Federal Agencies
Digital Guidelines Initiative

September 2016

Technical Guidelines for Digitizing Cultural Heritage Materials

Creation of Raster Image Files

Document Information

Title	***Editor***
Technical Guidelines for Digitizing Cultural Heritage Materials: Creation of Raster Image Files	Thomas Rieger
Document Type	Technical Guidelines
Publication Date	September 2016

Source Documents

Title	***Editors***
Technical Guidelines for Digitizing Cultural Heritage Materials: Creation of Raster Image Master Files *http://www.digitizationguidelines.gov/guidelines/FADGI_Still_Image-Tech_Guidelines_2010-08-24.pdf*	Don Williams and Michael Stelmach
Document Type	Technical Guidelines
Publication Date	August 2010
Title	**Authors**
Technical Guidelines for Digitizing Archival Records for Electronic Access: Creation of Production Master Files – Raster Images *http://www.archives.gov/preservation/technical/guidelines.pdf*	Steven Puglia, Jeffrey Reed, and Erin Rhodes U.S. National Archives and Records Administration
Document Type	Technical Guidelines
Publication Date	June 2004

PUBLICDOMAIN

This work is available for worldwide use and reuse under CC0 1.0 Universal.

METADATA

This section of the *Guidelines* serves as a general discussion of metadata rather than a recommendation of specific metadata element sets.

Although there are many technical parameters discussed in these *Guidelines* that define a high-quality master image file, we do not consider an image to be of high quality unless metadata is associated with the file. Metadata makes possible several key functions: the identification, management, access, use, and preservation of a digital resource, and is therefore directly associated with most of the steps in a digital imaging project workflow. Although it can be costly and time-consuming to produce, metadata adds value to master image files. Images without sufficient metadata are at greater risk of being lost.

Application Profiles

No single metadata element set or standard will be suitable for all projects or all collections. Likewise, different original source formats (text, image, audio, video, etc.) and different digital file formats may require varying metadata sets and depths of description. Element sets should be adapted to fit requirements for particular materials, business processes, and system capabilities.

Because no single element set will be optimal for all projects, implementations of metadata in digital projects are beginning to reflect the use of "application profiles," defined as metadata sets that consist of data elements drawn from different metadata schemes, which are combined, customized, and optimized for a particular local application or project. This "mixing and matching" of elements from different schemas allows for more useful metadata to be implemented at the local level while adherence to standard data values and structures is still maintained. Locally created elements may be added as extensions to the profile, data elements from existing schemas might be modified for specific interpretations or purposes, or existing elements may be mapped to terminology used locally.

Data or Information Models

Because of the likelihood that heterogeneous metadata element sets, data values, encoding schemes, and content information (different source and file formats) will need to be managed within a digital project, it is good practice to put all of these pieces into a broader context at the outset of any project in the form of a data or information model. A model can help to define the types of objects involved and how and at what level they will be described (i.e., are descriptions hierarchical in nature, will digital objects be described at the file or item level as well as at an higher aggregate level, how are objects and files related, what kinds of metadata will be needed for the system, for retrieval and use, for management, etc.), as well as document the rationale behind the different types of metadata sets and encodings used. A data model informs the choice of metadata element sets, which determine the content values, which are then encoded in a specific way (in relational database tables or an XML document, for example).

Levels of Description

Although there is benefit to recording metadata on the item level to facilitate more precise retrieval of images within and across collections, we realize that this level of description is not always practical. Different projects and collections may warrant more in-depth metadata capture than others. A deep level of description at the item level, for example, is not usually accommodated by traditional archival descriptive practices. The functional purpose of metadata often determines the amount of metadata that is needed. Identification and retrieval of digital images may be accomplished using a very small amount of metadata. However, management of and preservation services performed on digital images will require more finely detailed metadata – particularly at the technical level, in order to render the file; and at the structural level, in order to describe the relationships among different files and versions of files.

Metadata creation requires careful analysis of the resource at hand. Although there are current initiatives aimed at automatically capturing a given set of values, we believe that metadata input is still largely a manual process, and will require human intervention at many points in the object's lifecycle to assess the quality and relevance of metadata associated with it.

Common Metadata Types

Several categories of metadata are associated with the creation and management of master image files. The following metadata types are the ones most commonly implemented in imaging projects. Although these categories are defined separately below, there is not always an obvious distinction between them, since each type contains elements that are both descriptive and administrative in nature. These types are commonly broken down by what functions the metadata supports. In general, the types of metadata listed below, except for descriptive, are usually found "behind the scenes" in databases rather than in public access systems. As a result, these types of metadata tend to be less standardized and more aligned with local requirements. For an overview of different metadata types, standards, and applications, see the Diane Hillmann's presentations, available at http://managemetadata.org/msa_r2/

Descriptive

Descriptive metadata refers to information that supports discovery and identification of a resource (the who, what, when, and where of a resource). It describes the content of the resource, associates various access points, and describes how the resource is related to other resources intellectually or within an hierarchy. In addition to bibliographic information, it may also describe physical attributes of the resource such as media type, dimension, and condition. Descriptive metadata is usually highly structured and often conforms to one or more standardized, published schemes such as Dublin Core or MARC. Controlled vocabularies, thesauri, or authority files are commonly used to maintain consistency across the assignment of access points. Descriptive information is usually stored outside of the image file, often in separate catalogs or databases from technical information about the image file.

