

Federal Agencies Digitization Guidelines Initiative  
Still Image Working Group

Embedded Metadata Sub-group Charter  
Version 1.0

November 26, 2008

**Contents**

1	Background.....	2
2	Purpose / Objectives .....	3
3	Scope.....	4
3.1	Tools / Validation .....	4
4	Participants.....	4
4.1	Federal.....	4
4.2	Non-Federal Experts .....	4
5	Approach.....	5
5.1	Content Group.....	5
5.2	Embedding Group.....	5
5.3	Workflow Group.....	6
6	Process .....	6
7	Deliverables .....	6
8	Timeline .....	7
9	References.....	8
10	Version Control and Change History.....	9

## 1 Background

Metadata plays a critical role in the management and use of digital content. Comprehensive and consistent metadata allows users to find the content they are seeking, to identify, understand, and differentiate the content, and to know whether restrictions apply to its use. Technical metadata can provide information on the technical characteristics of the digital object (pixel dimensions, file encoding and compression, etc.), and how it was created. Technical information increases the interoperability of the content and provides a basis for evaluating the viability of a digital object or its suitability for reformatting.

Metadata about individual images and image sets commonly exists in a system external to the content file, such as online library catalogs and external databases. It can also be included as part of the digital content file itself. Metadata contained within the file is referred to as embedded metadata. The primary advantage of embedded metadata is that it travels with the content file, allowing the file to be, to some extent, “self-describing.” The Federal Agencies Digitization Guidelines Initiative Still Image Working Group formed the Embedded Metadata Sub-group specifically to deal with guidelines for embedded metadata.

With the advent of the web, United States Federal agencies and cultural institutions began converting cultural heritage materials to digital formats. Different agencies and institutions developed different requirements for these emerging imaging projects: some based their metadata on the original source document; others based it on the digital image itself. Some required significant amounts of technical metadata, others required less. The TIFF image file format, commonly used as a “master” digital image, included support for a flexible metadata set in tagged fields within the file itself. The TIFF baseline set of metadata tags provided the basic technical data needed for rendering and interoperability between a wide range of software tools on a variety of different platforms. This internal tag set was extendable to allow developers to add additional metadata needed for applications that were almost unimaginable at the time. Some of these added tags are “private” and only accessible and meaningful within proprietary applications.

Over time, the TIFF image format has become widely accepted and used, and software platforms and tools have stabilized. Additional imaging capabilities are available, such as the automatic creation and insertion of technical metadata during image capture, and new tools for the efficient insertion of technical, workflow and descriptive metadata. Other stakeholders, particularly digital photography commercial interests in manufacturing and media, have developed additional metadata standards and embedding methods, including EXIF, IPTC, Dublin Core, and Adobe’s Extensible Metadata Platform (XMP). Digitization projects use image formats other than TIFF, which have different metadata embedding capabilities.

This is an excellent time to review requirements for embedded metadata and develop baseline recommendations for federal agencies. Current embedded metadata standards for TIFF tags used by federal agencies are still, to a large extent, based on the issues and needs of those early imaging projects. Even within a single institution, embedded metadata practices have varied over time – particularly as projects evolved from bi-tonal scanning of

simple documents to full-color imaging of valuable rare books, manuscripts, photographs, and maps.

The Still Image Working Group identified embedded metadata as a high priority based on

- the important role of metadata in the management, use and sustainability of digital assets, and
- the lack of clear, comprehensive, and uniform guidelines in this area.

Digitized cultural heritage materials are frequently distributed on the Web, which increases the likelihood that they may become separated from the institutional infrastructure that provides users with information about these digital objects' identity, description, and context of creation. Embedded metadata can inform users of certain basic information about these objects, such as their identification, any rights associated with or restrictions on their usage, and technical characteristics should the digital image files become separated from their external description. Additionally, systems can use that metadata to know how to use the objects (such as proper display), to support interoperability, and to support continued access through the transformations of the digital object over time – thereby making the objects more sustainable.

