had what assignments and timelines for projects. DCM identified and joined in a series of ongoing meetings that focused on content releases for the Library's website. Library colleagues also used tools like Confluence for documentation and JIRA for tracking developer work. Section leaders recognised the need to adapt to this existing structure of communication and workflow tools. Equally important to understanding the Library's existing practices and procedures was the development of the section's own culture and operational framework.

DCM's new section head set forth to build a culture informed by a set of agreed values that would guide the section staff in how to approach not only assignments and projects but also how to work with colleagues across the Library.[10] The staff agreed upon nine values and descriptors, summarised below:

- *Care* — We are maintainers and supporters of the collections, collection users and each other.
- *Collaboration* — We support the work of others, and our work is supported by others.
- *Craft* — We apply the established professional principles and values of librarianship in our craft and resolve any friction or tension between them with discerning judgment anchored in reflective practice.
- *Ingenuity* — We continuously and creatively improve every aspect of our work and are always curious about new ways to do things.
- *Learning* — We are committed to continual learning and creating opportunities to learn together.
- *Safety* — We are committed to establishing and maintaining a safe space to do our work.
- *Service* — We serve staff across the Library who themselves serve the Library's full range of audiences.
- *Sustainability* — We ensure stable and enduring content, systems, practices, collaborations and selves by maintaining balance.
- *Trust* — We strive to be trusted as experts and leaders on digital content management in the Library, across the country and around the world.

The values are physically posted in DCM's office space and complement the section's mission, vision and the framework under which daily operations are carried out. With a set of values in place, the staff took operational inspiration from Agile Software Development approaches, including the framework known as Scrum, as a means of organising the team's collaborative and iterative work.

Scrum is defined as 'a lightweight framework that helps people, teams and organisations generate value through adaptive solutions for complex problems'.[11] Scrum concepts and adoption were not new at the Library, as colleagues in the Office of the Chief Information Officer (OCIO) were using Scrum practices. DCM's section head emphasised early on that the section's initial work would be the start of iterative endeavours to establish not only workflows but also collegial working relationships, which is in alignment with the Scrum framework that 'employs an iterative, incremental approach to optimise predictability and to control risk'.[12]

As there was no initial opportunity to train all DCM staff on Scrum, section leaders slowly integrated a few Scrum concepts to organise the section's work as a starting point. As one early effort, the section head's list of projects was documented using a JIRA board so that staff could easily see the backlog of work to be done. The section experimented with reviewing the backlog items as a group, but this initial approach soon failed because there were too many items and too many people in one room to have a focused discussion in a set amount of time. A sub-section dedicated to digital content care, ingest and access met instead to review the backlog tickets and prioritise work. This initial iteration made it easier to focus on next steps for projects that required processing and long-term preservation such as the release of ARPL and PALABRA recordings.

Scrum concepts — known as events and explained in more detail in the Scrum Guide (such as sprints, stand-ups and retrospectives) — became part of the section's operations.[13] The digital content care, ingest and access sub-section adapted these events to support a roster of multiple ongoing projects with various deadlines. For example, the sub-section works in two-week sprints and has weekly meetings to discuss goals and project status. Since 2019, most of the DCM staff have completed Scrum Master or Product Owner training. Certified Scrum Masters regularly lead retrospectives with DCM and other Library staff at the completion of major digital content projects. The retrospectives engender transparency and honesty about the work completed and challenges encountered that often lead to conversations about process improvements. For a new section, adapting elements of Scrum has resulted in strong working relationships across Library divisions and a cohesive and organised sub-section that regularly inspects its own work and measures progress over time, especially when reviewing the work to support releases of digital Library collections like ARPL and PALABRA.

