The Case for Implementing Core Descriptive Embedded Metadata at the Smithsonian

Stephanie Ogeneski Christensen, National Anthropological Archives, Smithsonian Institution; Washington, D.C., USA
Doug Dunlop, Smithsonian Institution Libraries; Washington, D.C., USA

Abstract

The long-term goal, as established by the institutional strategic plan, to digitize collections at the Smithsonian Institution along with the increasing need to share data and increase access to collections has made it essential to establish institution-wide metadata standards, including those for embedding metadata. This paper documents the ongoing process of establishing core embedded metadata within the institution through the work of the Smithsonian Embedded Metadata Working Group (EmDaWG), which is pan-institutional in nature and includes museums, libraries, archives, and research institutes. The focus of the working group described within this paper is the creation of core embedded metadata fields for use in still images.

History and Background

As the Smithsonian Institution (SI) begins to digitize and make its digital assets accessible, it begins to build a digital foundation for its collections and research. Accessing this digital foundation will be crucial to the institution. As system changes are implemented, the ability to search and readily locate these digital assets becomes a key component. As SI begins to move forward in creating these digital assets and pushing them out to the public, it is at this initial stage that we must also have the systems search and retrieval functions in mind. The Smithsonian Embedded Metadata Group (EmDaWG) was established in April 2009 in response to a growing request from staff who work daily with digital assets to clarify practices of embedding metadata in still images. It originally formed within one of the many Collection Information System (CIS) meetings across the institution. The group participation grew to include a range of representatives utilizing various content management and local database systems, therefore representing a broad range of users and types of data from across the institution. The group's goal was to establish a core set of minimal embedded metadata fields and provide best practices for implementation. It was created with the intention not to dictate practice, but rather to educate and provide guidance for those working with digital still image and born digital still image collections. The group felt that a commonly understood and implemented metadata terminology would assist in the implementation of a metadata management system that would minimize confusion in retrieving assets, readily identify which unit and collection within that unit the still image came from, and also inform the user that the still image may have a copyright restriction and that contacting the particular holding unit of the digital asset may be appropriate before use. The resulting document entitled: Basic Guidelines for Minimal Descriptive Embedded Metadata in Digital Images can be obtained through the Smithsonian Research Online http://hdl.handle.net/10088/9719.

Case for Core Embedded Metadata

Historically metadata describing the content and properties of an information object has been stored separately from the information object itself. The current information landscape is evolving with the deluge of digital materials and increasingly metadata travels with the information object in the form of embedded metadata. Embedded metadata is contained within the structure of a digital file usually at the beginning or end of a file and can be technical, administrative, or descriptive in nature. Having data embedded within a file using standards like International Press Telecommunications Council (IPTC) http://www.iptc.org/std/photometadata/specification/IPTC-PhotoMetadata%28200907%29_1.pdf and Extensible Metadata Platform (XMP) http://www.adobe.com/devnet/xmp/ makes a file self describing so that the file can be identified and described outside of its home system. XMP is a metadata platform developed by Adobe that employs the RDF (Resource Description Framework) http://www.w3.org/RDF/ data interchange model, and utilizes existing schema like Dublin Core and IPTC. IPTC as a standard dates back to the 1970s as a means of protecting the exchange of news images and information (IPTC, 2010). Using existing standards for embedded metadata, whether in the form of descriptive, technical, structural or administrative, can aid in searchability, provenance, rights management, interoperability, and data repurposing.

In the case of the Smithsonian, the use of embedded metadata (often in the form of IPTC added through the use of Adobe products) is increasing throughout the institution to support a number of workflows and needs ranging from born digital photographs taken by SI photographers for Folkways to digitized images of archival materials from the National Anthropological Archives. Until just recently, these and other still images have been collected in local databases and content management systems and described using local practices, with some units employing embedding. Increased digitization at SI, has created the consequent need for centralized searching of collections and the inherent need for long-term storage of digital materials. In order to meet these ever increasing needs, a Collections Search Center (CSC) http://collections.si.edu/search/ and a Digital Access Management System (DAMS) were created. The CSC is one stop searching based on a central index, with data pulled from and linked to a number of collections, which employs an index metadata model based on standards and internal data needs, while the DAMS is intended for the long-term storage of digital assets and uses a core set of metadata created through a pan-institutional committee process.