Although descriptive metadata may be stored elsewhere, it is recommended that some basic descriptive metadata (such as a caption or title) accompany the structural and technical metadata captured during production. The inclusion of this metadata can be useful for identification of files or groups of related files during quality review and other parts of the workflow, or for tracing the image back to the original.

Descriptive metadata is not specified in detail in this document. However, we recommend the use of the Dublin Core Metadata Element1[11] set to capture minimal descriptive metadata information where metadata in another formal data standard does not exist. Metadata should be collected directly in Dublin Core. If it is not used for direct data collection, a mapping to Dublin Core elements is recommended. A mapping to Dublin Core from a richer, local metadata scheme already in use may also prove helpful for data exchange across other projects utilizing Dublin Core. Not all Dublin Core elements are required in order to create a valid Dublin Core record.

Any local fields that are important within the context of a particular project should also be captured to supplement Dublin Core fields so that valuable information is not lost. We anticipate that selection of metadata elements will come from more than one preexisting element set – elements can always be tailored to specific formats or local needs. Projects should support a modular approach to designing metadata to fit the specific requirements of the project. Standardizing on Dublin Core supplies baseline metadata that provides access to files, but this should not exclude richer metadata that extends beyond the Dublin Core set, if available.

For large-scale digitization projects, only minimal metadata may be affordable to record during capture, and is likely to consist of linking image identifiers to page numbers and indicating major structural divisions or anomalies of the resource (if applicable) for text documents. For photographs, capturing caption information or keywords, if any, and a local identifier for the original photograph, is ideal. For other non-textual materials, such as posters and maps, descriptive information taken directly from the item being scanned as well as a local identifier should be captured. If keying of captions into a database is prohibitive, if possible scan captions as part of the image itself. Although this information will not be searchable, it will serve to provide some basis of identification for the subject matter of the photograph. Recording of identifiers is important for uniquely identifying resources, and is necessary for locating and managing them. It is likely that digital images will be associated with more than one identifier – for the image itself, for metadata or database records that describe the image, and for reference back to the original.

---

[11] Dublin Core Metadata Initiative, (http: /dublincore.org/usage/terms/dc/current-elements/). The Dublin Core element set is characterized by simplicity in creation of records, flexibility, and extensibility. It facilitates description of all types of resources and is intended to be used in conjunction with other standards that may offer fuller descriptions in their respective domains.

Administrative

The Dublin Core set does not provide for administrative, technical, or highly structured metadata about different document types. Administrative metadata comprises both technical and preservation metadata, and is generally used for internal management of digital resources. Administrative metadata may include information about rights and reproduction or other access requirements, selection criteria or archiving policy for digital content, audit trails or logs created by a digital asset management system, persistent identifiers, methodology or documentation of the imaging process, or information about the source materials being scanned. In general, administrative metadata is informed by the local needs of the project or institution and is defined by project-specific workflows. Administrative metadata may also encompass repository-like information, such as billing information or contractual agreements for deposit of digitized resources into a repository.

For additional information, see Harvard University Library's Digital Repository Services (DRS) User Manual for Data Loading, Version 2.04 at http://library.harvard.edu/lts, particularly Section 5.0, "DTD Element Descriptions" for application of administrative metadata in a repository setting; also the California Digital Library's Guidelines for Digital Objects at http://www.cdlib.org/inside/diglib/guidelines/. The Library of Congress has defined a data dictionary for various formats in the context of METS, Data Dictionary for Administrative Metadata for Audio, Image, Text, and Video Content to Support the Revision of Extension Schemas for METS, available at http://lcweb.loc.gov/rr/mopic/avprot/extension2.html.

Rights

Although metadata regarding rights management information is briefly mentioned above, it encompasses an important piece of administrative metadata that deserves further discussion. Rights information plays a key role in the context of digital imaging projects and will become more and more prominent in the context of preservation repositories, as strategies to act upon digital resources in order to preserve them may involve changing their structure, format, and properties. Rights metadata will be used both by humans to identify rights holders and legal status of a resource, and also by systems that implement rights management functions in terms of access and usage restrictions.

Because rights management and copyright are complex legal topics, legal counsel should be consulted for specific guidance and assistance. The following discussion is provided for informational purposes only and should not be considered specific legal advice.

Metadata element sets for intellectual property and rights information are still in development, but they will be much more detailed than statements that define reproduction and distribution policies. At a minimum, rights-related metadata should include: the legal status of the record; a statement on who owns the physical and intellectual aspects of the record; contact information for these rights holders; as well as any restrictions associated with the copying, use, and distribution of the record. To facilitate bringing digital copies into future repositories, it is desirable to collect appropriate rights management metadata at the time of creation of the digital copies. At the very least, digital versions should be identified with a designation of copyright status, such as: "public domain;" "copyrighted" (and whether clearance/permissions from rights holder has been secured); "unknown;" "donor agreement/ contract;" etc.