CRAFTING THE DIGITAL COLLECTIONS MANAGEMENT COMPENDIUM

As a trusted partner in the Library, one of DCM's most important accomplishments to date is the Digital Collections Management Compendium (DCMC), a publicly accessible resource of high-level guidance that forms the Library's digital content management practices. The DCMC represents one of DCM's five main areas of responsibility, which is to 'codify, communicate and support implementation of policy and procedures for the management of digital content for all custodial units'.[14] Work on the DCMC began in 2017 as an effort to identify, collocate, document or formulate policies, guidelines, best practices and specifications that govern digital collection management at the Library of Congress.

The Digital Collections Management Compendium, which is available on the Library's website, serves as:

> 'an overarching framework that expresses the Library's priorities and needs for planning policy and guidance that ensures long-term access to digital collection content. The statements in the DCMC function as policies that support and articulate the digital preservation needs for the Library's digital collections'.[15]

Former DCM staff member Jesse Johnston, who helped lead the development of the DCMC, commented:

> 'This resource is primarily a policy resource for staff at the Library of Congress, but we are also sharing it openly and publicly as a resource for colleagues at other institutions. As suggested in the Library's Digital Strategy, we aim to model openness in our practices, to share expertise and to "drive momentum in our [digital library] communities."'[16]

HOW THE DCMC GUIDES DCM'S WORK WITH PALABRA AND ARPL

In working with these historic literary archives to make them accessible to the public and perform digital preservation actions on the content, DCM utilised guidelines outlined in the DCMC. Most relevant was the guidance item 'Processing Digital Collection Content',[17] which defines criteria and levels of processing for digital content at the Library of Congress. This guidance element asserts that four criteria must be met for digital content to be considered fully processed: (1) inventoried in an approved inventory system; (2) organised in a meaningful structure; (3) described in an approved metadata system; and (4) accessible in an approved system that enables discovery. These criteria work in concert to ensure that all processed digital content is not at risk of loss, is described in a way that enables discovery and access internally and externally, and is available for users to engage with.

PROCESSING PALABRA AND ARPL DIGITAL COLLECTION MATERIALS FOR PUBLIC ACCESS

To process the digital audio content for long-term preservation and accessibility, the PALABRA and ARPL collection workflows are built on three foundational elements:

- *Cross-departmental collaboration*: DCM staff work with colleagues responsible for curatorial and previous custodial and processing responsibilities.

- *Development of organisational documentation*: DCM staff recognise the importance of documentation to ensure that processing and preservation actions are repeatable and intelligible, exemplified by DCM's commitment to high-quality and understandable coding per the adoption of agile software development approaches.
- *Learning and training*: DCM staff work in pairs to learn existing systems and workflows, innovating with scripting and command-line processes for audio files.

In 2018, DCM staff drafted the first PALABRA processing workflow, and since then have continued to update documentation to guide both experienced and new staff through the evolved process. The PALABRA archive workflow is largely similar to the ARPL workflow, but the two collections represent two different types of content streams: ARPL is solely comprised of materials digitised by MBRS, while the PALABRA archive includes digitised materials but is shifting predominantly to born-digital materials. These distinctions primarily dictate expected file formats; however, much of the necessary processing work is similar for both collections.

Digital files for both collections are received and inventoried by DCM in a Library-developed inventory system known as the Content Transfer Service (CTS). 'CTS 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'.[18] While CTS does not play a role in the digitisation for these collections, it serves as an endpoint to inventory the archival files and access derivatives within preservation and access storage once the files have been processed. MBRS digitises the analogue collection materials for the ARPL collection, which are then managed on its inventory system — the Packard Campus Workflow Application (PCWA). MBRS has been instrumental in the management of the analogue objects for these collections, their subsequent digitisation, and ensuring that copies of the digitised content are inventoried and organised in managed infrastructure that supports onsite access at specific terminals. These annual releases would likely be impossible without MBRS' efforts to digitise and preserve these materials.

The following provides a brief synopsis of the workflow to provide access to ARPL and PALABRA recordings from item selection to online accessibility.