## **2 Purpose / Objectives**

The work of the Embedded Metadata Sub-group falls within the scope and methodologies set forth in the Still Image Working Group Central Charter. The purpose of the Embedded Metadata Sub-group is to improve the production, usability, and sustainability of digital image assets by developing a consistent uniform standard for embedded metadata. The first objective of the group will be to carefully delineate the functions and uses of embedded metadata. To accomplish this, use cases will be developed for federal agencies and cultural institutions, image producers, end users, and those who serve and preserve the images.

Subsequent activities of the group will focus on three specific aspects of metadata creation:

- **Content** – Identify or establish common guidelines for embedded metadata fields or elements in digital image files as well as the format of the metadata in those fields. The recommendations regarding content will inform the method used for embedding the content.
- **Embedding Method**– Investigate and draft recommendations for a common method (format) for embedding a metadata model or schema that is able to represent diverse metadata, and is supported by all digital image file formats commonly used in cultural heritage, archival, or historical digitization activities. The format should be based on an open standard or openly documented and widely used specification. The internal data format and data transfer shall be Unicode compliant.
- **Workflow Tools** – Research applications and tools for:
  - Managing embedded metadata. The tool(s) should support the operations of batch embedding, editing, appending, deleting, or overwriting metadata.

Performing such operations should not compromise the validity or integrity of the data model.

- Validating the embedded metadata. The tools should validate the structure required by the metadata schema, the existence of required fields or elements of the schema, and the values and format of data in those fields. The validation tool should be flexible and customizable with regard to expected fields and data values.

### **3 Scope**

The scope of this work will be limited to metadata embedded within image files for work that falls within the scope of the primary charter of the Federal Agencies Digitization Guidelines Still Image Working Group (charter available at <http://www.digitizationguidelines.gov/stillimages/charter.html>). This group will focus on embedded metadata. Metadata stored outside the image file to which it relates, such as sidecar metadata or metadata in external data systems, will be out of scope, although the scope may include methods for, and guidelines concerning embedded “pointers” to external metadata describing the image file (*e.g.* Persistent Uniform Resource Locators). This activity will concentrate on embedded metadata that shall be usable in master image files, irrespective of format, and encompasses the migration of metadata from master to derivative files.

#### **3.1 Tools / Validation**

Guidelines are of limited value without the means to ascertain whether practices conform to those guidelines. For that reason, project scope will include methods to verify conformance to the guidelines within the scope of this charter.

The resource-intensive nature of interrogating metadata and the large quantity of image files being created makes it impractical to perform manual validation of metadata conformance. Therefore, a priority will be placed on identifying or developing automated tools and processes to validate metadata elements and structure as identified and documented under this charter.

### **4 Participants**

#### **4.1 Federal**

All United States Federal agencies and institutions (herein after referred to as "agencies") involved in the digitization of cultural heritage materials are encouraged to participate. All agencies are encouraged to designate one or more representatives possessing experience or expertise in the area of metadata as described within the scope of this document.

#### **4.2 Non-Federal Experts**

In addition to federal participants, experts with strong metadata backgrounds will be sought from the corporate and academic communities.

## 5 Approach

Due to differences among agencies' digitization programs, workflow, audiences and goals, the group will not develop comprehensive recommendations for embedded metadata in image files, but rather will identify common needs and establish baseline recommendations. This group will solicit use cases for embedded metadata to determine commonalities. Although the goal is to establish baseline recommendations, the model will anticipate the addition of extensions to accommodate deviations between different agencies' implementations.

It is expected that metadata embedded in digital images will need to be both human and machine readable and will provide a high-level description of the object being digitally represented, the technical characteristics of the digital object, and information on the permissions and restrictions governing the use of the object.