It was in this environment that the need for the Embedded Metadata Working Group was born. Several initial questions arose from the discussions of the CSC and DAMS including: could

images with only embedded metadata be used within both the CSC and the DAMS, and if so what would those core fields be? While the questions seemed straightforward the answers were not as easy to arrive at. One of the reasons that the questions were asked in the first place is that many SI image collections only have collection level metadata or basic item level metadata embedded within the images with the possibility of collections having images that number in the tens of thousands or greater. With only embedded metadata to retrieve the images, SI units wanted to make sure that they were including the most useful and consistent metadata. At the time that EmDaWG formed, it was known that embedded metadata could be used within the DAMS for image retrieval and within the CSC for record creation.

Core Fields

The effort of determining what core fields to use was a pan-institutional effort within the working group. This process included conducting a survey of the working group as to what embedded metadata fields respective units were using as well as conducting in depth meetings in which workflows, data needs, and data usage were discussed. Informing this discussion were the findings from the Federal Agencies Digitization Guidelines Initiative (FADGI) Still Image Working Group http://www.digitizationguidelines.gov/still-image/. Significant findings that emerged from this entire process, which informed the SI embedded working group, include: variations in use of embedded metadata in current practice, need for defining specific metadata fields to meet a variety of needs, importance of file name/identifier, copyright, and providing source/credit. The end result of the SI working group was the creation of two groupings of embedded metadata, one being the core group and the other being the suggested group. The core set contains fields that are consistent across Smithsonian units, while the suggested set of fields accommodates for the different, various needs across the institution's units. The suggested set might be useful for one unit, but not for another, therefore we determined these fields to be suggested and not core. In addition, the working group came up with best practices and recommendations for data input. The working group determined to use and follow IPTC fields and definitions as closely as possible, and to insure that the fields and data work well with the SI DAMS core metadata. During the process of crafting the document, some of the issues confronted by the group for employing embedded metadata include: the various institutional uses of the IPTC Title field either as a descriptive title or file identifier; inability of using commas within IPTC Keyword and Creator fields due to the fact that commas become delimiters when embedding using Adobe products; the conflicting use of Date Created to refer to the creation date of the digital object or as creation date of the original object; and the conflicting use of Creator as creator of the digital object or as creator of the original object. It was with these needs in mind, along with considerations for internal use and workflow, that the core fields were developed.

TABLE 1: Required: Core Set of Embedded Metadata

Element Name	Definition	Sample Data Values and Notes
Document Title	File number, Accession Number, Catalog Number, Digital File Name, Negative Number, Unique Identifier root level etc.	1) LB016021-a 2) 08596201 3) gn_03644 4) landes_photo_arizona_16 5) 23456.000 Notes: 5) example of a catalog number
Copyright Notice	Copyright Notice	1) The Smithsonian continues to research information on its collections. Contact Smithsonian for current status. 2) This image is in the public domain. 3) Copyright National Anthropological Archives, Smithsonian Institution. 4) Copyright William M. Groethe. Notes: Smithsonian staff should provide accurate copyright information particularly if the copyright status is known. The following default statement should only be used if the unit does not know the copyright status of the work. "The Smithsonian continues to research information on its collections. Contact Smithsonian for current status." The committee recognizes that copyright data may change, however it is of such importance that it is better to put in data if known.
Source	Name and Abbreviation of SI owning unit, Smithsonian Institution	1) NMAI-Natl. Museum of the American Indian, Smithsonian Institution 2) NAA- Natl. Anthropological Archives, Smithsonian Institution
Creator	Creator of digital	1) Smithsonian Institution Libraries

	object or Creator of original object	2) Department of Anthropology 3) National Anthropological Archives 4) Photographer Name 5) Cynthia Frankenburg 6) William Greene 7) Woody Guthrie Notes: IPTC Creator Job Title field can be used to define the role of the creator. 4) If photographer is not known then default to department name (refrain from using acronyms) *5)* If Creator = personal name then Creator Job Title field is populated e.g. Creator Job Title = Photographer 6) If Creator = personal name then Creator Job Title field is populated e.g. Creator Job Title = Scanner 7) Creator original object

