

Archiving Indigenous data and meta-data for a digital reproduction of the print book titled *Cherokee Earth Dwellers*

Turam Purty; Information School; University of Washington, Seattle, USA

Christopher Teuton; Department of American Indian Studies; University of Washington, Seattle, USA

Abstract

This case study outlines the meta-data curation strategies and workflows employed towards a digital reproduction of an Indigenous print book, *Cherokee Earth Dwellers*. The data corpus includes multimedia assets in the form of unstructured file formats such as multilingual text, images, audio, and video documents. Five stages guide the digitization process that assist the development of an interactive e-publication from the print book. First, a programmatic approach is utilized to retrieve names of biodiversity species from the pre-print. Second, meta-data for audio, video and image files associated with the print book are weaved together to form an intermediate meta-data package (IMDP). Third, independent meta-data reviewers programmatically validate the resulting IMDP outputs. Fourth, Reproducible Python Files (RPF) are generated to reformat IMDP into Simple Archive Format (SAF) for ingestion into DSpace. Fifth, RPF is published open source to streamline data curation workflows for unstructured multimedia assets. The study concludes with a discussion on the emerging interplay between Generative AI and human curation practices, and its implications for digital publishers, and archival repositories.

Introduction

Digitizing a multilingual print publication into a useful, purposeful, and engaging e-publication requires careful workflow management in metadata curation. This case study describes the digitization process associated with the production of the print book *Cherokee Earth Dwellers* [1]. The workflow begins with a pre-print Portable Document Format (PDF) version of the book together with a collection of multimedia assets, not included in the print book, but recorded by the book author during the book writing process. Metadata from the digitized assets are synthesized from both textual and multimedia sources using generative AI to aid human curation practices. A distinctive feature of the book is its multilingual catalogue of more than 600 biological species names in Cherokee and English syllabary, alongside cultural classifications, Indigenous meanings, vernacular pronunciations, and related illustrations and images.

The digitization process outlined in this case study integrates metadata from diverse file formats and

weaves them into a Simple Archive Format (SAF) package for ingestion into the open-source archival platform DSpace [2,3]. Prior research on DSpace implementations has highlighted the limited visibility of relationships across metadata files distributed between repositories and systems [4]. Generative AI tools have also been proposed as a means to improve the usability and efficiency of metadata curation workflows [5]. In this project, the central challenge is not only the conversion of files into archival formats, but also the creation of reliable referential associations between textual entries in the book and related image, audio, video, and illustration assets.

Digitizing data and metadata from Indigenous knowledge systems presents additional challenges that require rethinking interdisciplinary boundaries, responsibilities, and connections [6,7]. Metadata derived from Indigenous knowledge systems must include appropriate acknowledgements that extend relational accountability to Indigenous Peoples and their relationships to sovereign lands and more-than-human inhabitants [8,9].

Background

In this case study, we document the archival process of multimedia files in the form of images, audio recordings, video discussions, recordings, and illustrations that contributed to the print book, *Cherokee Earth Dwellers: Stories and Teachings of the Natural World* (2023). The print book includes a catalogue of names, stories, folklore, scientific facts, teachings, and mythology associated with different types of flora and fauna. The names, meanings, and illustrations of key wildlife species and edible plants are spread across the book in different chapters and various narrative sections. A majority of the Indigenous knowledge contribution lies in documenting environmental stewardship through socio-cultural narratives around birds, trees, plants, forest dwellers, amphibians, insects, and fishes. In many instances, we also encounter references to mythical creatures, which may be traced back to extinct species.

The print book provides a rich snapshot of the life works of Cherokee National Treasures Hastings Shade and Loretta Shade, and of their close family and

friends, in documenting Indigenous ecological knowledge in Cherokee and English [10]. The metadata curation and archival process involves documenting and keeping track of the names, affiliations, and relationships of various community-based contributors to the multimedia assets curated for the digitization project. The purpose of this case study is to share applied knowledge in metadata curation practices involved in digitizing multilingual and multimodal Indigenous datasets. The case study also illuminates the time, effort, and costs involved in aggregating qualitative data across different file formats and preparing it for archiving into DSpace.

