
Introduction to Metadata
Third Edition
Edited by Murtha Baca
MARC
Ontology
BIBFRAME
VRA Core

Introduction

Murtha Baca

Like metadata itself, the realm of online resources is constantly and rapidly growing and evolving. Much has changed in the digital information landscape since the first print edition of this book was published in 1998, followed by revised editions in 2000 and 2008. The time is right for an updated edition of this text, intended to give a general introduction to metadata and to explain some of the key tools, concepts, and issues associated with using metadata to build authoritative, reliable, and useful digital resources. In the last few years, phenomena such as linked open data have begun to play an important role in the Semantic Web; the standard for library cataloging, Anglo-American Cataloging Rules, has been largely replaced by the Resource Description and Access (RDA) standard; and the BIBFRAME linked open data standard is poised to become the successor to the venerable MARC format for encoding bibliographic metadata.

Metadata creation is—or should usually be—a collaborative effort, as is this publication. Anne Gilliland, the late Mary Woodley, and Maureen Whalen updated their chapters, and with the help of several colleagues, I updated Tony Gill's chapter on metadata and the web. The fact that this publication is the result of several people working together is significant—and indicative of how we work today.

In the first chapter Anne Gilliland provides an overview of metadata—its types, roles, and characteristics—as well as facts about metadata that belie several common misconceptions. She also addresses recent trends in metadata creation, particularly that of metadata created by users rather than by trained information professionals. Activities such as social tagging, social bookmarking, and the resulting forms of user-created metadata such as "folksonomies" are playing an increasingly important role in the realm of digital information.

Chapter 2 focuses on metadata as it relates to resources on the web. We explain how web search engines work and how they use metadata, data, links, and relevance ranking to help users find what they are seeking. We also discuss in detail the commercial search engine that, as of this writing, has dominated the web for several years: Google. A key concept in this chapter is the difference between the visible web and the hidden web and the important implications and issues related to making resources reachable from commercial, publicly available search engines versus systems that have one or more "barriers" to

access—either because they are fee based, password protected, or require a particular IP address, or simply because they are not technically exposed to commercial search engines. How library metadata behaves in the era of Google dominance is also addressed.

In the third chapter, Mary Woodley examines the methods, tools, standards, and protocols that can be used to publish and disseminate digital collections in a variety of online venues. She shows how "seamless searching"—integrated access to a variety of resources residing in different information systems and formulated according to a range of standard and nonstandard metadata schemes—is still far from a reality. Woodley contrasts the method of "federation" by means of building union catalogs of digital collections by aggregating metadata records from diverse contributors into a single database with "metasearching"—real-time searching of diverse resources that have not been aggregated but rather are searched in situ by means of one or more protocols. Each method requires specific skills and knowledge; particular procedures, protocols, and data standards; and the appropriate technical infrastructure. Creating union resources via physical aggregation of metadata records or via metadata harvesting is a good thing, but we should keep in mind that it does not necessarily solve the hidden web problem enunciated in chapter two. If resources are publicly available but users cannot reach them from Google and instead have to find the specific search page for a particular union resource, we cannot say that we have provided unfettered access to that resource. Woodley also stresses the importance of data value standards—controlled vocabularies, thesauri, lists of terms and names, and folksonomies—for enhancing end-user access. She points out that mapping metadata elements alone is not sufficient to connect all users with what they seek; the data values—that is, the vocabularies used to populate those metadata elements—should also be mapped.

Maureen Whalen's chapter, "Rights Metadata Made Simple," argues that the research and capture of standards-based rights metadata should be essential activities of memory institutions and offers practical, realistic options for determining and recording core rights metadata. If institutions would commit the effort and resources to following Whalen's advice, many of the obstacles to unfettered end-user access could be surmounted.

In the section on "Practical Principles for Metadata Creation and Maintenance," we emphasize that institutions need to change old paradigms and procedures. Libraries, archives, museums, and other memory organizations need to make a lasting commitment to creating and continually updating the various types of core metadata relating to their collections and the digital

surrogates of collection materials that we all seem to be in such a hurry to create and make available online.

Our updated glossary is not intended to be comprehensive; rather, its purpose is to explain the key concepts and tools discussed in this publication. The footnotes in each of the chapters provide additional references to publications and online resources relevant to the topic of metadata and digital libraries.

At the end of her chapter, Anne Gilliland compares metadata to an investment that, if wisely managed, can deliver a significant return on intellectual capital. I would venture to expand on her financial metaphor and say that metadata is one of our most important assets. Hardware and software come and go—sometimes becoming obsolete with alarming rapidity—but high-quality, standards-based, system-independent metadata can be used, reused, migrated, and disseminated in any number of ways, even in ways that we cannot anticipate at this moment (as in the case of linked open data, which is a relatively recent concept).

Digitization does not equal access. The mere act of creating digital copies of collection materials does not make those materials findable, understandable, or utilizable to our ever-expanding audience of online users. But digitization combined with the creation of carefully crafted metadata can significantly enhance end-user access—and our users are the primary reason we create digital resources.

In closing, I would like to dedicate this publication to my friend and colleague Mary Woodley, a consummate librarian and metadata expert. Mary's revised chapter, which she completed during what would be the last months of her life, is a testament to her deep knowledge of metadata and controlled vocabularies, her love of libraries, and her vocation to connect users to the information they seek.

Setting the Stage

Anne J. Gilliland

Metadata, literally "data about data," is today a widely used, yet frequently underspecified term that is understood in different ways by the diverse professional communities that design, create, describe, preserve, and use information systems and resources. Until the mid-1990s, *metadata* was a term used primarily by communities involved with the management and interoperability of geospatial data and with data management and systems design and maintenance in general. For these communities, the term referred to a suite of industry or disciplinary standards as well as additional internal and external documentation and other data necessary for the identification, representation, interoperability, technical management, performance, and use of data contained in an information system.

As a construct, however, metadata has been around for as long as humans have been organizing information, albeit transparently in many cases. Today, we create and interact with it in increasingly digital and overt ways. For more than a century, and particularly since the first developments of national and international descriptive standards, the creation and management of metadata was primarily the responsibility of information professionals engaged in cataloging, classification, and indexing; but as more information resources were created or put on line and networked—especially via the web—by the general public, metadata considerations were no longer solely the province of information professionals. Although metadata is arguably a less familiar term among creators and consumers of networked digital content who are not information professionals per se, those same individuals are increasingly adept at creating, exploiting, and assessing user-contributed metadata such as title, description, and keyword tags for web pages; terms from so-called folksonomies; and social bookmarks. Schoolchildren, college students, and adult learners are taught in information literacy programs to look for metadata such as provenance and date information in order to ascertain the authoritativeness of information they retrieve on line. Others are using tag clouds and tag graphs to visualize the terminology and structures being used in metadata for selective information resources. Thus it has become more important than ever that not only information professionals but also other creators and users of digital content understand the critical roles and potential uses of different types of metadata in ensuring accessible, authoritative,

interoperable, scalable, and preservable cultural heritage information and record-keeping systems.

Perhaps a more useful, "big picture" way of thinking about metadata is as the sum total of what one can say at a given moment about any *information object* at any level of aggregation.[1] In this context, an information object is anything that can be addressed and manipulated as a discrete entity by a human being or an information system. The object may be a single item, an aggregate of many items, or an entire database or record-keeping system. Indeed, in any given instance one can expect to find metadata relevant to any information object existing simultaneously at the item, aggregate, and system levels.

In general, all information objects, regardless of the physical or intellectual form they take, have three features—content, context, and structure—all of which can and should be reflected through metadata:

- *Content* relates to what the object contains or is about and is *intrinsic* to an information object.
- *Context* indicates the who, what, why, where, and how aspects associated with the object's creation and subsequent life and is *extrinsic* to an information object.
- *Structure* relates to the formal set of associations within or among individual information objects and can be *intrinsic*, *extrinsic*, or both.

All objects carry with them certain metadata that innately results from the circumstances of their creation, management, and use. However, cultural heritage information professionals such as museum registrars, library catalogers, and archival processors often apply the term *metadata* to the value-added information they create to arrange, describe, track, and otherwise enhance access to information objects and the physical items and collections related to those objects. Such metadata is frequently governed by community-developed and community-fostered standards and best practices in order to ensure quality, consistency, and interoperability. Our Typology of Data Standards (table 1) organizes these standards into categories and provides examples of each. Markup languages such as HTML and XML and a variety of schemas and metadata formats provide standardized ways to structure and express these standards for machine processing, publication, and implementation.

Table 1. A Typology of Data Standards

Type	Examples
Data *structure* standards (metadata element sets, schemas). These are "categories" or "containers" of data that make up a record or other information object.	MARC (Machine-Readable Cataloging) Fo Encoded Archival Description (EAD), BIE (Bibliographic Framework), Dublin Core M Element Set, Categories for the Description Works of Art, VRA Core
Data *value* standards (controlled vocabularies, thesauri, controlled lists). These are the terms, names, and other values that are used to populate data structure standards or metadata element sets.	Library of Congress Subject Headings, Nar Authority File, and Thesaurus for Graphic Materials; Getty Art & Architecture Thesau Union List of Artist Names (ULAN), and Thesaurus of Geographic Names; ICONCL Medical Subject Headings
Data *content* standards (cataloging rules and codes). These are guidelines for the format and syntax of the data values that are used to populate metadata elements.	Anglo-American Cataloguing Rules, Resou Description and Access, International Stan Bibliographic Description, Cataloging Cult Objects, Describing Archives: A Content S
Data *format/technical interchange* standards (metadata standards expressed in machine-readable form). This type of standard is often a manifestation of a particular data structure standard (see above), encoded or marked up for machine processing.	Resource Description Framework, MARC2 MARCXML, EAD XML DTD, METS, BIBFRAME, LIDO XML, Simple Dublin XML, Qualified Dublin Core XML, VRA XML

Note: This table is based on the typology of data standards articulated by Karim Boughida in his article "CDWA Cataloging Cultural Objects (CCO): A New XML Schema for the Cultural Heritage Community" in *Humanities Computers, and Cultural Heritage: Proceedings of the XVI International Conference of the Association for Hist Computing, 14–17 September 2005* (Amsterdam: Royal Netherlands Academy of Arts and Sciences,

Type	Examples

2005), http://www.dans.knaw.nl/nl/over/organisatie-beleid/publicaties/DANShumanitiescomputersandculturalheritageUK.pdf.

Library metadata development has been first and foremost about providing intellectual and physical access to collection materials. *Library metadata* includes indexes, abstracts, and bibliographic records created according to cataloging rules (i.e., data content standards, according to our typology) such as the Anglo-American Cataloguing Rules (AACR) and more recently Resource Description and Access (RDA) and data structure standards such as the MARC (Machine-Readable Cataloging) and BIBFRAME (Bibliographic Framework) formats, in combination with data value standards such as the Library of Congress Subject Headings (LCSH) or the Getty's *Art & Architecture Thesaurus* (AAT). Such bibliographic metadata has been systematically and cooperatively created and shared since the 1960s and made available to repositories and users through automated systems such as bibliographic utilities, online public access catalogs (OPACs), and commercially available databases. Today this type of metadata is created not only by humans but also in a variety of automated ways such as metadata mining, metadata harvesting, and web crawling.

Automation of metadata will inevitably continue to expand with the evolution and increased implementation of the Resource Description Framework (RDF), linked open data, and the Semantic Web, which are discussed later in this book.

A large component of archival and museum metadata creation activities has traditionally been focused on context. Elucidating and preserving context is what assists with identifying and preserving the evidential value of records and artifacts in and over time; it is what facilitates the authentication of those objects, and it is what assists researchers with their analysis and interpretation. *Archival and manuscript metadata* includes the products of value-added archival description such as finding aids, catalog records, and indexes. However, it also includes descriptive documentation generated in the course of creating, managing, preserving, using, and reusing both born-digital and digitized archival materials. Archival data structure standards that have been developed in the past three decades include the MARC Archival and

Manuscripts Control (AMC) format, published by the Library of Congress in 1984 (now integrated into the MARC21 format for bibliographic description); the suite of international descriptive standards anchored by the General International Standard Archival Description (ISAD [G]), first published by the International Council on Archives in 1994, that provide the basis for various national descriptive standards used around the world; Encoded Archival Description (EAD), adopted as a standard by the Society of American Archivists in 1999, and its companion data content standard, Describing Archives: A Content Standard (DACS), first published in 2004. The Metadata Encoding and Transmission Standard (METS), developed by the Digital Library Federation and maintained by the Library of Congress, is often used for encoding descriptive, administrative, and structural metadata and digital surrogates at the item level for objects such as digitized photographs, maps, and correspondence from the collections described by finding aids and other collection or group-level metadata records.

Many repositories make standardized descriptive metadata for library and archival collections available on line through resources such as WorldCat, the Digital Public Library of America, and ArchiveGrid.

Consensus and collaboration were slower to build in the museum community, where the benefits of standardization of description, such as shared cataloging and exchange of descriptive data, were less readily apparent until relatively recently. Since the late 1990s tools such as *Categories for the Description of Works of Art* (CDWA), the CIDOC Conceptual Reference Model (CRM), *Cataloging Cultural Objects: A Guide to Describing Cultural Works and Their Images* (CCO), the LIDO (Lightweight Information Describing Objects) XML schema, and more generic standards such as Dublin Core and METS have been considered and implemented by museums.

Although it would seem to be a desirable goal to integrate materials of different types that are related by provenance or subject but distributed across the repositories of museums, archives, and libraries, initiatives such as Museums and the Online Archive of California (MOAC) have met with limited success. As MOAC and the mid-1980s development of the now-defunct MARC AMC format have demonstrated, the distinctiveness of the various professional and object-based approaches (e.g., widely differing notions of provenance and collectivity as well as of structure), different institutional cultures, and divergent cultural approaches (e.g., those exemplified in indigenous protocols for archival and library materials) have left many professionals, and the communities they represent, feeling that their practices and needs have been shoehorned into structures that were developed by another community with

quite different epistemologies, practices, and users. As enunciated in principle 6 of "Practical Principles for Metadata Creation and Maintenance," there is no single metadata standard or set of standards that is adequate for describing all types of collections and materials. Selection of the most appropriate suite of metadata standards and tools—and creation of clean, consistent metadata according to those standards—will not only enable good descriptions of specific collection materials, but will also make it possible to map metadata created according to different community-specific standards, thus furthering the goal of interoperability discussed in subsequent chapters of this book.

An emphasis on the structure of information objects in metadata development by the library, archives, and museum communities has perhaps been less overt. However, structure has always been important in information organization and representation, even before computerization. Documentary and publication forms have evolved into industry standards and societal norms and have become almost transparent information management tools. For example, when users access a birth certificate they can predict its likely structure and content. When academics use a scholarly monograph, they understand intuitively that it will be organized with a table of contents, chapter headings, and an index. Archivists use the physical structure of their finding aids to provide cues to researchers about the structural relationships between different parts of a record series or manuscript collection. Archival description also exploits the hierarchical arrangement of records according to the bureaucratic structures, business practices, and personal systems of organization of the creators of those records. However, in recent years there has been increasing criticism that collection-level, hierarchical metadata as exemplified in archival finding aids, while valuable for retaining context and original order, represents an oversimplified view of the actual complexities of records-creation processes and provenance, privileges the scholarly user of the archive (and those who are familiar with the structure and function of archival finding aids) while leaving the non-expert user baffled, and unnecessarily perpetuates a paper-based descriptive paradigm.[2] In the online world, multiple descriptive relationships between objects can be supported simultaneously, and some of these, especially when used in addition to user-contributed metadata, may support new types of users and uses in an environment that is not mediated by a reference archivist. While concerned about reducing the amount of "overhead" involved in detailed metadata creation, archives and other collecting institutions are simultaneously exploring more granular methods of description, e.g., exploiting item-level metadata for digitized objects so that users can search for specific items, navigate through a collection "bottom up" as well as "top down," and collate

related collection materials through lateral searching across collections and repositories.

The role of structure in creating and exploiting machine-readable metadata has been growing as computer-processing capabilities become increasingly powerful and sophisticated. Information communities are aware that the more highly structured an information object is, the more that structure can be exploited for searching, manipulating, and interrelating with other information objects. Capturing, documenting, and enforcing that structure, however, can only occur if supported by specific types of metadata. In short, in an environment where a user can gain unmediated access to information objects over a network, metadata

- certifies the authenticity and degree of completeness of the content;
- establishes and documents the context of the content;
- identifies and exploits the structural relationships that exist within and between information objects;
- provides a range of intellectual access points for an increasingly diverse range of users; and
- presents some of the information that an information professional might have provided in a traditional, in-person reference or research setting.

But there is more to metadata than description and resource discovery. A more inclusive conceptualization of metadata is needed as we consider the range of activities that may be incorporated into digital information systems. Repositories also create metadata relating to the administration, accessioning, preservation, and use of collections. Acquisition records, exhibition catalogs, licensing agreements, and educational metadata are all examples of these other kinds of metadata and data. Integrated information resources such as virtual museums, digital libraries, and archival information systems include digital versions of actual collection content (sometimes referred to as digital surrogates) as well as descriptions of that content (i.e., descriptive metadata, in a variety of formats). Incorporating other types of metadata into such resources reaffirms the importance of metadata in administering collections and maintaining their intellectual integrity both in and over time. Paul Conway alluded to this capability of metadata when he discussed the impact of digitization on preservation: “The digital world transforms traditional preservation concepts from protecting the physical integrity of the object to

specifying the creation and maintenance of the object whose intellectual integrity is its primary characteristic."[3]

When applied outside the original repository, the term *metadata* acquires an even broader scope. An Internet resource provider might use *metadata* to refer to information that is encoded in HTML meta tags for the purposes of making a website easier to find. Individuals who are digitizing images might think of metadata as the information they enter into a header field for the digital file to record information about the image file, the imaging process, and image rights. A social science data archivist might use the term to refer to the systems and research documentation necessary to run and interpret a magnetic tape containing raw research data. A digital records archivist might use the term to refer to all the contextual, processing, preservation, and use information needed to identify and document the scope, authenticity, and integrity of an active or archival record in an electronic record-keeping or archival preservation system. Metadata is crucial in personal information management and digital archiving and for ensuring effective information retrieval and accountability in record-keeping—something that is becoming increasingly important with the rise of electronic commerce and the use of digital content and tools by governments. In all these diverse interpretations, metadata not only identifies and describes an information object; it also documents how that object behaves, its function and use, its relationship to other information objects, and how it should be and has been managed over time.