Preservation metadata dealing with rights management in the context of digital repositories will likely include detailed information on the types of actions that can be performed on data objects for preservation purposes and information on the agents or rights holders that authorize such actions or events.

For an example of rights metadata in the context of libraries and archives, a rights extension schema has also been added to the Metadata Encoding and Transmission Standard (METS), which documents metadata about the intellectual rights associated with a digital object. This extension schema contains three components: a rights declaration statement; detailed information about rights holders; and context information, which is defined as "who has what permissions and constraints within a specific set of circumstances." The schema is available at: https://www.loc.gov/standards/rights/METSRights.xsd.

For additional information on rights management, see:

Peter B. Hirtle, "Archives or Assets?" at http://www.archivists.org/governance/presidential/hirtle.asp;

June M. Besek, Copyright Issues Relevant to the Creation of a Digital Archive: A Preliminary Assessment, January 2003 at http://www.clir.org/pubs/reports/pub112/contents.html;

Karen Coyle, Rights Expression Languages, A Report to the Library of Congress, February 2004, available at http://www.loc.gov/standards/relreport.pdf;

MPEG-21 Overview v.5 contains a discussion on intellectual property and rights at http://www.chiariglione.org/mpeg/standards/mpeg-21/mpeg-21.htm;

Mary Minow, "Library Digitization Projects: Copyrighted Works that have Expired into the Public Domain" at http://www.librarylaw.com/DigitizationTable.htm;

For a comprehensive discussion on libraries and copyright, see: Mary Minow, *Library Digitization Projects and Copyright* at http://www.llrx.com/features/digitization.htm.

Technical

Technical metadata refers to information that describes attributes of the digital image (not the analog source of the image), and helps to ensure that images will be rendered accurately. It supports content preservation by providing information needed by applications to use the file and to successfully control the transformation or migration of images across or between file formats. Technical metadata also describes the image capture process and technical environment, such as hardware and software used to scan images, as well as file format-specific information, image quality, and information about the source object being scanned, which may influence scanning decisions. Technical metadata helps to ensure consistency across a large number of files by enforcing standards for their creation. At a minimum, technical metadata should capture the information necessary to render, display, and use the resource.

Technical metadata is characterized by information that is both objective and subjective – attributes of image quality that can be measured using objective tests as well as information that may be used in a subjective assessment of an image's value. Although tools for automatic creation and capture of many objective components are badly needed, it is important to determine what metadata should be highly structured and useful to machines, as opposed to what metadata would be better served in an unstructured, free-text note format. The more subjective data is intended to assist researchers in the analysis of a digital resource, or imaging specialists and preservation administrators in determining long-term value of a resource.

In addition to the digital image, technical metadata will also need to be supplied for the metadata record itself if the metadata is formatted as a text file or XML document or METS document, for example. In this sense, technical metadata is highly recursive, but necessary for keeping both images and metadata understandable over time.

Requirements for technical metadata will differ for various media formats. For digital still images, we refer to the *ANSI/NISO Z39.87 Data Dictionary - Technical Metadata for Digital Still Images* available from the NISO website http://www.niso.org/home. It is a comprehensive technical metadata set based on the Tagged Image File Format specification, and makes use of the data that is already captured in file headers. It also contains metadata elements important to the management of image files that are not present in header information, but that could potentially be automated from scanner/camera software applications. An XML schema for the NISO technical metadata has been developed at the Library of Congress called MIX (Metadata in XML), which is available at http://www.loc.gov/standards/mix/.

See also the TIFF 6.0 Specification at http://partners.adobe.com/public/developer/en/tiff/TIFF6.pdf as well as the Digital Imaging Group's DIG 35 metadata element set at http://www.bgbm.fu-berlin.de/TDWG/acc/Documents/DIG35-v1.1WD-010416.pdf; and Harvard University Library's Administrative Metadata for Digital Still Images data dictionary at http://preserve.harvard.edu/resources/imagemetadata.pdf.

Initiatives such as the Global Digital Format Registry (http://hul.harvard.edu/gdfr/) could potentially help in reducing the number of metadata elements that need to be recorded about a file or group of files regarding file format information necessary for preservation functions. Information maintained in the Registry could be pointed to instead of recorded for each file or batch of files.

Embedded Metadata

Although embedded metadata is mostly about "where" metadata is stored, it seems in some ways to be a subset of technical metadata as it primarily refers to attributes about the digital image and the creation of the digital image. See http://www.digitizationguidelines.gov/guidelines/digitize-tiff.html.

Structural

Structural metadata describes the relationships between different components of a digital resource. It ties the various parts of a digital resource together in order to make a useable, understandable whole. One of the primary functions of structural metadata is to enable display and navigation, usually via a page-turning application, by indicating the sequence of page images or the presence of multiple views of a multi-part item. In this sense, structural metadata is closely related to the intended behaviors of an object. Structural metadata is very much informed by how the images will be delivered to the user as well as how they will be stored in a repository system in terms of how relationships among objects are expressed.

