Two Open-Source Tools for Digital Asset Metadata Management

Alexander Duryee

Bertram Lyons

AVPreserve

In today's world of digital information, previously disparate archival practices are converging around the need to manage collections at the item level. Media collections require a curatorial approach that demand archivists know certain information about every single object in their care for purposes of provenance, quality control, and appraisal. This is a daunting task for archives, as it asks that they retool or redesign migration and accession workflows. It is exactly in gaps such as these that practical technologies become ever useful. This article offers case studies regarding two freely-available, open-source digital asset metadata tools—BWF MetaEdit and MDQC. The case studies offer on-the-ground examples of how four institutions recognized a need for metadata creation and validation, and how they employed these new tools in their production and accessioning workflows.

Introduction

Digital asset metadata is a critical aspect of preservation, access, and stewardship; without this information, objects become increasingly difficult to manage and use. Without knowing precisely what an object is (both as a technical asset and intellectual entity), questions such as "how do we present this?" and "are there obsolescence concerns?" cannot be answered. Supplementary to external datastores—databases, spreadsheets, and the like—technical and descriptive metadata about digital collections can be embedded within the assets themselves. Altogether, external and

embedded metadata comprise the total knowledge of a given digital object, and this information is critical in archival workflows and in preservation repositories. This study evaluates two tools that offer solutions for creating, reading, and making use of metadata that can be embedded within digital objects. Some uses for embedded metadata in an archival environment include automated quality control, object self-description, disaster recovery, and metadata sharing between systems. The two tools discussed in this study, MDQC (Metadata Quality Control) and BWF MetaEdit are open-source software utilities created to aid archivists in the creation and use of such metadata in daily workflows. This article explores four production implementations of these tools and their usefulness for overcoming problems and streamlining processes.

MDQC – Introduction

The sheer quantity of digital assets created as a result of digitization projects and the resulting large-scale ingests often overwhelm library staff. Outside of digitization on demand, objects are typically digitized at scale in order to capitalize on efficiencies of volume. In such cases it is not uncommon for small archival teams to handle many thousands of digital assets, each of which must go through an ingest workflow. The most important part of ingest workflows—performing quality control on incoming preservation masters—is often the most time consuming step for digital archivists. An archive may wish to ensure that its digital assets conform to naming standards, minimum quality specifications, or format compliance. These assets are typically reviewed manually at the item level, as evidenced in our case studies below. In such cases, a bottleneck emerges because the rate at which quality control is performed falls behind the rate at which newly digitized assets are created and acquired.

Quality verification also tends to be an ineffective use of staff time. Despite its importance, it is tedious and a poor use of skilled labor. Digitization projects and departments can sink unanticipated amounts of valuable time and resources into item-

level quality control, thus detracting from other services (both real and potential). All told, asset quality control is a step in archival workflows that is ripe for improvement.

MDQC – Tool Development

MDQC (Metadata Quality Control) is a software application developed by AVPreserve to address these bottlenecks and expedite digitization workflows. MDQC is a free and open-source tool based on existing utilities, Exiftool and MediaInfo, that allows users to set baseline rules for digital media asset quality (e.g., resolution, frame rate, or color space) and embedded metadata specifications (e.g., date formatting, completed fields, or standard values). Once a set of rules is created, it can be applied across an entire collection at once, reporting any assets that fail to meet the quality standard (e.g., wrong color space, below minimum resolution, gaps in descriptive metadata, or wrong sample rate). From these reports, which are generated using minimal staff time, an archivist can separate problematic assets from those that do meet the required specifications. As such, MDQC expedites the quality control of digital media collections, replacing a manual item-level task with an automated collection-level one.[i]

One important feature of MDQC is that it is designed around the sole task of technical metadata analysis and quality control. While it is a powerful tool during certain stages of the digitization workflow, it is not useful beyond these steps. For example, MDQC will not help in assessing the orientation of scanned images, the clarity of digitized audio, or in detecting frame-level corruption in digital video. Assessing the holistic qualities of a digital asset still requires human analysis, even in workflows where MDQC is applied. MDQC does not allow for writing to files—it will read metadata from digital assets, but is not an editor like BWF MetaEdit or Exiftool. While these shortcomings can serve as drawbacks in certain workflows, MDQC's focus on a single

task allows for it to be integrated into digitization processes with minimal disruption to existing practices.

MDQC – Case Studies

During the development of MDQC, AVPreserve worked with two organizations to test and implement MDQC in a production setting. The Digital Lab at the American Museum of Natural History applied MDQC in a large-scale image digitization project and successfully used it to expedite their processing workflow. Similarly, the Carnegie Hall Archives used MDQC to verify if vendor-generated assets were meeting the preservation quality specified in the statement of work.

The following short case studies outline how these two organizations implemented MDQC and how the tool subsequently affected their digital asset workflows.

