Metadata

Commonly defined as "data about data," metadata constitutes the documentation of all aspects of digital files essential to their persistence and usefulness and should be inextricably linked to each digital image. Metadata is captured in the form of a prescribed list of elements, or fields, known as a metadata schema. Each schema is tailored to a specific domain or purpose (see *Selecting a Metadata Schema*). Metadata schemas should be distinguished from metadata format: information in any given metadata schema can be formatted in a number of different ways—as a physical card-catalogue entry, a set of fields for a **database** or collection management system record, or an **XML** (Extensible Markup Language) document—and still convey the same meaning. The advantage of the increasingly popular XML format is that it provides a mechanism for encoding semantic data, as defined by a metadata schema, that facilitates data sharing and automated data processing. (For a more detailed discussion of the XML format, see *Metadata Format*.) However, the quality and consistency of metadata are more important than the particular format in which it is expressed or the software used to generate it: bad metadata in a sophisticated **native-XML** database will be less valuable than good metadata in a simple desktop spreadsheet, which can likely be migrated to new formats if need be. "Good" metadata was defined in 2001 by the Digital Library Forum as fulfilling the following criteria: it is appropriate to the materials digitized and their current and likely use; it supports interoperability; it uses standard **controlled vocabularies** to populate elements where appropriate; it includes a clear statement on the terms of use of the digital object; it supports the long-term management of digital objects; and it is persistent, authoritative, and verifiable. [1]

The depth and complexity of metadata captured will vary from one project to another depending on local policies and user needs, but images without appropriate metadata will quickly become useless—impossible to find, view, or migrate to new technology as this inevitably becomes necessary. It is metadata that allows collection managers to track, preserve, and make accessible digital images and enables end users to find and distinguish between various images. Metadata also allows digital image collections to be reused, built upon, and become part of larger cultural heritage offerings within and across institutions.

Metadata can be divided into three broad types, which may be simply defined as follows: descriptive, which describes content; administrative, which describes context and form and gives data-management information; and structural, which describes the relationships between parts and between digital files or objects.[2] The first is most akin to traditional cataloguing and would describe what a digital image depicts or its essence. This is essential for end-user access and to allow efficient search and retrieval. Administrative metadata records information such as how and why a digital object was created, and is used in the management of digital objects. Structural metadata documents information such as the fact that one image is a detail of the upper-right corner of another image, that a particular image depicts page two of a book of thirty-four pages, or that an image is one item in a given series.

Metadata may also be divided into more specific categories: for instance, rights metadata is used in **rights management** and documents with whom the intellectual property (**IP**) rights of a particular image or collection reside and describes access and usage restrictions. It may specify at what quality an image may be reproduced or the credit line that is required to accompany its display. Technical metadata documents aspects of files or collections that are distinct from their intellectual content or essence, such as production, format, and processing information. Preservation metadata documents the information necessary to ensure the longevity of digital objects. There is obvious crossover among these categories; preservation metadata is primarily a combination of administrative and structural metadata elements, or alternatively is primarily a subset of technical metadata.

It is important in this context to mention **CBIR**, or content-based information retrieval. CBIR technology is able to retrieve images on the basis of machine-recognizable visual criteria. Such indexing is able to recognize and retrieve images by criteria such as color, iconic shape, or by the position of elements within the image frame. Stock-photo houses that cater to the advertising industry have had some success in using this automatic indexing to answer such queries as "find images with shades of blue in the top part of the frame and shades of green in the bottom part" (i.e., landscapes). As this technology becomes more sophisticated, it is likely that automatic and manual indexing will be used together to describe and retrieve images.

Metadata Crosswalks and Controlled Vocabularies

To make different metadata schemas work together and permit broad cross-domain resource discovery, it is necessary to be able to map equivalent or nearly equivalent elements from different schemas to each other, something that is achieved by metadata **crosswalks**. The Getty Research Institute and the Library of Congress offer crosswalks between various metadata schemas, and UKOLN (UK Office for Library and Information Networking) maintains a Web page linking to a variety of crosswalk and metadata mapping resources. When integrated into search software, such crosswalks allow the retrieval of diverse records contained in different repositories and aid the migration of data to new systems. (See Metadata Format for a discussion of the use of **RDF**—Resource Description Framework—in Web crosswalks.)

Crosswalks are only a part of a coherent data structuring. Controlled vocabularies, **thesauri**, **authorities**, and **indices** provide accurate and consistent content with which to populate metadata elements. Their use improves searching precision and recall and enables automated interoperability. For example, a streamlined arrangement of the totality of data describing an image file might include a distinction between intrinsic and extrinsic information, the latter being ancillary information about persons, places, and concepts. Such information might be important for the description and retrieval of a particular work but is more efficiently recorded in separate "authority" records than in records about the work itself. In this type of system, such information is captured once (authoritatively) and may be linked to all appropriate work records as needed, thus avoiding redundancy and the possible introduction of error.

Examples of controlled vocabularies include the *Library of Congress Subject Headings* (**LCSH**), *Name Authority File* (**NAF**), and *Thesaurus for Graphic Materials I* and *II* (**TGM-I** and **TGM-II**), all maintained by the Library of Congress; the *Art & Architecture Thesaurus* (**AAT**), the *Getty Thesaurus of Geographic Names* (**TGN**), and the *Union List of Artist Names* (**ULAN**), all maintained by the Getty Research Institute; and **ICONCLASS**, a subject-specific international classification system for iconographic research and the documentation of images. These and other vocabularies and classification systems—many disciplines and professions have developed their own thesauri, tailored to their particular concerns—provide a wide range of controlled terminology to describe the people, places, things, events, and themes depicted in images, as well as the original objects themselves.