Miranda Import Schema, Annotated

Miranda accepts imports in JSON format. All imports are tested against a schema. The system was built with the ability to support multiple schemas for different content types.

This document outlines Miranda's handling of import files, from the JSON schema that validates them, to the details of the expected fields, to guidance on preparing files matching for import. More information on the import process is available in documents specific to that process.

Table of Contents

Schema Support

All imports are tested against a JSON schema. The system was built with the ability to support multiple schemas for different content types.

Miranda exposes her schemas at this URL:
https://server.collections.folger.edu/schema/
That will produce a list of the available schemas for managing imports. At present, there is only the default schema, which can be viewed here:
https://server.collections.folger.edu/schema/default

While the prototype supported multiple schemas, we found that only the default schema was necessary. More schemas can be defined in order to require the appropriate fields for a particular record type, but this default schema illuminates the data fields Miranda is prepared to consider.

Definitions

remoteIDsets
Provides a repeatable format for remote identifiers: it pairs a remote system and an ID code that is unique within that remote system. Wherever you see remoteUniqueID in this schema, you can expect it to use this definition.

> **remoteSystem**
> This is the name of the system that holds (or held) the record.
>
> **remoteID**
> This is the identifier for the item in the remote system given above.
>
> Any URI can work with this system.
> http://id.loc.gov/authorities/genreForms/gf2017026136.html would become
> remoteSystem: http://id.loc.gov
> remoteID: /authorities/genreForms/gf2017026136.html
> https://www.youtube.com/watch?v=gPlpphT7n9s would become
> remoteSystem: https://www.youtube.com
> remoteID: /watch?v=gPlpphT7n9s

Fields

remoteUniqueID
At the record level, used on initial import to ensure uniqueness. Also used as fallback linkage between items when dapID not available. ***This field is required.***

searchHandling
Accepts "exclude" or "include" and is used to determine if a record should be searchable. This field is used at import time and is not exposed via GraphQL.

lastUpdate
Used to check if an import should be processed. In the case that this datetime is older than the one on record, the record will not be imported.

identifiers
A collection of other identifiers in key-value pairs. Presumably useful for humans looking to find works in other systems.

> **key**
> **value**

title

In the initial version, we kept a single string for the title. Now, we're trying to capture more information about variations, including linking to an authority reference when available. As a result, we have sub-fields of title now:

displayTitle
The common short version of the title. This is preferred for display.

extendedTitle
A longer title, perhaps a transcribed long title, where appropriate.

alternateTitles
An array of key-value pairs: the alternate titles, and then labels to describe them. These include alternate official titles of the work, nicknames, alternate spellings for the title.

titleLabel
titleText

uniformTitle
This is used to track the work authoritatively. It contains a titleString that is the uniform title, and a titleURI, which is a link to the authority reference for that title.

titleString
titleURI

creator
This is a human-readable description of any creators of the work in question. For authority-URI-linked creator information, look to the relationships->agents field, and filter for agents who are tagged as creators.

dateCreated
This is used to handle the search and display of dates. We have two subfields: displayDate, for human consumption, and isoDate, which should use the ISO 8601 date format to facilitate search and computation.

displayDate
isoDate

caption
The caption is a one or two sentence statement that concisely describes the resource for search results (a little like the fragment in Google searches).

notes

A list of paired labels and notes on the item. The imagined use is to enable the logical separation of notes about various copies of a work as well as special notes about an aspect of the work, e.g. "Caveats about the Transcription".

label
note

extent
A string to describe the quantity of items in the record. For example: "72 letters" or "3 photographs"

folgerDisplayIdentifier *(previously folgerCallNumber)*
text to find the item in the Folger library, if appropriate. This is kept separately from the identifiers array to make it explicit and easy to work with.

folgerRelatedItems
This is a list of related items, each offering a little information over its base record. dapID or remoteUniqueID can be used to call out another record. This is a prime way of tying records together. Each item in the list can have the following fields

dapID
remoteUniqueID
Either of these can be used to link to another record in Miranda.

folgerRelationshipType
Provides information about the type of relationship between the record and this related item.

folgerObjectType
Tells the system what to expect for display. Current values include: 'image', 'oembed', 'mp4', 'soundcloud'. *Note: This field has transitioned from primary importance to vestigial -- the related record should specify this in fileInfo->encodingFormat.*

label
for display with the item

mpso
Stands for "Multi Page Sort Order" and used to order the display of items

format
This powers the format (aka media type) search as seen on the front page. ***This field is required.***

mirandaGenre *(previously folgerGenre)*
This powers the genre search as seen on the front page. It uses Folger-specific terms for genres.