As this discussion suggests, theory and practices vary considerably due to the differing professional and cultural missions of museums, archives, libraries, and other information and record-keeping communities. Information professionals have a bewildering array of metadata standards and approaches from which to choose. Many highly detailed metadata standards have been developed by individual communities—e.g., MARC, BIBFRAME, EAD, LIDO, the Australian Recordkeeping Metadata Schema, and some of the standards for geographic information systems—that attempt to articulate their mission-specific differences as well as to facilitate mapping between common data elements. If used appropriately and to their fullest extent, these standards have the potential to create extremely rich metadata that provides detailed documentation of record-keeping creation and use in situations in which such activities may be challenged or audited for their comprehensiveness and accuracy.[4] Creation and ongoing maintenance of such metadata, however, is complex, time consuming, and resource intensive and may only be justifiable when there is a legal mandate or other risk-management incentive, or when it is anticipated that the content and metadata may be reused or exploited in previously unanticipated ways, such as in digital asset management systems. By

contrast, the Dublin Core Metadata Element Set (DCMES) identifies a relatively small, generic set of metadata elements that can be used by any community, expert or nonexpert, to describe and search across a wide variety of information resources on the World Wide Web. Such metadata standards are necessary to ensure that different kinds of descriptive metadata are able to interoperate with one other and with metadata from non-bibliographic systems of the kind that the data management communities and information creators are generating. Relatively lean metadata records such as those created using the DCMES have the advantage of being cheaper to create and maintain, but they may need to be augmented by other types of metadata in order to address the needs of specific user communities and to adequately describe particular types of collection materials.[5]

User-created metadata, both individually contributed and crowd sourced, has been gathering momentum in a variety of venues on the web. Just as many members of the general public have participated in the development of web content, whether by blogging on Tumblr or by uploading photos onto Flickr or videos onto YouTube, they have also been creating, sharing, copying, and mapping metadata. Among the advantages of these developments is that individual web communities such as affinity groups or hobbyists may be able to create metadata that addresses their specific needs and vocabularies in ways that information professionals who apply metadata standards designed to cater to a wide range of audiences cannot. Individuals and particular communities may also be using this capacity to offer corrections to the existing metadata, to "talk back" to the record, or to suggest how an object should be interpreted. User-generated metadata is also a comparatively inexpensive way to augment existing metadata, with the cost and the sense of ownership shared among more parties than just those who create information repositories. The disadvantages of user-generated metadata relate to quality control (or lack thereof) and idiosyncrasies that can impede the trustworthiness of both metadata and the resource it describes and negatively affect interoperability between metadata and the resources it is intended to describe. Issues of interoperability are discussed in some detail in the third chapter of this book ("Metadata Matters").

Categorizing Metadata

All of these perspectives on metadata should be considered in the development of networked digital information systems, but they lead to a very broad and often confusing conception. To understand this conception better, it is helpful to separate metadata into distinct categories—administrative, descriptive, preservation, technical, and use metadata—that reflect key aspects of metadata

functionality. Table 2 defines each of these metadata categories and gives examples of common functions that each might perform in a digital information system.

Table 2. Different Categories of Metadata and Their Functions

Category	Definition	Example
Administrative	Metadata used in managing and administering collections and information resources	• Acquisition and appraisal information • Rights and reproduction tracking • Documentation of legal, cultural, and com access requirements and protocols • Location information • Selection criteria for digitization • Digital repatriation documentation
Descriptive	Metadata used to identify, authenticate, and describe collections and related trusted information resources	• Metadata generated by original creator and • Submission-information package • Cataloging records • Finding aids • Version control • Specialized indexes • Curatorial information • Linked relationships among resources • Descriptions, annotations, and emendation creators and other users
Preservation	Metadata related to the preservation management of collections and information resources	• Documentation of physical condition of re • Documentation of actions taken to preserv physical and digital versions of resource data refreshing and migration) • Documentation of any changes occurring d digitization or preservation

Category	Definition	Example
Technical	Metadata related to how a system functions or metadata behaves	• Hardware and software documentation • System-generated procedural information (routing and event metadata) • Technical digitization information (e.g., fo compression ratios, scaling routines) • Tracking of system-response times • Authentication and security data (e.g., encr keys, passwords)
Use	Metadata related to the level and type of use of collections and information resources	• Circulation records • Physical and digital exhibition records • Use and user tracking • Content reuse and multiversioning informa • Search logs • Rights metadata

In addition to its different types and functions, metadata exhibits many different characteristics. Table 3 presents some key characteristics of metadata, with examples. Metadata creation and management have become a complex mix of manual and automatic processes and layers created by many different functions and individuals at different points during the life cycle of an information object. Effective and efficient metadata management is essential to ensure that the metadata we rely on to validate digital resources is itself trustworthy and that the large volume of metadata that potentially can accumulate throughout the life of a resource is subject to a summarization and disposition regime.[6]

Table 3. Attributes and Characteristics of Metadata

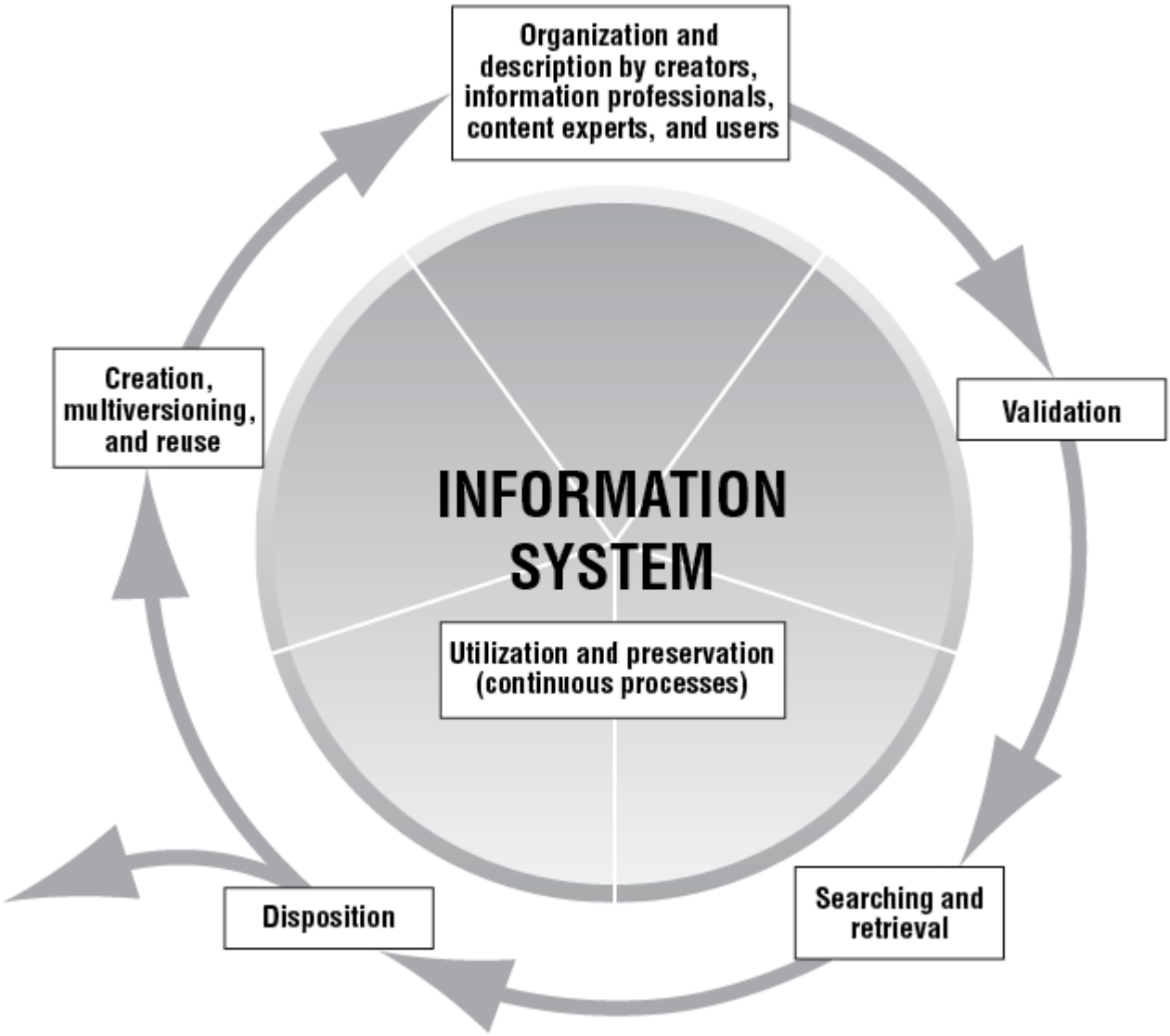
Attribute	Characteristics	Examples
Source of metadata	• Internal metadata generated by the creating agent for an information object at the time when it is first created or digitized • Metadata intrinsic to an item or work	• File names and header informatio • Directory structures • File format and compression sche • A title or inscription added to an a by its creator • A title or subtitle on the title page manuscript or printed book
	External metadata relating to an original item or information object; this is generated after the object is first created or digitized, often by someone other than the original creator	• URLs, URIs, PURLs, and other d statements of provenance and onl "location" • "Tracked" changes • Registrarial and cataloging record • Rights and other legal informatio
Method of metadata creation	Automatic creation, capture, or inferencing of metadata	• Keyword indexes • User-transaction logs • Audit trails • Descriptions of documentary interrelationships and intradocum relationships
	Manual creation of metadata by information specialists	Descriptive metadata such as catalog rec finding aids, and specialized indexes
	Manual or automatic creation of metadata during digitization processes	
	Individual user-contributed or crowd-sourced metadata	
Nature of metadata	Nonexpert metadata created by persons who are not subject or community specialists or information	• Title HTML tags and meta tags cr a personal web page

Attribute	Characteristics	Examples
	professionals (e.g., the original creator of the information object or a folksonomist)	• Personal filing systems • Folksonomies
	Expert metadata created by subject or community specialists and/or information professionals, often not the original creator of the information object	• Specialized subject headings • Bibliographic records • Archival finding aids • Catalog entries for museum objec • Ad hoc metadata created by subje experts (e.g., tags added to an inf object or catalog record by subjec experts)
Structure	Structured metadata that conforms to a predictable standardized or proprietary structure	MARC, BIBFRAME, TEI, EAD, LIDO, database formats
	Unstructured metadata that does not conform to a predictable structure	Unstructured note fields and other free-te annotations
Status	Static metadata that does not or should not change once it has been created	Technical information such as the date(s) creation and modification of an informati object, how it was created, file size
	Dynamic metadata that may change with use, manipulation, or preservation of an information object	• Directory structure • User-transaction logs
	Long-term metadata necessary to ensure that the information object continues to be accessible and usable	• Technical format and processing information • Rights information • Preservation management docum
	Short-term metadata, mainly of a transactional nature	Interim location information

Attribute	Characteristics	Examples
	Legacy metadata	Metadata created using an earlier system metadata scheme
Semantics	Controlled metadata that conforms to a standardized vocabulary or authority form and that follows standard content (i.e., cataloging) rules	• LCSH, LCNAF, AAT, ULAN, TG TGN • AACR, RDA, DACS, CCO
	Uncontrolled metadata that does not conform to any standardized vocabulary or authority form	• Free-text notes • User-created tags
Level	Collection-level or group-level metadata relating to collections or groupings of original items and/or information objects	• Collection- or group-level record bibliographic record for a group c collection of items; a finding aid f intact archival collection) • Series- or group-level information bibliographic record, finding aid, museum collection record
	Item-level or within-item-level metadata relating to individual items and/or information objects, often contained within collections	• Catalog records for individual bibliographic items or unique cult objects • Transcribed image captions and d • “Tombstone” information for wor and material culture • Format information

Figure 1 illustrates the different phases through which information objects typically move during their life cycles in today’s digital environment.[7] As they move through each phase, information objects acquire layers of metadata that can be associated with them in several ways.

Figure 1. The Life Cycle of an Information Object



Different types of metadata can become associated with an information object by a variety of processes, both manual and automated. These layers of accrued metadata can be contained within the same "envelope" as the information object itself—for example, in the form of header information for an image file or through some form of metadata bundling (e.g., via METS) that packages structural, descriptive, administrative, and other metadata with an information object or digital surrogate and indicates the types of relationships among the various parts of complex information objects (e.g., a digital surrogate consisting of a series of images representing the pages in a book or an album of illustrations or the constituent parts of a decorative arts object such as a tea service). Metadata can also be attached to the information object through bidirectional pointers or hyperlinks, while the relationships between metadata and information objects—and among different aspects of metadata—can be documented by registering them with a metadata registry. However, in any instance in which it is critical that metadata and content coexist, it is highly recommended that the metadata become an integral part of the information

object—that is, that it be “embedded” in the object and not stored or linked elsewhere.

As systems designers respond to the need to incorporate and manage metadata in information systems and to address how to ensure the ongoing viability of both information objects and their associated metadata through time, many additional mechanisms for associating metadata with information objects are likely to become available. Metadata registries and schema record-keeping systems are also more likely to develop as it becomes increasingly necessary to document schema evolution and to alert implementers to version changes.[8]

Primary Functions of Metadata

- *Creation, multiversioning, reuse, and recontextualization of information objects.* Objects enter a digital information system by being created digitally or by being converted into a digital format. Multiple versions of the same object may be created for preservation, research, exhibition, dissemination, or even product-development purposes. Some administrative and descriptive metadata may and indeed should be included by the creator or digitizer, especially if reuse is envisaged, such as in a digital asset management system.
- *Organization and description*. A primary function of metadata is the description and ordering of original objects or items in a repository or collection as well as of the information objects relating to the originals. Information objects are automatically or manually organized into the structure of the digital information system and may include descriptions generated by the original creator. Additional metadata may be created by information professionals through registration, cataloging, and indexing processes, or by others via folksonomies and other forms of user-contributed metadata.
- *Validation.* Users scrutinize metadata and other aspects of retrieved resources in order to ascertain the authoritativeness and trustworthiness of those resources.
- *Search and retrieval.* Good descriptive metadata is essential to users’ ability to find and retrieve relevant metadata and information objects. Information objects—both those that are locally stored and virtually distributed—are subject to search and retrieval by users, and information systems create and maintain metadata that tracks retrieval

algorithms, user transactions, and system effectiveness in storage and retrieval.

- *Utilization and preservation.* In the digital realm, information objects may be subject to many different kinds of uses throughout their lives, during which they may also be reproduced and modified. Metadata related to user annotations, rights tracking, and version control may be created. Digital objects, especially those that are born digital, also need to be subject to a continuous preservation regime and undergo such processes as refreshing, migration, and integrity checking to ensure their continued availability and to document any changes that might have occurred to the information object during preservation processes.
- *Disposition.* Metadata is a key component in documenting the disposition (e.g., accessioning, deaccessioning) of original objects and items in a repository as well as of the information objects relating to those originals. Information objects that are inactive or no longer necessary may be discarded.

Some Little-Known Facts about Metadata

- *Metadata does not have to be digital.* Cultural heritage and information professionals have been creating metadata for as long as they have been managing collections. Increasingly, such metadata is being incorporated into digital information systems, but metadata can also be recorded in analog formats such as card catalogs, vertical files, and file labels.
- *Metadata relates to more than the description of an object.* While museum, archive, and library professionals may be most familiar with the term in association with description or cataloging, metadata can also indicate the context, management, processing, preservation, and use of the resources being described.
- *Metadata can come from a variety of sources.* Metadata can be supplied by a human (by the creator of the digital file, by an information professional, and/or by an expert or non-expert user). It can also be generated automatically by a computer algorithm, or inferred through a relationship to another resource, such as a hyperlink.

- *Metadata continues to accrue during the life of an information object or system.* Metadata is created, modified, and sometimes even disposed of at many points during the life of a resource.
- *One information object's* metadata *can simultaneously be another information object's* data, *depending on the kinds of aggregations of and dependencies between information objects and systems.* The distinctions between what constitutes *data* and what constitutes *metadata* can often be very fluid and may depend on how one wishes to use a certain information object.

Why Is Metadata Important?

Metadata consists of complex constructs that can be expensive to create and maintain. How, then, can one justify the cost and effort involved? The development of the World Wide Web and other networked digital information systems has provided information professionals with many opportunities while at the same time requiring them to confront issues that they have not had occasion to explore previously. Judiciously crafted metadata, wherever possible conforming to national and international standards, has become one of the tools that information professionals are using to exploit some of these opportunities as well as to address some emerging issues, discussed below.

Increased accessibility: Effectiveness of searching can be significantly enhanced through the existence of rich, consistent, carefully crafted descriptive metadata. Metadata can also make it possible to search across multiple collections or to create virtual collections from materials that are distributed across several repositories—but only if the descriptive metadata records are in the same format or have been mapped across the various collections and formats. (Mary Woodley discusses this in more detail in chapter 3, "Metadata Matters.") Metadata standards that have been developed by different professional communities but include some common data elements (e.g. title, date, creator)—such as Dublin Core, EAD, MARC, BIBFRAME, the Metadata Object Description Schema (MODS), LIDO, and the Text Encoding Initiative (TEI)—are making it easier for users to negotiate between descriptive surrogates of information objects and digital versions of the objects themselves and to search at both the item and collection levels within and across information systems.

Retention of context: Museum, archival, and library repositories do not simply hold objects. They maintain collections of objects that have complex interrelationships and a variety of associations with people, places, movements

or styles, and events. In the digital world it is not unusual for a single object from a collection to be digitized and then for that digital surrogate to become separated from both its own cataloging information (descriptive metadata) and its relationship to the other objects in the same collection, resulting in a decontextualized information object. Metadata plays a crucial role in documenting and maintaining important relationships as well as in indicating the authenticity, structural and procedural integrity, and degree of completeness of information objects. In an archive, for example, by documenting the content, context, and structure of an archival record, metadata in the form of an archival finding aid is what helps to distinguish that record from decontextualized information.

Expanding use: Digital information systems for museum and archival collections make it easier to disseminate digital versions of unique objects to users around the globe who, for reasons of geography, economics, or other barriers, might otherwise not have an opportunity to view them. With new communities of users, however, come new challenges concerning how to make the materials most intellectually accessible. These new communities may have significantly different needs, cultural perspectives, language skills, and information-seeking behaviors from those of the traditional users for whom many existing information services were originally designed.

