Batch Ingest Package Format

This documentation is for Release 6.x. For the Release 1 version, see v.43. For the Release 2 version, see v.71. For Release 3.0.0, see v.86. For Release 3.1, see v.88. For Release 3.2, see v.108. For Release 3.3, see v.129. For Release 4.0, see v.159. For Release 5.x, see v.168.

Introduction

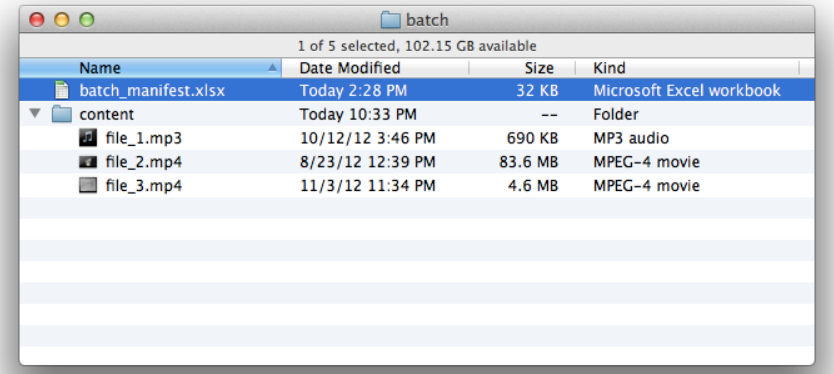
Avalon's *batch ingest* feature provides a method of building one or more media items at a time from uploaded content and metadata outside the user interface. A batch ingest is started by uploading an *ingest package* consisting of one *manifest file* and zero or more *content files* to the Avalon dropbox. For your convenience there is a demo ingest package available to download and import into test systems. Follow the instructions below to ensure a successful batch upload.

Ingest Packages

An *ingest package* is the combination of content and metadata that make up a single batch. Structural metadata documents in the form of XML may also be uploaded - one per a/v content file.

Package Layout

When a new collection is created, Avalon creates a subdirectory with the name of that collection (substituting underscores for any blanks), beneath the Avalon dropbox directory. The package (manifest file and associated content files) must be uploaded to that collection-named subdirectory or in a subdirectory beneath it. All items included in a single ingest package will be uploaded to the same collection. The following is a very simple package that has been uploaded:



Manifest File Format

The manifest file is a spreadsheet (`xls`, `xlsx`, `csv`, or `ods`) containing the metadata for the items to be created, as well as the names of the content files that make up each item. In this case, the manifest file is named `batch_manifest.xlsx`. See batch_manifest_template.xlsx for an Excel example file. Required fields are in bold. **Note: Neither the spreadsheet filename nor any folder/directory names above it can have blanks in them–substitute underscores.**

	A	B	C	D	E	F	G	H
1	Michael's First Test Batch	michael.klein@northwestern.edu						

2	Bibliographic ID	Title	Creator	Date Issued	File	Label	File	Label
3	123456	Test item 1	Klein, Michael B.	2012	content/file_1.mp3	Part 1	content/file_2.mp4	Part 2
4	789012	Test item 2	Northwestern	1951	content/file_3.mp4			

Row 1, Column A contains a reference name for the batch. This is mostly for your reference so we recommend naming the batch file according to what will help you remember the contents.

Row 1, Column B contains the submitter's email address (or username, depending on how your system is set up) to be used for notifications and exceptions. The submitter's email or user name must be listed as a manager, editor, or depositor for the collection in which this batch is deposited in the Avalon dropbox.

Row 2 specifies the names of the metadata fields supplied in the following rows. *Title*, *Date Issued,* and *File* are required. These fields are shown in bold in the Excel example file. Each subsequent row represents a single media item to be created. Metadata values are specified first, followed by a list of content files to be attached to each item. **Note: Make sure none of the field names in row 2 have leading or trailing blanks, or the field names will not be recognized by Avalon and will report an error.**

Content files listed in the manifest file must have the correct path noted for where those files are located in the Avalon dropbox, relative to the manifest file. Additionally, all content files must include a file extension. If necessary, include any directories or subdirectories (note the paths listed in columns E and G in the above example).