Item selection and transfer of files to DCM

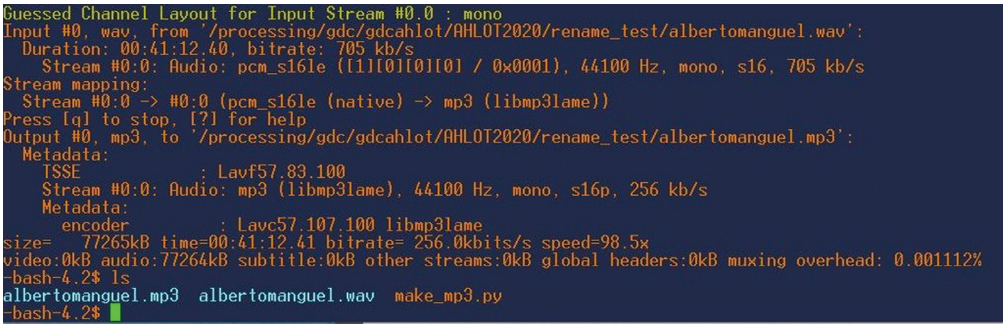
Annual releases are built on collaborations between DCM, the Hispanic Division, the Literary Initiatives Division and MBRS to inventory and manage both digitised and born-digital audio recordings. As a starting point for the eventual annual release of PALABRA or ARPL recordings, collection curators provide DCM staff with title lists of selected audio recordings at least a month in advance of the planned released date. These title lists include pertinent metadata such as recording titles, country of origin, likely file provenance, MBRS identifiers if any, and Library of Congress Control Numbers (LCCNs) that will be used for the eventual processing of the recordings for online access. At the start of this workflow, ARPL and/or PALABRA curators also request copies of any digitised recordings from MBRS they would like publicly released. The curators listen to recordings selected and make any edits for online presentation prior to transferring and/or delivering any audio files via network shared drive or external media to DCM for processing. Prior to 2020, the curators often transferred recordings via external media, but in the past year, DCM staff have greatly relied on shared network locations to receive the recordings. The PALABRA curator transfers WAV files with the expectation that DCM will generate the access MP3 derivatives, while the ARPL curator transfers MP3s that do not need additional derivative processing.

Processing files

In keeping with standardised workflows and content structures for audio content across the Library, the PALABRA data package is expected to include a master WAV file and an access MP3 file, which is made available via loc.gov. ARPL packages consist of MP3s only as the archival WAV files remain under the custodial control of MBRS in their own inventory system, PCWA. Over the last two years, DCM staff have routinised methods to rename files with the appropriate LCCN or MBRS ID number; generate MP3 derivatives through scripting if needed; and create directories for each item, which is reflective of one of the foundational elements, 'ingenuity', mentioned above as one of DCM's core values.

Based on ongoing experimentation to enable batch processing, DCM staff documented scripting workflows so that any staff member could generate MP3 files from master WAV files (Figure 3). After DCM staff have created the derivative access MP3 files, they are moved into appropriate directories that align with the identifiers discussed earlier — such as LCCNs or MBRS IDs — and that reflect one item in the ARPL or AHLOT collection. These directories correspond to distinct data packages to be inventoried in CTS, where the audio files are split along various storage file paths within the repository. DCM's established workflows utilise the BagIt standard, a 2008 Library of Congress-developed convention for structuring digital content, which generates checksums for files upon ingest, ensuring their integrity and authenticity.