Teaching and learning: K–12 teachers and students may want to search for and use information objects in quite different ways from those of scholarly researchers. Instructors may wish to develop lesson plans or to scaffold learning so that students build on prior knowledge or are introduced to technical terminology. Specialized forms of metadata have been developed to address these needs.[9] In addition, the judicious use of controlled vocabularies and folksonomies can enhance access for various types of user groups.

System development and enhancement: Metadata can document changing uses of systems and content, and that information can, in turn, feed back into systems-development decisions. Well-structured metadata can also facilitate an almost infinite number of ways for users to search for information, to present results, and even to manipulate and to present information objects without compromising their integrity.

Multiversioning: The existence of information about, and surrogates of, cultural objects in digital form has heightened interest in the ability to create multiple and variant versions of information objects. This process may be as simple as creating both a high-resolution copy of a digital image for preservation or scholarly research uses and a low-resolution thumbnail image that can be rapidly transferred over a network for quick reference purposes. Or it may

involve creating variant or derivative forms to be used, for example, in publications, exhibitions, or schoolrooms. In either case, there must be metadata to relate the multiple versions of a given information object and to capture what is the same and what is different about each version. The metadata must also be able to distinguish what is qualitatively different in the various digitized versions or surrogates from the original physical object or item.

Legal issues: Metadata allows repositories to track the many layers of rights, licensing, and reproduction information that exist for original items as well as for their related information objects and the multiple versions of those information objects. Metadata also documents other legal or donor requirements that have been imposed on original objects and their surrogates—for example, privacy concerns, restrictions on reproductions, and proprietary and commercial interests. (See chapter 4, "Rights Metadata Made Simple" by Maureen Whalen.)

Preservation and persistence: If digital information objects that are currently being created are to have a chance of surviving migrations through successive generations of computer hardware and software, or removal to entirely new delivery systems, they will need metadata that enables them to exist independently of the system that is currently being used to store and retrieve them. Technical, descriptive, and preservation metadata that documents how a digital information object was created and maintained, how it behaves, and how it relates to other information objects will be essential. It should be noted that for the information objects to remain accessible and intelligible over time, it will also be essential to preserve and migrate this metadata and to ensure that it does not become "disconnected" from the object it describes.

System improvement and economics: Benchmark technical data, much of which can be collected automatically by a computer, is necessary to evaluate and refine systems in order to make them more effective and efficient from a technical and economic standpoint. The data can also be used in planning for new systems.

A Note on Metadata, Version Control, Reuse, and Recontextualization

It is worth giving special mention to the roles that metadata increasingly needs to play in supporting some of the particular opportunities of the digital age. Historically, one goal of cataloging was to make it possible to distinguish one version of an object or work from another. One item might be different from

another, for example, because it was a second edition of the same work, because it contained printing anomalies distinct from other copies printed at the same time, because it was an abridged or translated version of the original title, or because its title had changed.[10] Various standardized practices exist to help catalogers alert potential users to such differences in versions of a work. Today metadata must still be able to elucidate such distinctions. However, it must also be able to help users distinguish between, and trace the changes in, the following:

- Original analog and digitized versions, noting any changes that might have occurred accidentally or deliberately during the digitization process (e.g., digital "repair" of a broken glass lantern slide).
- Digitized and born-digital objects that are created in a range of resolutions to facilitate a variety of distribution mechanisms and uses or that are periodically refreshed, migrated, or rendered into an alternate format for preservation and long-term storage or security purposes.
- Original and renamed, retitled, or reattributed objects. For example, museum objects may be renamed or reattributed or assigned a different creation date because new documentation has come to light. Metadata may also change due to cultural sensitivities or challenges regarding provenance; for example, place names or object names may be changed to their original Native American forms, with English-language names that were assigned after the objects' creation "demoted" to the status of variants or additional access points.
- Original born-digital materials and revised or updated versions (e.g., websites, reference databases).
- Original analog or born-digital materials that are reused in part or in whole in new digital resources (e.g., personal websites, digital art, or digital music compilations).
- Objects, especially but not only museum objects, that are described collectively in one context within their metadata (e.g., as objects that were all collected at the same time at the same archaeological excavation) but are then taken individually out of that collection and recontextualized (e.g., in a special exhibition of Greek vases from a particular period or an exhibition of paintings relating to a particular theme or subject).

Conclusion and Outstanding Questions

Metadata is like interest: it accrues over time. To extend the metaphor further, wise investments in metadata generate the best return on intellectual capital. Carefully crafted metadata results in the best information management—and the best end-user access—in both the short and the long term. If thorough, consistent metadata has been created, it is possible to conceive of it being used in an almost infinite number of new and even currently unforeseen ways to meet the needs of both traditional and nontraditional users for multiversioning and for data mapping and mining. But the resources and intellectual and technical design issues involved in good metadata development and management are far from trivial. Some key challenges that must be addressed by information professionals as they develop digital information systems and objects are

- identifying which metadata schema or schemas should be applied in order to best meet the needs of the information creator, repository, and users. As mentioned above, selection of an inappropriate schema (e.g., EAD for museum collections that do not share a common provenance) serves neither the collection materials themselves nor the users who wish to find, understand, and use those materials. Also, in many cases, especially with complex objects or hierarchically structured archival and other types of collections, a combination of schemas working together (e.g., MARC or BIBFRAME and/or EAD at the collection level; MARC, Dublin Core, MODS, VRA Core, or LIDO at the item level) may be the best solution.
- deciding which aspects of metadata are essential for the desired goal and how granular each type of metadata needs to be—in other words, how much is enough and how much is too much. There will likely always be important tradeoffs between the costs of developing and managing metadata to meet current needs and creating sufficient metadata that can be capitalized on for future, often unanticipated uses. Metadata creators should remember that good "core" metadata can be a valid approach in both economic and intellectual terms. (See principles 2 and 7 of "Practical Principles for Metadata Creation and Maintenance.")
- ensuring that the controlled vocabularies, thesauri, and taxonomies (including folksonomies) being applied are the most up-to-date, complete versions of those sets of data values and that they are the

appropriate terminologies for the materials being described and for the intended users.

What we do know is that the existence of many types of metadata will prove critical to the continued online and intellectual accessibility and utility of digital resources and the information objects that they contain as well as the original objects and collections to which they relate. In this sense, metadata provides us with the Rosetta stone that will make it possible to decode information objects and their transformation into knowledge in the cultural heritage information systems of the future.

1. An information object is a digital item or group of items, regardless of type or format, that can be addressed or manipulated as a single object by a computer. This concept can be confusing in that it can be used to refer both to digital "surrogates" of original objects or items (e.g., digitized images of works of art or material culture, a PDF of an entire book) and to descriptive records relating to objects and/or collections (e.g., catalog records or finding aids). ↩

2. Anne J. Gilliland-Swetland, "Popularizing the Finding Aid: Exploiting EAD to Enhance Online Browsing and Retrieval in Archival Information Systems by Diverse User Groups," *Journal of Internet Cataloging* 4, nos. 3–4 (2001): 199–225. ↩

3. Paul Conway, *Preservation in the Digital World* (Washington, DC: Commission on Preservation and Access, 1996), http://www.clir.org/pubs/reports/conway2/index.html. ↩

4. Sue McKemmish, Glenda Acland, Nigel Ward, and Barbara Reed, "Describing Records in Context in the Continuum: The Australian Recordkeeping Metadata Schema," *Archivaria* 48 (Fall 1999): 3–37. ↩

5. See Roy Tennant, "Metadata's Bitter Harvest," *Library Journal*, July 15, 2004, available at http://roytennant.com/column/?fetch=data/39.xml and the Digital Library Federation's Multiple Metadata Formats page at http://webservices.itcs.umich.edu/mediawiki/oaibp/index.php/MultipleMetadataFormats. ↩

6. See Anne J. Gilliland et al., "Towards a Twenty-first Century Metadata Infrastructure Supporting the Creation, Preservation and Use of Trustworthy Records: Developing the InterPARES2 Metadata Schema Registry," *Archival Science* 5, no. 1 (March 2005): 43–78. ↩

7. Figure 1 is modified from "Information Life Cycle" in C. L. Borgman et al., "Social Aspects Of Digital Libraries" (Final Report to the National Science Foundation, award number 95-28808, presented at the UCLA-NSF Social Aspects of Digital Library Workshop, Graduate School of Education and Information Studies, University of California, Los Angeles, February 15–17, 1996), p. 7, http://works.bepress.com/borgman/181/. ↩

8. See Gilliland et al., "Towards a Twenty-first Century Metadata Infrastructure." ↩

9. See Dimitrios A. Koutsomitropoulos, Andreas D. Alexopoulos, Georgia D. Solomou, and Theodore S. Papatheodorou, "The Use of Metadata for Educational Resources in Digital Repositories: Practices and Perspectives," D-Lib 16, nos. 1–2 (January–February 2010), http://www.dlib.org/dlib/january10/kout/01kout.html. ↩

10. According to the Functional Requirements for Bibliographic Records (FRBR) conceptual model, these are different “expressions” and/or “manifestations” of a work; see http://www.ifla.org/publications/functional-requirements-for-bibliographic-records. Note that the definition of a “work” (and the conceptual model) can differ considerably for unique works of art or architecture, as opposed to literary works or musical compositions, for which the FRBR model is ideal. See Murtha Baca and Sherman Clarke, “FRBR and Works of Art, Architecture, and Material Culture,” in *Understanding FRBR: What It Is and How It Will Affect Our Retrieval Tools*, ed. Arlene G. Taylor (Westport, CT: Libraries Unlimited, 2007), 103–10. ↩

Metadata and the Web

Tony Gill

Revised by Murtha Baca, with assistance from Joan Cobb, Nathaniel Deines, and Moon Kim

When the first edition of this book was published in 1998, the term *metadata* was comparatively esoteric, having originated in the information science and geospatial data communities at the end of the twentieth century. As of this writing, a Google search on "metadata" yields hundreds of millions of results; a search in WorldCat for publications with "metadata" in the title yields more than 13,000 results; and a keyword search on "metadata" in Amazon returns more than 5,000 results. Metadata has hit the big time; it is now a pervasive phenomenon, and even a consumer commodity. For example, almost all consumer-level digital cameras capture and embed exchangeable image file format (Exif)[1] metadata in digital images, and files created using Adobe's Creative Suite of software tools (e.g., Photoshop and Illustrator) contain embedded Extensible Metadata Platform (XMP)[2] metadata.

As the term *metadata* has been increasingly adopted and co-opted by more diverse communities, its definition has grown in scope to include almost anything that describes anything else. The standard concise definition is "data about data," a relationship that is frequently illustrated by the metaphor of a library card catalog, in which the individual entries describe the holdings. This metaphor is pedagogically useful because it is nonthreatening. Many people are familiar with the concept of a library catalog as a simple tool used to help readers find the books and other items they are seeking and to help librarians manage collections. However, the example may be seen as problematic from an ontological perspective, because neither catalog entries nor books are, in fact, data per se; they are *containers* or *carriers* of data. This distinction between information and its carrier is increasingly being recognized; for example, the International Council of Museums' International Committee for Documentation (CIDOC) Conceptual Reference Model (CRM), a domain ontology for the semantic expression and interchange of museum, library, and archival information, models the relationship between information objects—identifiable conceptual entities such as texts, images, algorithms, or musical compositions—and their physical carriers.[3]

The International Federation of Library Associations and Institutions Functional Requirements for Bibliographic Records (FRBR) model makes a somewhat

similar distinction between *works*, *expressions*, *manifestations*, and *items*: the first two entities are conceptual, while the last two are actual physical instances that are described by bibliographic records. (Note that as the new linked open data standard for bibliographic records, BIBFRAME, becomes more widely used and eventually replaces the MARC format, the FRBR conceptual data model may be superseded. In any case, it is clear that while FRBR can be useful for modeling literary and musical works, it does not provide an adequate model for unique works such as works of art or architecture or for "serial" works such as periodicals.) [4]

Of course, most library catalogs are now encoded as 0s and 1s in databases, and the "items" representing the "works" they describe (to continue with the FRBR nomenclature) are increasingly likely to be digital objects that reside on a server, as opposed to objects composed of ink, paper, cardboard, etc., that are located on shelves. This is even truer now in light of large-scale digitization initiatives that have been undertaken by many knowledge organizations, not to mention gigantic projects such as Google Books and the Internet Archive.

One property of metadata is that it is—or should be—structured to model the most important attributes of the type of object that it describes. For example, each component of a standard MARC record is clearly delineated by field labels that identify the meaning or type of each atomic piece of information that describes the bibliographic item—author, title, subject, and so on.

The structured nature of metadata is important. By accurately modeling the essential attributes of the class of objects being described, metadata in aggregate can serve as a catalog—a distillation of the attributes of the particular collection—thereby becoming a useful tool for using and managing that collection, be it a collection of books, other physical objects, or digital images or digital surrogates of books, musical scores, visual materials, and so on. In the context of this chapter, then, *metadata* can be defined as *a structured description of the essential attributes of an object*.

Web Search Engines

Web search engines such as Google are automated information retrieval systems that continuously traverse the web, visiting sites and saving texts (not including stop words), images, and locations (usually in the form of URLs) in order to build up a huge index or "list" of web pages. Search engines typically provide keyword searching and retrieve huge sets of results that are "relevance ranked" using a variety of proprietary algorithms. Search engines rely heavily on <title> HTML tags (a simple but very important type of metadata that

appears in the title bar and favorites/bookmarks menus of most web browsers), the actual text on the pages, and referring links (which are taken as an indication of the popularity of the resource).

The Web Continues to Grow

The World Wide Web is the largest collection of documents the world has ever seen, and its growth is showing no signs of slowing. Although it is impossible to determine the exact "size" of the web (both the visible web and deep web), some metrics are available. According to a Netcraft survey, in February 2004 there were approximately 47 million host names and 22 million active sites; ten years later (February 2014) there were 920 million host names and almost 180 million active sites.[5] Although the Netcraft survey clearly demonstrates the continuing upward trend in the growth of the web, it does not tell the whole story because it does not address how many sites are hosted on each server and how many accessible pages are contained in each site.

The Visible Web versus the Hidden Web

The problem of determining how many "pages" are really available on the web is complicated by the fact that a large and increasing amount of the content on the web is served dynamically from databases in response to user queries, is expressed in a non-web format, or requires some kind of user authentication or login. Web crawlers, also called spiders or robots (the software used by search engines to trawl the web for content and build their vast indices), can only search the so-called "visible web"; they cannot submit queries to databases, parse file formats that they do not recognize, click buttons on web forms, or log in to sites requiring authentication. As a result, all of this dynamically generated content (as opposed to "document-like" static HTML pages) is effectively invisible to the search engines and therefore is not indexed.

Collectively, this content beyond the reach of search engine crawlers is referred to as the "deep web," the "invisible web," or the "hidden web," and as these names suggest, estimating its size is even more difficult (and perhaps even more meaningless, since the number grows enormously every day) than measuring the visible web. Although much of the content on the deep web is deliberately kept out of the public sphere—either because it requires a password or institution-specific IP address or because some kind of fee or subscription must be paid to access it—there is a vast amount of information that is not accessible to search engines simply because it is located on sites that were not designed to be accessible to crawlers or robots. This is an especially common

problem for sites that generate pages dynamically from databases in response to users' queries (as with library catalogs and many other databases). Because commercial search engines like Google typically account for the vast majority of traffic on the web, building sites that are not accessible to crawlers can seriously limit the accessibility and use of the information they contain. Institutions seeking to make dynamically generated data as widely accessible as possible should design "crawler-friendly" sites; a good way to do this, which also facilitates access by human users (as opposed to web robots), is to provide access to information through hyperlinked hierarchies of categories in addition to search interfaces. Another option for the library, archive, and museum sectors is to contribute deep web collections information to union catalogs or other aggregated resources that are indexed by the commercial search engines. (This is a good strategy in any case, as the more "places" on the web from which information can be accessed, the more users can be reached.)

Search engine providers also provide tools to help webmasters expose otherwise hidden content; for example, Google's Sitemap feature allows webmasters to provide a detailed list of all the pages on their sites—even those that are dynamically generated—in a variety of machine-readable formats to ensure that every page gets crawled and indexed. Union catalogs and tools to expose deep web content to search engines are discussed in this chapter.

Finding Needles in a Huge and Rapidly Expanding Haystack

The web is the largest and fastest-growing collection of documents the world has ever seen, and it has undoubtedly revolutionized access to a formerly unimaginable amount of information, of widely variable quality, for the billions of users who have access to it. It is worth remembering, however, that this is still less than one person in five or six globally. The myth of "universal access" to the web remains just that—a myth.

Unfortunately, finding relevant, authoritative, high-quality information on the web is not always a straightforward proposition. There is no overarching logical structure to the web, and the core protocols do not offer any support for search and retrieval beyond the basic mechanisms provided by the HTTP for requesting and retrieving pages from a specific web address. Disappointment with the web was clearly evident in a comment by Ted Nelson (who first coined the term *hypertext* in 1965) in a speech delivered at the HyperText97 conference: "The reaction of the hypertext research community to the World

Wide Web is like finding out that you have a fully grown child. And it's a delinquent."[6]

Not surprisingly, tools designed to address the resource location problem and help make sense of the web's vast corpus of information resources started to appear soon after the launch of the first web browsers in the early 1990s. For example, Tim Berners-Lee founded the WWW Virtual Library,[7] a directory of sites maintained by human editors, shortly after introducing the web itself, and search engines such as Yahoo!, Lycos, and Webcrawler were launched in 1994.

As of this writing, the clear market leader in web search for almost two decades has been Google. According to its corporate site, "Google's mission is to organize the world's information and make it universally accessible and useful."[8] In the time since Google was registered as a web domain in 1997, it has grown to become a corporate and web giant, employing thousands of people. Clearly, helping people find information on the web is big business.

To maintain its position as the most popular search engine on the web, Google must routinely perform several Herculean tasks that are becoming increasingly difficult as both the web itself and the number of people using it continue to grow. First, it must maintain an index of publically available web pages that is both sufficiently current and sufficiently comprehensive to remain competitive. Currency is important, because as Google's Trends and Year in Search pages show,[9] many of the most popular searches are related to current events and popular culture. Any search engine that fails to deliver relevant results to queries about current events will rapidly lose a large share of the global search market.