Structural metadata often describes the significant intellectual divisions of an item (such as chapter, issue, illustration, etc.), and correlates these divisions to specific image files. These explicitly labeled access points help to represent the organization of the original object in digital form. This does not imply, however, that the digital form must always imitate the organization of the original – especially for non-linear items, such as folded pamphlets. Structural metadata also associates different representations of the same resource together, such as master image files with their derivatives, or different sizes, views, or formats of the resource.

Example structural metadata might include whether the resource is simple or complex (multi-page, multi-volume, has discrete parts, contains multiple views); what the major intellectual divisions of a resource are (table of contents, chapter, musical movement); identification of different views (double-page spread, cover, detail); the extent (in files, pages, or views) of a resource and the proper sequence of files, pages and views; as well as different technical (file formats, size), visual (pre- or post-conservation treatment), intellectual (part of a larger collection or work), and use (all instances of a resource in different formats – TIFF files for display, PDF files for printing, OCR file for full text searching) versions.

File names and organization of files in system directories comprise structural metadata in its barest form. Since meaningful structural metadata can be embedded in file and directory names, consideration of where and how structural metadata is recorded should be done upfront.

No widely adopted standards for structural metadata exist since most implementations of structural metadata are at the local level, and are very dependent on the object being scanned and the desired functionality in using the object. Most structural metadata is implemented in file naming schemes and/or in spreadsheets or databases that record the order and hierarchy of the parts of an object so that they can be identified and reassembled into their original form.

The Metadata Encoding and Transmission Standard (METS) is often discussed in the context of structural metadata, although it is inclusive of other types of metadata as well. METS provides a way to associate metadata with the digital files it describes, and to encode the metadata and the files in a standardized manner using XML. METS requires structural information about the location and organization of related digital files to be included in the METS document. Relationships between different representations of an object as well as relationships between different hierarchical parts of an object can be expressed. METS brings together a variety of metadata about an object all into one place by allowing the encoding of descriptive, administrative, and structural metadata. Metadata and content information can either be wrapped together within the METS document, or pointed to from the METS document if they exist in externally disparate systems. METS also supports extension schemas for descriptive and administrative metadata to accommodate a wide range of metadata implementations. Beyond associating metadata with digital files, METS can be used as a data transfer syntax so objects can easily be shared; as a Submission Information Package, an Archival Information Package, and a Dissemination Information Package in an OAIS-compliant repository (see below); and also as a driver for applications, such as a page turner, by associating certain behaviors with digital files so that they can be viewed, navigated, and used. Because METS is primarily concerned with structure, it works best with "library-like" objects in establishing relationships among multi-page or multi-part objects, but it does not apply as well to hierarchical relationships that exist in collections within an archival context.

See http://www.loc.gov/standards/mets/ for more information on METS.

Behavior

Behavior metadata is often referred to in the context of a METS object. It associates executable behaviors with content information that define how a resource should be utilized or presented. Specific behaviors might be associated with different genres of materials (books, photographs, Powerpoint presentations) as well as with different file formats. Behavior metadata contains a component that abstractly defines a set of behaviors associated with a resource as well as a "mechanism" component that points to executable code (software applications) that then performs a service according to the defined behavior. The ability to associate behaviors or services with digital resources is one of the attributes of a METS object and is also part of the "digital object architecture" of the Fedora digital repository system. See http://www.fedora.info/ for discussion of Fedora and digital object behaviors.

Preservation

Preservation metadata encompasses all information necessary to manage and preserve digital assets over time. Preservation metadata is usually defined in the context of the OAIS reference model (Open Archival Information System), and is often linked to the functions and activities of a repository. It differs from technical metadata in that it documents processes performed over time (events or actions taken to preserve data and the outcomes of these events) as opposed to explicitly describing provenance (how a digital resource was created) or file format characteristics, but it does encompass all types of the metadata mentioned above, including rights information. Although preservation metadata draws on information recorded earlier (technical and structural metadata would be necessary to render and reassemble the resource into an understandable whole), it is most often associated with analysis of and actions performed on a resource after submission to a repository. Preservation metadata might include a record of changes to the resource, such as transformations or conversions from format to format, or indicate the nature of relationships among different resources.

Preservation metadata is information that will assist in preservation decision-making regarding the long-term value of a digital resource and the cost of maintaining access to it, and will help to both facilitate archiving strategies for digital images as well as support and document these strategies over time. Preservation metadata is commonly linked with digital preservation strategies such as migration and emulation, as well as more "routine" system-level actions such as copying, backup, or other automated processes carried out on large numbers of objects. These strategies will rely on all types of pre-existing metadata and will also generate and record new metadata about the object. It is likely that this metadata will be both machine-processible and "human-readable" at different levels to support repository functions as well as preservation policy decisions related to these objects.