Because the PALABRA collection also comprises born-digital materials, it is possible to encounter novel formats. During the 2020 PALABRA release, the need for ingenuity was critical, as DCM needed to process MOV files, a common video file format, that had not been encountered in previous PALABRA collection releases. As an additional challenge, the MOV files contained only the audio of the interview, so DCM staff needed a method to extract the audio to follow documented workflows. This presented not only an opportunity to utilise new scripting expertise and experiment with new command-line functionality, but also an opportunity to develop/document a method for evaluating and processing new file formats. Current MP3 creation scripts utilise ffmpeg, an open source software for manipulating video and audio files, and DCM project staff are familiar with the large catalogue of functions for manipulating multimedia files. After research and experimentation, DCM staff found an ffmpeg command (ffmpeg -i filename.mov -ab 160k -ac 2 -ar 44100 -vn filename.wa) to extract the audio from the MOV file as an archival WAV file. From that point, DCM



```
Guessed Channel Layout for Input Stream #0.0 : mono
Input #0, wav, from '/processing/gdc/gdcahlot/AHLOT2020/rename_test/albertomanguel.wav':
  Duration: 00:41:12.40, bitrate: 705 kb/s
    Stream #0:0: Audio: pcm_s16le ([1][0][0][0] / 0x0001), 44100 Hz, mono, s16, 705 kb/s
Stream mapping:
  Stream #0:0 -> #0:0 (pcm_s16le (native) -> mp3 (libmp3lame))
Press [q] to stop, [?] for help
Output #0, mp3, to '/processing/gdc/gdcahlot/AHLOT2020/rename_test/albertomanguel.mp3':
  Metadata:
    TSSE            : Lavf57.83.100
    Stream #0:0: Audio: mp3 (libmp3lame), 44100 Hz, mono, s16p, 256 kb/s
    Metadata:
      encoder         : Lavc57.107.100 libmp3lame
size=   77265kB time=00:41:12.41 bitrate= 256.0kbits/s speed=98.5x
video:0kB audio:77264kB subtitle:0kB other streams:0kB global headers:0kB muxing overhead: 0.001112%
-bash-4.2$ ls
albertomanguel.mp3  albertomanguel.wav  make_mp3.py
-bash-4.2$ 
```

Figure 3: Python script to create MP3 access derivative file and subsequent output

can proceed with documented scripts to generate access derivative MP3 files.

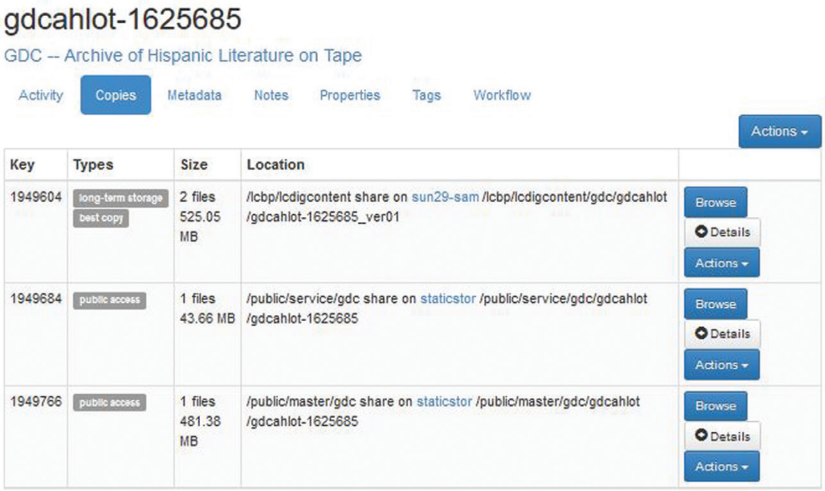
This process is not unlike the efforts described in the article, 'Never best practices: born-digital audiovisual preservation',[19] in which Erica Titkemeyer from the University of North Carolina Libraries illustrated the challenges of extracting audio from video files. Her case study and DCM's own experimental process indicate a need to further document workflows externally in a Library shared documentation space as well as within the ingested package of content. Titkemeyer mentioned that UNC Libraries repository staff included a TXT file within the BagIt bag, to explain and document staff processes. DCM realised that this added step would benefit its own processes as well. The ffmpeg command used is slightly different from the command Titkemeyer documented in her article, highlighting not only the nuances of ffmpeg command-line operations, but also the need to analyse a variety of methods to ensure the highest-quality audio content is being preserved and made accessible.