Unsupervised Image Digitization: American Museum Of Natural History

Background and practices

The Digital Lab at the American Museum of Natural History (AMNH) is working on an ambitious project digitizing historical photonegatives, with the goal of scanning each one—over one million in total—and making them accessible in a public digital asset management system for research use. Currently, the AMNH is digitizing these photonegatives using a volunteer force, which generates roughly 200–300 images per week in tandem with a small team that perform quality control and image processing. Due to the nature of volunteer labor, changing standards over time, and turnover, quality control is tremendously important to the Digital Lab's project. Traditionally, this was performed on a per-image basis, where scans were loaded into Adobe Photoshop

and visual/technical assessments were performed. This process was slow and repetitive, and created a bottleneck in the imaging workflow.

Selection and implementation

The Digital Lab's operational environment was ideal for testing and implementing MDQC. Using MDQC, the Digital Lab was able to set its imaging quality standard for resolution, color space, file format, compression, and bits per sample. Once the standards were set, MDQC could test each file automatically against the specified standards. While this does not capture every aspect of image quality control—MDQC's limitations mean that a brief visual inspection is still needed for alignment, cropping, and other activities—it supports rapid automated testing for basic technical standards. This expedites the image review step in the Digital Lab's digitization workflow. Images can now be assessed hundreds at a time for technical quality.

MDQC also proved to be successful in legacy asset management. The Digital Lab, when first embarking on its project, did not have established standards or workflows for its volunteer scanning efforts. As such, there were an overwhelming number of images—approximately sixty thousand—that were created without a standard specification in mind. These images may or may not meet the current standard, and may or may not need to be reprocessed. Manually performing quality control on these legacy images would be arduous because new images (requiring quality control) are being created every day. By automating technical quality control, MDQC allowed the Digital Lab to bring these legacy assets under control. The Digital Lab manager can set their current imaging standard into a rule template and apply it across thousands of images at once, and thus automatically sort between images that meet the specification and those that do not. As of this writing, MDQC has helped the Digital Lab bring fourteen thousand legacy assets forward into their workflow, saving the Lab weeks of labor.

Benefits to the organization

MDQC created considerable time savings for the Digital Lab. By verifying technical metadata and image quality automatically, it saves approximately 20 seconds of labor per image. Given that the Lab's volunteers generate approximately 300 images per week, this translates into nearly two extra hours of time created by MDQC. MDQC's ability to scale across large collections has also saved a tremendous amount of time in processing the legacy image collection: at 20 seconds per image, it has so far saved the Lab 78 hours of work, and will ultimately reduce the time needed on the project by eight weeks of full-time labor.

MDQC also allowed the Digital Lab manager to develop a training program on image processing for interns. Previous to implementing MDQC, the processing bottleneck was quality control; thus, there was little need for interns to work on post-QC tasks. Now that the rate of quality control is considerably faster, the point of backlog has moved forward to image processing. As such, the AMNH uses images pending processing to train Digital Lab interns, who then work at this point in the digitization pipeline. This is a more efficient use of both the interns' and manager's time, and helps to further expedite the digitization workflow.

Massive Vendor Digitization: Carnegie Hall

Background and practices

In 2012, the Carnegie Hall Archives (CHA) launched the Digital Archives Project (DAP), a comprehensive digitization program, to digitize and preserve a majority of their legacy archival holdings. Due to the scope and limited time period of the 3-year grant project, CHA used a vendor service to digitize manuscripts, audio, video recordings, and film, which were returned in bulk on hard disks. Because the vendor-supplied materials will be the digital masters for these assets, the archivists at CHA implemented a quality control workflow for returning assets.

Previous to implementing MDQC, the workflow involved a technician opening each file in Adobe Bridge and comparing technical metadata against a set standard in an Excel spreadsheet. This step is important in guaranteeing that the vendor meets the minimum standard for quality, but it is also time-consuming. The lead archivist at CHA estimated that the technician could manually process 70–100 images per hour out of a total of 35,816 images digitized. In order to perform item-level quality control on every single asset, approximately 400 hours of labor would be required. In addition to the images, CHA had 1,235 audio and 1,376 video assets in its digitization pipeline. Performing technical quality control on every single asset would take months of work. As such, CHA was performing manual quality control on 25–30% of their digital assets; while not optimal, the scale of the project prevented any further assessment.

Selection and implementation

CHA was developing a backlog of material to review, making MDQC a natural fit in their production and ingest workflow. The manual step of verifying technical quality could be automated via MDQC by establishing baseline rules (as outlined in the service contract with the vendor) and testing returned assets against those rules. This fit neatly into the CHA workflow, as returned assets could be scanned in-place on the hard drives before further action was taken.

Benefits to the organization

As a result of MDQC, CHA expedited their digitization workflow. Batches of newly digitized assets were assessed for technical quality (e.g., resolution, compression, format, and color space) within minutes instead of weeks or months. As MDQC is unable to perform content analysis, there is still a need for human analysis of assets for issues such as digital artifacts and playback problems. However, these can be performed more efficiently due to the reduction of steps necessary for thorough quality control—by having MDQC ensure that the vendor is meeting the technical specifications for DAP,

the staff can focus on content analysis. As such, CHA was able to accelerate their workflow and make progress on this aspect of DAP in a very short time.[ii]

BWF MetaEdit – Introduction

Another issue in digital archival workflows is the implementation of embedded metadata in audio assets—specifically, digitized preservation masters in uncompressed 24-bit/96 kHz WAV files.[iii] Embedding metadata in digital media is attractive for archives, both in terms of technical processes and long-term preservation. Embedding metadata in a file allows for very easy automation of workflows in the future—for example, a digital asset management system may read metadata directly from the file upon ingest, thus providing a rapid and automated method for generating metadata records. Using embedded metadata also makes the object-context relationship more robust, by adding an additional source of information about an asset. For audio assets, which can be nearly impossible to contextualize without external information, maintaining the media-metadata link is essential.