In its close link to repository functionality, preservation metadata may reflect or even embody the policy decisions of a repository; but these are not necessarily the same policies that apply to preservation and reformatting in a traditional context. The extent of metadata recorded about a resource will likely have an impact on future preservation options to maintain it. Current implementations of preservation metadata are repository- or institution-specific. A digital asset management system may provide some basic starter functionality for low-level preservation metadata implementation, but not to the level of a repository modeled on the OAIS.

See also *A Metadata Framework to Support the Preservation of Digital Objects* at http://www.oclc.org/research/projects/pmwg/pm_framework.pdf and

Preservation Metadata for Digital Objects: A Review of the State of the Art at http://www.oclc.org/research/projects/pmwg/presmeta_wp.pdf, both by the OCLC/RLG Working Group on Preservation Metadata, for excellent discussions of preservation metadata in the context of the OAIS model. The international working group behind PREMIS, or "Preservation Metadata: Implementation Strategies," has developed best practices for implementing preservation metadata and has published a recommended core set of preservation metadata in their Data Dictionary for Preservation Metadata, as well as an XML schema. Their work can be found at http://www.loc.gov/standards/premis/.

Tracking

Tracking metadata is used to control or facilitate the particular workflow of an imaging project during different stages of production. Elements might reflect the status of digital images as they go through

different stages of the workflow (batch information and automation processes, capture, processing parameters, quality control, archiving, identification of where/media on which files are stored). This is primarily internally-defined metadata that serves as documentation of the project and may also serve also serve as a statistical source of information to track and report on progress of image files. Tracking metadata may exist in a database or via a directory/folder system.

Meta-Metadata

Although this information is difficult to codify, it usually refers to metadata that describes the metadata record itself, rather than the object it is describing, or to high-level information about metadata "policy" and procedures, most often on the project level. Meta-metadata documents information such as who records the metadata, when and how it gets recorded, where it is located, what standards are followed, and who is responsible for modification of metadata and under what circumstances.

It is important to note that metadata files yield "master" records as well. These non-image assets are subject to the same rigor of quality control and storage as master image files. Provisions should be made for the appropriate storage and management of the metadata files over the long term.

Assessment of Metadata Needs for Imaging Projects

Before beginning any scanning, it is important to conduct an assessment both of existing metadata and metadata that will be needed in order to develop data sets that fit the needs of the project. The following questions frame some of the issues to consider:

- *Does metadata already exist in other systems (database, bibliographic record, finding aid, on item itself) or in structured formats (Dublin Core, local database)?*

If metadata already exists, can it be automatically derived from these systems, pointed to from new metadata gathered during scanning, or does it require manual input? Efforts to incorporate existing metadata should be pursued. It is also extremely beneficial if existing metadata in other systems can be exported to populate a production database prior to scanning. This can be used as base information needed in production tracking, or to link item level information collected at the time of scanning to metadata describing the content of the resource. An evaluation of the completeness and quality of existing metadata may need to be made to make it useful (e.g., what are the characteristics of the data content, how is it structured, can it be easily transformed?)

It is likely that different data sets with different functions will be developed, and these sets will exist in different systems. However, efforts to link together metadata in disparate systems should be made so that it can be reassembled into something like a METS document, an Archival XML file for preservation, or a Presentation XML file for display, depending on what is needed. Metadata about digital images should be integrated into peer systems that already contain metadata about both digital and analog materials. By their nature, digital collections should not be viewed as something separate from non-digital collections. Access should be promoted across existing systems rather than building a separate stand-alone system.

- *Who will capture metadata?*

Metadata is captured by systems or by humans, and is intended for system or for human use. For example, certain preservation metadata might be generated by system-level activities such as data backup or copying. Certain technical metadata is used by applications to accurately render an image. In determining the function of metadata elements, it is important to establish whether this information is important for use by machines or by people. If it is information that is used and/or generated by systems, is it necessary to explicitly record it as metadata? What form of metadata is most useful for people? Most metadata element sets include less structured, note or comment-type fields that are intended for use by administrators and curators as data necessary for assessment of the provenance, risk of obsolescence, and value inherent to a particular class of objects. Any data, whether generated by systems or people, that is necessary to understand a digital object, should be considered as metadata that may be necessary to formally record. But because of the high costs of manually generating metadata and tracking system-level information, the use and function of metadata elements should be carefully considered. Although some metadata can be automatically captured, there is no guarantee that this data will be valuable over the long term.

- *How will metadata be captured?*

Metadata capture will likely involve a mix of manual and automated entry. Descriptive and structural metadata creation is largely manual; some may be automatically generated through OCR processes to create indexes or full text; some technical metadata may be captured automatically from imaging software and devices; more sophisticated technical metadata, such as metadata that will be used to inform preservation decisions, will require visual analysis and manual input.

An easy-to-use and customizable database or asset management system with a graphical and intuitive front end, preferably structured to mimic a project's particular metadata workflow, is desirable and will make for more efficient metadata creation.

- *When will metadata be collected?*

Metadata is usually collected incrementally during the scanning process and will likely be modified over time. At least, start with a minimal element set that is known to be needed, and add additional elements later if necessary.