Ingesting processed files

DCM's established workflows generate checksums for files upon ingest, ensuring their integrity and authenticity. The CTS bag structure (shown in Figure 4) includes requisite storage locations and file paths. These storage systems are accessible to CTS and managed by OCIO for long-term storage, presentation and processing of the Library's digital cultural heritage materials. The CTS workflows determine that WAV and MP3 files are copied to 'long-term' tape storage while each file is copied separately to public/master and public/service file paths on presentation storage where content served through loc.gov is stored. ARPL collection CTS bags that include only MP3 files copy these audio files to public/service file paths on presentation storage to be served online. To further improve ingest workflows, DCM staff utilise Python scripting to communicate with the CTS application programming interface (API) that automates ingest and service request processes. This step limits manual errors when inputting data, while the scripts themselves provide the ability to 'test' ingests. Testing the ingest step generates an output CSV file for DCM staff to review before formally initiating ingest.

Media ingest updates and preparation for the ETL process

The Library has standardised workflows for providing access to audio collection materials, which includes creating a link between the



gdcahlot-1625685

GDC -- Archive of Hispanic Literature on Tape

Key	Types	Size	Location
1949604	long-term storage best copy	2 files 525.05 MB	/lcbp/lcdigcontent share on sun29-sam /lcbp/lcdigcontent/gdc/gdcahlot /gdcahlot-1625685_ver01
1949684	public access	1 files 43.66 MB	/public/service/gdc share on staticstor /public/service/gdc/gdcahlot /gdcahlot-1625685
1949766	public access	1 files 481.38 MB	/public/master/gdc share on staticstor /public/master/gdc/gdcahlot /gdcahlot-1625685

Figure 4: A CTS inventory record for a PALABRA (formerly known as AHLOT) audio recording that displays its bag structure

inventoried files on presentation storage in CTS, and the media services database, a managed inventory of audiovisual files for public presentation. The data submitted for ingest into the media services database link the file path and associated metadata of the inventoried audio files to the bibliographic records in the integrated library system (ILS).

Making these digital collections publicly accessible on the Library's website requires the extract, transform and load (ETL) process:[20]

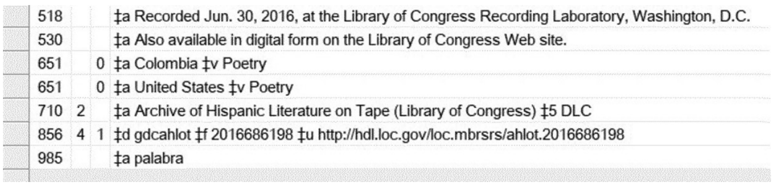
- '*Extract* is the process of reading data from a database.
- *Transform* is the process of converting the extracted data from its previous form into the form it needs to be in so that it can be placed into another database. Transformation occurs by using rules or lookup tables or by combining the data with other data.
- *Load* is the process of writing the data into the target database'.

This ETL process is reliant on specifically formatted fields in an item's MARC bibliographic record in the ILS, particularly the 856 and 985 MARC fields, an example of which is shown in Figure 5 utilising an example recording from the PALABRA archive.

The 985 MARC field, a local field, in this case 'palabra', indicates a 'part of' that will group all items with this identical field together in the loc.gov online presentation. This 985 field also serves as an indicator for the ETL extraction processes that are associated with the content. The 856 field, known as the electronic location and access field, is a repeatable field used in a bibliographic record to provide information needed to locate and access an electronic resource. DCM staff use 856 fields to facilitate ETL by designating the aggregate, the bag containing the content and the handle pointing to the loc.gov item page. As shown in Figure 5, the 856 field comprises five parts:

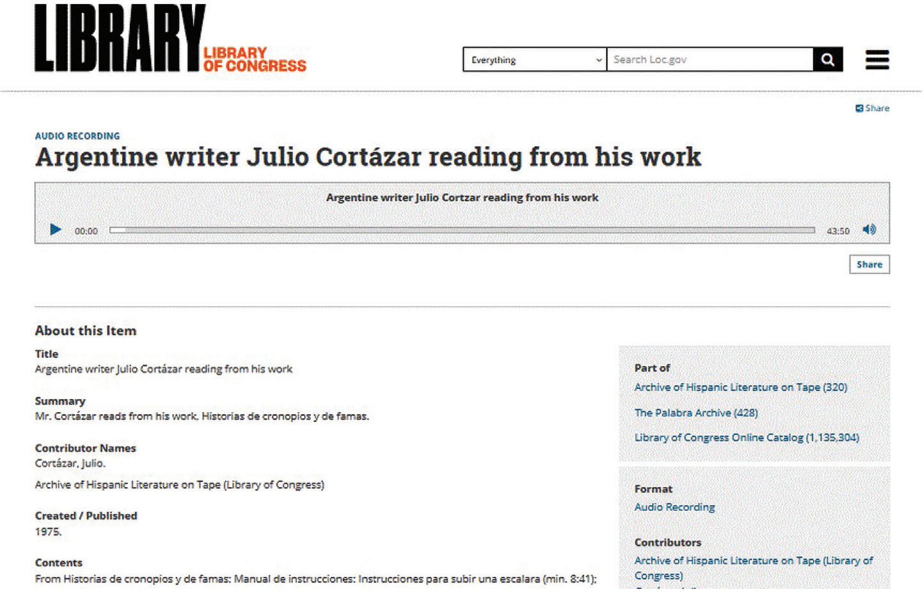
- *Indicator 4:* HTTP access;
- *Indicator 1:* resource — what is described by record as a whole;
- *‡d:* aggregate value, which corresponds to the CTS Project and collection location on storage;
- *‡f:* a digital ID, such as an LCCN, which corresponds to the CTS Bag ID and, combined with the aggregate, enables ETL to locate the content in storage; and
- *‡u:* URI — for Library of Congress purposes, a registered handle (a unique and persistent URI) pointing to resource.

The 'gdcahlot' subfield ‡d serves as a link back to the content's storage in CTS, where the content is being served from. After content is ingested and the media database has been updated, Hispanic Division staff make ILS record updates in accordance with the specified ETL procedures. DCM staff use JIRA to track project progress and create tickets to assign specific projects tasks. When DCM and Hispanic Division staff have completed all necessary steps to prepare the digital content for ETL, a ticket is created and assigned to OCIO staff to initiate the ETL code. The combined presentation of the cataloguing from the ILS record and the associated content inventoried in CTS can then be reviewed by DCM and curatorial staff in a test environment that mimics the functionality of the live loc.gov site. When the digital content display has passed this quality assurance stage, OCIO can push the recordings to the live site (Figure 6).



518			‡a Recorded Jun. 30, 2016, at the Library of Congress Recording Laboratory, Washington, D.C.
530			‡a Also available in digital form on the Library of Congress Web site.
651		0	‡a Colombia ‡v Poetry
651		0	‡a United States ‡v Poetry
710	2		‡a Archive of Hispanic Literature on Tape (Library of Congress) ‡5 DLC
856	4	1	‡d gdcahlot ‡f 2016686198 ‡u http://hdl.loc.gov/loc.mbrsrs/ahlot.2016686198
985			‡a palabra

Figure 5: Voyager module MARC record view for a PALABRA recording



AUDIO RECORDING

Argentine writer Julio Cortázar reading from his work

Argentine writer Julio Cortzar reading from his work

00:00 43:50

About this Item

Title
Argentine writer Julio Cortázar reading from his work

Summary
Mr. Cortázar reads from his work, Historias de cronopios y de famas.

Contributor Names
Cortázar, Julio.
Archive of Hispanic Literature on Tape (Library of Congress)

Created / Published
1975.