Multivalued fields are specified by multiple columns with the same header, e.g. *Topical Subject* in the following example:

	A	B	C	D	E	F
1	Michael's Second Test Batch	michael.klein@northwestern.edu				
2	Title	Creator	Date Issued	Topical Subject	Topical Subject	File
3	Nachos: A Memoir	Klein, Michael B.	2012-12-22	Meat	Cheese	content/tasty_tasty_nachos.mp4

Supported Field Names

*Required fields include "Title", "Date Issued" and "File" (bolded below and in .xls template)**.** Exception: If "Bibliographic ID" ingest is used, only "file name" is then required.*

- Bibliographic ID
  - MODS mapping: relatedItem@type="original"/identifier
  - Not repeatable
  - Editable after ingest in "Bibliographic ID" field of Resource Description form.
  - Used to identify the original MARC catalog record from which metadata was generated and import data from the catalog record into Avalon.
  - Inclusion of a bibliographic ID will cause any other descriptive metadata (including Required descriptive fields) entered into the spreadsheet besides the **Bibliographic ID Label** to be ignored. The Required file fields will, however, be read and used.
  - Information about how the corresponding MARC record is mapped to Avalon MODS is found in the MARC record ingest section below.
- Bibliographic ID Label
  - MODS mapping: relatedItem@type="original"/identifier@type
  - Not repeatable
  - Editable after ingest in "Bibliographic ID" field of Resource Description form.
  - Identifies the type of bibliographic ID supplied in the **Bibliographic ID** column. Valid types depend on system configuration and by default include "local", "oclc", lccn", "issue number", "matrix number", "music publisher","video recording identifier", and "other".
  - The value of "local" maps to "Catalog Key" in the Resource Description Form.
  - Batch will fail if value is not in the configured list of valid values.
  - Will be ignored if no Bibliographic ID value is present
- Other Identifier

        - MODS mapping: relatedItem@type="original"/identifier
        - Repeatable
        - Editable after ingest in "Other Identifier" field of Resource Description form.
        - Used to identify an external record that can connect the Avalon item to a catalog record or other record for the original item. This identifier differs from Bibliographic Identifier in that it is not used to retrieve a record from another system.
        - Must be paired with a value for Other Identifier Type
    - Other Identifier Type
        - MODS mapping: relatedItem@type="original"/identifier@type
        - Not Repeatable within Other Identifier
        - Editable after ingest in "Other Identifier Label" field of Resource Description form.
        - Identifies the type of external record identifier supplied in the Other Identifier column.
        - Valid types depend on system configuration and by default include "local", "oclc", "lccn", "issue number", "matrix number", "music publisher","video recording identifier", and "other"
        - Batch will fail if value is not in the configured list of valid values.
        - Will be ignored if no paired Other Identifier value is present.
    - **Title -** *required*
        - MODS mapping: titleInfo/title
        - Not repeatable
        - **Required descriptive field.** Title is used for display in search results and single item views. Only the first 32 characters of a title are included in search results listings. Recommended use is to reflect the content captured in digitized media files (such as the title of the piece performed or a short description of the content of a home movie).
        - Editable after ingest in "Title" field of Resource Description form.
        - If title is not available or missing, create a title that describes something about the content of the item. This is necessary for identifying items in search results.
    - Creator
        - MODS mapping: name@usage="primary"/namePart (role/roleTerm set to "Creator")
        - Repeatable
        - No ability to specify Corporate Body in batch at this time
        - Main contributors are the primary persons or bodies associated with the creation of the content. Main contributors will be included in search results display and aggregated for browsing access. At this time there is no ability to specify a main contributor as a corporate body. When possible, use the Library of Congress Name Authority File.
        - Editable after ingest in "Main contributor(s)" field of Resource Description form.
    - Contributor
        - MODS mapping: name/namePart (role/roleTerm set to "Contributor")
        - Repeatable
        - Contributors are persons or bodies associated with the item but not considered primary to the creation of its content. Examples of this would be performers in a band or opera, conductor, arranger, cinematographer, and choreographer. At this time this is no ability to specify a contributor as a corporate body. When possible, use the Library of Congress Name Authority File.
        - Editable after ingest in "Contributor(s)" field of Resource Description form.
    - Genre
        - MODS mapping: genre
        - Repeatable
        - Genre can be used to categorize an item by form, style, or subject matter. For consistency and to allow for sorting and aggregating, use terms from the Open Metadata Registry labels for PBCore: pbcoreGenre.
        - Editable after ingest in "Genre(s)" field of Resource Description form.
    - Publisher
        - MODS mapping: originInfo/publisher
        - Repeatable
        - Publisher of the content of the item.
        - Editable after ingest in "Publisher(s)" field of Resource Description form.
    - Date Created
        - MODS mapping: originInfo/dateCreated@encoding=”edtf”
        - Not repeatable
        - Creation date should only be used if Date Issued is a re-issue date. Then Creation date would contain the original publication date. Enter date information in a format consistent with the options shown in Extended Date/Time Format (EDTF) 1.0.
        - Editable after ingest in "Creation date" field of Resource Description form.
    - **Date Issued -** *required*
        - MODS mapping: originInfo/dateIssued@encoding=”edtf”
        - Not repeatable
        - **Required descriptive field.** Date should be the main publication date associated

