



University of Colorado Digital Library

Metadata Best Practices

Version 1.0

Contents:

- I. [ABOUT THIS DOCUMENT](#)
 - II. [INTRODUCTION](#)
 - III. [METADATA: AN OVERVIEW](#)
 - IV. [MINIMUM REQUIREMENTS FOR DESCRIPTIVE & ADMINISTRATIVE METADATA AT THE UNIVERSITY OF COLORADO](#)
 - V. [TECHNICAL METADATA](#)
 - VI. [MINIMUM REQUIREMENTS FOR TECHNICAL METADATA AT THE UNIVERSITY OF COLORADO](#)
 - VII. [CONTROLLED VOCABULARIES & METADATA CROSSWALKS](#)
 - VIII. [OPEN ARCHIVES INITIATIVE \(OAI\)](#)
 - IX. [CU DIGITAL LIBRARY'S METADATA BEST PRACTICES WORKING GROUP](#)
-

I. ABOUT THIS DOCUMENT

Collections in the University of Colorado Digital Library must follow appropriate metadata standards. This document offers an introduction to metadata, provides links to resources containing more information, and describes the required and recommended elements for collections in the CU Digital Library. The document is maintained by the University of Colorado's Metadata Best Practices Working Group: Christopher Cronin (University Libraries' Cataloging & Metadata Services Department), Katie Lage (University Libraries' Map Library), Lynn Lickteig (College of Architecture & Planning), and Elaine Paul (Department of Art and Art History).

II. INTRODUCTION

Metadata facilitates the discovery and use of information, and is an essential component of planning and implementing a large digital collection. The consistent provision of metadata is crucial for the organization, identification, selection, access, retrieval, reuse, preservation, and management of physical and digital assets. It also ensures a greater likelihood of accurate, meaningful and consistent results when performing cross-collection searches.

Metadata are often divided into broad, conceptual groups, such as descriptive, administrative, and structural metadata:

- *Descriptive metadata* represents the intellectual content and the information about the creation of that content, such as subject, author, and date.
- *Administrative metadata* records information used for the long-term management and distribution of the physical and digital assets. Administrative metadata can include data about technical qualities and processes, rights, preservation, use, and provenance.
- *Structural metadata* describes the configuration of complex objects comprising multiple parts or sections.

Metadata standards provide structure and rules for the consistent provision of data. These standards vary according to disciplines and the types of resources being used. They include guidelines for data structure, data content, markup languages, and data values. A data value is the choice of words and/or numbers that populate fields in a database. Authorities such as controlled vocabularies and thesauri provide the terms for consistent data values in a database. Data structures, also called metadata element sets, provide a set of categories that create the basis for a fielded structure in a database. And data content standards guide the formatting and syntax of the data values. Examples of these structures and standards include:

Descriptive metadata structure standards

- [Metadata Object Description Schema \(MODS\)](#)
- [MARC Bibliographic](#)
- [MARCXML](#)
- [Dublin Core \(DC\)](#)
- [Collaborative Digitization Program's Dublin Core Metadata Best Practices, Version 2.1.1](#)
- [Visual Resources Association \(VRA\) Core 4.0](#)
- [Categories for the Description of Works of Art \(CDWA\) Lite](#)
- [Publisher's Broadcasting Metadata Dictionary \(PBCore\), Version 1.1](#)

Descriptive metadata content standards

- [Anglo-American Cataloguing Rules, 2nd edition \(AACR2\)](#)
- [Describing Archives: A Content Standard \(DACS\)](#)
- [Cataloging Cultural Objects \(CCO\)](#)

Authority data standards

- [MARC Authority](#)
- [Metadata Authority Description Schema \(MADS\)](#)
- [Encoded Archival Context \(EAC\)](#)

Technical metadata standards

- [Technical Metadata for Digital Still Images in XML \(MIX\)](#)
- [Library of Congress AV Prototyping Project Video Metadata Schema](#)

Markup languages

- [Encoded Archival Description \(EAD\)](#)
- [Text Encoding Initiative \(TEI\)](#)
- [MusicXML](#)

Other metadata standards

- [Metadata Encoding and Transmission Standard \(METS\)](#)
- [PREservation Metadata: Implementation Strategies \(PREMIS\)](#)

Selected Resources

Gill, Tony, Anne J. Gilliland, Mary S. Woodley. *Introduction to Metadata: Pathways to Digital Information*. Online Edition, Version 2.1. Ed. Murtha Baca.

http://www.getty.edu/research/conducting_research/standards/intrometadata/index.html

National Information Standards Organization. *Understanding Metadata*. Online Edition. Bethesda, MD: NISO Press, 2004.