We take an instrumental view in documenting the problem of preserving digital data and its metadata in ways that are cost effective, standardized, and interoperable [11]. Metadata are enriched descriptions of digital assets that allow for information identification, selection, organization, and for search optimization. In information science, a primary key may be assigned to each unique file and its related metadata for its access and retrieval.

A referential key is assigned to multiple files and metadata items in order to preserve a unique known relation or association between two or more digital assets. In this study, referential keys preserve the relationships across different types of text and multimedia assets and their associated metadata. This referential integrity ensures that metadata derived from text sources are accurately linked to text, images, audio, video, and illustrations as they are prepared for archival ingestion.

Problem

The primary problem encountered in the digitization process is the initial lack of referential associations between different types of media assets eligible for metadata cataloging and archiving into DSpace. In the context of this project, this problem emerged during the effort to build a digital catalogue of multilingual species names and its associated multimedia files. Each species name has to be identified from the pre-print PDF and assigned a unique identifier. This identifier then has to be linked to digital assets that includes the relevant audio pronunciations in Cherokee and English, related images, scientific nomenclature, and in some cases related video recordings or hand-drawn illustrations. For example, a species name appearing as text in the PDF document could not simply be treated as an isolated textual record. In order to support a meaningful digital publication and archival package, text records from source materials

have to be connected to one or more related multimedia files and meta-data fields. If these associations, or relations, are missing, inconsistent, or ambiguous across named files, then the resulting metadata package would not preserve the relational integrity between the print text and its supporting digital assets.

A secondary problem encountered during this digitization process was challenges to the referential integrity of metadata files as they are reformatted into normalized structures for bulk ingestion into DSpace archival repository. Even when associations, or relations between files and metadata are identified during the curation process, still they have to be preserved and validated consistently across multiple stages of file transformation, including intermediate metadata packaging, and final conversion into SAF packages ready for bulk ingestion into DSpace.

Methodology

The digitization process for this print book involved processing text and multimedia files that resulted in a data corpus of over 2000 digital assets, including more than 600 text and image files supported by audio and video files. The metadata curation workflow for this multilingual and multimodal dataset involved five stages.

Stage 1: Source text search, retrieval and formatting

First, a programmatic process was implemented to identify text keywords, particularly the names of the species in Cherokee and English syllabary. The pages in the pre-print PDF version of the print book, where the names of species were listed, were identified and extracted into a separate processing file. A Python script was developed with the assistance of generative AI tools, and the script was run locally to extract the Cherokee and English text entries together with their related page numbers from the pre-print version. Each row entry was written in a Comma-Separated Values (CSV) output file. The resulting CSV file was then reformatted in a Microsoft Excel spreadsheet using macro formulas that separated multilingual text fields row entries based on delimiters visible in retrieved text data, such as bullet markers and spacing conventions.

A unique numeric identifier, referred to as the root-ID, was appended to each text entry retrieved from the print book with the same name. This root-ID was also included as a persistent reference key to connect the text entries to their corresponding images, multilingual

audio pronunciations, and page number references to illustrations and stories included in the print book.

Stage 2: IMDP curation and assembly

Second, a human curation effort was undertaken through the Intermediate Meta-data Package (IMDP) to map the extracted text keywords to related multimedia assets being considered for catalogue creation and archival ingestion. The root-ID assigned to each text entry was extended to additional digital files associated with that species. For example, the audio files for species names were recorded by Tom Belt and Christopher Teuton to preserve the native pronunciations that could not be included in the print book [12]. For the curation of IMDP, the audio pronunciation files were first split into individual clips of up to three to five seconds each using the open-source software Audacity and then renamed using the corresponding root-ID. Because the print book did not include images for all listed species names, additional Creative Commons (CC) images were sourced from open-access repositories such as Wikimedia Commons, along with associated attribution metadata and scientific nomenclature based on Binomial classification [13]. These CC image records were then grouped according to the classification ontologies described in the print book, such as birds, trees, plants, edible plants, fish, reptiles and amphibians, forest dwellers, and insects. In addition, audio and video recordings produced during the book-writing process were transcribed using AI tools, and the transcripts were manually reviewed for references to any stories or species names appearing in the print book. Where a specific reference was found, it was recorded in the metadata for that multimedia asset and linked to the corresponding root-ID and page number in the print book. In this way, one root-ID connected the English text, the Cherokee text, the English pronunciation audio, the Cherokee pronunciation audio, an image, or an illustration. In some cases, the root-IDs were also mapped to pages in the print book which contained Indigenous storytelling narratives.