Despite these advantages, it has traditionally been difficult to create, view, and edit metadata written directly to an audio file. Metadata often exists in databases or spreadsheets, where an item-by-item migration process would be difficult. Depending on the format of assets, there are also questions of the ease of embedding metadata; while database and office software are common in archival organizations, specialized tools for writing data to media files may not be. Writing metadata to a file will also change its checksum, which may cause issues in digital preservation environments unless properly handled.

Embedded metadata also carries a level of fragility with it. While it is rather secure when embedded into a static object (for example, a preservation master in a repository), there is a level of risk when the asset goes through a processing workflow.

Many editing applications will overwrite or damage embedded metadata, and transcoding will almost always cause the loss of metadata.[iv] Coupled with the difficulty of reading and writing embedded metadata, there are risks of creating metadata that will be lost during routine processing.

BWF MetaEdit – Tool Development

BWF MetaEdit was developed by the Federal Agencies Digitization Guidelines Initiative (FADGI), supported by AVPreserve, in order to expedite the creation and management of embedded Broadcast WAVE Format (BWF) metadata (including bext, LIST-INFO, axml, ixml, and xmp) in WAV files. BWF metadata is embedded as well-defined data either before or after the audio section of a WAV file, in conformance with the EBU 3285 standard.[v] As an embedded standard, very few tools can read and write this metadata—most are production-caliber software suites that are beyond the scope and reach of many organizations. There are also few tools designed from the ground up for working with embedded BWF metadata; due to its origin as a broadcast standard, much of the software supporting BWF metadata treats it as an ancillary aspect of other workflows. FADGI created BWF MetaEdit in order to overcome these barriers to the adoption of BWF in archival environments. BWF MetaEdit was designed with the singular goal of providing a powerful and simple tool for managing metadata embedded in WAV files as part of archival workflows across organizations and stakeholders. The tool was also released by FADGI as a free and open-source solution on multiple platforms, thus removing the financial and technical burden of acquiring software.

BWF MetaEdit provides a graphical interface for viewing, adding, and editing embedded metadata in WAV audio files, as well as bulk operation tools. In the main window of the GUI, a technician can work with embedded metadata similarly to a spreadsheet, with each cell representing one field for one file. In this manner, files can have metadata added to them directly, either via single entry or setting all of one field

to a single value (e.g., setting the digitization technician's name in every file). BWF MetaEdit also allows for importing large amounts of metadata as a CSV file, which can be mapped to corresponding RIFF and BWF fields and written as embedded metadata. This feature in particular allows for rapid migration of legacy metadata to embedded BWF metadata. BWF MetaEdit's spreadsheet-like view allows for easy quality control and adjustment before writing out data. Just as it allows importing of data, BWF MetaEdit provides bulk and item-by-item metadata export (in CSV and XML). Additionally, BWF MetaEdit can utilize otherwise unused space in the WAV header to calculate and write MD5 checkums of the audio data of the file, allowing it to have a known fixity value even after new metadata is written.

BWF MetaEdit is similar to MDQC in that it was designed with a single purpose in mind. Whereas metadata editing capabilities are often built into more comprehensive audio engineering applications, BWF MetaEdit was built as a standalone tool strictly for handling WAV metadata. As such, while it offers unique capabilities for metadata creation and extraction, it does nothing else. While this can be a benefit—it can be integrated into existing processes—it is also only a small part of a complete digitization and management workflow.[vi]

BWF MetaEdit – Case Studies

Embedding and Extracting Audio Metadata: THE SMITHSONIAN CENTER FOR FOLKLIFE AND CULTURAL HERITAGE

Background and Practices

The Smithsonian Center for Folklife and Cultural Heritage (CFCH) oversees hundreds of thousands of digitized audio assets. These assets were previously described in Microsoft Access databases or SIRSI records, which were difficult to access outside of CFCH. The lack of centralized description also made it difficult to merge CFCH's catalog

with the Smithsonian's Digital Asset Management System (DAMS), because the migration process would be unduly arduous. Furthermore, since assets were separate from their descriptive metadata, they were difficult to move around the Smithsonian Institution—unless accompanied by sidecar files, the assets had no meaningful context.

Selection and Implementation

By using BWF MetaEdit's CSV import feature, the Ralph Rinzler Folklife Archives and Collections (RRFAC), a division of CFCH, quickly and effectively improved nearly 300,000 audio assets. Descriptive and technical metadata that were located in databases and spreadsheets were exported into comma-separated text documents and aligned with BWF metadata fields. These CSV documents were then imported into BWF MetaEdit, where the values were matched with their corresponding files and reviewed for accuracy. RRFAC was then able to embed the metadata into the audio files as a batch operation within BWF MetaEdit. Using this workflow, RRFAC was able to embed item-level BWF metadata into nearly 300,000 audio files in a matter of months, dramatically improving the quality and flexibility of their collections.