<http://www.niso.org/standards/resources/UnderstandingMetadata.pdf>

III. METADATA: AN OVERVIEW

The [California Digital Library's](#) (CDL) [Glossary](#) defines metadata as “structured information about an object, a collection of objects, or a constituent part of an object such as an individual content file. Digital objects that do not have sufficient metadata or become irrevocably separated from their metadata are at greater risk of being lost or destroyed. Ephemeral, highly transient digital objects will often not require more than descriptive metadata. However, digital objects that are intended to endure for long periods of time require metadata that will support long-term preservation” – [CDL Glossary](#).

Metadata are usually categorized according to three types: descriptive, administrative, and structural. Structural metadata are particularly important for complex digital objects. Complex digital objects are those that include two or more content files and their associated metadata. The content files function as parts of a whole and each file usually requires its own specific metadata (e.g., pages of a book; images of a single object from multiple angles; video, audio and transcript of an oral history interview).

Descriptive Metadata

“Metadata used for the discovery and interpretation of the digital object. Descriptive metadata may be referred to externally or indirectly by pointing from the digital wrapper to a metadata object, a MARC record, or an EAD instance located elsewhere. Or, descriptive metadata may be embedded in the appropriate section of the digital wrapper” – [CDL Glossary](#).

Administrative Metadata

“Used for managing the digital object and providing more information about its creation and constraints governing its use” – [CDL Glossary](#). Administrative metadata includes technical metadata, rights management metadata, source metadata, and digital provenance/preservation metadata.

Technical Metadata

“Administrative metadata that describes the technical attributes of the digital file” – [CDL Glossary](#).

Rights Management Metadata

“Administrative metadata that indicates the copyrights, user restrictions, and license agreements that might constrain the end-use of the content files” – [CDL Glossary](#).

Source Metadata

“Administrative metadata for describing the source from which the digital content files were produced. Sometimes this will be the original material; other times it will be an intermediary such as a photographic slide, or another digital content file” – [CDL Glossary](#).

Digital Provenance / Preservation Metadata

“Administrative metadata that is the history of migrations, transformations, or translations performed on a digital library object's content files from their original digital capture or encoding. It should contain information regarding the ultimate origin of the content files” – [CDL Glossary](#).

Structural Metadata

Structural metadata are “used to indicate the logical or physical relationship of the content files comprising the complex digital object, e.g., the sequence of pages for a group of images of a diary or of detailed images of a larger image. The structural metadata specifies a coherent presentation of the digital content and its pertinent associated metadata” – [CDL Glossary](#).

IV. MINIMUM REQUIREMENTS FOR DESCRIPTIVE & ADMINISTRATIVE METADATA AT THE UNIVERSITY OF COLORADO

The University of Colorado’s Metadata Best Practices Working Group strongly encourages that all metadata records present in the CU Digital Library include the following six classes of descriptive and administrative metadata elements: Title, Rights Management, Holding Institution, Collection Name, SubCollection Name, and Date of Original. Metadata in these classes of elements should be recorded in the appropriate field, according to the metadata schema being used.

1. **Title (*Required*)**. This element is important not just for resource description, but also as a way for some systems and users to sort records in results displays. If an object does not have a title already assigned to it by its creator, supply one according to established standards. Multiple titles should be expressed in repeated fields (e.g., Alternative Title).
2. **Rights Management (*Required*)**. The intent of this class of element is to provide users with information on restrictions on usage of the resource. It is important to include a rights statement even when the resource is in the public domain and free of restrictions. Language in this element should be clear and explicit. Best practice is to include information on whom to contact at the holding institution, especially in cases where users must request permissions for use. Example rights statements are currently being developed by the University of Colorado's Rights Management Working Group.
3. **Holding Institution (*Required*)**. Record the name of the department, school, college, campus, or faculty member that holds the digital resource. Examples include: (University of Colorado at Boulder Map Library).
4. **Collection Name (*Required*)**. Every record in the digital library is automatically assigned a Collection Name by the Insight software.
5. **SubCollection Name (*Mandatory if applicable*)**. Add SubCollection names when an organizational hierarchy exists within a single collection.
6. **Date of Original (*Mandatory if applicable*)**. The date that a resource was originally produced is important metadata for resource discovery, and can be essential when trying to distinguish between similar resources. Enter a date for the original even if it is born digital (i.e., not a digitized reproduction of an analog resource). When a date is either approximate or not known at all, follow the guidelines of your descriptive metadata content standard for recording that information.