genre
This links records to genres as defined by authority records. It has two subfields: name, for display and human consumption, and uri, for authoritative matching. Several different authority references might be used, but URIs should help avoid collisions. Reference:
http://id.loc.gov/search/?q=cs:http://id.loc.gov/authorities/genreForms

name
uri

groupings
This field, taken at face value, functions as a list of tags that can be used to cluster records together. All the letters in the Bagot Family Papers could be placed into one group. Future expanded use includes building a separate index of these groupings and any relationships between them. (Bagot Letters are part of Bagot Family Papers are part of … , etc)

holdingInstitution
This field will be used to track the institution that holds the work and contributed it to Miranda.

name
The name of the institution, suitable for display. For example, "Guilford College" or "Blockbuster Video".

contactPerson
The person at the institution to be contacted about the record.

exhibitionCode
If this record is part of an exhibition at the partner Institution, keep that information here.

notes
Any notes on the record as it relates to the institution.

language
This powers the language search feature. For each language, use its ISO 639-2 or ISO 639-3 code, which is a three-letter code.
(https://iso639-3.sil.org/code_tables/download_tables)

license

Who owns copyright, what license is available to use work. This will be very important to enable sharing.

locationCreated
From whence the item in question comes. We also have location information in the relationships section, but having this definitively pulled out for attention helps manage search and display, as well as nudging data providers to consider this an important field. We use three subfields derived from schema.org, and one special one.

> **locationDescriptor**
> This is the vernacular name of the particular place or building. "Buckingham Palace" or "201 E Capitol St SE".
>
> **addressLocality**
> This is the vernacular name of the particular locality, usually a town, village, or city. It might read "Verona" or "Greensboro", or, "Washington".
>
> **addressRegion**
> This is the vernacular name of the particular place. It might read "Suffolk", "North Carolina", or "District of Columbia".
>
> **addressCountry**
> This is the vernacular name of the country housing the particular place. It might read "Italy" or "U.S.A"

locus
Page number (for an image); range of pages (for an article); URL (for a web resource). Where the sortOrder field is an integer for machine consumption, this locus field is for human consumption.

preferredCitation
A field to show the Folger Shakespeare Library's recommended citation format for the work or record in question.

abstract
Concise summary of contents

sortOrder
Intended to support fallbacks for sorting sibling records. Future-facing version of MPSO from folgerRelatedItems.

size
A human-readable string describing the physical size of the item, if appropriate.

subjects
Provides authority references and human-readable strings for the subject of the item. Used to link to authoritative concepts. Examples of existing data include Health, Vegetarian Cookies, and "Theatre Royal, Drury Lane (London, England)". The URI field is intended to link to an authority record. This could be a source of entity relationship information.

uri
description

fileInfo
We're targeting one binary asset per record, and keep all the various fields for binary files as subfields here. *Note for future work: There is some question about associating multiple binary files per record in future development.*

fileURL
The URL from which to import the binary asset (if any) associated with this record.
This field is required for file imports to be properly processed.

encodingFormat
Used to show file type and hint at the necessary player. eg RAM, MP4, JPG, WAV, etc.
This field is required for file imports to be properly processed.
On import, Miranda will use this field to manage ingested files. Ensure that your values here match the appropriate type for the file:

- "audio" for mp3
- "video" for mp4
- "image" for jpg
- "JPEG2000" for jp2

Otherwise, the file's extension will match this encodingFormat value
So, to handle TIF or TIFF files, for example, use the encodingFormat "TIFF".
For links to embeddable resources, please use "oembed", which will let the system know how to proceed.

contentSize
Used for a human-readable size of binary file.

fieldList
For CSV/database style files, a list of strings to provide the headers and column names. We didn't use this in the prototype, but anticipated its usage.*/

numberOfRows
For CSV/database style files

duration
Human readable time used for A/V content to show its duration

height
Human readable linear distance used for A/V content to show its height, eg 800px

width
Human readable linear distance used for A/V content to show its width, eg 640px

simplifiedTranscription
A field used to include a simplified version of the transcription in the record itself. Where a TTML (Timed Text) transcription would have time codes for each line, this simplifiedTranscription would only include the textual content. To link to transcription files, use folgerRelatedItems.