Assignment of unique identifier or naming scheme should occur upfront. We also recommend that descriptive metadata be gathered prior to capture to help streamline the scanning process. It is usually much more difficult to add new metadata later on, without consultation of the originals. The unique file identifier can then be associated with a descriptive record identifier if necessary.

A determination of what structural metadata elements to record should also occur prior to capture, preferably during the preparation of materials for capture or during collation of individual items. Information about the hierarchy of the collection, the object types, and the physical structure of the objects should be recorded in a production database prior to scanning. The structural parts of the object can be linked to actual content files during capture. Most technical metadata is gathered at the time of scanning. Preservation metadata is likely to be recorded later on, upon ingest into a repository.

- *Where will the metadata be stored?*

Metadata can be embedded within the resource (such as an image header or file name), or can reside in a system external to the resource (such as a database), or both. Metadata can be also encapsulated with the file itself, such as with the Metadata Encoded Transmission Standard (METS). The choice of location of metadata should encourage optimal functionality and long-term management of the data.

Header data consists of information necessary to decode the image, and has somewhat limited flexibility in terms of data values that can be put into the fields. Header information accommodates more technical than descriptive metadata (but richer sets of header data can be defined depending on the image file format). The advantage is that metadata remains with the file, which may result in more streamlined management of content and metadata over time. Several tags are saved automatically as part of the header during processing, such as dimensions, date, and color profile information, which can serve as base-level technical metadata requirements. However, methods for storing information in file format headers are very format-specific and data may be lost in conversions from one format to another. Also, not all applications may be able to read the data in headers. Information in headers should be manually checked to see if data has transferred correctly or has not been overwritten during processing. Just because data exists in headers does not guarantee that it has not been altered or has been used as intended. Information in headers should be evaluated to determine if it has value. Data from image headers can be extracted and imported into a database; a relationship between the metadata and the image must then be established and maintained.

Storing metadata externally to the image in a database provides more flexibility in managing, using, and transforming it and also supports multi-user access to the data, advanced indexing, sorting, filtering, and querying. It can better accommodate hierarchical descriptive information and structural information about multi-page or complex objects, as well as importing, exporting, and harvesting of data to external systems or other formats, such as XML. Because metadata records are resources that need to be managed in their own right, there is certainly benefit to maintaining metadata separately from file content in a managed system. Usually a unique identifier or the image file name is used to link metadata in an external system to image files in a directory.

We recommend that metadata be stored both in image headers as well as in an external database to facilitate migration and repurposing of the metadata. References between the metadata and the image files can be maintained via persistent identifiers. A procedure for synchronization of changes to metadata in both locations is also recommended, especially for any duplicated fields. This approach allows for metadata redundancy in different locations and at different levels of the digital object for ease of use

(image file would not have to be accessed to get information; most header information would be extracted and added into an external system). Not all metadata should be duplicated in both places (internal and external to the file). Specific metadata is required in the header so that applications can interpret and render the file; additionally, minimal descriptive metadata such as a unique identifier or short description of the content of the file should be embedded in header information in case the file becomes disassociated from the tracking system or repository. Some applications and file formats offer a means to store metadata within the file in an intellectually structured manner, or allow the referencing of standardized schemes, such as Adobe XMP or the XML metadata boxes in the JPEG 2000 format. Otherwise, most metadata will reside in external databases, systems, or registries.

- *How will the metadata be stored?*

Metadata schemes and data dictionaries define the content rules for metadata creation, but not the format in which metadata should be stored. Format may be determined partially by where the metadata is stored (file headers, relational databases, spreadsheets) as well as the intended use of the metadata – does it need to be human-readable, or indexed, searched, shared, and managed by machines? How the metadata is stored or encoded is usually a local decision. Metadata might be stored in a relational database or encoded in XML, such as in a METS document, for example.

Adobe's Extensible Metadata Platform (XMP) is another emerging, standardized format for describing where metadata can be stored and how it can be encoded, thus facilitating exchange of metadata across applications. The XMP specification provides both a data model and a storage model. Metadata can be embedded in the file in header information or stored in XML "packets" (these describe how the metadata is embedded in the file). XMP supports the capture of (primarily technical) metadata during content creation and modification and embeds this information in the file, which can then be extracted later into a digital asset management system or database or as an XML file. If an application is XMP enabled or aware (most Adobe products are), this information can be retained across multiple applications and workflows. XMP supports customization of metadata to allow for local field implementation using their Custom File Info Panels application. XMP supports a number of internal schemas, such as Dublin Core and EXIF (a metadata standard used for image files, particularly by digital cameras), as well as a number of external extension schemas. XMP does not guarantee the automatic entry of all necessary metadata (several fields will still require manual entry, especially local fields), but allows for more complete customized and accessible metadata about the file.

See http://www.adobe.com/products/xmp/index.html for more detailed information on the XMP specification and other related documents.