It is recommended that the work of the Embedded Metadata Sub-group be divided into three interdependent, yet separate, working groups focusing on Content, Embedding Method, and Workflow Tools. These working groups will have the following objectives:

### 5.1 Content Group

The Content Group will have overall responsibility for defining categories of embedded metadata, and recommending categories of and specific data elements that will comprise a "minimum set" in the form of a data dictionary. The group will describe options for formatting the metadata elements and recommend the best formatting structures for the Still Image Working Group to consider. These may include the definition of a single structure that encompasses all of the data elements recommended above, or the use of multiple structures, each fitted to a specific metadata category, but bound together in an appropriate way. The use of multiple structures may be recommended when pre-existing, standardized metadata expressions are well suited to the metadata subcommittee's purposes. Examples for such pre-existing structures include EXIF and IPTC, and the refinements and reconciliations being proposed as guidelines by the Metadata Working Group (<http://www.metadataworkinggroup.org/>). The group will describe requirements for the "pointer" that is embedded in the image file that connects/resolves (in a persistent manner) to the authoritative or master metadata for the file (including those that contain structural metadata that groups image files, such as files comprising a book), decide whether versioning will be used for embedded metadata, and, if so, describe the method and rules for version control. The group will make recommendations pertaining to the movement or inheritance of metadata when content is migrated from master files to derivative files. These recommendations should include direction on what metadata is repeated, excluded, added or altered when added to derivative files.

### 5.2 Embedding Group

The Embedding Group will be responsible for investigating and recommending models used to embed metadata in still image files. The group will describe current formats/models (e.g. TIFF Header Tags, EXIF, IPTC, XMP) used for this, including a brief overview, history and current usage (in cultural heritage digitization as well as born

digital). The group will provide a list of image file formats supported by metadata models (e.g. TIFF, JPG, JP2, PNG, etc.), a list of metadata commonly stored in each format (e.g. administrative or descriptive), and a description of how the metadata is stored/encoded in the image file. The group will describe the ability of each metadata model to store multiple types of metadata and the formatting of that metadata as defined by the Content Group, and draft recommendation on technologies to use in the short-term and mid-term as well as suggest technologies to research for long-term solutions.

### 5.3 Workflow Group

The Workflow Group will be responsible for drafting requirements and designing model workflows that incorporate the information and requirements from the Content and Embedding Groups. The group will draft requirements for extracting metadata from various sources and writing that metadata into still image files to accommodate various production models, including in-house, outsourced, and mixed manufacturing. The model workflows must accommodate the embedded linking from within the image file to an authoritative or master record (if such a record exists and can be linked to in a practical manner). The group will define requirements for a system which will have the capacity to confirm that required metadata exists and recognize whether that metadata is properly formatted and encoded. The group will also define requirements for tools to perform the following (all refer to both single-file or batch mode): embed, delete, correct/reformat, append/concatenate, and overwrite metadata, and to migrate metadata from master image files to derivative image files. Finally, the group will derive metrics on metadata embedded, supplemented, updated or corrected.

## 6 Process

The first phase of the project will focus on identifying existing guidelines and practices, primarily among participating agencies, but also among other recognized institutions in the field, and evaluating the relevance and strength of those guidelines.

The discovery process will begin with a comparative analysis of existing guidelines published by the agencies participating in this project. Differences between these guidelines documents will be identified and documented.

In addition to this cross-agency comparison, the existing guidelines documents will be evaluated against a hypothetical ideal, with gaps identified and documented. Documented differences and gaps will be evaluated against scope, and then assigned priorities by the working group as project tasks to be shared among agencies.

The Embedded Metadata Sub-group will follow the same processes for Conceptual Framework, Guidelines Development, Review, Finalization, Updates, and Communication outlined in Still Image Working Group Charter.

## 7 Deliverables

- Vocabulary – Identify or establish vocabulary to describe and differentiate terminology describing metadata that falls within the scope of this activity
- Use Cases – Solicit use cases to more clearly identify objectives and needs