Contents
From Historias de cronopios y de famas: Manual de instrucciones: Instrucciones para subir una escalara (min. 8:41);

Part of
Archive of Hispanic Literature on Tape (320)
The Palabra Archive (428)
Library of Congress Online Catalog (1,135,304)

Format
Audio Recording

Contributors
Archive of Hispanic Literature on Tape (Library of Congress)

Figure 6: Library of Congress web presentation of a PALABRA audio recording

Many of the ETL processes are specific to a type of content, a reflection of the incredible ingenuity of the Library's OCIO development team, who are instrumental in getting content online. ETL processes are defined by explicit aggregates that have associated CTS workflows with automated rules for how particular types of content are processed and organised as well as how the content is identified in its file-naming schemes. In the case of PALABRA audio content, WAV files are copied to public/master storage, while the MP3 files are copied to public/service presentation storage with both files named with the appropriate LCCN or MBRS ID. ARPL collection releases comprise only MP3 files, which are copied to public/service presentation. The PALABRA and ARPL content thus satisfies all the necessary requirements laid out by DCMC guidance for fully processed content in that it is secure, organised, discoverable and usable. This cross-departmental collaboration between OCIO and DCM continues to be a fruitful partnership to advance the Library's mission and improve access to its collections.

LESSONS LEARNED AND BEST PRACTICES

Like many digital projects, working on the ARPL and PALABRA collections provides DCM staff with a multitude of opportunities to evaluate current practices and standardise or improve its workflows. DCM staff are fortunate to be well acquainted with standards such as the Library's own BagIt standard, which helps to ensure long-term preservation of its digital content. As referenced earlier, Titkemeyer's work with audio and video content highlights the continuous learning opportunities when working with these types of materials and developing solutions that are suitable for a given workflow. This embodies one of the foundational elements of DCM's work to learn and innovate. These innovations and new practices will be instrumental for future processing efforts and will lead DCM to consider additional processes to inventory MOV (and other audio formats) files as part of the CTS bag. A best practice that continues to prove vital is the cross-departmental approach that is critical to providing access to these materials. Furthermore, DCM staff continue to build their knowledge base of scripting and quality assurance review to improve existing documentation so that multiple staff members can support future releases.

After the latest PALABRA release, DCM staff held an informal retrospective to review the work on last year's release. In addition to discussing updates needed for documentation, staff also mentioned a desire to understand better which recordings and file formats have not been inventoried via CTS. Additionally, during the 2020 release, a new colleague shadowed the workflow and will most likely take the lead on the 2021 release. DCM will continue to document its workflows and best practices to streamline future processing efforts and provide a guide to troubleshoot future problems. Many of the lessons learned from the recent processing work were immediately incorporated into current project documentation. While the documentation is specific to PALABRA processing, additional DCM workflows, such as improved CTS functionality and generalised scripting processes, are also incorporated into this project documentation, indicative of the interconnectedness of DCM's work, which was essential to cross-training a new colleague on PALABRA processing workflows.

CONCLUSION

The growing relationship between long-established Library of Congress divisions like the Hispanic Division and the Poetry Office (now part of Literary Initiatives) and new sections such as DCM represent emerging digital partnerships that will ensure long-term preservation of born-digital and digitised audio for years to come. These valuable collaborations continue to evolve, especially as the repositories material from these archives continues to grow (the PALABRA Archive is entering a new era where many of its born-digital recordings will be remotely captured with the assistance of collaborative partners around the USA, Spain and Latin America). While it is likely that new systems and technologies might emerge that will hopefully enhance and improve the current workflows and processes for preserving and making ARPL and PALABRA recordings accessible, DCM has developed the culture, relationships and methodologies that will enable it to respond and adapt over time to ensure that 80 more years of culturally valuable recordings will be accessible to public audiences.