The technicians at RRFAC also used BWF MetaEdit to expedite their digitization workflow. Much of the metadata for describing their audio assets is universal across the collection; for example, LIST-INFO tags IARL (archival location) and ICOP (copyright statement) are consistent for most incoming assets. Since these metadata fields are constant, RRFAC was able to use BWF MetaEdit's command line mode to create a reusable script for embedding these values. By using BWF MetaEdit to embed multiple metadata fields with a single key command, the digitization process was accelerated by reducing the number of fields requiring manual data entry to the specific fields that are unique to each asset.

One significant use of BWF metadata at the Smithsonian was the acceleration of asset migration. One obstacle to the migration of RRFAC assets into DAMS was keeping

metadata and data together through the process. Given that the metadata existed in a mixture of databases and spreadsheets, this would have required considerable effort—metadata would have to be exported into a temporary sidecar format (which DAMS would have to recognize and parse), and digital assets would have to be monitored as they were ingested and married to their records. During the development of the ingest workflow for CFCH material, it was realized that DAMS could automatically extract embedded metadata from BWF files (as per the EBU 3285 specifications) and use it to generate its descriptive records. Thus, the migration process could be simplified. Instead of ingesting database records and assets separately, the DAMS could acquire much of its metadata directly from the asset itself. As the metadata would be embedded into the preservation master file, it removed the need for temporary sidecar records; in addition, it ensured metadata interoperability by adhering to an internationally accepted standard. This also improved the accessibility of audio assets, as new methods of searching could be applied across them (e.g., across keywords embedded in metadata).

Since the digital assets were being ingested into a preservation environment as master assets, there was no risk of the embedded metadata being inadvertently removed via transcoding or audio editing software.

Benefits to the Organization

Using BWF MetaEdit allowed CFCH to embed technical and descriptive metadata in their digitized audio assets. The benefits of embedding metadata in audio objects are many—by keeping metadata and asset joined in a single file, assets are more easily managed as they move through the greater Smithsonian organization. The smooth workflow established by the Smithsonian for CFCH-to-DAMS ingest would be arduous and time-consuming without the assistance of a tool such as BWF MetaEdit.

Embedding Audio Metadata in Post-digitization Workflows: Recorded Sound Section, Library of Congress

Background and practices

The Recorded Sound Section of the Motion Picture, Broadcast and Recorded Sound Division of the Library of Congress is responsible for the acquisition, care, management, description, and preservation of the vast majority of the audio holdings of the Library of Congress. Situated at the Packard Campus of the National Audiovisual Conservation Center (NAVCC) in Culpeper, Virginia, audio engineers in the Recording Laboratory (RL) work diligently to reformat (digitize) these sound recordings for preservation and access. The process of digitizing sound recordings at the Library of Congress is a highly technical activity requiring the expertise, equipment, and supplies necessary to handle every type of audio format from the earliest wax cylinders to the most recent digital bit stream.

NAVCC is also home to a state-of-the-art technical infrastructure for digital storage and management, including an ever-growing layer of middleware preservation tools and appliances to support the services needed for long-term digital preservation, including checksum generation, fixity checks, preservation metadata management, and rule-driven automation, among many others. After engineers complete the digitization of a given recorded sound object at NAVCC, the newly created digital files make their way through a series of manual and automated steps into the preservation storage environment and/or onto spinning disk servers for immediate access.

Between these two very complex processes—digitization for preservation and ingest into long-term digital storage—are a number of steps that relate to the composition of the digital file itself and to the documentation of various aspects of the file for access, administrative, and preservation purposes. One of these steps relates to the specific question of what human-readable information is embedded into the file

itself. Until recently, most archives and libraries that were creating digital sound recordings were not embedding consistent metadata in the headers of preservation sound files. However, the work of FADGI, as mentioned earlier in this paper, established a recognized set of guidelines for conforming to existing EBU standards for embedding metadata in audio files in the broadcast environment. The audio engineers at the Library of Congress now had a mandate and a desire to implement a new workflow into their current post-digitization processes: embedding metadata in the header of each WAV file created as a result of preservation reformatting.

Selection and implementation

The development of BWF MetaEdit gave the audio engineers at NAVCC a chance to look at the header of a WAV file and edit it. They did not have a way to do so before. The Digital Audio Workstation being used at NAVCC was embedding some information in the header, but it was doing it in a way that was opaque to the engineers. While metadata may have been present, the lack of clear specifications and unclear provenance made it difficult for future use. They needed a way to determine what was being embedded in the file so they could make comparisons to the FADGI recommendations and determine what they needed to do to conform to the recommendations. Because BWF MetaEdit offered the ability to read the headers and to edit them, NAVCC audio engineers decided to implement the tool in their post-processing workflows.