relationships
This is the holder for the various types of relationships we'd like to track. By clustering them here, we can add to this list without cluttering up the main space in the schema.

parents
A list of dapID or remoteUniquelDs used to indicate a parent record. The goal of this field is to enable e.g. pages of a book to inherit metadata from the book's record. We are allowing multiple parents, which may cause some issues with data inheritance, but which would also allow an image used in multiple sources to be linked to them all.

agents
These refer to people or institutions involved with the record in some way. examples include: "A & B Booksellers", "Adelphi Theatre (Edinburgh, Scotland)", and "Clothiers (LEEDS)". At this time, unique identifiers are not present in our import data.

agentURI
An authority URI ensures the system (and those who use it) are referencing the intended individual. http://id.loc.gov/authorities/names/n2009073240.html, for example, refers to a particular Warren Master.

agentName
String representation of the agent's name for display.

relationship
This describes the relationship, e.g. Creator, recipient, signatory, nemesis, etc.

works
Works aren't actual works in the sense of copies or editions of plays, but are the abstract concepts relating to the various copies and editions. e.g. "Hamlet" is the abstract work from which various plays, productions, and even Mel Gibson's film is derived.

workTitle
Human readable reference to title. This will not need to be the perfect official title, but rather one for ease of consumption. The workURI field is used to make connections.

workInstance
Human readable reference to instance of the work. e.g. Quarto I. This is required as the workURI is not specific enough. Thus, workInstance and workURI will function as a pair to provide specific connections.

workURI
A link to an authority reference for the particular work. Note that this is not as specific as an instance or edition of a work. This might look like http://id.loc.gov/authorities/names/no97058371.html, which is the reference for "Série Estudos alagoanos. Historiografia".

relationship
This describes the relationship to the work in question

locations
Places related to the item in question. The DAP Data Model document had suggested the use of an authority URI for locations, but I was unable to find a reference for this, so I've followed the approach we used with locationCreated.

locationDescriptor
This is the vernacular name of the particular place or building. "Buckingham Palace" or "201 E Capitol St SE".

addressLocality
This is the vernacular name of the particular locality, usually a town, village, or city. It might read "Verona" or "Greensoro", or, "Washington".

addressRegion
This is the vernacular name of the particular place. It might read "Suffolk", "North Carolina", or "District of Columbia".

addressCountry
This is the vernacular name of the country housing the particular place. It might read "Italy" or "U.S.A"

locationURI
This is used for authority references to locations. It might look like: https://data.cerl.org/thesaurus/cnl00016230

relationship
This describes the relationship to the work in question. It might read "set in" or "authored in" or "starts in" or "Ophelia kills herself here".

Future Items

isUpdate *(not yet enforced)*
We have discussed adding a flag to instruct the import system to overwrite only the provided fields, rather than replacing the whole document.

permissions *(not yet enforced)*
Not part of the record metadata itself, but for administrative purposes. This set of fields will be used to instruct the system which users or groups can edit a record, and when it should be available to the public. See our permissions documentation for more details.

readPermitted
writePermitted
startTime
endTime

Concept Database, Groupings Database
Referenced in the concepts and groupings fields, these would be additional indices used to track the relations between concepts or groupings, thus elevating them from flat tags to a graph of related ideas.

Import File Preparation

Miranda assumes that the files provided for import are valid JSON. She'll try to give feedback about errors, but sometimes malformed JSON files will just flummox her.

Each file should be an array of one or more JSON objects matching our schema for imports.

To ensure that your files are valid JSON, use a tool like JSON Lint, which can be installed locally (https://github.com/zaach/jsonlint) or used online (https://jsonlint.com/)

For dealing with larger import files, the jq utility can also be quite handy. (https://stedolan.github.io/jq/)

Sample Records

We have a sample JSON file available here:
https://drive.google.com/file/d/1JWfdEx-lF4psWCtYI2U0n3-DOhacpY0A/view?usp=sharing

Before using it, please search for the phrase REPLACE_THIS_URL and update the fileURL values it is a part of to point to valid image resources.

Here are the sample images referred to:

- https://drive.google.com/open?id=1L2gofWmYqzvsQfR6ElnZpzfuP51ac04G
- https://drive.google.com/open?id=15jVj3N02gD5ISrJD4B--W1G8V9_AN040
- https://drive.google.com/open?id=1875ViGUxTG4VXKKQvRb8fzbbQEB5CNW-

Upload them to a publicly-accessible place on the web and update the fileURL references to point to them.