Best practice is to record all dates in a standard, machine-readable format, such as ISO8601. Recording dates in a standard machine-readable format allows for sorting of search results chronologically, as well as searching by date across multiple collections (provided those collections have also used the standard format). The ISO standard calls for dates to be recorded in YYYY-MM-DD format.

The ISO standard also allows for recording of dates down to the second: YYYY-MM-DDThh:mm:ss. The character [T] shall be used as time designator to indicate the start of the representation of the time of day component in these expressions. The hyphen [-] and the colon [:] shall be used as separators within the date and time of day expressions, respectively (ISO 8601:2004(E), p. 19).

V. TECHNICAL METADATA

The primary purpose of technical metadata is to record the technical attributes of digital objects and their production or creation, not the analog source of the digital object (if there is one). Technical metadata also includes information on the digital

capture process, such as the hardware and software used to acquire the digital object. Capturing technical metadata may help to ensure uniformity for large digital projects by enforcing standards for the creation or acquisition of digital files.

For still image, audio, and multi-media digital files, it is important to include information in the technical metadata on whether a “codec” program has been used for the compression or decompression of the digital files, and if used, to list the type and version. Codecs are used to encode and decode (or compress and decompress) various types of digital files--especially large files that would normally use up inordinate amounts of disk space such as movies and large still image files.

Common codecs include those for converting analog video signals into compressed video files (such as MPEG) or analog sound signals into digitized sound (such as RealAudio). Codecs can be used with streaming (live video or audio) or files-based (AVI, WAV) content, as well as for data compression of still images (LZW).

While dozens of file formats exist for the various types of digital objects, what follows are some of the more common file formats and the technical characteristics for still, audio, and multimedia digital files.

Digital Still Images

For digital still images, the technical metadata for the image quality and capture process may include information about the:

- file format [may feature either lossless (i.e., no) compression or lossy (i.e., with) compression]
- file resolution (pixels per inch, or ppi)
- dimensions (image dimension or size in inches or centimeters)
- bit-depth (e.g., 8-bit, 16-bit, 24-bit, etc.)
- color mode (e.g., RGB, CMYK, or grayscale)
- scanner or digital camera brand, name, and model number
- software used to manipulate or compress the image, including the software name and version (e.g., Adobe Photoshop CS2 version 8.0, LZW, etc.)

For digital still images, examples of file formats include:

- .BMP (Bit-Mapped; can be 1, 4, 8, or 24-bit storage; Windows image format)
- .GIF (Graphical Interchange Format; 8-bit storage; lossless)
- .JPEG or .JPG (Joint Photographic Experts Group; 24 bit storage; lossy)
- JPEG 2000 or .jp2 (a high-encoded wavelet based format that does not save the raster data of the original file or the actual bits that represent the pixels, instead it recreates the data representing the pixels)
- .PICT (default format used on Macintosh computers)
- .PNG (Portable Network Graphic; lossless)
- .PSD (PhotoShop Document)
- .QXD (QuarkXpress Document)
- .SVG (Scalable Vector Graphics; for XML files)
- .TIF or .TIFF (Tagged Image File Format; 24-bit storage; lossless)

Digital Audio Files

For digital audio files (sound only) the technical metadata for the image quality and the image capture process may include the:

- file format, including whether the format is lossless or lossy
- sample rate
- resolution
- number of channels
- software used to manipulate or compress the audio file, including the software name and version
- brand, name, and model number of the recording equipment

For digital audio files (sound only), examples of file formats include:

- Advanced Audio Coding File (.aac)
- Audio Codec 3 File (.ac3; standard in DVD audio)
- MIDI (.mid or .midi; Musical Instrument Digital Interface)
- MP3 (MPEG-1 layer 3)
- RealAudio (.ra; for streaming audio) or Real Audio Media (.ram)
- Waveform or Windows Wave Sound File (.wav)
- Windows Media Audio (.wma)
- Audio Interchange File Format; (.aif, widely used on the Mac OS platform and equivalent to Waveform)

Digital Multi-Media Files

For digital multi-media files (including video, sound, and animations) the technical metadata for the image quality and capture process may include the:

- file format
- bitrate
- software used to manipulate or compress the multimedia file, including the software name and version
- brand, name, and model number of the equipment used

For digital multi-media files, examples of file formats include:

- Advanced Systems Format (.ASF)
- Apple QuickTime Movie (.mov or .qt)
- Audio Video Interleave (.avi)
- DV and Mini-DV (.dv)
- Microsoft ASF Redirector File (.asx)
- Moving Pictures Expert Group or .mpg (has multiple standards: MPEG-1 video file or .m1v; MPEG-2 video or .m2v; MPEG-4 video file or .m4e; MPEG-4 video file or .mp4; iTunes video file or .m4v)
- Macromedia Flash Movie (.swf)
- RealMedia (.rm)

- RealVideo (.rv)
- Shockwave (.dcr)
- Windows Media Video (.wmv)

Selected Resources for Technical Metadata

Source for File Format Extension Information:

<http://fileinfo.net>

Additional file formats are listed at:

<http://www.ace.net.nz/tech/TechFileFormat.html>

U.S. National Archives and Administration. Technical Guidelines for Digitizing Archival Materials for Electronic Access: Creation of Production Master Files – Raster Images. (June 2004). Online Edition

<http://www.archives.gov/research/arc/digitizing-archival-materials.html>

National Information Standards Organization and AIIM International. NISO Data Dictionary – Technical Metadata for Digital Still Images

http://www.niso.org/standards/resources/Z39_87_trial_use.pdf

California Digital Library

<http://www.cdlib.org/inside/diglib/guidelines/bpgimages/>

Collaborative Digitization Program (select section on digital imaging)

<http://www.cdpheritage.org>

VI. MINIMUM REQUIREMENTS FOR TECHNICAL METADATA AT THE UNIVERSITY OF COLORADO

The University of Colorado's Metadata Best Practices Working Group requires that all metadata records present in the Digital Library minimally include the following two classes of technical metadata elements: file format and file size. Metadata in these classes of elements should be recorded in the appropriate field, according to the metadata schema being used.

1. **File Format.** Recommended best practice is to select a value from a controlled vocabulary, such as the list of [Internet Media Types \[MIME\]](#) that defines computer media formats (also known as MIME types). For example: image/jpeg; audio/mp3.
2. **File Size.** For most formats, the recommended best practice is to record file size as bytes (e.g., 3,000,000 bytes) and not as kilobytes (KB), megabytes (MB), etc., because it is the most specific measurement of file size. It is also best practice to include duration time for multi-media. For example: 4,200,000 bytes; 5 minutes, 34 seconds. Follow the technical metadata requirements best suited for the format being described.

Note: U.S. National Archives and Records Administration states that for raster image files it is also "always necessary to provide both the resolution and the image dimensions;

ex. 300 ppi at 8" x 10" or even 300 ppi at original size" (U.S. National Archives and Records Administration (NARA) *Technical Guidelines for Digitizing Archival Materials for Electronic Access: Creation of Production Master Files – Raster Images* June 2004 page 21).

VII. CONTROLLED VOCABULARIES & METADATA CROSSWALKS

Best practice is to use a controlled vocabulary to provide subject access to your collection. A controlled vocabulary is a set of authorized subject terms used to describe items in a database. The Library of Congress Subject Headings, MeSH (medical subject headings), and the Art & Architecture Thesaurus (AAT) are examples of controlled vocabularies. A local controlled vocabulary can be used also.

The Getty Research Institute has a useful explanation of why controlled vocabularies are important:

"Why do we need vocabularies? It is because language is ever-changing, nuanced, and complex. These very characteristics that make language so wonderfully expressive can cause ambiguity and confusion in documentation, and ultimately, hamper access to materials in databases."

(http://www.getty.edu/research/conducting_research/vocabularies/introvocabs/what.html#bases).

Metadata Crosswalks

A metadata crosswalk is used to translate between metadata standards. For instance, Insight software maps the different metadata standards used by the collections in the digital library using a version of the Getty Crosswalk in order to facilitate cross collection searching (<http://www.luna-imaging.com/insight/featuretour/crosscollectionsearch.html>). For example, an item catalogued using Dublin Core would have a field called, "Creator", a MARC record would have the "Personal Author Main Entry" field, and a VRA Core 4.0 record would have "AGENT name." All these fields are mapped together in the crosswalk so that someone searching across all these collections would not have to know in which standard a particular collection was catalogued and what that author/agent/creator/main entry field was called in each.

