# INTRODUCTION

Since about 2016, the New York Public Radio (NYPR) Archive uses a workflow for audio-file ingestion that uses at its core the generation of a "golden record" which combines data from several sources. The process usually starts with adding the files to be ingested to a Windows folder. The files are then processed via a combination of software and xslt scripts, which generate xml documents to ingest into the Archives’ systems.

# OVERVIEW

The workflow accepts metadata from Broadcast WAV files as an xml RDF document. It then generates a new set of metadata as a "golden record" xml RDF document, taking into account the following possible sources:

## INPUTS

1. Embedded metadata
   1. Parseable filename or BEXT Description
   2. LIST chunk
   3. BEXT chunk
   4. Technical metadata
   5. XMP metadata
2. NYPR sources
   1. “Cavafy” Archives catalog
   2. DAVID System
   3. Publisher (NYPR’s web publishing system)
   4. NewsBoss (yearly)
3. External sources
   1. Library of Congress (id.loc.gov)
   2. AAPB data (https://americanarchive.org/)

**MASTER OUTPUT DOCUMENT**

The new set of metadata is an xml RDF document (top level: <rdf:RDF>) in a format similar to an xml file generated by exiftool when analyzing a wav file. This master document includes at least 19 fields (listed below).

From this “master file”, various scripts generate xml documents to use in the following processes:

## ADDITIONAL OUTPUTS

1. BWF MetaEdit (top level: <conformance\_point\_document>)
2. DAVID (top level: <ENTRIES>) (file extension: .DBX)
3. Cavafy (top level: <pbcoreCollection>)
4. Html promotional document
5. BWF MetaEdit (top level: rdf:RDF) (extension: .xmp)(to be implemented)
6. Complete log
7. Error log

INPUTS

External MD loc.gov

americanarchive

Embedded MD Filename

LIST

BEXT

Tech MD

xmp

NYPR Data "Cavafy"

D.A.V.I.D.

Publisher (CMS)

NewsBoss

Master RDF Filename

LIST

BEXT

Tech MD

xmp

BWF MetaEdit

(LIST, BEXT)

DAVID

(.DBX)

"Cavafy"

(pbcore)

Promo (html)

logs

BWF MetaEdit

(xmp)

 OUTPUTS

# SOFTWARE USED

## Exiftool

Exiftool (https://exiftool.org/) appears to read the most metadata from wav (and many other types of) files. The command we use is:

*exiftool –X –a –ext wav “your drive/directory/files.wav”*

## BWF MetaEdit

BWF MetaEdit can read and write metadata from wav files. See https://mediaarea.net/BWFMetaEdit

## OxYgen/Saxon

OxYgen can implement xml transformations and generate new documents. See <https://www.oxygenxml.com/>, http://saxon.sourceforge.net/

## Adobe Audition

This sound editing software includes a good metadata window. See https://www.adobe.com/products/audition.html

# THE PROCESS

The Archives catalog ("cavafy", <https://cavafy.wnyc.org>) uses the pbcore standard, which differentiates between conceptual "assets" and their associated "instantiations" (recordings).

Generally, the ingestion of new digital audio files follows the following steps:

1. New assets are identified (e.g. a donation of new tapes)
2. Archives personnel creates new cavafy assets with at least the following two fields either manually or via spreadsheet (recommended):
   * Collection name (e.g. WNYC, MUNI...)
   * Unique asset number

It is usually advisable to add as many additional fields as possible from the information at hand.

1. The transfer technician generates audio files according to the Archives naming convention, namely  
    COLL-SERIES-YYYY-MM-DD-12345.ext[ Free text].wav  
   (see below)
2. Using BWF MetaEdit, the transfer technician embeds additional metadata in the audio file. At a minimum:
   1. Medium (IMED)
   2. Technician (ITCH)

Encouraged:

* 1. Source (ISRC) (cavafy catalog URL)

The following fields will also need to be present either embedded or in cavafy:

* 1. Title
  2. Description (abstract/subject)
  3. One Library of Congress subject heading (keyword)

1. Archives personnel creates an rdf file via exiftool
2. The rdf file runs through an xslt transform, which generates several xml and html files
3. Metadata is embedded in the wav file via BWF MetaEdit
4. Files are uploaded to DAVID through a .DBX file
5. Files are added to cavafy through a pbcore file

# INPUTS

## Filename

The NYPR Archives naming convention is as follows:

COLL-SERIES-YYYY-MM-DD-12345.ext[ Free text].wav

Where

* COLL: Collection name (e.g. MUNI, WNYC…)
* SERIES: Series acronym (e.g. OTM, BLSH…)  
  YYYY-MM-DD: significant earliest date (e.g. 1978-12-03); unknown numbers are replaced with ‘u’ (e.g. 1978-12-uu means 'sometime on December 1978')
* 12345: cavafy asset number
* ext: Instantiation suffix, where each instantiation within an asset has a unique extension. Parts are designated with letters (e.g. .1a, .1b…)
* [ Free text]: An optional free text that may include information about :
  + other IDs (e.g. LT6978 for MUNI)
  + generation information (first two may be combined):
    - * WEB EDIT: file to be published
      * SEGMENT: partial file
      * ACLIP: short promotional “hot tape”
  + Descriptive information