Second, a search engine must have an adequately comprehensive index of the web, because without such an index it will fail to deliver relevant results that a competitor with a more comprehensive index could provide. Index size is one of the key metrics on which search engines compete and measure success, and, as we know, the size of the web is continuously growing, with more sites and pages and words to index appearing daily.

Third, in addition to maintaining a current and comprehensive index of the rapidly expanding web, a search engine must be able to search that index, ranking the search results and presenting them to the user as quickly as possible—ideally, in less than half a second. In fact, when users do a search on Google, they are not searching the web "live"—they are searching a vast corpus of tags and text that the search engine has stored in the form of an index; this is why a search via a commercial search engine sometimes results in a "page not found" message—that is, the page that was indexed by the web crawler is no

longer available, but the data taken from it still exists in an index on one of the search provider's servers.

Much of Google's rise to dominance in the search engine market can be attributed to its sophisticated PageRank™ algorithm, which assesses the importance of retrieved web pages according to the number of links from other pages that point to them.[10] The text contained in the <title> HTML tag and the PageRank value of each page are the only metadata that Google seems to use to any meaningful and consistent extent in providing its search service—as described above, the search itself is performed on an index of the actual data content of the HTML pages themselves. In other words, what Google enables its users to search is an index or huge "list" of every word that appears on every HTML page on the visible web; and because the search is performed on the index and not in real time, results are retrieved very quickly, which is what users have come to expect.[11]

Fourth, a market-leading search engine such as Google must be able to respond to hundreds of millions of search requests from users all around the world every day. To meet these gargantuan and constantly increasing information retrieval challenges, Google has developed one of the largest and most powerful computer infrastructures on the planet. Unlike most of its competitors, which typically use relatively small clusters of very powerful servers, Google has developed a massive parallel architecture comprising large numbers of inexpensive networked personal computers, which Google claims is both more powerful and more scalable than using a smaller number of more powerful servers.[12]

Google provides little information about its hardware infrastructure, but given the explosive growth of both the amount of information on the web and the number of users, coupled with the wide range of other services offered by the Internet giant (Google Scholar, Google Books, Google Images, Gmail, and Google Earth, to name some of the best-known), the number of server nodes is undoubtedly huge. There is widespread speculation that the Google server cluster could comprise anywhere from hundreds of thousands to millions of nodes and that it could in fact be the most powerful "virtual supercomputer" in the world.

Can the Search Engines Keep Up?

Can commercial search engines continue to scale up their operations as both the amount of content on the web and the number of users continue to grow? Since before the new millennium, analysts have been predicting that the web would

outgrow the search providers' abilities to index it, but as of this writing, the tipping point has not been reached.

Steve Lawrence and C. Lee Giles of the NEC Research Center conducted a scientifically rigorous survey of the leading search engines' coverage of web content in February 1999. Their findings, published in the peer-reviewed journal *Nature,* indicated that at that time no search engine indexed more than about 16 percent of the web: "Our results show that the search engines are increasingly falling behind in their efforts to index the Web," they wrote.[13] However, if we compare this with the January 2005 study by Gulli and Signorini, which estimated that Google had indexed about 76 percent of the 11.5 billion pages on the web at that time, it seems that the search engines learned to provide better coverage than they did in the web's infancy. Clearly, search engines in general and Google in particular have been able to scale up their technology better than had been predicted at the end of the twentieth century.

But common sense suggests that there may be a limit to the search engines' current ways of dealing with the continuous and rapid growth of the web. Even if Google's massively networked computer architecture is technically capable of indefinite expansion, other kinds of constraints may prove insurmountable at some point in the future. As long ago as 2005, an article by one of Google's principal hardware engineers warned that unless the ratio of computer performance to electrical power consumption improves dramatically, power costs for commercial search engines and other service providers could become a larger component of the total cost of ownership than the initial hardware costs.[14] This could become a significant barrier to continued expansion of the Google technical platform in the future, particularly if energy costs continue to rise. A million interconnected servers consume a tremendous amount of electrical power and require a tremendous amount of energy for "collateral" expenses such as climate control.

Metadata to the Rescue?

In the early days of the web, many people, particularly those in the emerging digital library community, saw metadata as the long-term solution to the problem of resource discovery on the web. The reasoning behind this was very logical and goes back to the classic example of metadata: library catalogs had proved their efficacy in providing both access to and control of large bibliographic collections, so why should the web be different?

Research and development projects to catalog useful web resources sprang up around the globe. One of the first lessons learned from these early pilot projects was that the economics of cataloging web resources was very different from the economics of cataloging books. Whereas the creation of a carefully crafted (and often very expensive) MARC record—complete with subject headings and other controlled terminology and conforming to standard cataloging rules—could be justified in the traditional bibliographic world because the record would be used by many different libraries for many years, web resources are both more dynamic and more transient than traditional published materials; unlike books, websites often change, and sometimes they disappear altogether.

As a result, metadata standards for describing Internet resources began to appear, ranging from relatively simple embedded metadata in the form of meta tags to Dublin Core, a metadata element set purportedly developed specifically for web resources, to the Resource Description Framework (RDF), a complex model for data interchange. It should be noted, however, that many search engines, including Google, make little or inconsistent use of embedded metadata for a variety of reasons (primarily, it seems, because of lack of trustworthiness of that "hidden" metadata).

Meta Tags

The early, now-defunct search engine AltaVista was the first to popularize the use of two simple metadata elements, "description" and "keywords," that can be easily and invisibly embedded in the <head> section of web pages using the HTML meta tag. Here are two examples:

<metaname="description" content="Version 4.0 of the site devoted to metadata: what it is, its types and uses, and how it can improve access to web resources.">

<metaname="keywords" content="data standards, metadata, meta data, World Wide Web, WWW, digital resources, metatags, Dublin Core, RDF, Semantic Web, crosswalks, metadata mapping.">

The description tag is intended to display in search engine results lists to provide an accurate, concise, authoritative summary of the particular web resource so that users can decide whether or not to click and go to the resource itself. If the description tag is longer than about 120 characters, it usually gets cut off in search results displays. Google and other search engines do use the description tag, but not consistently; when it is absent, however, text from the web page itself is displayed, seemingly chosen at random but generally coming

from the top of the page. This often produces confusing and even senseless results displays.

The keyword tag is a sort of "container" for subjects, names, and other access points, intended to provide more effective retrieval and relevance ranking. The keywords tag enables the creators of a web resource to include terms that do not appear on the web page itself—in short, it enables the creators of web pages to "catalog" their resources. This tag seems to be used by commercial search engines less than the description tag; most often, it is ignored. But it can be useful for enhancing searching within an institution's or company's own website, where the search engine is controlled internally.

Dublin Core

The Dublin Core Metadata Element Set[15] is a set of fifteen elements that can be used to describe a wide variety of resources for the purpose of cross-disciplinary and cross-system resource discovery. Although originally intended solely as the equivalent of a simple "electronic catalog card" for networked resources, Dublin Core is now used to describe almost any kind of information object or asset. The fifteen Dublin Core elements are *contributor*, *coverage*, *creator*, *date*, *description*, *format*, *identifier*, *language*, *publisher*, *relation*, *rights*, *source*, *subject*, *title*, and *type*.

The Dublin Core elements and their meanings were developed and refined by an international group of librarians, information professionals, and subject specialists through an ongoing consensus-building process that has included numerous conferences and workshops, working groups, and electronic mailing lists. The element set has been published as both a national and international standard (NISO Z39.85-2001[16] and ISO 15836:2009,[17] respectively). There are a significant number of large-scale deployments of Dublin Core metadata around the globe,[18] and it has become the preferred schema for metadata mapping and harvesting.

Resource Description Framework

The Resource Description Framework (RDF)[19] is a standard developed by the World Wide Web Consortium (W3C) for encoding resource descriptions in machine-readable form, so that computers can "understand," share, and process the information in meaningful and useful ways. RDF metadata is normally encoded using standard syntaxes such as XML and Turtle.[20] As the name indicates, RDF provides a *framework* for resource description; that is, it

provides the formal syntax or structure of the resource description, but it does not furnish the actual data values to be expressed. The semantics or meaning must be specified for a particular domain or community in order for computers to be able to make sense of the encoded metadata. The semantics are specified by an RDF vocabulary,[21] which is a knowledge representation or model of the metadata that unambiguously identifies what each individual metadata element means and how it relates to the other elements in the domain. RDF vocabularies can be expressed as RDF schemas[22] or, when they convey more complex relationships among data elements, as Web Ontology Language (OWL) ontologies.[23]

The CIDOC Conceptual Reference Model (CRM) is an example of an ontology that provides the semantics for a specific domain—the interchange of library, archive, and museum collection documentation. By expressing the classes and properties of the CIDOC CRM as an RDF schema or one or more OWL ontologies, information about cultural heritage collections can be expressed in a semantically unambiguous way, thereby facilitating information sharing and interchange across different computer systems. In the age of linked data (see below), the CIDOC CRM has great potential because it explicitly models the relationships among entities, agents, and events.

Using the highly extensible and robust framework of RDF, RDF schemas, and OWL ontologies, rich metadata descriptions of digital resources can be created that draw on a theoretically unlimited set of semantic vocabularies. Interoperability for automated processing is maintained because the strict underlying syntax requires that each vocabulary be explicitly specified.

RDF, RDF schemas, and OWL ontologies are all fundamental building blocks of the W3C's so-called "Semantic Web" activity.[24] The Semantic Web is the vision of Tim Berners-Lee, director of the W3C and the "inventor" of the original World Wide Web.[25] Berners-Lee's vision is for the web to evolve into a seamless network of interoperable, meaningfully linked data that can be shared and reused across software, enterprise, and community boundaries.

Linked Data

Linked data is data that encodes semantic relationships by following a set of best practices for publishing and interlinking structured data that uses RDF syntaxes and HTTP URIs.[26] Linked data can be published on the open World Wide Web or behind a firewall. If linked data is made available for use, reuse, and redistribution on the visible web, it is called linked open data (LOD). Examples of large LOD datasets are DBPedia,[27] the Library of Congress Subject

Headings, and the Virtual International Authority File (VIAF); the Getty's electronic thesauri are also available as LOD.[28]

As of this writing, LOD offers great promise for semantically rich, easier, and more widespread use, reuse, and sharing of both metadata records and the controlled vocabularies that are used to populate those records and provide meaningful connections among resources. LOD has the potential to revolutionize the way data can be disseminated and integrated in ways that will significantly enhance the process of information- and resource-seeking and utilization.

Metadata Harvesting

Harvesting is the process of gathering metadata or data from the Internet in order use it in a variety of ways. Often metadata is harvested in order to create a central index of searchable records from a variety of repositories.

The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)[29] provides a method for making deep web metadata more accessible. Rather than embedding metadata in the actual content of HTML pages, the OAI-PMH uses a set of protocols that allows metadata records to be exposed in a predictable way so that other OAI-PMH–compatible computer systems can access and retrieve them.

The OAI-PMH supports interoperability (which can be thought of as the ability of systems to communicate meaningfully) between two different computer systems: an OAI data provider and an OAI harvester, which in most cases is also an OAI service provider. As the names suggest, an OAI data provider is a source of metadata records, whereas the OAI harvester retrieves or "harvests" records from one or more data providers. Since both the data provider and harvester must conform to the same basic information-exchange protocols, metadata records, if properly formatted, can be reliably retrieved from the provider(s) by the harvester.

Although the OAI-PMH can in theory support records expressed in any XML metadata schema, the protocol mandates that all OAI data providers must be able to deliver Dublin Core XML metadata records as a minimum requirement. In this way, the OAI-PMH supports interoperability of metadata originating in different systems. It should be noted that, while the OAI-PMH is the prevalent protocol for metadata harvesting and aggregation as of this writing, when the time comes that LOD is prevalent, the OAI protocol may become obsolete.

Meta-utopia or Meta-garbage?

In an oft-quoted diatribe from 2001 (ancient history for the Internet, but the content is still valid today—although the HTML page where the text appears is formatted in a very old-fashioned way), the Canadian journalist and blogger Cory Doctorow enumerated what he calls the "seven insurmountable obstacles between the world as we know it and meta-utopia."[30] In this piece, Doctorow, a great proponent of making digital content as widely available as possible, puts forth his arguments for the thesis that metadata created by humans will never have widespread utility as an aid to resource discovery on the web. These arguments are paraphrased here:

- *"People lie."* Metadata cannot be trusted because there are many unscrupulous content creators who publish misleading or dishonest metadata in order to draw traffic to their sites.
- *"People are lazy."* Most content publishers are not sufficiently motivated to do the labor involved in carefully cataloging the content they publish.
- *"People are stupid."* Most content publishers are not intelligent enough to effectively catalog the content they produce.
- *"Mission impossible—know thyself."* Metadata on the web cannot be trusted because there are many content creators who inadvertently publish misleading metadata.
- *"Schemas aren't neutral."* Classification schemes are subjective.[31]
- *"Metrics influence results."* Competing metadata standards bodies will never agree.
- *"There's more than one way to describe something."* Resource description is subjective.

Although obviously intended as a satirical piece, Doctorow's short essay nevertheless contains several grains of truth when considering the web as a whole. His most compelling argument is the first one: people lie. It is very easy for unscrupulous web publishers to embed "meta tag spam"—deliberately misleading or dishonest descriptive metadata—in their web pages. This tactic is intended to increase the likelihood that a web page will appear in search engine results and to improve the site's visibility and ranking on search engines. Increased visibility and higher ranking can dramatically increase the amount of user traffic to a site, resulting in potentially greater profits in the case of

commercial sites and greater success for nonprofit organizations seeking to reach a broader audience. However, the search engine companies have long been wise to this practice, and as a result they often treat embedded metadata with skepticism—or ignore it altogether. It is rumored that some search engines may even penalize sites that contain suspect metadata by artificially lowering their page ranking. Because many search engines do not utilize embedded metadata, using instead the text in the Title HTML tag and the text on the actual page itself (Google appears to use both embedded and explicit metadata, but not in a consistent way), there may seem to be little incentive for honest web publishers to expend the time and effort required to add this potentially useful information to their own pages—unless the particular search engine that they use to index their own site makes use of the embedded keyword and description meta tags.

The other points on Doctorow's list are less convincing, particularly if we look at the subset of web content created by libraries, museums, and archives. Librarians, museum documentation specialists, and archivists are typically diligent, trained information professionals, and they are not usually dishonest, lazy, or stupid. They have a long tradition of using standard metadata element sets such as MARC, the Metadata Object Description Schema (MODS), Encoded Archival Description (EAD), VRA Core, LIDO (Lightweight Information Describing Objects), and Dublin Core; classification schemes and controlled vocabularies such as Library of Congress authority files, its *Thesaurus for Graphic Materials,* and the Getty vocabularies; and community-specific cataloging rules such as Resource Description and Access (RDA), *Describing Archives: A Content Standard* (DACS), and *Cataloging Cultural Objects: A Guide to Describing Cultural Works and Their Images* (CCO). They use these tools to describe resources in standardized ways that have been developed over decades of collaborative consensus-building efforts. In effect, they have been demonstrating the value and power of descriptive metadata created by skilled human beings for many decades.

Playing Tag

A relatively recent development in the field of metadata on the web that significantly weakens Doctorow's argument is exemplified by the so-called folksonomies. A folksonomy is a sort of "uncontrolled vocabulary" that is built when many people use a shared system to label online content such as web pages or images with descriptive terms and names, known as tags. Many people are motivated to tag web content because it allows them to organize and find certain content; they are effectively building their own personal "catalogs" of

online content. In folksonomies, any terms or names can be used; unlike true taxonomic classification systems and controlled vocabularies, in which synonyms are explicitly linked to one another, concepts are often organized in such a way as to encode their hierarchical relationships, and carefully constructed rules exist for the application of terms or names to describe an item.

The folksonomy aspect of "uncontrolled" tags comes into play when all the tags applied to a specific resource by multiple users are aggregated and ranked. For example, if one person tags an image with the term "impressionist" it doesn't carry a great deal of weight in terms of searching. But if hundreds of users use this term, and it is the most frequently applied tag for a particular image or other online resource, it is a pretty safe bet that the resource is about or related to Impressionist art.

Two well-known examples of folksonomy/social tagging sites are LibraryThing[32] and Flickr.[33] LibraryThing enables users to assign tags to records from library catalogs and commercial sites; it also enables users to search the Library of Congress, Amazon, and hundreds of library catalogs enhanced by the tags added by users from around the world. Libraries, archives, and museums can use LibraryThing for Libraries[34] to leverage third-party metadata, including tags, reviews, ratings, and so on, to enhance access to and discovery of their collections. In effect, the "uncontrolled" tags and other data that are added to the controlled terms in a standard library record by LibraryThing users "complement" the library record by providing many more potential access points.

Flickr is a digital image-sharing site that enables users to tag images for easier retrieval. The fact that the Library of Congress uses Flickr to allow users to tag and comment on images from selected photographic collections[35] is another example of how libraries are adding access points and providing broader dissemination of their visual materials by taking advantage of user-generated metadata. In essence, the user-created keywords, in a variety of languages, are appended to a user-friendly version of the underlying Library of Congress MARC or Dublin Core metadata records.

In Metadata We Trust (Sometimes)

Metadata is not a universal solution for resource discovery in the digital environment. The underlying issues of trust, authenticity, and authoritativeness continue to impede the widespread use of structured, standards-based metadata for web pages, and this situation is unlikely to change as long as the search engines can continue to satisfy (or seem to satisfy) the search needs of most

users with their current methods—indexing the <title> HTML tag, the actual words on web pages, and ranking the "popularity" of pages based on the number of referring links.

But human-created metadata has a well-established and extremely important role in specific communities and applications, especially in the library, archive, and museum communities, where "metadata" is equivalent to "cataloging." Many standards and technology components aimed at facilitating resource discovery and information sharing and dissemination have been in place for some time. These key building blocks include:

- data structure and data format standards for different types of resource descriptions, such as MARC[36] (to be replaced by BIBFRAME), Dublin Core,[37] MODS,[38] EAD,[39] VRA Core,[40] and LIDO;[41]
- data value standards such as Library of Congress authority files, Getty vocabularies, Medical Subject Headings (MESH), ICONCLASS, and many others;
- tools and methods for encoding metadata in machine-readable form: for example, XML, RDF, SKOS (Simple Knowledge Organization System), CIDOC CRM, and FOAF (Friend of a Friend);
- protocols for distributed search and metadata harvesting: for example, the Z39.50 family of retrieval protocols, web service protocols such as SOAP (Simple Object Access Protocol) and REST (Representational State Transfer), and the OAI-PMH.