Although the use of the tool at NAVCC is still evolving, currently audio engineers make use of a combination of batch and manual options for editing and inserting metadata into WAV files. After NAVCC audio engineers create a preservation master file, an additional piece of software is used to generate a derivative access file. Once the pair of files has been created, the engineer opens BWF MetaEdit and uses a customized CSV template to insert standard information into selected fields in the bext and INFO chunks of the headers of the two files. The template also triggers the generation of an md5

checksum for the bit stream of the audio file. This functionality is built into BWF MetaEdit and supports fixity checks on the audio stream to ensure files are not corrupted by the use of the tool. The engineer follows this step by inserting source-specific information about the given files, including unique control numbers for the original recording and information about the encoding history of the file. Once complete, the engineer saves the files through BWF MetaEdit and the files now contain the specified embedded metadata in the header.

As mentioned above, the files are now ready to be ingested into NAVCC's preservation storage environment.

Benefits to the organization

Embedding metadata into files provides valuable data to digital preservation systems and workflows. In case of disaster, it allows for the positive identification of files and contents—if, for example, a filename were to be changed, the BWF metadata can be used to re-identify it; the information within the file remains unchanged and tools such as BWF MetaEdit can be used to recover that information as well as to create such information in the first place.

At NAVCC, BWF MetaEdit has allowed the audio engineers to have full control of what metadata is or is not included within the preservation audio files they are creating on a daily basis. The engineers can be certain that they are following international recommendations for audio preservation practices. Additionally, the use of BWF MetaEdit at NAVCC has generated new lines of communication and inspired audio engineers and cataloging staff to work together to determine the right balance for what descriptive and administrative information should be contained within a file.

Conclusions

MDQC and BWF MetaEdit allowed these four organizations to accelerate their digitization workflows and to effectively manage existing assets. By automating technical quality control and asset augmentation, they allowed these organizations to focus their time and labor on more fruitful tasks. With MDQC, technicians can focus on processing and ingest instead of technical standards, and interns/volunteers can be trained on more valuable tasks than the rote checking of technical metadata. Meanwhile, BWF MetaEdit allows for the creation of embedded metadata in preservation audio files, which can have a tremendous impact in effectively managing assets. Additionally, by expediting the previously slow process of quality control and metadata creation, assets can move quickly through production workflows. The continued development of tools such as MDQC and BWF MetaEdit will increase digitization throughput and productivity.

The most surprising and exciting development from these implementations was how dramatically they could affect an organization. By automating tedious and time-intensive tasks, they opened the door to new services and simultaneously expedited existing ones. The AMNH was able to use MDQC to offer new research services by applying it to patron-generated assets, thus creating a new source of materials for their digital archive. BWF MetaEdit allowed the CFCH to migrate their assets into the emerging Smithsonian DAMS, which greatly improves access without dramatically increasing labor. These successes showcase how software appliances help relieve the burden of managing high volumes of digital assets. Verifying and enhancing batches of received or created content now require minimal additional work by the archivist and can be done easily as part of a daily workflow.

Acknowledgements

The authors would like to thank Anna Rybakov and Jennifer Cwiok of the AMNH Digital Lab and Miwa Yokoyama of the Carnegie Hall Archives for their use, feedback, and patience in beta testing MDQC, as well as Dan Charette, formerly of the Center for Folklife and Cultural Heritage, Smithsonian Institution, and Brad McCoy of the Library of Congress for their input and correspondence regarding the BWF MetaEdit case studies. We also thank Phil Harvey, Jerome Martinez, and FADGI for developing and supporting Exiftool, MediaInfo, and BWF MetaEdit respectively.

Useful Resources

Tools covered:

- MDQC – http://www.avpreserve.com/avpsresources/tools/
- BWF MetaEdit – http://bwfmetaedit.sourceforge.net/

Tools used by MDQC:

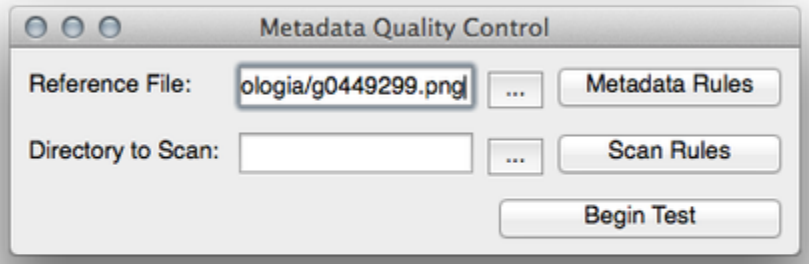
- ExifTool – http://www.sno.phy.queensu.ca/~phil/exiftool/
- MediaInfo – http://mediaarea.net/en/MediaInfo

Resources on BWF Metadata:

- Guidelines for Federal Agency Use of Broadcast WAVE Files
  http://www.digitizationguidelines.gov/audio-visual/documents/Embed_Guideline_20120423.pdf
  http://www.digitizationguidelines.gov/guidelines/digitize-embedding.html
- "Embedded Metadata In WAVE Files: A Look Inside Issues and Tools" – http://www.avpreserve.com/wp-content/uploads/2014/04/EmbeddedMetadata.pdf