with the item to be used for sorting browse and search results. Enter date information in a format consistent with the options shown in Extended Date/Time Format (EDTF) 1.0.
    - Editable after ingest in "Publication date" field of Resource Description form.
    - If date issued is not available or missing, enter a date that is narrowed down as much as possible (by range of years) or enter a date for century (18uu, 19uu, 20uu), in accordance with EDTF specifications.
- Abstract
    - MODS mapping: abstract
    - Not repeatable
    - Abstract provides a space for describing the contents of the item. Examples include liner notes, contents list, or an opera scene abstract. This field is not meant for cataloger's descriptions but for descriptions that accompany the item. The first 15-20 words are included in search result listings.
    - Editable after ingest in "Summary" field of Resource Description form.
- Language
    - MODS mapping: language/languageTerm
    - Repeatable
    - Language should describe the language of the content. Only terms or codes from the MARC Code List for Languages list may be used. Entering a language term not from the list will display an error when the page is saved.
    - Editable after ingest in "Language(s)" field of Resource Description form.
- Physical Description
    - MODS mapping: relatedItem@type="original"/physicalDescription/extent
    - Not repeatable
    - Physical Description provides a description of the original carrier for content that has been digitized from analog content.
    - Editable after ingest in "Physical Description" field of Resource Description form.
- Related Item URL
    - MODS mapping: relatedItem@displayLabel/location/url
    - Repeatable
    - Related Item URL provides a URL to related content, such as an adaptation or original version.
    - Editable after ingest in "Related Item(s)" field of Resource Description form.
    - Must be paired with a value for Related Item Label
- Related Item Label
    - MODS mapping: relatedItem@displayLabel
    - Not repeatable within Related Item
    - Related Item Label provides a descriptive label for the Related Item URL field.
    - Editable after ingest in "Related Item(s)" field of Resource Description form.
    - Must be paired with a value for Related Item URL
- Topical Subject
    - MODS mapping: subject/topic
    - Repeatable
    - Subject should be used for the topical subject of the content. For consistency and to allow for sorting and aggregating, use terms from the Library of Congress Subject Headings. For temporal subjects (time periods), use Temporal Subject and for geographic subjects (locations), use Geographic Subject. See below.
    - Editable after ingest in "Subject(s)" field of Resource Description form.
- Geographic Subject
    - MODS mapping: subject/geographic
    - Repeatable
    - Geographic Subject should be used for the location associated with the content. For consistency and to allow for sorting and aggregating, use terms from the Getty Thesaurus of Geographic Names.
    - Editable after ingest in "Location(s)" field of Resource Description form.
- Temporal Subject
    - MODS mapping: subject/temporal
    - Repeatable
    - Temporal Subject should be used for the time period of the content (for example, years or year ranges). Enter date information in a format consistent with the options shown in Extended Date/Time Format (EDTF) 1.0.
    - Editable after ingest in "Time period(s)" field of Resource Description form.
- Terms of Use
    - MODS mapping: accessCondition@type="use and reproduction"
    - Not repeatable
    - Terms of Use describes the conditions under which content may be used.
    - Editable after ingest in "Terms of Use" field of Resource Description form.
- Table of Contents
    - MODS mapping: tableOfContents
    - Repeatable
    - Editable after ingest in "Table of Contents" field of Resource Description form.