## LIST chunk

ASCII-encoded fields, as expressed by Exiftool in xml. See “Master Output Document” below. See also <https://www.recordingblogs.com/wiki/list-chunk-of-a-wave-file>

## BEXT chunk

ASCII-encoded fields, as expressed by Exiftool in xml. See “Master Output Document” below. See also  
<http://www.digitizationguidelines.gov/audio-visual/documents/btext.html>

## Technical metadata

Exiftool can extract technical metadata as one document along with the other chunks. See <http://www.digitizationguidelines.gov/audio-visual/documents/tech_view.html>

## XMP metadata

Exiftool can extract XMP metadata, ISO standard16684, as one document along with the other chunks. See <https://www.adobe.com/products/xmp.html>

## “Cavafy” Archives catalog

A pbcore-based, MySQL database that allows UTF-8 xml output of its metadata via a limited RESTful API. Connected to wav files via the asset/instantiation number in the filename. See <https://cavafy.wnyc.org/login>

## DAVID System

NYPR’s scheduling system, developed by DAVID Systems GmbH. Allows ISO-8859 xml output of individual files with a ‘.DBX’ extension. See https://www.davidsystems.com/

## Publisher

NYPR’s web publishing system allows json output of result sets via a RESTful API. May connect to wav files in DAVID via the “theme” (aka “MOTIVE”) field. See https://github.com/nypublicradio/wnyc-web-client

## NewsBoss

Text-prompting software used by NYPR. Connects to WAV files via the “theme” (aka “MOTIVE”) field. Allows batch html output of internal result sets. See https://www.newsboss.com/

## Library of Congress

Vast authority files for names and subject headings (and many others). Outputs in xml (and others). Includes limited RESTful API via SRQ. See id.loc.gov

## AAPB (American Archive of Public Media)

NYPR has several thousand files in the American Archive of Public Media; some of them include transcripts. See <https://github.com/WGBH-MLA/AAPB2#api>

# THE MASTER OUTPUT DOCUMENT

The master document includes LIST and BEXT chunks in ASCII-encoded xml. Due to compatibility issues with current NYPR Systems, the NYPR Archives use the fields listed below; fields marked with an asterisk allow multiple entries.

## RIFF Chunk

|  |  |
| --- | --- |
| **Info ID** | **The corresponding text describes** |
| IARL | The location where the subject of the file is archived |
| IART\* | The artist of the original subject of the file |
| ICMS\* | The name of the person or organization that commissioned the original subject of the file |
| ICMT | General comments about the file or its subject |
| ICOP | Copyright information about the file (e.g., "Copyright Some Company 2011") |
| ICRD | The date the subject of the file was created (creation date) |
| IENG\* | The name of the engineer who worked on the file |
| IGNR | The genre of the subject |
| IKEY\* | A list of keywords for the file or its subject |
| IMED | Medium for the original subject of the file |
| INAM | Title of the subject of the file (name) |
| IPRD | Name of the title the subject was originally intended for |
| ISBJ | Description of the contents of the file (subject) |
| ISFT | Name of the software package used to create the file |
| ISRC | The name of the person or organization that supplied the original subject of the file |
| ISRF | The original form of the material that was digitized (source form) |
| ITCH | The name of the technician who digitized the subject file |

## BEXT Chunk:

Description

CodingHistory

## System, File, other technical MD

## XMP:

Image

Coverage (Location)

Lyrics (transcript)

# THE MASTER OUTPUT DOCUMENT FIELDS

Some fields in the master document allow for multiple entries, marked above with an asterisk\*. Multiple entries are separated by ‘ ; ‘ (space, semicolon, space).

## IARL (Archival location)

Correspondences in hierarchical order:

|  |  |
| --- | --- |
| Filename | ‘US, ’ + COLL |
| LIST | IARL |
| Cavafy | US, ‘ + pbcoreTitle[@titleType=Collection] |
| DAVID | Description: COLL |
| Publisher | Web domain |
| NewsBoss | 'WNYC' |
| Default | ‘US, NYPR’ |

## IART\* (Artists)

Merge the following that have LOC URL:

|  |  |
| --- | --- |
| LIST | IART ELSE ICMS ELSE IARL – ‘US, ‘ |
| Cavafy | pbcoreContributor (asset) ELSE pbcoreContributor (series) |
| DAVID | PERFORMER |
| Publisher | contributors |
| NewsBoss | Creators |
| Default | ‘http://id.loc.gov/authorities/names/n81047053’ |