By using these various components in intelligent and appropriate ways in order to provide access to the rich information content generated by libraries, archives, and museums, it should become possible to build a global Semantic Web of digital cultural content and integrated search tools to help users find the content they are seeking.

Libraries and the Web

The web has dramatically changed the global information landscape—a fact that has had a particularly significant impact on libraries, which were the traditional gateways to information for two millennia. Whereas previous generations of researchers relied almost entirely on libraries for their information-seeking needs, members of the current generation of advanced researchers, students, and the general public are much more likely to start (and often end) their research at a web search engine like Google, Bing, or Yahoo.

Faced with this reality, libraries and related service organizations have worked hard to bring information from their online public access catalogs (OPACs)—resources traditionally hidden in the deep web, beyond the reach of search engines' web crawlers—out onto the visible web. The Online Computer Library Center (OCLC),[42] the largest library cooperative and service provider in the world, has made its vast union catalog, WorldCat,[43] available free of charge on the web; individual WorldCat records are retrievable from commercial search engines so that users are not obliged to start their searches from the WorldCat search page, the existence of which many users may not be aware.[44]

One of the most striking examples of collaboration between libraries and a commercial search engine company is the Google Books project.[45] This is a service provided by the search engine giant that enables users to search the full text of books that Google has scanned, converted to machine-readable text using optical character recognition, and stored in a digital data-base. The books are provided by publishers and authors who choose to participate in the Google Books Partner Program[46] and by Google's library partners, through the Google Books Library Project.[47]

By purportedly making available the full text of millions of printed volumes, Google Books offers users the possibility to search not just the metadata or bibliographic records for items in libraries and elsewhere but also every word in the books themselves. The reality is that, depending upon a book's copyright status, only excerpts of it may be available for searching and viewing. For books that are in the public domain, Google provides a brief bibliographic record, links to places where it can be purchased on line, in print form, and as an ebook (if available), and the full text of the book itself (but not in downloadable form). The Google Books "Find in a library" link takes users to the relevant record in WorldCat, where he or she will find bibliographic records with an indication of which libraries own copies of the book. While this may be useful to users who have the ability to request books via interlibrary loan, for the majority of users the ability to obtain free online access to full digitized copies of many books remains an illusion. Metadata and full-text searching hold great promise for a "democratization" of access to knowledge in written form, but we still have a long way to go before the World Wide Web is truly a "library."

1. See http://www.cipa.jp/std/documents/e/DC-008-2012_E.pdf. ↩
2. See http://www.adobe.com/products/xmp/. ↩
3. See http://www.cidoc-crm.org/. ↩
4. Olivia Madison et al., *Functional Requirements for Bibliographic Records* (Munich: K. G. Saur; International Federation of Library Associations and Institutions, 1998), http://www

.ifla.org/publications/functional-requirements-for-bibliographic-records. See also Arlene G. Taylor, ed., *Understanding FRBR: What It Is and How It Will Affect Our Retrieval Tools* (Westport, CT; London: Libraries Unlimited, 2007). ↩

5. Of course, these numbers increase constantly. For the latest Netcraft surveys, see http://news.netcraft.com/. ↩

6. Ted Nelson quoted in Nick Gibbins, "The Eighth ACM International Hypertext Conference," *Ariadne*, no. 9 (May 19, 1997), http://www.ariadne.ac.uk/issue9/hypertext. ↩

7. See http://vlib.org. ↩

8. See http://www.google.com/about/company/. ↩

9. See http://www.google.com/trends/. ↩

10. See David Austin, "How Google Finds Your Needle in the Web's Haystack," American Mathematical Society website, http://www.ams.org/samplings/feature-column/fcarc-pagerank. ↩

11. Performance—or lack thereof—is almost certainly the chief reason for the lack of success up to now of true metasearching. Because that type of search is done "live" on a number of databases and using a variety of protocols, it can be excruciatingly slow. ↩

12. See Luiz André Barroso, Jeffery Dean, and Urs Hölzle, "Web Search for a Planet: The Google Cluster Architecture," *IEEE Micro* 23, no. 2 (April 2003), http://static.googleusercontent.com/media/research.google.com/en/us/archive/googlecluster-ieee.pdf. ↩

13. Steve Lawrence and C. Lee Giles, "Accessibility of Information on the Web," *Nature* 400 (July 9, 1999): 107–09. ↩

14. Luiz André Barroso, "The Price of Performance: An Economic Case for Chip Multiprocessing," *ACM Queue* 3, no. 7 (October 18, 2005), http://queue.acm.org/detail.cfm?id=1095420. ↩

15. See http://dublincore.org/documents/dces/. ↩

16. See http://www.niso.org/apps/group_public/download.php/6577/z39-85-2001_dublin_core.pdf. ↩

17. See http://www.iso.org/iso/home/store/catalogue_ics/catalogue_detail_ics.htm?csnumber=52142. ↩

18. See http://www.dublincore.org/projects/. ↩

19. See http://www.w3.org/RDF/. ↩

20. See http://www.w3.org/XML/. ↩

21. Note that an RDF "vocabulary" is not the same as a "controlled vocabulary"; see Patricia Harpring, *Introduction to Controlled Vocabularies* (Los Angeles: Getty Research Institute, 2010), http://www.getty.edu/research/publications/electronic_publications/intro_controlled_vocab/index.html. The 2014 revised edition is only available in print. ↩

22. See http://www.w3.org/TR/rdf-schema/. ↩

23. See http://www.w3.org/TR/owl2-overview/. ↩

24. See http://www.w3.org/2013/data/. ↩

25. See http://www.w3.org/People/Berners-Lee/. ↩

26. See http://www.w3.org/standards/semanticweb/data. ↩

27. See http://wiki.dbpedia.org/Datasets for details. ↩

28. See http://www.getty.edu/research/tools/vocabularies/lod/. ↩

29. See http://www.openarchives.org/pmh/. ↩

30. Cory Doctorow, "Metacrap: Putting the torch to seven straw-men of the meta-utopia," http://www.well.com/~doctorow/metacrap.htm. ↩

31. Doctorow confusingly uses *schema* to refer to classification schemes (i.e., ways of describing content with words) rather than the more common meaning of a metadata structure as used in this publication. See "A Typology of Data Standards" in chapter 1. ↩

32. See https://www.librarything.com/. See also *What Makes LibraryThing LibraryThing* (blog), April 3, 2013, http://blog.librarything.com/main/2013/04/what-makes-librarything-librarything/. ↩

33. See https://www.flickr.com/. ↩

34. See https://www.librarything.com/forlibraries. ↩

35. See https://www.flickr.com/photos/8623220@N02/ . ↩

36. See http://www.loc.gov/marc/. ↩

37. See http://dublincore.org/documents/dces/. ↩

38. See http://www.loc.gov/standards/mods/. ↩

39. See http://www.loc.gov/ead/. ↩

40. See http://www.loc.gov/standards/vracore/. ↩

41. See http://network.icom.museum/cidoc/working-groups/data-harvesting-and-interchange/what-is-lido/. ↩

42. See http://www.oclc.org/. ↩

43. See http://www.worldcat.org/. ↩

44. However, unless the keyword "WorldCat" is included in the Google search, the record in WorldCat may not appear on the first page of search results. ↩

45. See http://books.google.com/. ↩

46. See http://www.google.com/googlebooks/partners/. ↩

47. See http://www.google.com/googlebooks/library/. ↩

Metadata Matters: Connecting People and Information

Mary S. Woodley

Revised by Murtha Baca

In the current environment of global access to the universe of electronic resources, the importance of metadata has only increased. Metadata standards and their structures are in a state of flux as they aim to accommodate "futuristic" models of information sharing. These standards reflect the functionality of how information and knowledge are stored and expressed for machine processing and how search engines can serve as better filters for discovery. In recent years we have witnessed a transition from print-based content to content that is born in digital form or made available simultaneously in multiple formats. All of this is accompanied by a blurring of the lines between articles in journals, chapters in books, books that have been digitized in their entirety, accompanying data, and structured or unstructured data that is archived as content.

There are still no magical means for perfect, seamless access to the right information in the right context. Institutions of all kinds have transitioned to automated systems to provide access to their collections and manage their assets or repositories for all their content, while institutions or communities support multiple repositories that may or may not be interoperable. Individual institutions, or communities of similar institutions, have created shared metadata standards that organize this content. These standards might include metadata elements or fields[1] with their definitions, codified rules or best practices for recording the information, and controlled lists of terms to populate access fields. We need to remember that these legacy or preexisting systems still may serve only specialized knowledge communities. Each community maintains its own structure and rules for fields of access, description, and vocabulary control (if any) that best serve it. It is when a community shares content with others or wants to reuse the information for other purposes that problems of interoperability arise. Achieving seamless and precise retrieval of information objects is not a simple process. Well-structured and carefully mapped metadata plays a fundamental role in reaching this goal.

The development of sophisticated tools to find, access, and share digital content, such as link resolvers, the Open Archives Initiative Protocol for

Metadata Harvesting (OAI-PMH) harvesters, and the emergence of the so-called Semantic Web have increased users' demand for the ability to search simultaneously across many different metadata structures. This has motivated institutions either to convert their legacy content developed for in-house use to standards more readily accessible for public display or sharing or to provide a single interface to search many heterogeneous databases or web resources at the same time. Crosswalks are at the heart of our ability to make this possible, whether they support conversion of data to a new or different standard, to harvest data from multiple resources and repackage it, to search across heterogeneous resources, or to merge information.

Definitions and Scope

For the purposes of this chapter, we will refer to *mapping* as the intellectual activity of comparing and analyzing two metadata schemas, and to *crosswalks* as the visual product of mapping. A crosswalk is a table or chart that maps the relationships and equivalencies between two or more metadata formats.[2] The first section of this chapter will deal with the different situations where crosswalks are used. The second section will focus on metasearching and harvesting metadata for reuse and enhancement by service providers. The chapter closes with case studies that serve to illustrate the issues.

Over the years many different terms have been used for what is known as metasearching: "federated searching," "broadcast searching," and "parallel searching," to name just three. We hear about search portals and find expressions like "screen scraping." In this publication we will follow the National Information Standards Organization (NISO) Metasearch Initiative's definition of metasearching as "search and retrieval to span multiple databases, sources, platforms, protocols, and vendors at one time."[3] Harvesting is not a search protocol; it is a protocol that allows the gathering or collecting of metadata records from a variety of repositories or databases in order to create union catalogs or "federated" resources. As of this writing, the OAI-PMH is the prevalent protocol being used to harvest metadata.[4]

Metadata Mapping and Crosswalks

In comparing two or more metadata element sets or schemas, distinctions and similarities must be understood on multiple levels to evaluate the degree to which they are interoperable. One definition of interoperability is "the ability of different types of computers, networks, operating systems, and applications to work together effectively, without prior communication, in order to exchange

information in a useful and meaningful manner. There are three aspects of interoperability: semantic, structural, and syntactical."[5]

Semantic mapping is analyzing the definitions of the elements or fields so as to determine whether they have the same or a similar meaning. Crosswalks are the visual representations or "maps" that show these relationships.[6] A crosswalk supports the ability of a search engine to query fields with the same or similar content in different databases; in other words, it supports "semantic interoperability." Crosswalks are not only important for supporting the demand for "one-stop shopping," or cross-domain searching, they are instrumental for converting data in one format to another format that is more widely accessible.

Structural interoperability refers to the presence of data models or wrappers that specify the semantic scheme being used. For example, the Resource Description Framework (RDF) is a model that allows metadata to be defined and shared by communities.[7]

Table 1. Example of a Crosswalk of a Subset of Elements from Different Metadata Schemes

CDWA	MARC	EAD
Object/Work-Type	655 Genre/form	<controlaccess><genreform>
Titles or Names	24Xa Title and Title—Related Information	<unittitle>
Creation–Date	260c Imprint—	<unitdate>

CDWA	MARC	EAD
	Date of Publication	
Creation-Creator-Identity	1XX Main Entry 7XX Added Entry	<origination><persname><origination><corpname><origination><famname><controlacces
Subject Matter	520 Summary, etc.6xx Subject Headings	<abstract><scopecontent><controlaccess><subject>
Current Location	852 Location	<repository><physloc>

Table 2. Example of a Crosswalk: MARC21 to Simple Dublin Core

MARC fields	Dublin Core elements
130, 240, 245, 246	Title

MARC fields	Dublin Core elements
100, 110, 111	Creator
100, 110, 111, 700, 710, 711	Contributor
600, 610, 630, 650, 651, 653	Subject / Keyword
Notes 500, 505, 520, 562, 583	Description
260 $b	Publisher
581, 700 $t, 730, 787, 776	Relationship
008/ 07-10 260 $c	Date

Mapping metadata elements between standards is only one level of crosswalking. At another level of semantic interoperability are the content standards or cataloging rules for populating metadata elements or fields, such as the form for personal names, encoding standards for dates, and thesauri used for

topical or subject headings. A weakness of crosswalks of metadata elements alone is that the results of a query will be less successful if the name or concept is expressed differently in each resource. By using controlled vocabularies for identifying people, places, corporate bodies, and concepts, it is possible to greatly improve retrieval of relevant information associated with a particular concept; it is hoped that if linked open data makes the Semantic Web a reality, this kind of vocabulary-enhanced search and retrieval will become much more prevalent than it is at present.

Some databases (such as library and other discovery systems) provide access to controlled terms along with cross-references for variant forms or words that point the searcher to the preferred form. This optimizes the searching and retrieval of digital objects (bibliographic records, images, sound files, etc.). There is currently no universal authority file that catalogers, indexers, and users can consult. Each cataloging or indexing domain tends to develop its own thesauri or lists of terms designed to support the research needs of a particular community. Crosswalks have been used to migrate the *data structure* of a vocabulary from one format to another, but only recently have there been projects to map the *data content* that actually populates that structure.[8]

To meet the need to share records between international libraries, the Online Computer Library Center (OCLC) and the Library of Congress have spearheaded an international initiative to develop the Virtual International Authority File (VIAF).[9] As of this writing the VIAF is a file of millions of authority records for personal and corporate names from libraries and other institutions from all over the world that supports data sharing and exchange and enhances retrieval. This is important when searching the records of many databases simultaneously, where precision and relevancy become even more crucial. This is especially true if one is searching single-search query bibliographic records, records from citation databases, and full-text resources at the same time. Integrated authority control significantly improves both retrieval and interoperability in searching resources like these that are aggregations of disparate metadata records.[10]

The Gale Group solved the problem of multiple-subject thesauri by creating a single thesaurus and mapping the controlled vocabulary from the individual databases to that thesaurus. It is unclear to what extent this merging of data compromised the depth and coverage of the controlled terms in the individual databases.[11]

The Simple Knowledge Organization System (SKOS), a project developed by the World Wide Web Consortium's Semantic Web Best Practices and Deployment Working Group, is a set of specifications for organizing,

documenting, and publishing taxonomies, classification schemes, and vocabulary schemes such as thesauri, subject lists, and glossaries or terminology lists within an RDF framework.[12] SKOS is a specific application that is used to express mappings between knowledge organization schemes. The ability to map vocabularies as well as metadata element standards will strengthen the ability of search engines to search across heterogeneous databases more effectively.[13]

Crosswalks for Repurposing and Transforming Metadata

The notion of repurposing metadata covers a broad spectrum of activities: converting or transforming metadata from one standard to another; migrating metadata from one legacy standard to a different one; integrating two metadata standards; and harvesting or aggregating metadata created using a shared community standard or different metadata standards. Dushay and Hillmann note that the library community has an extensive and successful history of aggregating bibliographic metadata records encoded in the MARC (Machine-Readable Cataloging Record) format created by many different libraries that share content standards (Anglo-American Cataloging Rules, Library of Congress authorities and classifications, and so on). However, aggregating metadata records from different repositories may create confusing display results, especially if some of the metadata was automatically generated or created by institutions or individuals that do not follow best-practice standards.[14]

Conversion projects transfer metadata fields or elements from one standard to another. Institutions have converted data for a variety of reasons; for example, when upgrading to a new system because the legacy system has become obsolete, or when the institution has decided to provide public access to some or all of its content. Conversion is accomplished by mapping the structural elements in the older system to those in the new system. In practice, there is rarely the same degree of granularity among all of the fields in the two systems, which makes the process of converting data from one system to another more complex (see table 1). Data fields in the legacy database may not have been well defined or may contain a mixture of types of information. In the new database, this information may reside in separate fields. Identifying the unique information within a field in order to map it to another field may not always be possible; manipulating the same data several times before migrating it may be necessary.

Some of the common misalignments that occur when mapping between metadata include:[15]

1. *Fuzzy matches*: A concept in the original database does not have a perfect equivalent in the target database. For example, when mapping the *creation-creator-identity-nationality/culture/race* elements in Categories for the Description of Works of Art (CDWA)[16] to the *subject* element in Dublin Core, which does not have the same exact meaning. Since the Dublin Core *subject* element is a much broader category, this is a "fuzzy" match.

2. *Hybrid records*: Although some metadata standards (e.g., Dublin Core) follow the principle of a one-to-one relationship between a metadata record and an "item"—be it analog or digital—in practice many memory institutions use a single metadata record to record information about an original object as well as its digital surrogate, thus creating a sort of "hybrid" record. When migrating and harvesting data, this may pose problems if the harvester cannot distinguish between the elements that describe the original item and those that describe the digital surrogate.

3. *One element into many*: Data that exists in one element in the original schema may exist in separate elements in the target database. For example, the CDWA *creation-place* element may appear in the *subject* and the *coverage* elements in Dublin Core.

4. *Many elements into one*: Data in separate elements in the original schema may be in a single element in the target schema. For example, in CDWA the birth and death dates for a "creator" are recorded in the *creator-identity-dates* element as well as in other elements, all apart from the element for the creator's name. In the MARC format, birth and death dates are a "subfield" in the string for the "author's" name.