        - Used to provide the titles of separate works or parts of a resource. Information provided may also contain statements of responsibility or other sequential designations. Titles of separate works or parts should be separated by " – " (space-hyphen-hyphen-space).
    - Statement of Responsibility
        - MODS mapping: note@type="statement of responsibility"
        - Repeatable
        - Editable after ingest in "Statement of Responsibility" field of Resource Description form.
        - Used to provide information about primary persons or bodies associated with the creation of the content, along with details about their roles. This information can be transcribed from the credits listed in the resource itself or on its packaging.
        - Recommended use is to provide a separate Contributor field for each person or body listed in the Statement of Responsibility. Statement of Responsibility may be left empty if the use of Contributor fields alone is preferred.
        - Statement of Responsibility is displayed in the user interface appended to the Title field, following a " / ".
        - Also may be included as a Note/Note Type pair with Note Type='statement of responsibility'.
    - Note
        - MODS mapping: note
        - Repeatable
        - Editable after ingest in "Note" field of Resource Description form.
        - Used to describe aspects of the resource not accounted for in any of the other fields, such as creation or production credits, performers, venue/event date, historical or biographical information, language details, awards given to the performance or the work performed.
        - Recommended use is to provide a separate Contributor field for each person or body associated with the creation of the content and to use a Note to provide more information about such contributions or to provide information about secondary persons or bodies associated with the creation of the content.
        - Must be paired with a value for Note Type
    - Note Type
        - MODS mapping: note@type
        - Not repeatable
        - Editable after ingest in "Note Label" field of Resource Description form.
        - Identifies the type of note and is used as a label in the user interface.
        - Valid types depend on system configuration and by default include: general, awards, biographical/historical, creation/production credits, language, local, performers, statement of responsibility, venue
        - Must be paired with a value for Note

In addition to the descriptive fields, there are operational fields for the items(s) being ingested:

- Publish
    - Whether the item should be automatically published after ingest.
    - Default is "No".
    - To auto-publish, enter value of "Yes".
- Hidden
    - Whether the item will appear in search/browse results for end users. Use this field to prevent users from discovering items that would be confusing outside some externally-determined context (such as video figures for a research paper or audio clips contextualized in an Omeka exhibit). Hidden items can also provide "security by obscurity" when it is desirable to provide easy access but you don't want to publicize the availability of the items.
    - Default is "No".
    - To trigger hiding, enter value of "Yes".
    - Hidden items will still appear in search/browse results for those with ingest privileges.

There are also several fields that describe the media file(s) that are part of the ingested item. These fields must be repeated for each attached file:

- **File**
    - **Required file field.** Content files listed in the manifest file must have the correct path noted for where those files are located in the Avalon dropbox, relative to the manifest file. Additionally, all content files must include a file extension. If necessary, include any directories or subdirectories (note the paths listed in columns D and F in the above example).
    - Repeatable
    - Label, Offset, and Skip Transcoding can be listed in any order following the file they are describing. Absolute Location can only be used following Skip Transcoding if Skip Transcoding is included and its value is set to "yes".

- Label
    - Label is used for display in single item views. Recommended use is to reflect the content captured in digitized media files (such as the Part 1 and Part 2 of the piece performed or titles of songs).
    - Only repeatable following a file entry.
    - Editable after ingest in "Label" field of Manage Files page
- Offset
    - Offset is used to set the thumbnail and poster image for the display in search/browse results and single item views. Must be entered between 00:00:00.000 and length of file.
    - Excel will automatically format hh:mm:ss into time. To circumvent this, begin time offset with a single quote, for example: '0:10 for 00:00:10 and '1:06 for 00:01:06.
    - Only repeatable following an additional file.
    - Default is 2 seconds into playback.
    - Only applicable to video files. Audio files have a default thumbnail, offset will be ignored.
    - If a record contains multiple files, the first offset listed will set the thumbnail and poster image for the Avalon record.
    - Editable after ingest in "Poster Offset" field of Manage Files page or on the item preview page.
- Skip Transcoding
    - Skip Transcoding is used if a pre-encoded derivative of the file is what is being uploaded to Avalon instead of the master version of the file. This presumes that the derivative(s) match the requirements explained in Avalon Derivatives. Master file location information should be included for complete object ingest. See Absolute Location (below) for further information.
    - Only repeatable following a file entry.
    - Valid values: “yes” or “no”
    - See section below for skipping transcoding with multiple quality levels of derivative.
- Absolute Location
    - Absolute Location is used with Skip Transcoding to indicate the location of the master version of a video or audio file when the file uploaded to Avalon is a pre-encoded derivative.
    - Only repeatable following Skip Transcoding if Skip Transcoding is included and its value is set to “yes”.
    - If Skip Transcoding is set to “no” or not included, Absolute Location will be ignored.
    - Absolute Location should be the full URI path of the server housing the master version of the file.
- Date Ingested
    - This represents the date the item was ingested into Avalon Media System.
    - This date will not be visible within the user interface to normal users.
    - For system administrators and collection managers, a Limit By facet with these values will be available for search/browse.
    - If this column is not included, Date Ingested will automatically be set to the day on which the ingest process is completed by Avalon.
    - Include a valid date with format 2015-12-31 in this column to override the value being automatically set by the system.