## ICMS\* (Commissioned by)

Merge the following that have LOC URL:

|  |  |
| --- | --- |
| Filename | IARL |
| LIST | IART |
| Cavafy | pbcoreCreator, pbcorePublisher (asset) ELSE pbcoreCreator , pbcorePublisher (series) |
| DAVID | Description: COLL |
| Publisher | Creators, Producing Organizations ELSE Creators |
| NewsBoss | Writer |
| Default | IART ELSE ‘http://id.loc.gov/authorities/names/n81047053’ |

## ICMT (Comments)

Merge the following:

|  |  |
| --- | --- |
| LIST | ICMT |
| Cavafy | pbcoreAnnotation |
| DAVID | Notes |
| Default | [Various annotations] |

## ICOP (Copyright)

Correspondences in hierarchical order:

|  |  |
| --- | --- |
| LIST | ICOP |
| Cavafy | copyrightInfo |
| DAVID | Description: COLL |
| Publisher | Audio may download, audio may embed ELSE Web domain |
| Filename | COLL |
| NewsBoss | 'WNYC' |
| Default | COLL |

## ICRD (Create date)

Choose the earlier one of:

|  |  |
| --- | --- |
| Filename | YYYY-MM-DD |
| LIST | ICRD |
| Cavafy | pbcoreDate |
| DAVID | DATE |
| Publisher | newsdate |
| NewsBoss | date |
| Default | ‘1901-01-01’ |

## IENG\* (Original engineer)

Merge:

|  |  |
| --- | --- |
| LIST | IENG |
| Cavafy | pbcoreContributor[@contributorType=’\*engineer’] |
| DAVID | EDITOR |
| Publisher | Engineer |
| NewsBoss | Writer |
| Default | COLL + ‘ engineer’ |

## IGNR (Genre)

Correspondences in hierarchical order:

|  |  |
| --- | --- |
| LIST | IGNR |
| Cavafy | pbcoreGenre (ONE ONLY) |
| DAVID | SUBJECT |
| Publisher | type |
| NewsBoss | ‘News’ |

## IKEY (Keywords)

Merge the following that have LOC URL:

|  |  |
| --- | --- |
| LIST | IKEY |
| Cavafy | pbcoreSubject |
| DAVID | KEYWORDS (reverse LOC lookup) |
| Publisher | tags |

## IMED (Original medium)

Correspondences in hierarchical order:

|  |  |
| --- | --- |
| Filename | MEDIUM |
| LIST | IMED |
| Cavafy | pbcoreAnnotation[@annotationType=Provenance] |
| DAVID | SOURCE |
| Default | COLL + ‘ audio material’ |

## INAM (Title)

Correspondences in hierarchical order:

|  |  |
| --- | --- |
| LIST | INAM |
| Cavafy | pbcoreTitle[@titleType=Episode] |
| DAVID | GENERICTITLE |
| Publisher | title |
| NewsBoss | Filename |
| Filename | [Free text] |

## IPRD (Product)

Correspondences in hierarchical order:

|  |  |
| --- | --- |
| Filename | SERIES |
| Cavafy | pbcoreTitle[@titleType=Series] |
| LIST | IPRD |
| DAVID | PROGRAM |
| Publisher | Show |
| NewsBoss | 'News' |

## ISBJ (Description)

Correspondences in hierarchical order:

|  |  |
| --- | --- |
| Cavafy | pbcoreDescription[@descriptionType=Abstract] |
| LIST | ISBJ |
| Publisher | body |
| DAVID | Notes (partial) |
| NewsBoss | [body] |
| Default | INAM + ‘ on ‘ + IPRD + ‘ on ‘ + ICRD |

## ISFT (Software)

Correspondences in hierarchical order:

|  |  |
| --- | --- |
| LIST | ISFT |
| DAVID | Description (D.A.V.I.D GmBH…) |
| Default | ‘Unknown software’ |

## ISRC (Source)

Correspondences in hierarchical order:

|  |  |
| --- | --- |
| Cavafy | ‘cavafy.wnyc.org/asset/’ + pbcoreIdentifier[@identifierType=cavafyUUID] |
| LIST | ISRC |
| Publisher | URL |
| DAVID | CATALOG |

## ISRF (Source form)

Correspondences in hierarchical order:

|  |  |
| --- | --- |
| Filename | ‘US, ’ + COLL |
| LIST | ISRF |
| Cavafy | pbcoreAnnotation[@annotationType=Provenance] |
| Default | ‘Transfer from ‘ + IMED |

## ITCH\* (Transfer technician)

Merge:

|  |  |
| --- | --- |
| LIST | ITCH |
| Cavafy | pbcoreContributor[@contributorType=’\*engineer’] |
| DAVID | EDITOR ELSE AUTHOR |
| Publisher | Engineer |
| Default | ‘Unknown transfer technician’ |