- *Will the metadata need to interact or be exchanged with other systems?*

This requirement reinforces the need for standardized ways of recording metadata so that it will meet the requirements of other systems. Mapping from an element in one scheme to an analogous element in another scheme will require that the meaning and structure of the data is shareable between the two schemes in order to ensure usability of the converted metadata. Metadata will also have to be stored in or assembled into a document format, such as XML, that promotes easy exchange of data. METS-compliant digital objects, for example, promote interoperability by virtue of their standardized, "packaged" format.

- *At what level of granularity will the metadata be recorded?*

Will metadata be collected at the collection level, the series level, the imaging project level, the item (digital object) level, or file level? Although the need for more precise description of digital resources exists so that they can be searched and identified, for many large-scale digitization projects, this is not realistic. Most archival or special collections, for example, are neither organized around nor described at the individual item level, and cannot be without significant investment of time and cost. Detailed description of records materials is often limited by the amount of information known about each item, which may require significant research into identification of subject matter of a photograph, for example, or even what generation of media format is selected for scanning. Metadata will likely be derived from and exist on a variety of levels, both logical and file, although not all levels will be relevant for all materials. Certain information required for preservation management of the files will be necessary at the individual file level. An element indicating level of aggregation (e.g., item, file, series, collection) at which metadata applies can be incorporated, or the relational design of the database may reflect the hierarchical structure of the materials being described.

We recommend that standards, if they exist and apply, be followed for the use of data elements, data values, and data encoding. Attention should be paid to how data is entered into fields and whether controlled vocabularies have been used, in case transformation is necessary to normalize the data.

Relationships

Often basic relationships among multi-page or multi-part files are documented in a file naming scheme, where metadata is captured as much as possible in the surrounding file structure (names, directories, headers). However, we consider that simple, unique, meaningless names for file identifiers, coupled with more sophisticated metadata describing relationships across files stored in an external database, is the preferred way forward to link files together. This metadata might include file identifiers and metadata record identifiers and a codified or typed set of relationships that would help define the associations between digital files and between different representations of the same resource. (Relationships between the digital object and the analog source object or the place of the digital object in a larger collection hierarchy would be documented elsewhere in descriptive metadata). Possible relationship types include identification of principal or authoritative version (for master image file); derivation relationships indicating what files come from what files; whether the images were created in-house or come from outside sources; structural relationships (for multi-page or –part objects); sibling relationships (images of the same intellectual resource, but perhaps scanned from different source formats).

Permanent and Temporary Metadata

When planning for a digital imaging project, it may not be necessary to save all metadata created and used during the digitization phase of the project. For example, some tracking data may not be needed once all quality control and redo work has been completed. It may not be desirable, or necessary, to bring all metadata into a digital repository. An institution may decide not to explicitly record metadata that can easily be recalculated in the future from other information, such as image dimensions if resolution and pixel dimensions are known, or certain file format properties that might be derived directly from the file itself through an application such as JHOVE. Also, it may not be desirable or necessary to provide access to all metadata that is maintained within a system to all users. Most administrative and technical metadata will need to be accessible to administrative users to facilitate managing the digital assets, but does not need to be made available to general users searching the digital collections.

Identifiers and File Naming

File Naming

A file-naming scheme should be established prior to capture. The development of a file naming system should take into account whether the identifier requires machine- or human-indexing (or both – in which case, the image may have multiple identifiers). File names can either be meaningful (such as the adoption of an existing identification scheme which correlates the digital file with the source material), or non-descriptive (such as a sequential numerical string). Meaningful file names contain metadata that is self-referencing; non-descriptive file names are associated with metadata stored elsewhere that serves to identify the file. In general, smaller-scale projects may design descriptive file names that facilitate browsing and retrieval; large-scale projects may use machine-generated names and rely on the database for sophisticated searching and retrieval of associated metadata.

A file naming system based on non-descriptive, non-mnemonic, unique identifiers usually requires a limited amount of metadata to be embedded within the file header, as well as an external database which would include descriptive, technical, and administrative metadata from the source object and the related digital files.

One advantage of a non-descriptive file naming convention is that it eliminates non-unique and changeable descriptive data and provides each file with a non-repeating and sustainable identifier in a form that is not content-dependent. T his allows much greater flexibility for automated data processing and migration into future systems. Other benefits of a non-descriptive file naming convention include the ability to compensate for multiple object identifiers and the flexibility of an external database, which can

accommodate structural metadata including parts and related objects, as well as avoiding any pitfalls associated with legacy file identifiers.