Multiple File Ingest of Different Quality Files For a Single Avalon Item

Avalon supports ingest of multiple derivatives that may be selected with the High/Medium/Low gear-buttons of the video player during playback (or High/Medium for audio). The “File” field in the manifest and the naming convention of the files in the Avalon dropbox directory must be formatted correctly for the batch ingest to be successful. Avalon will know what filename to look for from the manifest file, find the quality levels specified in the dropbox directory, and ingest the formatted files accordingly. It is not required to have all three quality tiers for multiple file ingest.

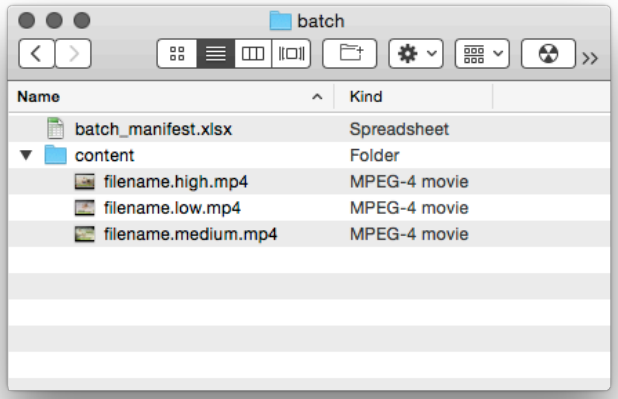
For a single Avalon item, input a filename in the “File” field and input “Yes” in the “Skip Transcoding” field of the manifest file. Add multiple files for this Avalon item to the dropbox directory. The “File” field as well as the file names of your different quality files in the Avalon dropbox directory must be formatted with the following convention:

File Name in Manifest File	filename.mp4
Files in Dropbox Directory	filename.high.mp4; filename.medium.m filename.low.mp4

Please note that files must match this convention strictly; extra periods are not allowed. **filename.test.high.mp4** is invalid; **filename.high.mp4** is valid.

Example manifest file for multiple file ingest of different quality files for a single avalon item:

	A	B	C	D	E
1	Michael's Third Test Batch	michael.klein@northwestern.edu			
2	Title	Creator	Date Issued	File	Skip Transcoding
3	Multiple Quality Ingest	Klein, Michael B.	2015	content/filename.mp4	Yes



Adding structure files via batch

The Batch Ingest Package can include XML structure files. One structure XML file can be attached per media file. See the demo ingest package at the top of this page for an example structural XML file included in a batch.

If the manifest lists a file named test.mp4, it will look for a structure file named test.mp4.structure.xml - you can edit the xml later via the user interface "Structure" tab in Avalon.

For more information about structure files (schema expectations and examples), see Adding Structure to Files Using the Graphical XML Editor.

Adding caption files via batch

The Batch Ingest Package can include WebVTT or WebSRT captions files. One captions file per media file. If the manifest lists a file named test.mp4, it will look for a captions file named test.mp4.vtt. If one is found, it will be attached to the media file as captions. This captions file can be updated or removed later via the user interface "Structure" tab in Avalon.

Batch Processing Notes

Each batch will generate 2 emails to the user listed at the top of the manifest.

Once Avalon detects the presence of an unprocessed manifest file, it will first verify that the necessary metadata columns are present in the manifest and that the file is not broken. At this stage, only the manifest file itself and **not** the metadata listed in the manifest has been validated.

If the manifest is incomplete or includes errors, such as invalid metadata values, only items that are valid (metadata which is valid, media file paths which are valid) will be created. An email will be sent to the email address specified in the manifest detailing the outcome, whether successful or not, listed in the manifest.

If a Bibliographic ID is provided for a resource but fails to process, the error email will only indicate that required fields are missing and will not indicate that the Bibliographic ID failed or was invalid.

To re-run a completed batch, follow the instructions in the email sent by the system after the batch is fully processed. It will contain a special filename that can be used to run the batch job again.

MARC record ingest

When a Bibliographic ID is provided for a resource the corresponding MARC record is mapped to a MODS record for use in Avalon. The MARC to MODS mapping is based on the Library of Congress mapping to MODS 3.5: http://www.loc.gov/standards/mods/mods-mapping.html

The Avalon mapping differs mainly:

- in placing elements describing the original physical resource inside the relatedItem element with attribute type=”original”;
- general, temporal, or geographical subdivisions of subject headings are split into separate elements for better faceting; and
- the typeOfResource element is determined by Avalon based on the media type uploaded.

Detailed mappings of MARC fields and subfields to MODS records for the Resource Description Form and the Batch Ingest Form can be found at the Metadata Crosswalks page.