5. *No correspondence*: There is no element in the target schema equivalent to the element in the original schema, and unrelated data may be forced into a single "bucket" with unrelated or loosely related content.

6. *Evolving standards*: In some cases, the original "standard" is actually a mix of standards that evolved over time. Kurth, Ruddy, and Rupp have pointed out that even in transforming metadata from a single metadata standard, it may not be possible to use the same conversion mapping for all records. The Cornell University Library metadata project revealed the difficulties in "transforming" MARC library records to TEI (Text Encoding Initiative). Not only had the use of MARC in the library evolved over time, but the rules guiding how content was added had changed from pre–Anglo-American Cataloguing Rules to AACR2rev.[17] Some MARC subelements were dropped, and others were added, so that various communities could more easily reuse library records for their own purposes.[18] The conversion of library records created according to AACR2rev in order to make them conform with the new cataloging standard, Resource Description and Access (RDA), is beyond the scope of this chapter. In the years to come, as more and more libraries create original bibliographic records according to RDA, interesting challenges are sure to emerge.

7. *Incomplete correspondence*: In only a few cases does the mapping work equally well in both directions, due to differences in granularity and community-specific practices (see number 2 above.) The large Getty metadata crosswalk[19] was created by mapping in a single direction: CDWA was analyzed, and other data systems were mapped to its elements. However, there are some types of information recorded in MARC that are lost in this process; for example, the *publisher* and *language* elements are important in library records but are less relevant to CDWA.

8. *Differing structures*: One metadata set may have a hierarchical structure with complex relationships while the other may have a flat file organization—EAD (hierarchical) versus MARC (flat), for example.[20]

Metasearching

The number of metadata standards is growing, and it is unrealistic to think that every standard can be mapped or converted to a common standard that will satisfy both general and domain-specific needs. (At one time it was believed that Dublin Core would be the "Holy Grail" in this sense, but practical

experience in a variety of metadata communities has proved otherwise). An alternative is to maintain the separate databases that have been developed to support the needs of specific communities in their original schemas and to offer a search interface that allows users to search across the various heterogeneous databases simultaneously. This can be achieved through a variety of methods.

The best-known and most widely used metasearch engines in the library world are based on the Z39.50 protocol.[21] The development of this protocol was initiated to create a "virtual" union catalog that would enable libraries to share their cataloging records (then all in the MARC format). With the advent of the Internet, the protocol was extended to enable searching of abstracting and indexing services and full-text resources when they were Z39.50 compliant. Some information professionals touted the Z39.50 protocol as a "one-stop-shopping" solution that would provide users with seamless access to all authoritative information about a particular search query. At the time of its initial implementation, the Z39.50 protocol had no competitors, but it was not without its detractors.[22]

The library community is divided over the efficacy of metasearching. When are "good enough" search results really good enough? Often the results created through a keyword query have high recall and low precision, leaving the user at a loss on how to proceed. Users who are familiar with web search engines will often take the first hits generated by a search regardless of their suitability. Authors have pointed to the "success" of Google to reaffirm the need for federated searching without referring to any studies that evaluate the satisfaction of researchers.[23] A preliminary study conducted by Lampert and Dabbour on the efficacy of federated searching laments the fact that studies have focused on technical aspects without considering users' search and selection habits and the impact of federated searching on information literacy.[24] Unfortunately, this is a familiar phenomenon in the world of library technology: frequently, the emphasis is on technical issues and solutions, while user needs and behavior and user interface and usability issues tend to be neglected.

What are some of the main issues associated with metasearching? In some interfaces, results may be displayed in the order retrieved or by computer-determined relevance (which may have little or nothing to do with how relevant the results actually are to the user's query), either sorted by categories or integrated. Having the choice of searching a single database or multiple databases allows users to take advantage of the specialized indexing and controlled vocabulary resources of a single database or to cast a broader net. There are several advantages of a single gateway or portal to information. Users

are not always aware which of the many databases that exist for a particular domain will provide them with the best information, or they may not be aware of them at all. Many libraries have attempted to list domain-specific databases by categories and provide brief descriptions of them, but users tend not to read lists, and this type of "segregation" of resources neglects the interdisciplinary nature of research. Few users have the tenacity to read lengthy A–Z lists of databases or to ferret out databases relevant to their queries when they are "buried" in lengthy menus. On the other hand, users can be overwhelmed by large search result sets and may have difficulty finding what they need, even if the results are sorted by relevancy ranking.[25]

Some of the commercial "metasearch" engines for libraries are still using the Z39.50 protocol to search across multiple repositories simultaneously.[26] In simple terms, this protocol allows two computers to communicate in order to retrieve information; a client computer will query another computer, called a server, which provides a result. Libraries employ this protocol to support searching other library catalogs along with abstracting and indexing services and full-text repositories. This approach restricts the searches and results to those databases that are Z39.50 compatible. The results users see from searching multiple repositories through a single interface and those achieved when searching repositories in their native interfaces may differ significantly due to the following factors:

- How the server interprets the query from the client. This is particularly clear when using multiple keywords. Some databases will search a keyword string as a phrase; some will automatically add the Boolean operator "and"; and some will automatically add the Boolean operator "or."
- How a specific person, place, event, object, idea, or concept is expressed in one database may not be how it is expressed in another (the vocabulary issue).
- How results are displayed in various metasearch engines: in the order they were retrieved; by the database in which they were found; sorted by date; or integrated and ranked by computer-determined relevancy. The greater the number of results, the more advantage there is in sorting by relevancy and date.[27]

The limitations of Z39.50 have encouraged the development of alternative solutions to federated searching to improve results. One approach is the Metasearch XML Gateway (MXG), which allows queries in an XML format

from a client to generate result sets from a server in an XML format.[28] Another approach used by metasearch engines when the database does not support Z39.50 is relying on HTTP parsing or "screen scraping." In this approach, the search retrieves an HTML page that is parsed and submitted to the user as a set of search results. Unfortunately, this approach requires a high level of maintenance, as the target databases change continually and the level of accuracy in retrieving content varies from one database to another.

Table 3. Methods for Gathering, Aggregating, and Making Metadata Available

Method	Description	Examples
Batch processing of contributed data	A contributing institution/data provider makes its metadata records available in a standard format for batch loading/ingest by a service provider.	• Contribution of MARC reco OCLC's WorldCat • Contribution of batch data Getty vocabularies • Contribution of MARC reco EAD finding aids to OCLC ArchiveGrid
Real-time access to data via web services or linked-data APIs	Data is made available in real time via web APIs (application programming). The data is provided in documented standardized serializations (e.g., JSON, JSON-LD, XML, linked open data), and can be used/reused in local applications, combined with local search results, etc.	• The WorldCat search API • The Getty vocabularies web APIs, which make it poss retrieve regularly update from the Art & Architect Thesaurus and the other vocabularies • The Getty vocabularies link data SPARQL endpoint • The Europeana APIs • The Ex Libris Alma APIs
Harvesting	Records expressed in a standard metadata schema (e.g., Dublin Core) are made available by data providers on specially configured servers. The records are harvested, batch processed, and made available by service providers from a single database or index. The service provider preprocesses the contributed data and stores it locally before it is made	• The OAISTER database • The Digital Public Library of • Europeana

Method	Description	Examples
	available for searching by users. In order for records to be added or updated, data providers must post fresh metadata records, and service providers must reharvest, batch process, and integrate the new and updated records into the union catalog.	
Screen scraping	The extraction of display data (usually unstructured) and hidden embedded data from web pages.	• How search engines like Go robots to extract data fr pages for creating their l search indexes
Metasearch of distributed metadata records	Diverse databases on different platforms and often with different metadata schemas are searched in real time via one or more protocols (e.g. Z39.50, screen scraping, APIs). The service provider does not preprocess or store data but rather processes data only when a user launches a search. Search results are usually displayed target-by-target, as integrating retrieved records from each different database into a single ranked result set is difficult and time consuming.	

The key to improvement may lie in implementing multiple protocols rather than a single protocol. Currently, some vendors[29] are combining Z39.50 and XML gateway techniques to increase the number of "targets" or servers that can be queried.

The Summon discovery tool, developed by Serials Solutions, a subsidiary of Proquest, purports to be an approach that goes beyond Z39.50.[30] Summon does not link out to other databases to retrieve content, but rather ingests metadata and full texts from a variety of resource producers into a single repository. Jeffrey Daniels and Patrick Roth of Grand Valley State University, in Allendale, Michigan, described their implementation of Summon and the mapping between catalog records for books and the Summon fields.[31] Unlike Dublin Core, the Summon metadata model provides far more granular access, reflecting the wide variety of publication types and metadata standards of the resources in the federated repository.[32]

Metadata Harvesting

Searching across multiple heterogeneous databases in real time causes significant performance issues; retrieval of search results is so slow that users are likely to lose patience and abandon their queries. Another approach is to create single repositories by "harvesting" metadata records from various resources and putting them into a single database or index. The challenge is how to ensure that the harvested records "play well" and are understandable in their new environment while maintaining their original integrity.

Within a single community, union catalogs can be created where records from different institutions can be centrally maintained and searched with a single interface. This is possible when the community shares the same rules of description and access as well as the same protocol for encoding the information. OCLC, the major bibliographic utility in North America, provides what is essentially a huge union catalog representing the holdings of many libraries around the world.[33] Local union catalogs, on the other hand, are typically based on geography and/or a shared system; for example, the University of California and California State University systems each maintain their own union catalogs. Interoperability tends to be high in such shared systems because of the shared rules for creating the catalog records.[34]

To simplify the process for federating diverse resources and to preserve interoperability, the OAI-PMH adopted Simple Dublin Core as its minimum standard. Data providers who expose their metadata for harvesting are required to provide records in Simple Dublin Core expressed in XML and to use UTF-8 character codes in addition to any other metadata formats they expose. The data providers may expose all of their metadata records or selected sets of records for harvesting. Data services operating downstream of the harvesting source may enhance the value of the metadata in the form of added fields (for example, additional audience or grade-level metadata elements for educational resources). Data services have the potential to provide a richer contextual environment where users can find related and relevant content. OAI harvesters request the data through HTTP. Repositories using a richer metadata standard than Dublin Core need to map their content to Simple Dublin Core before exposing it for OAI harvesting. Part of the challenge of creating a crosswalk is understandings the pros and cons of mapping all of the content from the larger metadata set or deciding which subset of that content should be mapped.

The limitations of mapping between metadata standards have been outlined above. Bruce and Hillmann established a set of criteria for measuring the quality of metadata records harvested and aggregated into a larger collection. The criteria may be divided into two parts, one that evaluates the metadata content as a whole for completeness, currency, accuracy, and provenance; and

another that evaluates the technical solutions: conformance (or lack thereof) of the metadata sets and application profiles, and consistency and coherence of the metadata standards.[35] In the context of harvesting data for reuse, Dushay and Hillmann have identified four categories of metadata problems in the second criterion for quality:

- *missing data*: data that the metadata contributor considered unnecessary to expose for harvesting;
- *incorrect data*: data entered in the wrong metadata element or encoded improperly;
- *confusing data*: data that uses inconsistent formatting or punctuation;
- *insufficient data*: incomplete data concerning the encoding schemes or vocabularies that were used.[36]

A recent study evaluating the quality of harvested metadata found that while collections of metadata records from a single institution did not vary significantly in terms of the criteria above, the amount of "variance" increased dramatically when the aggregations of harvested metadata came from many different institutions.[37]

Roy Tennant echoes the argument that this problem may be largely due to mapping richer metadata records to Simple Dublin Core. He suggests that both data providers and service providers consider exposing and harvesting metadata that is expressed in schemas that are richer than Simple Dublin Core. He argues that the metadata harvested should be as granular as possible and that the service provider should transform and normalize content such as date information, which can be expressed in a wide variety of encoding schemes (or following no scheme or standard at all).[38] This approach creates a single silo for searching rather than decentralized or distributed searching. In order to facilitate searching, an extra "layer" is added to the repository to manage the mapping and searching of heterogeneous metadata standards within a single repository. Godby, Young, and Childress suggested a model for creating a repository of metadata crosswalks that could be exploited by OAI harvesters. Documentation about the mapping would be associated with the standard used by data providers and the standard used by the service providers encoded in the Metadata Encoding and Transmission Standard.[39] This would provide a mechanism that supports repositories with OAI-harvested metadata in dealing with transforming metadata. As of this writing, it remains to be seen whether metadata harvesting (and indeed metasearching as it is currently understood)

will eventually be made obsolete as a result of the widespread adoption of the Semantic Web of linked open data.

CASE STUDIES

Each instance of conversion, transformation, metasearching, or harvesting brings its own unique set of issues. Below are examples of projects that illustrate the complexities and pitfalls of creating crosswalks in order to transfer data to a new schema or to support cross-domain searching.

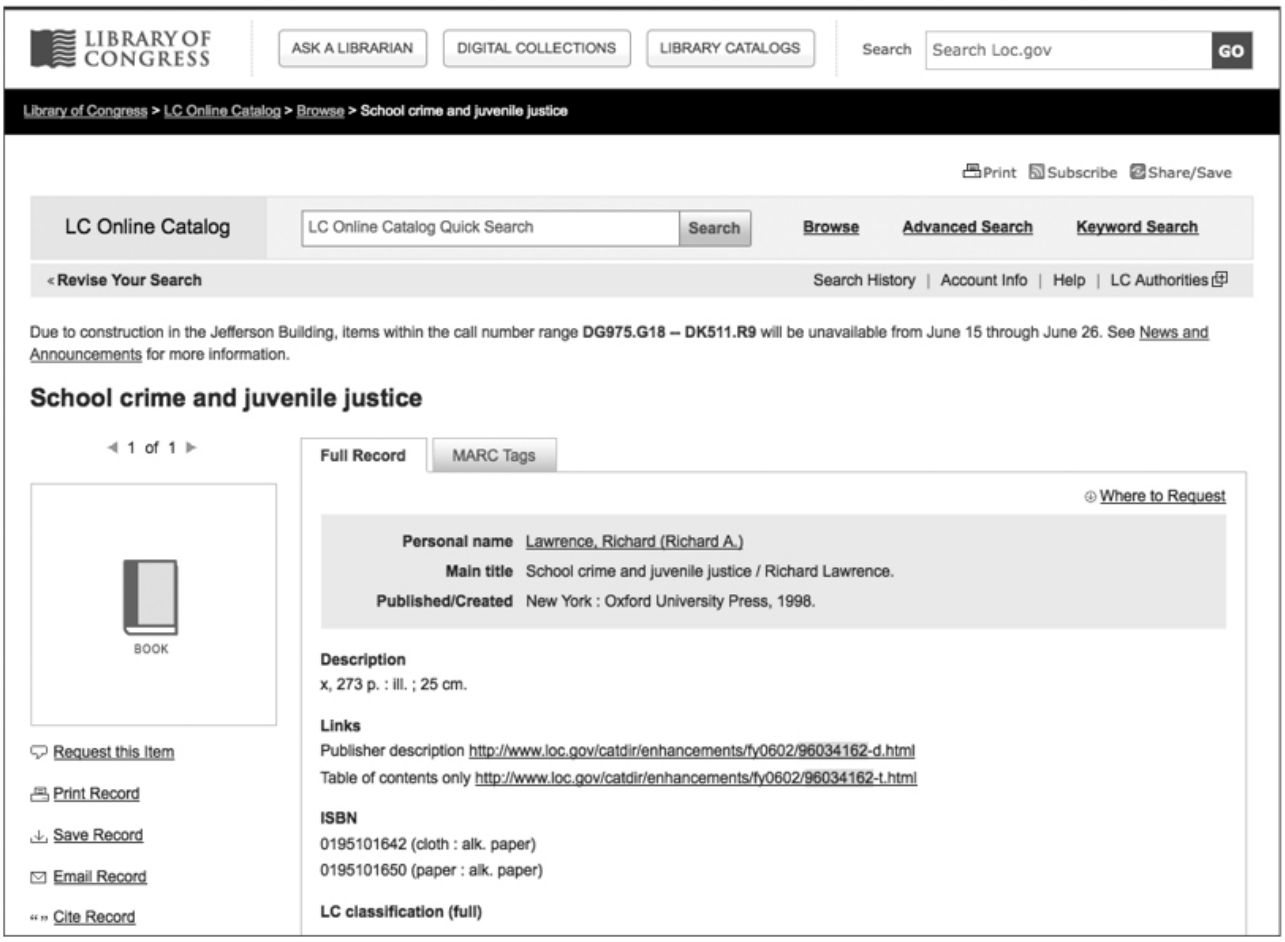
Case Study 1: Repurposing Metadata—ONIX and MARC21/RDA

In 2001 a task force was created by the Cataloging and Classification: Access and Description Committee, an Association for Library Collections and Technical Services committee of the American Library Association, to review a standard developed by the publishing industry and to evaluate the usefulness of data in its records that could potentially enhance the bibliographic records used by libraries. The task force reviewed and categorized the ONIX (Online Information Exchange) metadata element set and found that fields developed to help bookstores increase sales also had value for library users.[40] In response, the Library of Congress directed the Bibliographic Enrichment Advisory Team to repurpose three categories of data supplied by the publishers: table of contents, descriptions, and sample texts. The publisher-generated content is saved on servers at the Library of Congress and appears in the bibliographic records as links (see fig. 1).[41] Although the metadata was originally created to manage books as business assets and to provide information to bookstores that would help increase book sales, the same metadata (accessed by a hyperlink) has been incorporated into the bibliographic records used by libraries to provide additional information for users so that they can more easily evaluate the particular item.

In 2006 the Joint Steering Committee for the Development of RDA proposed a new crosswalk that would map RDA and ONIX.[42] The International Federation of Library Associations and Institutions Cataloging Section developed a RDA-ONIX mapping of only two areas: content form and media type.[43] The mapping is a moving target, since in the meantime not only has the library world transitioned from AACR2rev to RDA and from MARC to other metadata schemes (notably MODS, though as of this writing MARC remains the prevailing metadata schema for library production systems), but ONIX has moved to version 3.0. Carol Jean Godby at OCLC has written a report with a

crosswalk that maps ONIX 3.0 to MARC21.[44] This replaces earlier crosswalks used by OCLC to incorporate data from ONIX records to enhance bibliographic records. Godby quotes Karen Coyle to the effect that the migration to RDA as the new data content standard and the use of identifiers rather than descriptive strings should make it easier to automate the repurposing of data from ONIX records for incorporation into library catalog records.[45]

Figure 1. MARC Record (Brief Display) with Embedded Links to ONIX Metadata (Publisher Description and Table of Contents)



In an interesting side note, Godby recognizes that the complexity of the XML structure for both standards makes it difficult to visualize the relationship between them in a standard table. Instead, she compares separate sections of the records rather than creating one long table. At the end of the report, she addresses the possibility that future implementation of RDA may provide some improvement in the ability to share data between standards.[46] The takeaway is that no crosswalk is static. As standards change, the mappings between and among them must be continually updated.