- TIFF Tags – Published draft and final guidelines for minimum core TIFF metadata represented in TIFF header tags. The standard will define required tags and format of values written to those tags
- Descriptive Metadata – Published draft and final guideline detailing the minimum type and set of descriptive metadata content to be embedded in still image, including embedded passive or active reference to master or authoritative metadata stores, referred to as "persistent reference pointers." Define categories of embedded metadata
- Common Metadata Schema – Research results and recommendations whether to adopt a common image metadata schema, and if so, the recommended schema.
- Describe current formats/models (e.g. TIFF Header Tags, EXIF, IPTC, XMP) used to embed metadata in still image files, including:
  - A brief overview, history and current usage (in cultural heritage digitization as well as born digital)
  - A list of image file formats supported by metadata models (e.g. TIFF, JPG, JP2, PNG, etc.)
  - A list of metadata commonly stored in each format (e.g. administrative or descriptive)
  - A description of how the metadata is stored/encoded in the image file
- Describe the ability of each metadata model to store multiple types of metadata and the formatting of that metadata as defined by the Content Group
- Draft recommendation on technologies to use in the short-term and mid-term as well as suggest technologies to research for long-term solutions
- Applications/Tools – Research and recommendations. Support must take into account the technical infrastructure (operating systems) of all NDSAB participating agencies and institutions
- Draft requirements for extracting metadata from various sources and writing that metadata into still image files as defined by the Content and Embedding Groups
- Design model workflow(s) that incorporate the information and requirements from the Content and Embedding Groups to accommodate various production models, including in-house, outsourced, and mixed manufacturing. The model workflows must accommodate the embedded linking from within the image file to an authoritative or master record (if such a record exists and can be linked to in a practical manner).
- Define requirements for a system which will have the capacity to:
  - Confirm that required metadata exists and
  - Recognize whether that metadata is properly formatted and encoded
- Define requirements for tools to perform the following (all refer to both single file or batch mode):
  - Embed metadata
  - Delete metadata
  - Correct/reformat metadata
  - Append/concatenate metadata
  - Overwrite metadata
  - Migrate metadata from master files to derivative files
- Derive metrics on metadata embedded, supplemented, updated or corrected.

## 8 Timeline

- TIFF Tag Guidelines Completed January 9, 2009
- Standardized Vocabulary Completed February 10, 2009
- Use Cases Completed March 5, 2009
- Base Embedded Content and Format Completed April 17, 2009

Target dates for other milestones will be developed during the course of activities through common agreement of participating agencies.

## 9 References

Adobe Systems Incorporated. "Extensible Metadata Platform (XMP) Specification." September, 2005. [http://www.adobe.com/devnet/xmp/pdfs/xmp\\_specification.pdf](http://www.adobe.com/devnet/xmp/pdfs/xmp_specification.pdf)

Federal Agencies Still Image Digitization Working Group. "Still Image Digitization Working Group Charter." July, 2008. <http://www.digitizationguidelines.gov/stillimages/charter.html>

----. "TIFF Metadata: Recommended Elements and Format (Draft)." September, 2008. [http://www.digitizationguidelines.gov/stillimages/documents/TIFF\\_Metadata\\_DRAFT-ms2.pdf](http://www.digitizationguidelines.gov/stillimages/documents/TIFF_Metadata_DRAFT-ms2.pdf)

International Press Telecommunications Council. "IPTC Core Schema for XMP, Version 1.0." March, 2005. [http://www.iptc.org/std/Iptc4xmpCore/1.0/specification/Iptc4xmpCore\\_1.0-spec-XMPSchema\\_8.pdf](http://www.iptc.org/std/Iptc4xmpCore/1.0/specification/Iptc4xmpCore_1.0-spec-XMPSchema_8.pdf)

Japan Electronic Industry Development Association. "Digital Still Camera Image File Format Standard (Exchangeable image file format for Digital Still Camera : Exif), Version 2.1." December, 1998. <http://it.jeita.or.jp/document/publica/standard/exif/english/Exife.pdf>

Metadata Working Group. "Guidelines For Handling Image Metadata, Version 1.0." September, 2008. [http://www.metadataworkinggroup.org/pdf/mwg\\_guidance.pdf](http://www.metadataworkinggroup.org/pdf/mwg_guidance.pdf)

## 10 Version Control and Change History

Version	Date	Change	Change documentation
1.0	November 26, 2008	Baseline charter	