TABLE 2: Not Required: Suggested Set of Embedded Metadata

Element Name	Definition	Sample Data Values and Notes
Date	Date of creation of original object or Date of creation of digital object	1) 07/23/1967 2) 02/14/2009 Notes: Date structure for IPTC is MM/DD/YYYY. Description field can be used if date range or other date structure is being used.
Description	Free narrative text about the object	1) 123457.000 (right); 123458.000 (left) 2) Cultural Resources Center 2007 Powwow Open House, CRC Open House, CRC Exterior, Chief Joseph, Nez Perce 3) scan from 4x5 CT, slide or B&W negative 4) media of original art (oil on canvas, watercolor, etc.) Notes: 1) This field is used to locate individual objects within an image that contains multiple objects.
Keywords	Free text field but should be used to store a list of standard term(s) separated by a common delimiter such as semicolon.	1) Lighting Archive; Electrification; Lighting; Lighting Fixtures; Architectural History; History of Architecture 2) Object; Publication; Our Lives 3) Viento de Agua; plena; bomba 4) rugby; Wales; sports Notes: 5) Taken from defined look-up list in database: Object=image of object in the collection; Publication connotes quality suitable for publication; Our Lives connotes image for Our Lives exhibition.
Credit/Provider	What you would like to accompany the image in a publication. Ex: Image Number, SI owning Unit, Smithsonian Institution	1) Image Number, National Anthropological Archives, Smithsonian Institution 2) Papers of Ruth Landes, National Anthropological Archives, Smithsonian Insitution 3) Ken Rahaim, Smithsonian Institution 4) Photographer's name, museum (i.e. National Portrait Gallery)
Job Identifier	Instructions or unit id for a job	MSC07-04608
Headline	(Formally called Caption)descriptive title or a caption	Caldwell and Company Collection

Institutional Use

Part of the Smithsonian's mission is the "increase and diffusion of knowledge." As digitization increases across the institution, new ways of learning about and modes of discovery of the collections and vast resources that the Smithsonian holds will begin to take place. Users in today's world will expect to have greater access and search functions on the institution's digital assets. They may want to relate one collection in the National Anthropological Archives with holdings in the Smithsonian Institution Libraries. A publication in the Smithsonian Institution Libraries may provide a researcher with a lead to a collection at the Smithsonian Institution Archives. The list of possibilities and roadmaps is endless. As these discoveries take place, the ability to readily determine which digital assets are available for use, which have restrictions, or which assets belong to a particular unit will become even more necessary. While the CSC and DAMS were important driving forces for the development of core embedded metadata, other institutional needs were also at play including the dispensing of still image resources to researchers, the general public, and for licensing purposes; and the increased dissemination of still images via social networking sites. Determining core metadata became essential so that the distributed and disseminated still images could be: described as to what they were of, identified as to the owning institution, and provide necessary rights information.

Future Developments

While this has proven to be a productive collaboration, we have recognized that there still are challenges within our own institution. Defining a core set of fields that can accommodate a broad range of materials was not an easy undertaking. Our core model suggests elements that are *consistent* across SI, while our suggested set of elements was established to accommodate the needs that one unit may have over another. As the group begins to move forward with the core and suggested set, we hope to run an XMP pilot. Also, some units that are currently embedding metadata are beginning to test the capabilities of the CSC and DAMS systems with new ways to extract and deliver embedded metadata. For example, the National Anthropological Archives will be working on extracting embedded metadata from images within the DAMS to help assist in creating MARC records in the CSC system. As we begin to establish a consensus on still images, the analog holdings of the institution's materials ranging from scientific data, video, audio, to PDF will require its own attention. While some fields may be consistent across formats, we recognize that the needs for core embedded metadata will be disparate. The group hopes to address each format over the course of the year, modifying the group membership so that units holding or working with these formats will actively contribute to the development of basic guidelines. Also recognizing that technology changes rapidly and what we decide on today may not be what we need in the future, the group recommends that core descriptive embedded metadata be reviewed every two years to stay current with best practices. At the of the publication of the embedded metadata document, Adobe had just released the latest version of their Creative Suite software package (CS5), which includes recent updates to the IPTC schema that in large part address some of the data input issues the various members of the working group encountered (i.e. Creator and Date Created). The working group determined that these changes would not be reflected in the document until CS5 is more widely adopted and evaluated institutionally.

Acknowledgements

The following members of EmDaWG were instrumental in the creation of the SI embedded metadata core: Stephanie Ogeneski Christensen (co-leader), Doug Dunlop (co-leader), Lowell Ashley, Ricc Ferrante, Cindy Frankenburg, Ducky Nguyen, Kay Peterson, Suzanne Pilsk, Ken Rahaim, Marguerite Roby, Erin Rushing, Stephanie Smith, Rebecca Snyder, Amy Staples, Sarah Stauderman, Patti Williams

References

[1] Adobe. (2008). Part 1, Data and Serialization Models. Retrieved from http://www.adobe.com/devnet/xmp/.