Selected Resource

Lanzi, Elisa. *Introduction to Vocabularies: A Guide to Enhancing Access to Art and Material Culture Information*. Revised by Patricia Harpring (2000).
http://www.getty.edu/research/conducting_research/vocabularies/introvocabs/.

VIII. OPEN ARCHIVES INITIATIVE (OAI)

The Open Archives Initiative (<http://www.openarchives.org/>) evolved out of the scholarly community's efforts to provide interoperability to metadata describing digital resources across institutions and systems. The protocol is based on a

universe of “data providers” and “service providers.” Data providers comply with the protocol as a means to expose their metadata to service providers. Service providers, in turn, aggregate the metadata of many OAI repositories in one search interface. It is imperative to understand that OAI is strictly a metadata protocol; it does not manage digital content. The metadata records in the service provider’s database simply link back to the institution from which the metadata had originally been harvested, where the user can access the resource itself. The University of Colorado’s Digital Library will function, as appropriate, as a data provider. Therefore, it is important that collections cataloged in the Digital Library be compliant with the OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting). The current version of the protocol (2.0), and associated implementation documents, may be found at:
<http://www.openarchives.org/documents/index.html>.

“Open” metadata vs. “Open access” resources

It should be noted that the word *open* in the protocol’s name refers to the availability of the metadata, not the digital resource itself. Data providers can still restrict access to and use of the resource being described, and can even charge fees if desired. Furthermore, the metadata harvesting is not done in “real time.” That is, a user searching the service provider’s aggregated database is viewing static metadata that may have been harvested five minutes ago, or five years ago. This makes URL maintenance in any service provider’s aggregated database rather challenging. Once the metadata has been harvested, there is no dynamic connection to the data provider’s master record and therefore no assurance of currency. In an effort to increase currency, many service providers establish scheduled and regular “re-harvesting” of the data provider’s records.

Open access, on the other hand, is a term that most often refers to resources that have no monetary charge attached to usage. Open access is neither a protocol for resource description nor metadata harvesting. OAI-compliant metadata can (and, in theory, should) be used to describe open access materials. However, just because a resource is OAI-compliant does not mean it is *open access*.

What does OAI mean for the University of Colorado’s Digital Library?

The Digital Library’s Luna Imaging XML Gateway platform (<http://www.lunaimaging.com/insight/XMLGateway.pdf>) is compliant with the OAI-PMH: “Insight has a utility to generate a static OAI repository for any or all of your collections. A static OAI repository is an XML document formatted in accordance with OAI guidelines. Static repositories are useful for small and relatively static metadata collections” (<http://www.lunaimaging.com/insight/featuretour.html#share>). Therefore, it is recommended that collections not be opened for harvesting until they are relatively static (i.e., fully cataloged).

Adhering to recognized and established standards for descriptive and administrative metadata (especially the technical and preservation components of administrative metadata) is perhaps the most important step in creating metadata that is interoperable and shareable. While the current protocol for OAI uses simple Dublin Core as the common metadata scheme at the service provider/aggregator

level, data providers are encouraged to use other established, more granular, and domain-specific metadata schema as well, as suited to the collection.

Please refer to the [*Best Practices for OAI Data Provider Implementations and Shareable Metadata*](#), which includes general recommendations for producing OAI-compliant metadata, as well as some useful recommendations on how to populate the following classes of elements:

Titles	Geographic Places
Names	Identifiers
Dates	Rights for Resources
Subjects / Topics	Bibliographic Citations
Language	Types of Resources

IX. CU DIGITAL LIBRARY'S METADATA BEST PRACTICES WORKING GROUP

The following people have contributed to these best practices. Please contact any one of them with questions.

[Christopher Cronin](#)

Head of Digital Resources Cataloging
Cataloging & Metadata Services Department
University of Colorado at Boulder Libraries

[Katie Lage](#)

Map Librarian
Jerry Crail Johnson Earth Sciences and Map Library
University of Colorado at Boulder Libraries

[Lynn Lickteig](#)

Director of Visual Resources
College of Architecture & Planning
University of Colorado at Denver and Health Sciences Center

[Holley Long](#)

Systems Librarian for Digital Initiatives
Systems Department
University of Colorado at Boulder Libraries

[Elaine Paul](#)

Visual Resources Curator
Department of Art and Art History
University of Colorado at Boulder