Stage 3: Human Review and Validation

Third, the curated IMDP corpus was reviewed and verified for accuracy, cost efficiency, and transferability into DSpace. A file inventory was developed in spreadsheet form in close collaboration with editors at RavenSpace, University of British Columbia, who manually checked and validated the

records derived from the pre-print book. For each text entry, the inventory consolidated information such as file location, file name, media type, title, description, source, page number in the print book, attribution information, and related identifiers. This stage helped verify that the referential links established in the IMDP remained accurate across the multilingual and multimodal dataset. This human review process was especially critical for identifying any missing relationships, correcting any naming inconsistencies, and confirming that the overall descriptive metadata across the dataset preserved the intended relations between the text and its supporting digital assets.

Stage 4: RPF generation for SAF transformation

Fourth, generative AI programming tools were utilized to assist in generating Reproducible Python Files (RPF) for reformatting IMDP records into Simple Archive Format (SAF) for DSpace ingestion. These Python scripts were developed to read the spreadsheet-based IMDP working files, preserve the root-ID relationships across different media types, and transform the reviewed metadata into a format suitable for archival processing and ingestion into DSpace. A generative AI tool was utilized to generate a reproducible programming template that would run locally on the user's machine, without sharing the data with the generative AI tool. Therefore, another programmer, or a digital archivist, could take the same RPFs, make some modifications to the existing code, and run them locally on their own machines. In practice, almost all RPFs for this project normalized the rows in IMDP format into explicit and redundant metadata for ingestion into DSpace archives.

Therefore, the RPF stage served as the bridge between the human-curated IMDP environment and the machine-actionable SAF structure required for bulk archival ingest. Human oversight remained necessary at this stage to confirm that the generated scripts performed local data transformations, preserved the intended referential relationships, and did not collapse or mislabel important distinctions across files. In addition, selected multimedia assets, particularly Creative Commons images, were made more accessible through OpenAI image-recognition token calls. These token calls were pre-configured with prompts to generate descriptive alt-text for each image, thereby supporting accessibility and supplementing the metadata available for archival and digital publication purposes.

Stage 5: Open-source knowledge sharing

Fifth, the resulting RPF workflow has been prepared for publication in github, as an open-source software repository to support future adaptations. The linked open-source repository includes dummy datasets to demonstrate the functionality of each stage without exposing the copyright-restricted project materials. The purpose of this repository is not only to share code, but also to document the practical digitization and archival workflow through which multilingual text extraction, multimedia asset linking, metadata review, and SAF packaging were carried out. The open-source repository is intended to function as a reproducible template or demonstration environment for other archival and digital publishing projects working with unstructured multimedia collections. In this sense, the repository extends the methodological contribution of the paper by showing how the IMDP to RPF to SAF workflow may be adapted, tested, and reused beyond the present case study.

Outcomes

Once the SAF packages are prepared, the counts and names of species were cross-validated across different clusters of data between the IMDP version and the metadata sheet retrieved from the DSpace instance after data ingestion. Datasets were validated in clusters of 50 and 100, and any exceptions, inconsistencies in file names, or mismatches in expected counts were recorded for further review. This cross-validation technique helped confirm that the referential integrity established during the IMDP stage were preserved through the RPF and SAF transformation stages and remained interpretable even after archival ingestion.

The open-source repository associated with this workflow is published here:

<https://github.com/saahil634/imdp-package>

The repository is organized to mirror the overall metadata curation and archival workflow described in this study. At the top level, it includes folders for assets, metadata, scripts, exports, workflow documentation, and workflow-specific processing folders. The assets and metadata directories reflect the broader conceptual organization of the project, while the scripts directory includes reusable code for text retrieval, IMDP assembly, RPF generation, and SAF packaging. The docs directory includes workflow notes, methodology drafts, and codebook gap analysis, allowing future users to understand not only the scripts

themselves but also the logic through which this particular workflow has been designed.