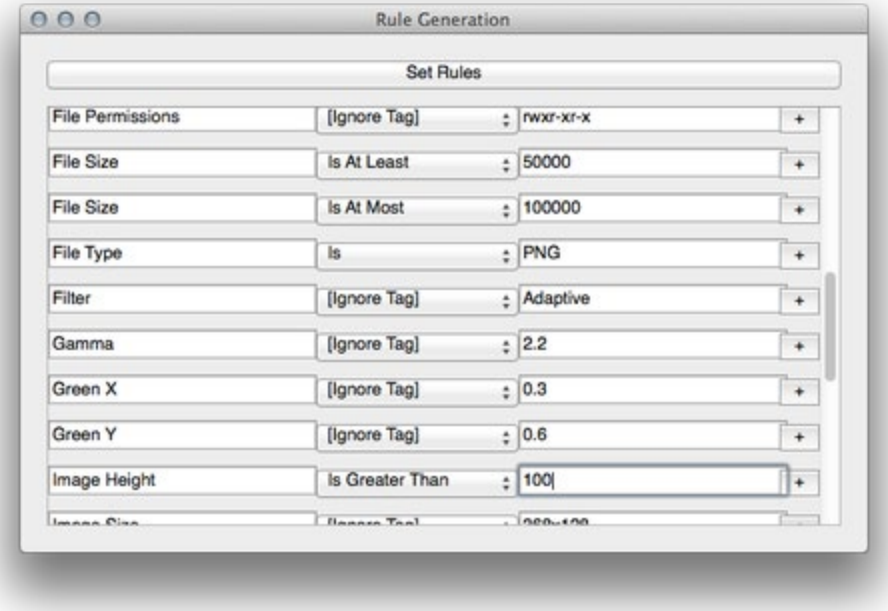
APPENDIX A: MDQC Walkthrough

1. Upon starting MDQC, the main window will appear. Here, it asks to set a reference file, metadata rules, a base directory, and file filters. The process begins with selecting a file to serve as a basis for building your metadata rules—typically, this will be a file that is known to meet quality standards.

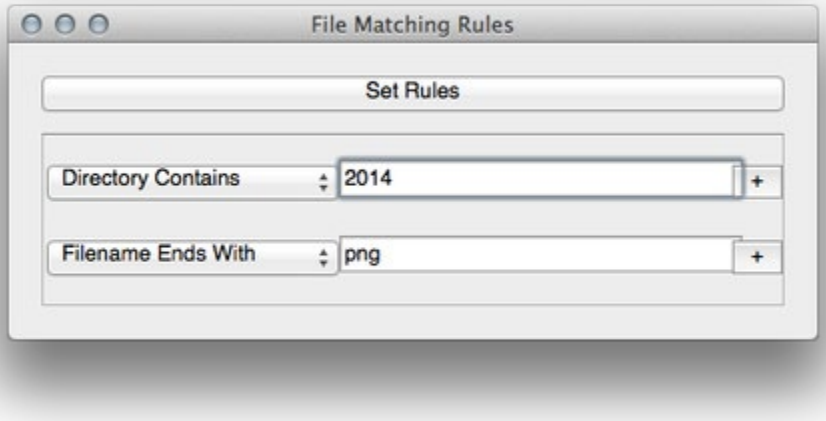


2. Once a file is selected, the next step is opening the Metadata Rules window. By default, Exiftool is selected as the metadata extraction tool to scan files (this can be changed under the Tools menu). MDQC will take each metadata field from Exiftool and present it for building your QC rules. In the center column, value operators are selected from a list—for example, a rule can be set that a value must be present, or that an element is the same as a certain value. These operators allow for powerful comparisons of digital asset metadata, such as verifying formats (by comparing the MIME type) and or ensuring that digitized image DPI meets recommended standards.

In this sample, every file in the collection should be at least 50 kilobytes and at most 100 kilobytes, of file type PNG, and larger than 100 pixels high. Note that the + button on the right allows for duplication of rules, allowing the setting of multiple constraints.

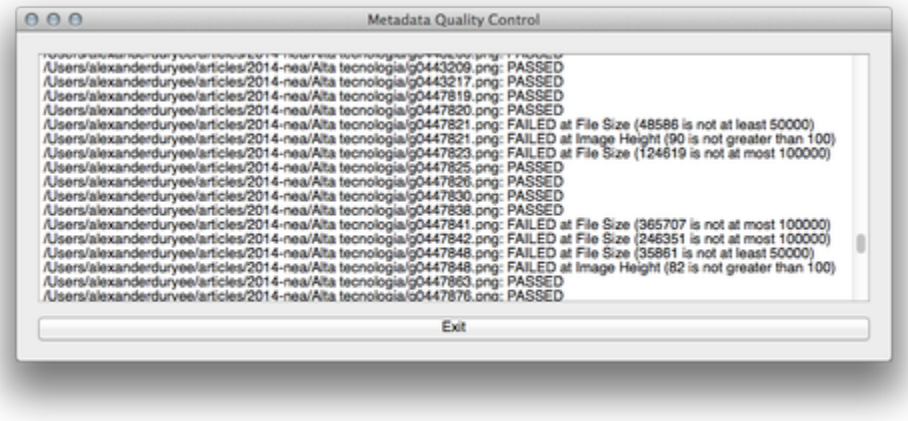


3. Following this, MDQC will need a base directory to scan. MDQC will test every file in this directory, along with every file in every subdirectory, against the metadata rules set in Step 2. This is not always desirable—for example, a collection may contain mixed asset types, and it does not make sense to test text documents for image width. This is solved by Scan Rules, which set filters that filenames and file paths must meet in order to be tested by MDQC. In this sample, the filter is set to only test files with the extension png, in directories containing 2014.