Case Study 2: Developing Standards for the Cultural Heritage Community—CIMI, VRA Core 4.0?

Compared to the library community, the cultural heritage community is a latecomer to creating data standards to share content and to facilitate search and retrieval. An early project, the Consortium for the Computer Interchange of Museum Information (CIMI), was founded in 1990 to promote the creation of standards for sharing cultural information electronically.[47] In 1998 CIMI designed a project to map museum data to Simple Dublin Core.[48] The main goal of CIMI was to test the efficacy of automating the conversion of nonstandard legacy museum data to a "web-friendly" standard—that is, to Dublin Core—with as little human intervention as possible.

The CIMI test bed was successful in that it demonstrated the pitfalls of migrating between two different sets of metadata whose granularity and purposes differ so greatly. It is a good example of how difficult it is to map data that resides in very specific and narrowly defined fields to a schema that lacks the same depth or coverage. During the transformation process, information from existing museum records ended up being entered into inappropriate metadata elements or duplicated in two separate elements. For example, since museum systems traditionally recorded subject information as a single field without subfield coding, a subject string like "baroque cathedral" was duplicated in both the Dublin Core *coverage.temporal* metadata element and the *coverage.topical* element. There are two ways of looking at this dilemma: there is no program that is sufficiently sophisticated to deconstruct a topical string into its component parts (temporal, topical, and geographical) for migrating to the appropriate separate element; or, that migrating to separate elements was not appropriate and duplicating the information was unnecessary.

In the mid-1990s, paralleling the work of the CIMI project, the visual resources community—led by the Visual Resources Association (VRA), the leading organization in North America for visual resources information professionals—developed a schema for describing image collections. After reviewing the elements of description employed by more than sixty institutions, the community developed a core group of thirteen elements based on the *Categories for the Description of Works of Art* (CDWA).[49] The VRA schema has five metadata elements that describe the content of an image and five elements that describe a digital or other surrogate. The Library of Congress hosts the VRA Core 4.0 standard, which is expressed as an XML schema.[50] Mapping between VRA Core and Dublin Core is fairly straightforward, so long as the underlying data has been consistently recorded.

The cultural heritage organizations and projects that use the CONTENTdm collection management system for their digital collections have the choice of using Dublin Core or VRA Core as their metadata standard. Dublin Core has a longer history with library projects than VRA Core and is a standard with which many libraries have more familiarity. One challenge is that Dublin Core purportedly follows the principle that each record should describe either the original *or* a digital surrogate, but not both.[51] In reality, it is not unusual for institutions and projects to violate this principle for the sake of expediency: many libraries lack a separate preservation database or digital asset management system in which they can record details concerning the digital surrogate they must manage. In contrast, a VRA Core record—like many records in photo archives and other image repositories—has a "hybrid" structure that supports the description of both the original work of art, architecture, or material culture and its visual surrogates, digital or otherwise. The VRA Core schema has the added advantage of describing the cultural aspects of the object; for example, *culturalContext* and *creatorRole* are elements in the VRA Core schema; Dublin Core could never accommodate this kind of specificity.

Providing a metadata structure to record categories of content (a "data *structure* standard" according to our typology of data standards; see table 1 in chapter 1) is not the same as providing the rules to follow in populating the individual metadata categories ("data *content* standard," according to our typology). Cataloging Cultural Objects (CCO)[52] is a data content standard establishing the rules for cataloging cultural materials and their visual (including digital) surrogates for the cultural heritage community. CCO was conceived in 1999 and was published in the summer of 2006 by the American Library Association. The need of a transmission standard to support data using CCO led to the creation of the CDWA Lite XML schema.[53]

Case Study 3: Preparing Metadata Records in the CDWA Lite XML Schema for Harvesting by Artstor

Experience from the CIMI experiment and other projects showed that Dublin Core and the Metadata Object Description Schema (MODS) were not sufficient to handle the kinds of information needed by the cultural heritage community. The Getty Trust proposed another approach that would enhance the process of making the legacy content found in library and museum collections management systems accessible to the public.

In 2005 the Getty Trust partnered with Artstor to test the efficacy of harvesting data from legacy databases for inclusion in the Artstor Image Library. The Getty Museum and Getty Research Institute (the data providers) worked together with Artstor (the service provider) to develop an XML schema that could be used with the OAI protocol to harvest both data and images (known as "resources" in OAI parlance). The schema developed, CDWA Lite, is a subset of the huge CDWA element set expressed in XML. This XML schema is comprised of 22 of the more than 300 metadata elements included in the complete CDWA specification. The objective of the project was to offer museums and image repositories a less labor-intensive approach to sharing their content. In order to expedite the process, the schema was optimized to work with OAI-PMH, then, as now, the prevailing protocol for metadata harvesting.[54]

Two collections were chosen for the project: paintings on public display in the Getty Museum and historical tapestries in the photo archive of the Getty Research Institute. The working group determined that the museum records contained more information than was necessary or appropriate for sharing in a "union" repository like Artstor. Examples of information that was not considered appropriate for inclusion in the Artstor contribution included provenance, exhibition history, specific location information within the museum, and some metadata elements that contained administrative or confidential information (such as the price paid for an object). The group worked under the assumption that the URL embedded in the Artstor record that links to the web page and image(s) for a particular object on the Getty's website would provide the user with more detailed information in the object's own institutional context. Therefore, a subset of information elements was selected for the project. Fortunately, the Getty Museum uses a schema that maps to CDWA for the basis for their collection management system. The in-house content guidelines are similar to the CCO guidelines, but some of the data needed to be manipulated during the export process. For instance, the *object-type* element in the Getty Museum system uses the plural form (*paintings*, not *painting*) and therefore does not comply with the CCO standard. The project team members responsible for the metadata mapping determined which information to select for mapping and which information to exclude altogether.

Migrating the tapestry records from the Getty Research Institute was more complex, since the records were created in a proprietary database using a nonstandard, collection-specific schema. In this case the working group approached the mapping differently, choosing to map the nonstandard metadata records in their entirety to the CDWA Lite XML schema. As part of the conversion, the nonstandard diacritics in the photo archive databases had to be converted to Unicode UTF-8, which is required by the OAI protocol. Although

content values were successfully migrated, the more than fifty-five elements in a Getty Research Institute tapestry record had to be “forced into” the twenty-two CDWA Lite metadata elements. Once harvested, the data and the images were converted to the Artstor standard, which is based on CDWA Lite.

The experience that the Getty Museum and Getty Research Institute gained in crosswalking and preparing legacy metadata records for harvesting and reuse led to the realization that it would be desirable to have a single, simple schema to facilitate sharing metadata relating to cultural heritage objects from museums, including those not dedicated to the fine arts. The decision was made to modify the CDWA Lite schema for harvesting metadata with as little loss of information as possible and to improve the mapping between the various metadata standards used by the cultural heritage communities.

Case Study 4: LIDO

The standards communities that had developed CDWA Lite, museumdat, SPECTRUM, and the CIDOC CRM worked collaboratively to create a new standard to support sharing content among the scientific, bibliographic, and cultural heritage communities. The result was an XML schema called Lightweight Information Describing Objects (LIDO).[55] LIDO is not a substitute for more robust standards like CDWA or the CIDOC CRM but can serve as a common standard to map metadata from various repositories and as a harvesting standard to work with OAI-PMH.[56] It contains fourteen groups of metadata elements, only three of which are required: *object/work type*, *title/name*, and *record*. A record has seven areas, four of which are descriptive and three of which are administrative. Like many current metadata standards, LIDO is expressed in XML, which supports structural data that bundles elements in order to create sets of related elements. The alphabetical list of LIDO elements[57] is a combination of a data dictionary—where each element has its description, tags, and restrictions—and a crosswalk of the four main cultural heritage metadata standards to LIDO.

Who has implemented LIDO?[58] An important example is Europeana,[59] a federated repository of cultural heritage metadata records that numerous European repositories use to share their content through harvesting and ingestion activities. The Yale Center for British Art has also adopted the standard for the inclusion of its rich metadata in a discovery portal being developed there.[60]

Conclusion

The technological universe of crosswalks, mapping, federated searching over heterogeneous databases, and aggregating full-text resources and metadata sets into single repositories is rapidly changing. Crosswalks are still at the heart of supporting conversion projects and enabling the semantic interoperability that makes it possible to search across heterogeneous distributed databases. Inherently, there will always be limitations to crosswalks; there is rarely a one-to-one correspondence between metadata standards, even when one standard is a subset of another. Mapping the elements or fields of metadata systems is only one part of the picture. Crosswalks of controlled terms or thesauri will further enhance searchers' ability to retrieve the most precise search results. As the number and size of online resources increases, the ability to refine searches and to use controlled vocabularies and thesauri both at the metadata creation stage and at the moment of searching will become increasingly important.

1. What determines the granularity or level of detail in any element will vary from standard to standard. In various systems, a single instance of metadata may be referred to as a field, a label, a tab, an identifier. Margaret St. Pierre and William P. LaPlant Jr., "Issues in Crosswalking, Content Metadata Standards" (white paper, National Information Standards Organization, Baltimore, 1998), http://www.niso.org/publications/white_papers/crosswalk/. ↩

2. A metadata standards crosswalk that includes many of the standards discussed here is available at http://www.getty.edu/research/publications/electronic_publications/intrometadata/crosswalks.html. ↩

3. See http://www.niso.org/workrooms/mi. ↩

4. See http://www.openarchives.org/. Of specific interest, under documents, is the documentation on the protocol for harvesting, a tutorial, and a link to the National Science Digital Library's Metadata Primer. ↩

5. See http://www.dublincore.org/documents/usageguide/glossary.shtml#I. ↩

6. Ibid. Crosswalks can be expressed or coded for machines to automate the mapping between standards. ↩

7. See http://www.w3.org/RDF. ↩

8. Sherry L. Vellucci, "Metadata and Authority Control." *LRTS* 44 (1) (January 2000): 33–43. ↩

9. See http://viaf.org/. ↩

10. Biligsaikhan Batjargal et al., "Linked Data Driven Dynamic Web Services for Providing Multilingual Access to Diverse Japanese Humanities Databases," in *DC-2103—Proceedings of the International Conference on Dublin Core and Metadata Applications, September 2–6, 2013, Lisbon, Portugal* (Dublin Core Metadata Initiative, 2013), http://dcevents.dublincore.org/IntConf/dc-2013/paper/view/150/72. ↩

11. Jessica L. Milstead, "Cross-File Searching," *Searcher* 7, no. 1 (May 1999): 44–55. ↩

12. See http://www.w3.org/2004/02/skos/. ↩

13. For an introduction to SKOS, see Alistair Miles et al., "SKOS Core: Simple Knowledge Organization for the Web," in *DC-2005—Proceedings of the International Conference on Dublin Core and Metadata Applications, September 12–15, 2005, Madrid, Spain* (Dublin Core Metadata Initiative, 2005), http://dcpapers.dublincore.org/pubs/article/view/798. ↩

14. Naomi Dushay and Diane Hillmann, "Analyzing Metadata for Effective Use and Re-use," in DC-2003—Proceedings of the DCMI International Conference on Dublin Core and Metadata Applications, September 28–October 2, 2003, Seattle, Washington (Dublin Core Metadata Initiative, 2003), http://dcpapers.dublincore.org/pubs/article/view/744. ↩

15. St. Pierre and LaPlant, "Issues in Crosswalking." ↩

16. CDWA is a standard for cataloging cultural objects maintained by the J. Paul Getty Trust and the College Art Association; see http://www.getty.edu/research/publications/electronic_publications/cdwa/. ↩

17. Martin Kurth, David Ruddy, and Nathan Rupp, "Repurposing MARC Metadata: Using Digital Project Experience to Develop a Metadata Management Design," *Library High Tech* 22, no. 2 (2004): 153–65. ↩

18. Numerous presentations and publications explain the differences between the two standards. The Library of Congress has created an excellent crosswalk between MARC and RDA at http://www.loc.gov/catdir/cpso/RDAtest/training2word9.doc. ↩

19. See http://www.getty.edu/research/publications/electronic_publications/intrometadata/crosswalks.html. ↩

20. For other examples of crosswalk issues, see "Challenges and Issues with Metadata Crosswalks," *Online Libraries & Microcomputers* (April 2002). ↩

21. For a fuller history of the development of the standard, see Clifford A. Lynch, "The Z39.50 Retrieval Standard—Part 1: A Strategic View of its Past, Present and Future," *D-Lib* 3, no. 4 (April 1997), http://www.dlib.org/dlib/april97/04lynch.html. ↩

22. Roy Tennant, "Interoperability: The Holy Grail," *Library Journal*, July 1, 1998, available at http://roytennant.com/column/?fetch=data/95.xml. Tennant argued that interoperability was the holy grail; to others it is Z39.50 and its successor, ZING (Z39.50 International: Next Generation). ZING is no longer a standard maintained by the Library of Congress. It has been folded into SRU (Search/Retrieval via URL); see http://www.loc.gov/standards/sru/. ↩

23. Judy Luther, "Trumping Google? Metasearching Promise," *Library Journal*, October 1 2003, 36–39. ↩

24. Lynn Lampert and Kathy Dabbour, "Librarian Perspectives on Teaching Metasearch and Federated Search Technologies," *Internet Reference Services Quarterly* 12, nos. 3–4 (September 2007): 253–78. ↩

25. Terence K. Huwe, "New Search Tools for Multidisciplinary Digital Libraries," *Online* 23, no. 2 (March 1999): 67–70. ↩

26. The protocol is an NISO standard, http://www.niso.org/standards/resources/Z39.50_Resources, which is maintained by the Library of Congress; see http://www.loc.gov/z3950/agency and the International Organization for Standardization standard, http://www.iso.org/iso/en/CatalogueDetailPage.CatalogueDetail?CSNUMBER=27446. A good history of Z39.50 was published in William E. Moen, "Interoperability and Z39.50 Profiles: The Bath and US Profiles for Library Applications" (2001) in *From Catalog to Gateway:*

Charting a Course to Future Access, ed. Bill Sleeman and Pamela Bluh (Chicago: Association for Library Collections and Technical Services, American Library Association, 2005), 113–20. ↩

27. Tamar Sadeh. "The Challenge of Metasearching," *New World Library* 105, nos. 1198–1199 (2004): 104–12. ↩

28. NISO Metasearch Initiative, Standards Committee BC, Task Group 3, "Metasearch XML Gateway Implementers Guide, Version 1.0" (Bethesda, MD: National Information - Standards Organization, 2006), http://www.niso.org/publications/rp/RP-2006-02.pdf. ↩

29. Ex Libris's MetaLib product and Endeavor's now-defunct Encompass system adopted this approach. ↩

30. See http://www.dc4.proquest.com/en-US/products/brands/pl_ss.shtml. ↩

31. Jeffrey Daniels and Patrick Roth, "Incorporating Millennium Catalog Records into Serials Solutions' Summon," *Technical Services Quarterly* 29, no. 3 (2012), 193–99. ↩

32. See http://api.summon.serialssolutions.com/help/api/search/fields. ↩

33. See "What Is WorldCat?" at http://www.worldcat.org/whatis/. ↩

34. Interoperability issues caused by changes to the standards still remain; see number 6 in the list of common misalignments above. ↩

35. Thomas R. Bruce and Diane I. Hillmann, "The Continuum of Metadata Quality: Defining, Expressing, Exploiting," in *Metadata in Practice*, ed. Diane Hillmann and Elaine L. Westbrooks (Chicago: American Library Association, 2004), 238–56. ↩

36. Dushay and Hillmann, "Analyzing Metadata for Effective Use and Reuse." ↩

37. Sarah L. Shreeves et al., "Is 'Quality' Metadata 'Shareable' Metadata? The Implications of Local Metadata Practices for Federated Collections" (paper presented at the Twelfth National Conference of the Association of College and Research Libraries, Minneapolis, April 9, 2005), http://www.ala.org/acrl/sites/ala.org.acrl/files/content/conferences/pdf/shreeves05.pdf. ↩

38. Roy Tennant, "Bitter Harvest: Problems and Suggested Solutions for OAI-PMH Data and Service Providers," last modified May 14, 2004, http://roytennant.com/bitter_harvest.html. ↩

39. Carol Jean Godby, Jeffrey A. Young, and Eric Childress, "A Repository of Metadata Crosswalks," D-Lib 10, no. 12 (December 2004), http://www.dlib.org/dlib/december04/godby/12godby.html. ↩

40. The full report can be found at http://www.libraries.psu.edu/tas/jca/ccda/tf-onix1.html. The ONIX standard is available at http://www.editeur.org/. The crosswalk between ONIX and MARC21 can be found at http://www.loc.gov/marc/onix2marc.html. ↩

41. See http://www.loc.gov/catdir/beat/. The announcement of the ONIX project is available at http://www.loc.gov/catdir/beat/beat_report.1.2001.html. ↩

42. See http://www.rda-jsc.org/rdaonixann.html. ↩

43. See http://www.ifla.org/files/assets/cataloguing/isbd/OtherDocumentation/ISBD2ROFMapping v1_1.pdf. ↩

44. Carol Jean Godby, *A Crosswalk from ONIX Version 3.0 for Books to MARC 21* (Dublin, OH: Online Computer Library Center; OCLC Research, 2012), http://www.oclc.org/content/dam/research/publications/library/2012/2012-04.pdf. The crosswalk is available at http://www.oclc.org/content/dam/research/publications/library/2012/2012-04a.xls. ↩

45. Godby, *A Crosswalk*, 11. ↩

46. Ibid., 37. ↩

47. The CIMI project ceased as of December 15, 2003. Some of the original documentation can still be found at http://old.cni.org/pub/CIMI/framework.html, and older documents are archived at http://web.archive.org/web/*/http://www.cni.org. ↩