A distinctive feature of this code repository is the workflows directory, which is structured around individual processing steps. Each workflow folder includes a src folder for the original or adapted Python script, an input folder where a user places the file or files to be processed, an output folder where processed files are written, an expected folder containing small dummy outputs for validation, and a tests folder for lightweight checks. This structure was designed to make the repository legible and reusable for future archivists and digital publishers who may wish to adapt the particular case study to inform their practices in digitizing multilingual or multimedia collections.

This open source repository is the research documentation provided to guide future educators and practitioners who may want to implement similar workflows in their projects by changing only a small number of parameters. In most SAF generation scripts, these include the local file path of the metadata CSV, the content directory containing the digital files, the output directory for the SAF package, and the naming convention used for zip outputs or bulk file chunk sizes. In the image alt-text workflow, additional parameters such as token keys, cloud bucket names, server region names, and spreadsheet column name variables may be replaced as needed. The repository documentation therefore serves not only as a software deposit, but also as a procedural guide for adapting the IMDP to RPF to SAF workflow across new archival projects and settings.

Finally, the repository includes dummy data and validation scripts so that the input and output structure of each workflow can be tested without exposing copyright-restricted or culturally sensitive project materials. In this sense, the repository functions both as a reproducible demonstration environment and as an extensible template for future archival adaptation.

Discussion

Content analysis through IMDP highlights critical gaps in attempts to automate metadata curation and underscores the continuing importance of human-mediated oversight in generative AI-assisted metadata practices. While programmatic tools were useful for structuring files, retrieving records, and generating reproducible workflows, they were not

sufficient for interpreting the full range of linguistic, cultural, and relational meanings embedded in the dataset. For example, some Cherokee names were associated with more than one species, even where the English labels appeared distinct. In other cases, some Cherokee words did not have a direct equivalent in current scientific literature and were therefore provisionally classified as “Mythical.” Such classifications should not be read as absence or error, but rather as evidence that certain names persist in collective memory and continue to hold significance within language, storytelling, and cultural identity. These examples demonstrate the limits of automated classification and the need for relational care when digitizing Indigenous knowledge materials [8,9].

A unique identifier, referred to as the root-ID, was appended to each species name as a way of maintaining referential integrity across distributed files. This strategy was especially important because some species had more than one name, and some names had more than one pronunciation or related media file. The root-ID ensured that each text record retrieved from the print book could be mapped to its corresponding multimedia files through a shared relational identifier. In this sense, the root-ID differs from the unique identifier commonly associated with a single digital asset, such as a UUID or a system-generated file identifier. Rather than identifying one file in isolation, the root-ID functions as a persistent referential key that repeats across multiple closely related files within a networked information system. It provides a persistent reference point that connects text, audio, image, illustration, transcripts, and later archival packaging stages of metadata curation.

The use of a root-ID also makes it visible the layered and relational character of the Indigenous knowledge represented in the collection. Each species could have more than one name, each name could have more than one pronunciation, and each pronunciation or textual reference could be linked to other digital files such as images, stories, transcripts, or illustrations. Together, these digitized artifacts weave an intricate relationship of Indigenous knowledge that exceeds the current scientific boundaries of any single file, or format. In this sense, the described metadata workflow is not only technical, but also an exercise of expressing relational accountability [14].

The IMDP format demonstrates how to preserve relations between names, media, and meanings that might otherwise be flattened in conventional

normalized data-processing. This relational view of knowledge resonates with Indigenous understandings of interdependence and with the broader idea of “braiding” distinct ways of knowing, where language, classification, memory, and ecological knowledge remain interconnected rather than isolated into discrete units [15].

Document analysis further shows how IMDP enabled the integration and consolidation of varying file formats for images, audio, video, and illustrations into a coherent intermediary structure. The resulting RPF workflows then demonstrated how this intermediary structure could be translated into SAF for DSpace ingestion. In practical terms, the repository produced through this study shows how programmatic structures can be generated using AI tools for wider reuse, while the actual analysis and transformation of culturally grounded data may still be carried out locally. This distinction is particularly important when working with Indigenous data, where ethical and institutional constraints may limit the circulation of source materials [10]. Where local analysis is not possible, future curation workflows may rely on secure enterprise environments or locally controlled storage infrastructures, while using synthetic or simulated data to develop and test programmatic structures before applying them to restricted, or sensitive collections.