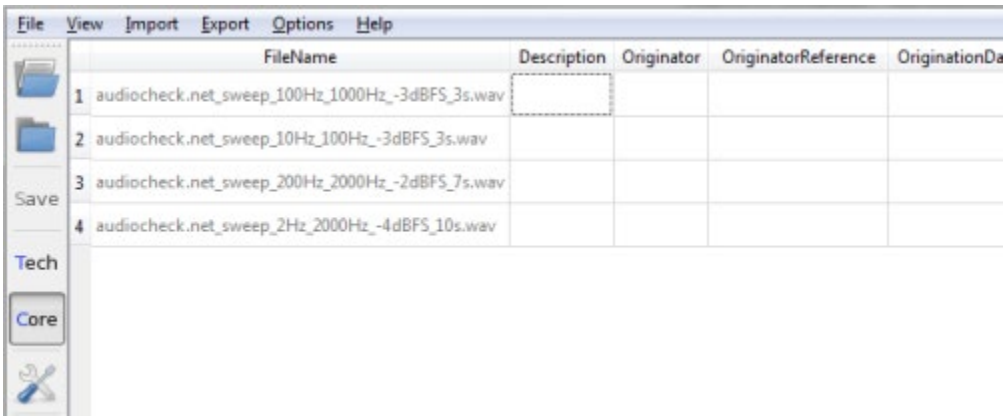
4. Often, it is helpful to save a rules set and scan rules to reuse—for example, if a complex rules set is highly detailed, or in order to create a toolbox of rules sets to use for different formats. Under File -> Save Template, templates can be saved for later use; they are loaded again via File -> Load Template.

5. Once the metadata rules and scanning rules are in place, it is time to scan by selecting Begin Test. This initiates the discovery of files meeting the rules set in Scan Rules, which are then processed by Exiftool or MediaInfo. The metadata is then compared against the rules set in Metadata Rules, and if it meets the constraints set there, the file passes. If it does not—in the sample, some files were too small—MDQC will report the filename, element in question, the value from the file, and the constraint it did not meet. The output is written to a report file (a CSV containing the rules and the result of each file scanned) as well as to the report window, allowing for later analysis.



APPENDIX B: BWF Metaedit Walkthrough

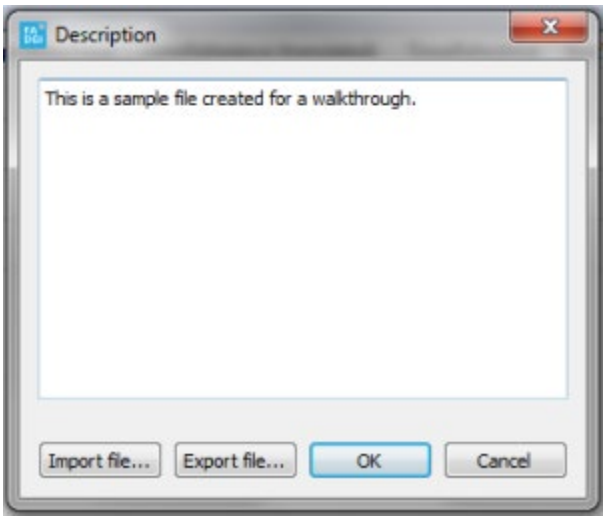
1. Starting BWF MetaEdit opens a blank window. From here, you can open a single file via File -> Open File(s), or load in an entire directory of WAV files via File -> Open Directory. In this screenshot, there are four sample WAV files. Note that BWF MetaEdit separates metadata into Tech and Core windows. Tech metadata encompasses information such as file size, bit and sample rates, duration, and MD5 checksums (most of this is metadata that you should not change, and BWF MetaEdit, therefore, does not all you to change the majority of this metadata). Core metadata includes more descriptive metadata, such as freetext descriptions, information on the origination of the asset, and provenance documentation about the digitization of the asset.



2. BWF MetaEdit automatically extracts embedded metadata from WAV files and displays it in a table format. By exploring the interface, you can easily see which fields are populated and what metadata they contain. Hovering over a field brings up a tooltip providing information on the field, along with standards and recommendations for its value. If you wish to edit a single field in a file, double-clicking the field will bring up a

window for editing that field. In the sample, the description of the first file in the list is being edited.

Note that BWF MetaEdit does not write any information to a file until the Save button on the main window is pressed. Until then, changes can be made without altering the WAV file.



