



Digitization Metadata



Appropriate use of metadata is a critical component of a successful digitization project. Metadata provides the structure to identify key information about a collection of records (organization, department, collection name etc.) as well as the individual records within the collection (title, date created, file format, access rules, etc.) Metadata is used for the identification, management, access, use and preservation of individual digital records and provides context for the collection as a whole. Digitization projects that combine the collection of sufficient metadata in conjunction with the use of OCR technologies to enable the ability to search within a document stand the greatest chance of success for management over time.

While no single metadata set will apply to all records series since record sets vary, it is recommended that you select a basic list of metadata fields that will be collected regardless of the records being digitized. Collection level metadata is applied to the overall record series being digitized. This is the “big picture” information that describes what the records are, who manages them, what happens to them at the end of their records lifecycle, and how they are arranged. This data tends to be fairly straightforward and well defined.

File level metadata is more challenging and perhaps more critical as it will help you locate a specific file among potentially thousands within a records series. It is important to evaluate the records prior to scanning and determine what data is needed to separate one item from another and how to find it when searched. A basic set of text documents around a project might be fine with minimal metadata (title, date, author, doctype). A scanned photo collection might require additional metadata such as photo location, event, and subjects. A map collection may require that metadata include a map ID, county code, street address, or GPS coordinates for items to be located in the future.

Part of your evaluation should also review the record series to see if the same data can be found on all records over time. For records that span decades, things like the format of document IDs may have changed over time, or fields in forms where metadata can be extracted from may have been added, deleted, changed or not filled in at all. You need to determine how you will reconcile these anomalies and collect sufficient data to find the digital object when searched. It may make sense to use your QA process to review and manually enter fields where only a minimal amount of data is missing or to clean up misspellings which would hamper future search efforts.

A note about Access Restrictions

Each list below has a place for documenting access restrictions due to legal reasons (PII or Confidentiality) at both the collection level and file level. Documenting access restrictions at the collection level is the least desirable in a records series where restricted and non-restricted documents live side-by-side because it often has the effect of restricting the entire collection or requiring an intermediary each time a document is requested to ensure no restricted items are released. If done correctly, the co-location and identification via metadata of restricted digital objects during a digitization project can actually make this task easier through the creation of security groups on a network drive or in a content management system that allow select individuals to access the items. Depending on the records, it may be most effective to gather restricted documents into a subfolder and apply the restriction to the subfolder which would then propagate down to the items within the folder rather than on individual documents scattered throughout a collection of records. An evaluation and determination of how restricted items will be managed should be part of your project plan prior to starting the digitization process.

The below sections list recommended metadata fields for digital collections and individual records. Your organization should identify and document key items that will be required for every project. This will help ensure consistency across projects and enable them to be searchable and accessible well into the future. Other fields are going to be “nice to have”; items that would be beneficial but not mandatory for the long-term management of the item. These are not meant to be comprehensive lists of all possible metadata fields. If there are other fields that make sense for your unit or your specific records, please use them.

Collection Level Metadata

The following metadata fields can be applied to the record series as a whole. Items in **BOLD** text are fields that, if populated, will provide the greatest opportunity for successful long-term management of the records.

Metadata	Description
Division/Department	Division/Department name that manages the records
Project Title	Common name for the group of records in the digitization project
Event	Event that initiates the start of the retention time period
Retention	Retention time period for the records
Disposition	Specify: Destroy, Destroy Confidential, Transfer to WHS
Access Restrictions (Y/N)	Yes or No – whether there are access restrictions on these records (PII, Confidential)
Description	Description summarizing the specific records being digitized so that others can understand the materials in context.
Arrangement	Description of the materials' organizational scheme. (ex: docket number, alphabetically by name)
Keywords	Name of a person, organizational unit, place, or topic that provides information about the records being digitized or which might be searched
Location	Names of places that are relevant to the collection which can be searched (community, county, state)

File Metadata

The following metadata fields are applied to individual records. Items in **BOLD** text are recommended fields. Given the variety of records, it is highly likely additional metadata fields **not** included on this list will be identified through the evaluation process to support the long term management of your digitized items.

Metadata	Description
Title	File name of the record
Author	Author of record (if known)
Date Created	Creation date of original item
Date Digitized	Date when record was digitized
Disposition Date	Date when WHS should be notified. “Permanent” if it stays with the unit
Format	Format of digitized record (.tif, PDF...)
File Size	File size of digitized object
Pages	Number of pages (helpful in multi-page documents)
Description	Abstract or summary of items contents
Location	Names of places that are relevant to the record which can be searched (community, county, state)
Access Restrictions (Y/N)	Yes or No – whether there are access restrictions on this record (PII, Confidential)
Keyword	Terms that might be searched, but may not appear in the body of the text or other metadata fields.
Unique ID #	Unique ID that was assigned to the paper document at the time of creation. (This may or may not exist depending on the process the paper document was created under)
Document Type	This is assigned during the digitization process and will help in locating the object while searching (ex: report, map, image, presentation)