The implications of this study also extend to education, language revitalization, and cultural naming conventions. The multilingual species catalogue in *Cherokee Earth Dwellers* can also be represented as a pedagogical and cultural resource. By linking names, pronunciations, illustrations, and related stories, the archival workflows help preserve the conditions through which Indigenous language and knowledge is accessed, learned, heard, interpreted, and connected to place and time. The digitization and archival process implemented for this study supports longer-term care and visibility of Indigenous naming conventions and ecological knowledge systems. This digitization workflow contributes to the documentation of living language revitalization practices and to the continuity of cultural knowledge across generations.

Conclusions

This case study documents the steps involved in digitizing a multilingual and multimodal dataset using programmatic tools for metadata curation before ingestion into archival repositories such as DSpace. Similar work has contributed to the publication of robust archival records [16], and open-source archival digitization workflows and toolkits have also been developed for preserving community histories in digital libraries and archival productions [17, 18, 19].

The present study builds on these efforts by showing how a multilingual print book and its associated multimedia assets may be organized through an intermediary metadata workflow that preserves referential relationships across distributed file formats.

The paper has discussed the implications for archivists and for digital publishers in considering IMDP as an integral toolkit for streamlining metadata associated with unstructured multimedia assets spread across multiple file types and storage locations. The RPFs produced in this study may be adapted to different data curation workflows that can better prepare materials for DSpace ingestion using SAF. More specifically, the study shows an intermediary layer of information (IMDP) that helps mitigate challenges of referential integrity and advance relational accountability in digitization projects involving Indigenous data.

IMDP is therefore proposed as a flexible metadata packaging framework for archiving data and fostering more descriptive and relational metadata sharing across distributed information systems. The RPFs generated in this study illustrate the emerging interplay between generative AI and human curation practices [20]. Such AI-supported programmatic tools can support archival practices that include retrieval, search, transformation, and packaging of multimodal and multilingual assets. The study also highlights that human oversight remains essential for interpreting multilingual archival records, preserving cultural naming conventions, and maintaining the relational significance of Indigenous knowledge across diverse forms of text and multimedia. Future work envisions participatory and collaborative workshops with skilled archival practitioners to solicit feedback on an integrated IMDP interface for managing metadata and file formats across unstructured text and multimedia collections.

References

- [1] C. B. Teuton and H. Shade, *Cherokee earth dwellers: stories and teachings of the natural world*. University of Washington Press, 2023, ISBN: 9780295750187.
- [2] J. Kinner and M. Bass, "The History Component of the DSpace(TM) Institutional Repository," *archiving*, vol. 1, no. 1, pp. 71-76, Jan. 2004, doi: 10.2352/issn.2168-3204.2004.1.1.art00017.
- [3] J. Ahmed and S. Kataria, "Batch Bitstreams and Metadata import using SAFBuilder in Dspace: A Practical Experience," *Library philosophy and practice*, pp. 1-4, 2023.
- [4] B. Anderson and D. Galvin, "The Problem of Preserving Database-Driven Content," *archiving*, vol. 8, no. 1, pp. 162-164, Jan. 2011, doi: 10.2352/issn.2168-3204.2011.8.1.art00037.
- [5] T. Räisänen, M. Lipsanen, and A. Föhr, "AI Powered Tools for Improving Usability in Digital Archiving," *archiving*, vol. 20, no. 1, pp. 71-74, Jun. 2023, doi: 10.2352/issn.2168-3204.2023.20.1.14.
- [6] J. Abdelnour-Nocera, T. Clemmensen, and M. Kurosu, "Reframing HCI Through Local and Indigenous Perspectives," *International Journal of Human-Computer Interaction*, vol. 29, no. 4, pp. 201-204, 2013, doi: 10.1080/10447318.2013.765759.
- [7] T. Balogun, "Data management of digitized indigenous knowledge system in repositories," *Information Development*, vol. 39, no. 3, pp. 425-438, 2023, doi: 10.1177/026666669231186777.
- [8] C. L. Palmer, K. Sebastian, M. Belarde-Lewis, S. Littletree, and N. Guerrero, "Stewarding Contextual Integrity in Data Services for Indigenous Scholarship," *Proceedings of the Association for Information Science and Technology*, vol. 61, no. 1, pp. 626-631, 2024, doi: 10.1002/pa2.1070.
- [9] M. Belarde-Lewis, S. Littletree, I. R. Braine, K. Srader, N. Guerrero, and C. L. Palmer, "Centering Relationality and CARE for Stewardship of Indigenous Research Data," *Data Science Journal*, vol. 23, p. 32, May 2024, doi: 10.5334/dsj-2024-032.
- [10] C. L. Palmer, M. Belarde-Lewis, T. Hohn, and C. B. Teuton, "Indigenous Research Data Case Study: Toward Contextual Integrity for Indigenous Data," Aug. 2022, Accessed: Feb. 25, 2025. [Online]. Available at: <http://hdl.handle.net/1773/49173>
- [11] M. K. Buckland, *Information and society*. in The MIT Press essential knowledge series. Cambridge, Massachusetts ; London, England: The MIT Press, 2017, ISBN: 9780262533386.
- [12] C. B. Teuton, MB. Timothy, T.N. Belt, Larry Shade, Loretta Shade, and H. Shade, *Cherokee Natural World: Stories, Language, and Teachings*. RavenSpace Publishing, 2025. Available at: <https://doi.org/10.59962/kpuj7830>
- [13] J. E. Winston, *Integrative and Comparative Biology*, Twenty-First Century Biological Nomenclature-The Enduring Power of Names, 2018. Available at: <https://doi.org/10.1093/icb/icy060>