3. Collections often include many audio assets, and embedding metadata one field at a time will not scale to them. BWF MetaEdit allows for importing and exporting metadata from many files at a time via comma separated text files. You can create a CSV file in Excel or a text editor, which aligns metadata in a table similar to BWF MetaEdit's main screen. Here, metadata can be edited and created using powerful tools external to BWF MetaEdit—for example, you can copy a single cell down an entire column, or perform a find-and-replace to update metadata. The CSV format also makes it simple to export metadata from other sources, such as spreadsheets and databases, and manipulate them before importing into BWF MetaEdit.

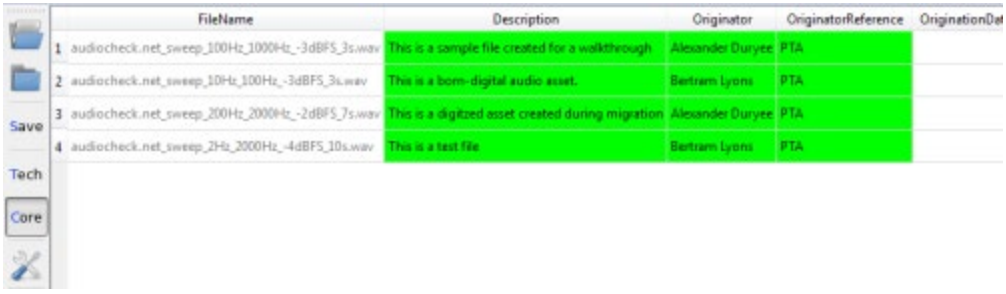
One useful trick to build a basic CSV template is to export metadata from files before editing them. This will create a file with the empty metadata fields properly

aligned and ready to populate. Metadata can also be exported from BWF MetaEdit in bulk, which can then be used to populate a database or to be transformed into metadata records.

It is also critical to be mindful of how spreadsheet applications store metadata: Excel, for example, stores times and dates in a format that is incompatible with the Broadcast WAVE standard. As such, any spreadsheets that are being prepared for import should be opened as text data to ensure data uniformity.

	B	C	D	E
	Description	Originator	Originator	Origin
0Hz_1000Hz_-3dBFS_3s.wav	This is a sample file created for a walkthrough	Alexander Duryee	PTA	
Hz_100Hz_-3dBFS_3s.wav	This is a born-digital audio asset.	Bertram Lyons	PTA	
0Hz_2000Hz_-2dBFS_7s.wav	This is a digitzed asset created during migration	Alexander Duryee	PTA	
z_2000Hz_-4dBFS_10s.wav	This is a test file	Bertram Lyons	PTA	

4. Once the metadata is satisfactory, it can be imported into BWF MetaEdit. The files that are going to be modified do not need to be loaded before importing—as long as they are at the location provided in the CSV, BWF MetaEdit can work on them. During the import process, if there are any problems, the program will display a log of any issues that arose—for example, if Excel transformed time and date data into its own format, BWF MetaEdit will alert you to this. If the import was successful, BWF MetaEdit will return to the main screen, where it will display the newly imported metadata in green cells. As before, the main window offers the opportunity for editing metadata, providing a final opportunity to adjust values. By selecting Save, the metadata will be embedded to the files.



About the Authors

Bertram Lyons is a certified archivist who works as Senior Consultant with AVPreserve and as Digital Assets Manager at the American Folklife Center at the Library of Congress. Prior to the Library of Congress Bert was Archivist and Collection Manager for the Alan Lomax Archive, one of the most significant collections of field recordings in 20th century history. A subject specialist in folklife collections, oral histories, and field and musical recordings, Bert has deep experience in all aspects of managing such collections, from the digitization of legacy materials, description, schema and database development, and accessioning/processing. He is active nationally and internationally with professional archival organizations such as the International Association of Sound and Audiovisual Archives (Member of the Executive Board and Editor of IASA publications) the Society of American Archivists (Chair of the Career Development Subcommittee and Vice-chair of the Oral History Section), the Association of Recorded Sound Collections, and the Association of Moving Image Archivists.

Alexander Duryee is a digital preservationist and metadata analyst with AVPreserve experienced in web archiving, optical and magnetic storage migration, and data preservation. Holding an MLIS from Rutgers University, his background includes archival tool development, data access, preservation infrastructure analysis, and metadata management.

---

[i] See Appendix A for a walkthrough of how to use MDQC.

[ii] See Appendix A for more information about how to use MDQC.

[iii] International Association of Sound and Audiovisual Archives. Guidelines on the Production and Preservation of Digital Audio Objects (IASA TC-04). London: IASA, 2009

[iv] Association of Recorded Sound Collections. “A Study of Embedded Metadata Support in Audio Recording Software: Summary of Findings and Conclusions.” ARSC: 2011. Accessed 2014-05-08 at http://www.arsc-audio.org/pdf/ARSC_TC_MD_Study.pdf.

[v] European Broadcasting Union. EBU-TECH 3285: Specification of the Broadcast Wave Format. Geneva: EBU, 2011

[vi] See Appendix B for a walkthrough of how to use BWF MetaEdit.