Recommended Characteristics of File Names

- Are unique - no other digital resource should duplicate or share the same identifier as another resource. In a meaningful file-naming scheme, names of related resources may be similar, but will often have different characters, prefixes, or suffixes appended to delineate certain characteristics of the file. An attempt to streamline multiple versions and/or copies should be made.
- Are consistently structured - file names should follow a consistent pattern and contain consistent information to aid in identification of the file as well as management of all digital resources in a similar manner. All files created in digitization projects should contain this same information in the same defined sequence.
- Are well-defined - a well-defined rationale for how/why files are named assists with standardization and consistency in naming and will ease in identification of files during the digitization process and long afterwards. An approach to file naming should be formalized for digitization projects and integrated into systems that manage digital resources.
- Are persistent – files should be named in a manner that has relevance over time and is not tied to any one process or system. Information represented in a file name should not refer to anything that might change over time. The concept of persistent identifiers is often linked to file names in an online environment that remain persistent and relevant across location changes or changes in protocols to access the file.
- Observant of any technical restrictions – file names should be compliant with any character restrictions (such as the use of special characters, spaces, or periods in the name, except in front of the file extension), as well as with any limitations on character length. Ideally, file names should not contain too many characters. Most current operating systems can handle long file names, although some applications will truncate file names in order to open the file, and certain types of networking protocols and file directory systems will shorten file names during transfer. Best practice is to limit character length to no more than 32 characters per file name.

General Guidelines for Creating File Names

- We recommend using a period followed by a three-character file extension at the end of all file names for identification of data format (for example, .tif, .jpg, .gif, .pdf, .wav, .mpg, etc.) A file format extension must always be present.
- Take into account the maximum number of items to be scanned and reflect that in the number of digits used (if following a numerical scheme).
- Use leading 0's to facilitate sorting in numerical order (if following a numerical scheme).
- Do not use an overly complex or lengthy naming scheme that is susceptible to human error during manual input.
- Use lowercase characters and file extensions.
- Record metadata embedded in file names (such as scan date, page number, etc.) in another location in addition to the file name. This provides a safety net for moving files across systems in the future, in the event that they must be renamed.
- In particular, sequencing information and major structural divisions of multi-part objects should be explicitly recorded in the structural metadata and not only embedded in filenames.
- Although it is not recommended to embed too much information into the file name, a certain amount of information can serve as minimal descriptive metadata for the file, as an economical alternative to the provision of richer data elsewhere.
- Alternatively, if meaning is judged to be temporal, it may be more practical to use a simple numbering system. An intellectually meaningful name will then have to be correlated with the digital resource in an external database.

Directory Structure

Regardless of file name, files will likely be organized in some kind of file directory system that will link to metadata stored elsewhere in a database. Master files might be stored separately from derivative files, or directories may have their own organization independent of the image files, such as folders arranged by date or collection identifier, or they may replicate the physical or logical organization of the originals being scanned.

The files themselves can also be organized solely by directory structure and folders rather than embedding meaning in the file name. This approach generally works well for multi-page items. Images are uniquely identified and aggregated at the level of the logical object (i.e., a book, a chapter, an issue, etc.), which requires that the folders or directories be named descriptively. The file names of the individual images themselves are unique only within each directory, but not across directories. For example, book 0001 contains image files 001.tif, 002.tif, 003.tif, etc. Book 0002 contains image files 001.tif, 002.tif, and 003.tif. The danger with this approach is that if individual images are separated from their parent directory, they will be indistinguishable from images in a different directory.

Versioning

For various reasons, a single scanned object may have multiple but differing versions associated with it (for example, the same image prepped for different output intents, versions with additional edits, layers, or alpha channels that are worth saving, versions scanned on different scanners, scanned from different original media, scanned at different times by different scanner operators, etc.). Ideally, the description and intent of different versions should be reflected in the metadata; but if the naming convention is consistent, distinguishing versions in the file name will allow for quick identification of a particular image. Like derivative files, this usually implies the application of a qualifier to part of the file name. The reason to use qualifiers rather than entirely new names is to keep all versions associated with a logical object under the same identifier. An approach to naming versions should be well thought out; adding 001, 002, etc. to the base file name to indicate different versions is an option; however, if 001 and 002 already denote page numbers, a different approach will be required.

Naming Derivative Files

The file naming system should also take into account the creation of derivative image files made from the master files. In general, derivative file names are inherited from the masters, usually with a qualifier added on to distinguish the role of the derivative from other files (i.e., “pr” for printing version, “t” for thumbnail, etc.) Derived files usually imply a change in image dimensions, image resolution, and/or file format from the master. Derivative file names do not have to be descriptive as long as they can be linked back to the master file.

For derivative files intended primarily for Web display, one consideration for naming is that images may need to be cited by users in order to retrieve other higher-quality versions. If so, the derivative file name should contain enough descriptive or numerical meaning to allow for easy retrieval of the original or other digital versions.

Quality Management

Quality control (QC) and quality assurance (QA) are the processes used to ensure digitization and metadata creation are done properly. QC/QA plans and procedures should be initiated, documented and maintained throughout all phases of digital conversion. The plan should address all specifications and reporting requirements associated with each phase of the conversion project, including issues relating to the image files, the associated metadata, and the storage of both (file transfer, data integrity). Also, QC/QA plans should address accuracy requirements for and acceptable error rates for all aspects evaluated. For large digitization projects it may be appropriate to use a statistically valid sampling procedure to inspect files and metadata. In most situations QC/QA are done in a 2-step process- the scanning technician will do initial quality checks during production followed by a second check by another person.