REFERENCES

1. Library of Congress (2021) 'Legal', available at: https://www.loc.gov/legal (accessed 22nd January, 2021).
2. Rawson, E. (2009) 'The Library of Congress Archive of Recorded Poetry and Literature, 1941–1961, and, Famous Birds, a collection of poems', PhD dissertation, University of Southern California, available at: http://digitallibrary.usc.edu/cdm/ref/collection/p15799coll127/id/568825 (accessed 1st February, 2021).
3. McGuire, W. (1988) 'Poetry's Catbird Seat', Library of Congress, Washington, DC.
4. Aguilera, F. (1974) 'The Archive of Hispanic Literature on Tape: A Descriptive Guide', available at: https://babel.hathitrust.org/cgi/pt?id=uc1.31210008 (accessed 2nd February, 2021).
5. Library of Congress (2021) 'Archive of Recorded Poetry and Literature', available at: https://www.loc.gov/collections/archive-of-recorded-poetry-and-literature (accessed 2nd February, 2021).
6. Library of Congress (2021) 'The PALABRA Archive', available at: https://www.loc.gov/collections/the-palabra-archive (accessed 2nd February, 2021).
7. Library of Congress (2017) 'Fiscal 2017 Budget Justification', available at: https://www.loc.gov/static/portals/about/reports-and-budgets/documents/budgets/fy2017.pdf (accessed 2nd February, 2021).
8. Cooper, M., Derochers, A., Owens, T., Salas, C. and Johnston, J. (2019) 'Extensive extensions: exploring file extensions in Library of Congress collections', in Proceedings of iPres, 16th–20th September, available at: https://ipres2019.org/static/pdf/iPres2019_paper_39.pdf (accessed 2nd February, 2021).
9. Library of Congress (2018) 'Digital Strategy for the Library of Congress', available at: https://www.loc.gov/digital-strategy/ (accessed 19th January, 2021).
10. DesRochers, A. (2019) 'Defining shared values in our growing digital content management section', available at: https://blogs.loc.gov/thesignal/2019/02/defining-shared-values-in-our-growing-digital-content-management-section/ (accessed 11th January, 2021).

11. Schwalbe, K. and Sutherland, J. (2020) 'The 2020 Scrum Guide: Purpose of the Scrum', available at: https://www.scrumguides.org/scrum-guide.html#purpose-of-the-scrum-guide (accessed 26th January, 2021).
12. Schwaber, K. and Sutherland, J. (2020) 'The 2020 Scrum Guide: Scrum Theory', available at: https://scrumguides.org/scrum-guide.html#scrum-theory (accessed 26th January, 2021).
13. Schwaber, K. and Sutherland, J. (2020) 'The 2020 Scrum Guide: Scrum Events', available at: https://www.scrumguides.org/scrum-guide.html#scrum-events (accessed 26th January, 2021).
14. Library of Congress (2021) 'Digital Collections Management Compendium', available at: https://www.loc.gov/programs/digital-collections-management/about-this-program/ (accessed 22nd January, 2021).
15. Library of Congress (2021) 'Digital Collections Management Compendium: Frequently Asked Questions', available at: https://www.loc.gov/programs/digital-collections-management/about-this-program/frequently-asked-questions/ (accessed 19th January, 2021).
16. Johnson, J. (2019) 'Launching the Digital Collections Management Compendium', available at: http://blogs.loc.gov/thesignal/2019/10/launching-the-digital-collections-management-compendium/ (accessed 11th January, 2021).
17. Library of Congress (2021) 'Digital Collections Management Compendium: Processing Digital Collection Content', available at: https://www.loc.gov/programs/digital-collections-management/access/processing-digital-collection-content/ (accessed 22nd January, 2021).
18. Cooper *et al.*, ref. 8 above.
19. Kim, J., Fraimrow, R. and Titkemeyer, E. (2019) 'Never best practices: born-digital audiovisual preservation', *code4lib Journal*, No. 43, 14th February, available at: https://journal.code4lib.org/articles/14244 (accessed 4th January, 2021).
20. Beal, V. (n.d.) 'ETL — Extract, Transform, Load', available at: https://www.webopedia.com/definitions/etl/ (accessed 15th January, 2021).