- [14] S. Wilson, *Research is ceremony: Indigenous research methods*, Fernwood Publishing, 2008, ISBN: 1552662810.
- [15] R. W. Kimmerer, *Braiding sweetgrass: Indigenous wisdom, scientific knowledge and the teachings of plants*. Milkweed Editions, 2015, ISBN: 9781571313560
- [16] A. Collie, D. Higgins, L. Mak, and S. Nicholson, “Accumulating Metadata from Tributary Workflows for More Robust Archival Records,” *archiving*, vol. 10, no. 1, pp. 84–87, Jan. 2013, doi: 10.2352/issn.2168-3204.2013.10.1.art00019.
- [17] A. Wilson, “An Open Source Tool for Migrating Digital Records for Long-term Preservation,” *archiving*, vol. 2, no. 1, pp. 119–123, Jan. 2005, doi: 10.2352/issn.2168-3204.2005.2.1.art00025.
- [18] A. Miranda and Z. Hruska, “Developing ARCLib - An Open-Source Solution for a Bit-Level and Logical Long-Term Preservation,” *archiving*, vol. 14, no. 1, pp. 74–77, May 2017, doi: 10.2352/issn.2168-3204.2017.1.0.74.
- [19] B. Presnall, “Archive Digitization Application (ADA),” *archiving*, vol. 11, no. 1, pp. 115–117, Jun. 2014, doi: 10.2352/issn.2168-3204.2014.11.1.art00025.
- [20] J. E. Davet, B. Hamidzadeh, P. C. Franks, and J. Bunn, “Tracking the Functions of AI as Paradata & Pursuing Archival Accountability,” *archiving*, vol. 19, no. 1, pp. 83–88, Jun. 2022, doi: 10.2352/issn.2168-3204.2022.19.1.17.

Author Biography

Turam Purty is a member of the Ho indigenous tribe from South East Asia. Turam develops AI assisted data curation practices at the intersection of Indigenous Knowledge Systems and Participatory Sciences. Previously, an Anecdotal Research Fellow (2020-21), Turam has developed online tools for the self-assessments of data privacy and usability of digital technologies utilized in crowdsourcing data from participatory science approaches.

Dr. Christopher Teuton is a citizen of the Cherokee Nation and member of Echota-Tanasi Ceremonial Grounds in Park Hill, Oklahoma. Dr. Teuton is the author of the print book, *Cherokee Earth Dwellers*. The print book is co-created with the late Cherokee National Treasures Hastings Shade and Loretta Shade and presents a Cherokee ecology explored through Cherokee creature names, environmental relationships, traditional stories, and philosophical discussions with fluent Cherokee speakers and knowledge keepers.