[2] Adobe. (2008). Part 2, Standard Schemas. Retrieved from http://www.adobe.com/devnet/xmp/.

[3] Adobe. (2008). Part 3, Storage in Files. Retrieved from http://www.adobe.com/devnet/xmp/.

[4] Federal Agencies Digitization Guidelines Initiative. (2010). Still Image Working Group, Embedded Metadata. Retrieved from http://www.digitizationguidelines.gov/stillimages/subcommittees.html.

[5] International Press Telecommunications Council. (2009). IPTC Photo Metadata Standard specification released, including IPTC Core 1.1 and IPTC Extension 1.1. Retrieved from http://www.iptc.org/cms/site/index.html?channel=CH0086.

[6] Smithsonian Embedded Metadata Group. (2010). Basic Guidelines for Minimal Descriptive Embedded Metadata in Digital Images. Retrieved from http://hdl.handle.net/10088/9719.

[7] Smithsonian Institution. (2010). Collections Search Center. Retrieved from http://collections.si.edu/search/.

[8] World Wide Web Consortium. (2004). Resource Description Framework (RDF). Retrieved from http://www.w3.org/RDF/.

Author Biography

Stephanie Ogeneski Christensen is the Digital Imaging Manager at the National Anthropological Archives, Smithsonian Institution. Prior, she served as the manager of the digital imaging facility at the Chicago Albumen Works. She is a member of the Federal Agencies Digitization Guidelines Initiative Still Image Working Group. She holds a Certification in Photographic Preservation and Archival Practice from the George Eastman House (1998) and an MFA from Indiana University (1997).

Doug Dunlop is a Metadata Librarian at the Smithsonian Institution Libraries. Prior to that he was a Metadata Cataloger at the University of Central Florida, Orlando and worked on the Central Florida Memory project. He is a member of the Federal Agencies Digitization Guidelines Initiative Audio Visual Working Group. He holds a MLIS from the University of North Texas (2005) and a MA in Art History from the University of North Texas (1997).

Appendix

TABLE 3: Mapping of SI Embedded Fields to Dublin Core and IPTC

SI Embedded Metadata (Core Required Fields)	IPTC	Dublin Core	Sample Data Value
Document Title	Title	Identifier	1) LB016021-a 2) 08596201 3) gn_03644 4)landes_photo_arizona_16 5) 23456.000
Copyright Notice	Copyright Notice	Rights	1) The Smithsonian continues to research information on its collections. Contact Smithsonian for current status. 2) This image is in the public domain. 3) Copyright National Anthropological Archives, Smithsonian Institution. 4) Copyright William M. Groethe.
SI Embedded Metadata (Core Required Fields)	IPTC	Dublin Core	Sample Data Value
Source	Source	Publisher	1) NMAI-Natl. Museum of the American Indian, Smithsonian Institution 2) NAA- Natl. Anthropological Archives, Smithsonian Institution
Creator	Creator	Creator	1) Smithsonian Institution Libraries 2) Department of Anthropology 3) National Anthropological Archives 4) Photographer Name 5) Cynthia Frankenburg 6) William Greene 7) Woody Guthrie
SI Embedded Metadata (Suggested Not Required Fields)			
Date	Date Created	Date Created	1) 07/23/1967 2) 02/14/2009
Description	Description	Description	1) 123457.000 (right); 123458.000 (left) 2) Cultural Resources Center 2007 Powwow Open House, CRC Open House, CRC Exterior, Chief Joseph, Nez Perce 3) scan from 4x5 CT, slide or B&W negative 4) media of original art (oil on canvas, watercolor, etc.)
Keywords	Keywords	Subject	1) Lighting Archive; Electrification; Lighting; Lighting Fixtures; Architectural History; History of Architecture 2) Object; Publication; Our Lives 3) Alice Fletcher; Francis La Flesche 4) Viento de Agua; plena; bomba 5) rugby; Wales; sports 6) name of exhibit object is photographed for inclusion in
Credit /Provider	Provider	Publisher	1) NMAI-Natl. Museum of the American Indian, Smithsonian Institution 2) NAA- Natl. Anthropological Archives, Smithsonian Institution
Job Identifier	Job Identifier	Identifier	1) LB016021-a 2) 08596201 3) gn_03644 4) landes_photo_arizona_16 5) 23456.000
Headline	Headline	Title	Caldwell and Company Collection