48. The Dublin Core element set is NISO standard Z39.85: http://www.niso.org/apps/group_public/download.php/10256/Z39-85-2012_dublin_core.pdf. ↩

49. Murtha Baca and Patricia Harpring, *Categories for the Description of Works of Art* (Los Angeles: J. Paul Getty Trust; College Art Association, 2009), last modified March 2014, http://www.getty.edu/research/publications/electronic_publications/cdwa/. The CDWA is an exhaustive set of metadata elements that may be used to describe art objects. It is not the goal of CDWA to be a standard used in its entirety; rather, it is intended to serve as a framework or guidelines for creating and/or mapping descriptive metadata records for cultural objects. The CCO data content standard provides guidelines for creating the content to fill the elements set forth in CDWA. ↩

50. See http://vraweb.org/resources/cataloging-metadata-and-data-management/data-standards-faqs/. ↩

51. See http://dublincore.org/documents/usageguide/glossary.shtml; a new version is in the process of being revised at http://wiki.dublincore.org/index.php/Glossary. ↩

52. Documentation is available at http://cco.vrafoundation.org/index.php/aboutindex/. The standard is hosted by the Library of Congress: http://www.loc.gov/standards/vracore/. ↩

53. *CDWA Lite: XML Schema Content for Contributing Records via the OAI Harvesting Protocol*, version 1.1 (Los Angeles: J. Paul Getty Trust; Artstor, 2006), http://www.getty.edu/research/publications/electronic_publications/cdwa/cdwalite.pdf. ↩

54. See Karim B. Boughida, "CDWA Lite for Cataloguing Cultural Objects (CCO): A New XML Schema for the Cultural Heritage Community" in *Humanities, Computers, and Cultural Heritage: Proceedings of the XVI International Conference of the Association for History and Computing, 14–17 September 2005* (Amsterdam: Royal Netherlands Academy of Arts and Sciences, 2005). ↩

55. Erin Coburn et al., *LIDO—Lightweight Information Describing Objects, Version 1.0*. (ICOM-CIDOC, November 2010), http://www.lido-schema.org/schema/v1.0 /lido-v1.0-specification.pdf. ↩

56. See http://network.icom.museum/cidoc/working-groups/data-harvesting-and-interchange/what-is-lido/. ↩

57. Coburn et al., *LIDO*, 27. ↩

58. A select list of implementations of LIDO is available at http://network.icom.museum/cidoc/working-groups/data-harvesting-and-interchange/lido-community/use-of-lido/. ↩

59. See http://www.europeana.eu. ↩

60. See http://britishart.yale.edu/sites/default/files/files/2011_GS_CI_Implementing_Lightweight_Information_Describing_Objects_LIDO_at_the_Yale_Center_for_British_Art.pdf. ↩

Rights Metadata Made Simple

Maureen Whalen

Since writing this chapter for the 2008 edition of *Introduction to Metadata*, I have found that people are now more aware of the importance of rights metadata and the need to collect and share it. More institutions are implementing digital asset management systems and seeking ways to expand the distribution of their collections through websites and social media. Underlying all such efforts are intellectual property laws governing copyright, privacy, publicity, and trademarks. Keeping track of what rights an institution has, who the rights holders are, and what their contact information is, is essential for institutions that want to participate actively and quickly in online environments. This chapter includes some tips and insights learned through experience over the last several years, so that mistakes need not be repeated and improvements can be considered for incorporation into ongoing rights metadata efforts. In addition, more and more institutions are now including rights metadata along with other information about works in their collections and efforts to improve standardization of terms and definitions continues. The Digital Public Library of America[1] is one of several organizations seeking to help institutions find simple and flexible solutions to rights metadata challenges. It is our hope that in the next few years, rights metadata will become an expected, routine component of any metadata record about a work and that its existence will improve public online access to digital surrogates of a wide variety of cultural materials.

Introduction

There are three common reactions when the issue of rights metadata arises:

1. “It’s too complicated and overwhelming.”
2. “We don’t have the staff or the money.”
3. “It’s not the library’s [or archive’s, or museum’s] job; it’s up to users to figure out rights information if they want to publish something from our collections.”

Here are some reasoned responses:

1. Yes, rights metadata can be complicated and overwhelming, but so is knitting a cardigan sweater until one simplifies the project by mastering a few basic techniques and following the instructions step by step.
2. Your institution is probably already spending staff time and money on rights research. Capturing rights metadata in a shared information system as a routine, programmatic activity with structured data rules and values and an established workflow should not cost any more than ad hoc rights research—and it will provide longer-lasting benefits.
3. In a world where "if it's not digital, it doesn't exist," libraries, archives, and museums have new roles with respect to their users as well as the creators and authors of the works in their collections. Moreover, cultural heritage institutions need rights information for their own uses of the works in their collections. Rights metadata is not just about compliance with intellectual property laws. Rights metadata is about being responsible stewards of the works in our collections and their digital surrogates—and in a digital world, it is crucial to the institution's broader mission of collection, preservation, and access.

Rights Metadata Dictionary

A major breakthrough in rights metadata efforts for the Getty Research Institute was the creation and implementation of a rights metadata dictionary for Special Collections.[2] Two important improvements came from this project: (1) clarification of which work was being described in the record, and (2) the addition of terms to the drop-down menus that allow users to better understand some of the ambiguities or unknowns about the rights information provided.

Core Work

For various reasons, staff in Special Collections tended to be confused about which work was being described in the metadata record. This confusion was resolved with a clear definition that the work being described is the work in the library's collections (neither the work depicted in a visual work nor the digital surrogate thereof), which we describe with the term *core work*. The core work may be a digital work.

People and Works Depicted

Frequently, the core work includes images of people or copyright-protected works. Metadata records in a digital asset management system or other information system may not provide fields for these kinds of "layers" of rights information. To ensure that rights information is collected about people or works depicted in a core work, the rights metadata dictionary instructs users to identify people and works depicted and to identify them as "potential claimants." Detailed information for these potential claimants may be included in the rights metadata records, if known.

Unknown, Additional Research Required, and Not Researched

Not all the rights information will be available when people are creating the metadata records. Some granularity may be desired to give those using the records over time a better sense of the status of rights research. In addition, due-diligence research about rights holders for designated orphan works will be necessary to document what was done and when. For that reason, precise definitions for "unknown," "additional research required," and "not researched" should be used; depending on the collection, more nuanced definitions may be added to the rights metadata dictionary.

Priority Information

Usable, shareable, repurposable rights metadata can be obtained by capturing the following core information:[3]

1. The **name of the creator** of the work or image, including the **nationality**, **date of birth**, and, if applicable, **death date**. Ideally, this information should be copied automatically from an authority file. *(Generally, the "work" is the original work in the institution's collection and NOT a digital surrogate. If the institution wants to create a rights metadata record for the digital surrogate, the rights metadata approach described herein would be valid, provided the digital "work" is described and differentiated from the original work.)*
2. **The year the work was created.** The year of creation may not be the same as the year of publication. Where two different dates exist, they should be identified separately. If the publication date is known, it should be recorded in the "publication status" field.
3. **Copyright status** *(one of these five options can be selected from a controlled pick list by staff tasked with recording rights metadata)*:

- **Copyright owned by the institution** means that the copyright is assumed valid and is owned by the institution that holds the work.
- **Copyright owned by a third party** assumes that the copyright is valid and is owned by someone or some entity other than the institution that holds the work; if known, the name of the third party should be captured in a database field or metadata element designated for that purpose.[4]
- **Public domain.** If the work is determined to be in the public domain, it is helpful to identify the year in which the work entered (or will enter) the public domain, if known. Some institutions, depending on the nature of their collections, may want more information about why the work is in the public domain—for example, if it is a work of the federal government, the copyright term has expired, or the work was published without a copyright notice before 1978 and did not qualify for copyright restoration.
- **Orphan work** is a work that may be protected by copyright law but for which the copyright owner or claimant cannot be identified or located.[5] Given the two-prong definition, it is recommended that the reason why a work is characterized as an orphan work should be included in the rights metadata. Therefore, two terms should be used: **Orphan work—rights holder cannot be identified** and **Orphan work—rights holder identified but cannot be located**.

4. **Publication status** (*one of these four options can be selected from a controlled pick list by staff tasked with recording rights metadata*):
    - **Published**. Include date, if known. Publication is defined in the Copyright Act as "the distribution of copies … of a work to the public by sale or other transfer of ownership, or by rental, lease, or lending." Note that the offer to distribute copies—including the original work, even if there is only one copy of it—constitutes publication.[6] Because of different treatment of foreign works under copyright law, some institutions may want to clarify where the work was published—in the United States only, in a foreign country only, or both.

- **Unpublished.** Some materials such as manuscripts and correspondence may be easily determined to be unpublished. Other works, however, such as a speech or painting that is known to the public, can still be considered "unpublished" under the Copyright Act definition.
- **Unknown.** It is sometimes difficult to determine whether or not a work has been published, particularly for photographs of which there may be multiple prints or for manuscripts from which a work was later published.

5. **Date that rights research was conducted** (*if there are multiple dates on which rights research was conducted, best practice would be to include all of those dates, along with the initials of the researcher[s]).*

Gathering rights metadata and including it in an institutional information system[7] will allow users with some basic understanding of copyright to make thoughtful judgments about how the law may affect use of the work in accordance with a legal exception.[8] It may also help guide determinations about how easy or difficult it might be to obtain permission, if needed.

Here are some examples of how the above rights metadata elements can be applied in day-to-day decision making:[9]

- Knowing the birth and death dates of the creator, or the year(s) in which the work was created and published, will allow for some quick calculations about the copyright term for the work. To do the analysis and arithmetic, follow Peter Hirtle's excellent chart "Copyright Term and the Public Domain in the United States."[10] Note: There are slightly different rules for works of foreign (non-US) origin, including restoration of copyrights in works of foreign origin that may have been in the public domain for a period of time before restoration; that is why it is good practice to identify the nationality of the creator, if known.
- Unpublished works tend to have longer copyright terms than published works; therefore, if the work is assumed to be unpublished, the term of copyright protection should be calculated in accordance with the formula for unpublished works.
- While the Copyright Act specifically states that unpublished works qualify for fair use, courts tend to protect the creator's right to decide about first publication, so the standard for fair use of unpublished works is usually higher than for published materials.[11] If the rights

metadata states that a work is unpublished, the user can assess how that status affects the fair use.

- For works published in the United States between 1923 and 1963, renewal of the original copyright registration was required.[12] Therefore, a work published in 1945 with the correct copyright notice and registration would require a renewal of the original copyright in 1973 (1945 + 28 = 1973) in order for that copyright to be valid today. One study indicates that 15 percent or fewer of the works in their original copyright terms between 1923 and 1963 were renewed.[13] This means the majority of works initially protected by copyright during this period are now in the public domain. Of course, the more famous the work, the greater likelihood the original copyright registration was renewed. By contrast, renewals of registrations for more obscure works may be less likely.
- Creation date may determine when the copyright term begins and ends; it is especially important when the author is unknown, the work is a work made for hire, or the work is one of corporate authorship (i.e., a work created by a "corporate body" such as a movie studio or record company).
- In 2006 the US Copyright Office issued its report on orphan works.[14] Hearings on orphan works were held in both the House of Representatives and the Senate, and legislation amending the Copyright Act to reduce the legal liabilities relating to use of orphan works was introduced in the House. While this legislation did not pass, many experts think that orphan-works legislation could be enacted in the next few years. Indeed, the US Copyright Office issued a Notice of Inquiry seeking updated comments about current issues relating to orphan works. If new legislation is introduced and passed, many hope that penalties and remedies for use of orphan works will be reduced or eliminated altogether. For that reason, it makes sense to identify works in institutional collections as orphan works now. Moreover, regardless of whether or not orphan-work legislation passes, it seems reasonable that if an institution attempts to identify and/or locate the copyright claimant and cannot do so despite diligent efforts, and this is explained to the court, there may be some recognition of this good-faith activity by the judge if an infringement claim is brought by the emergent copyright claimant.
- Prior to 1978 the law required that a copyright notice be affixed to published works. Failure to include a legally sufficient notice put into

the public domain American works that were published in the United States (without the notice). Therefore, an institution may decide to classify works as in the public domain if they were purchased before January 1, 1978, or were believed to have been offered for sale to the public before that date, and there is no copyright notice affixed to the work.

- Obviously, if one knows a work is in the public domain or if the institution owns the copyright, permission to use of the work is not required by law, although local policy may require internal authorization.

In order for catalogers and rights metadata analysts to be able to populate the recommended metadata elements, the institution will need some basic rules or assumptions to apply when copyright and publication status may not be clear and some suggestions for resources to help locate the sought-after information. There are numerous recommendations for where to look for the information requested. Currently, there is no resource that sets forth commonly accepted practices about what is legally reasonable to assume about copyright or publication when only limited information is available, so institutions will need to draft their own guidelines.[15] Of course, local policy regarding use of material presumed to be protected by copyright and the institution's risk tolerance for infringement claims that arise in case the assumption is wrong will govern use decisions.[16] With a little bit of effort, however, the basic information needed to make informed decisions about rights for many works in an institution's collections could be easily available and accessible if the suggested rights information is captured.

Any rights metadata effort should be viewed as dynamic and ongoing. New information may come from various sources: a user, a curator, a librarian, or even the creator of the work. Rights information needs to be updated and augmented, and additional information will need to be captured for works with more complicated rights, such as audiovisual materials. Therefore it is important that staff tasked with inputting rights metadata be identified to all those involved in cataloging and digitization efforts so that when new rights information is discovered, it can be input into the appropriate information system.

Now is the time to get started and not to be overwhelmed. Rights metadata can be made simple if everyone in the institution is aware of its long-term importance and there is a concerted, coordinated effort to research it, record it according to standards and best practices, and share it in fulfillment of the institutional mission in the digital age.

Table 1. Example of Core Elements for Rights Metadata

Metadata element	Valid data values for this element	Example: public domain work	Examp work r the pu domai
Title	The data values for this element should be copied (preferably in an automated manner) from the title element from the descriptive metadata record for the work or item. Per Cataloging Cultural Objects (CCO), this element, which is repeatable, can contain translated titles, brief titles, display titles, etc., in addition to the title that is inscribed on the item or object, if one exists. Include a subelement for the parent object/work ("title larger entity") when applicable.	*Puzza in the Likeness of Isis, Seated on a Lotus Flower/Puzza sous une forme parallele à Isis, assise sur la fleur de lotos* from *Cérémonies et coutumes religieuses de tous les peuples du monde*	*San Di Stadiu Diego, Califo* from *J Shulm photog archiv*
Creator	The name of the creator of the original object or work, taken from a published controlled vocabulary (e.g., Library of Congress Name Authority File, Library of Congress Subject Headings, the Getty's Union List of Artist Names) or local authority file whenever possible.	*Picart, Bernard*	*Shulm*
	The life dates in the case of individual creators, including the death date if applicable. Dates should be expressed according to a standard format, (e.g., ISO 8601).	*b. 1673–11–06* *d. 1733–08–05*	*b. 191*
Creation dates	The date(s) of the creation of the work.* Dates should be expressed according to a standard format (e.g., ISO 8601).	*1723–1743*	*1967*

Metadata element	Valid data values for this element	Example: public domain work	Examp work r the pu domai
Creator nationality	The nationality or culture of the creator of the work, if known	*French*	*Americ*
Copyright status	Valid values for this element should be selected from a controlled list. For example: • Copyright owned by the institution that holds the original object/work or item • Copyright owned by a third party—include a subelement for the name of the third party, taken from a published controlled vocabulary whenever possible • Public domain • Orphan work • Not yet researched	*public domain*	*copyrig owned institut* *© J. Pa Trust*
Publication status	Valid values for this element should be selected from a controlled list. For example: • Published—include a subelement with the date of publication, if known, in a standard format (e.g., ISO 8601). Note that date of creation and date of publication are not necessarily identical. • Unpublished (in which case, the creator dates and/or date of creation are extremely important) • Unknown, after research and due diligence • Not yet researched	*published 1723–1743*	*not res*

Metadata element	Valid data values for this element	Example: public domain work	Examp work r the pu domai
Date of rights-metadata research	This should be a repeating element, since metadata research is often necessarily an incremental process to which more than one individual contributes. The individual's name or initials should be provided by the information system and associated with the relevant dates of research. Dates should be expressed according to a standard format (e.g., ISO 8601).	*2008–10–07 MTW*	*2007–0 MTW*

* Note that under current US copyright law, a work is protected for the life of an individual author/creator plus 7 regardless of the date of creation. The copyright term for corporate works and works made for hire is 125 years f date of creation, or 95 years from the date of publication.

Author's Note

The rights metadata proposal and examples provided herein are not legal advice. To answer specific questions of law or address policy matters with legal implications, professional advice from an attorney is always recommended.

1. See http://dp.la/. ↩

2. For a detailed explanation of the process and results of the effort to create the rights metadata dictionary, see my article "Developing a Rights Metadata Dictionary for Digital Surrogates," *Journal of Library Metadata* 9, nos. 1–3 (2009): 15–35. ↩

3. These suggestions for a simplified rights metadata approach are based on required rights metadata recommendations for copyrightMD, an XML schema for rights metadata developed by the California Digital Library. The copyrightMD schema is designed for incorporation with other XML schemas for descriptive and structural metadata (e.g., CDWA Lite, MARCXML, METS, MODS). See http://www.cdlib.org/groups/rmg/.

   Note that the title of the work is not identified herein as a rights metadata element per se; it is assumed that the title would be included in any metadata schema used to describe the work and, thus, that element could be copied into the rights metadata schema from the descriptive metadata record in an automated manner. ↩

4. There may be certain conditions under which a license for certain specified uses of the work may have been granted to the institution. A license is not the same as ownership. If desired, when the copyright is known to be owned by a third party, the pick list could include an option for "license granted to the institution"; such a notation by itself, however, would not be adequate to describe the various rights granted, or denied, or the specific term during which the license is valid, so a review of the specific licensed rights would be necessary. ↩

5. "An 'orphan work' is an original work of authorship for which a good faith, prospective user cannot readily identify and/or locate the copyright owner(s) in a situation where permission from the copyright owner(s) is necessary as a matter of law." "Notices: Library of Congress, Copyright Office [Docket No. 12-2012], Orphan Works and Mass - Digitization," *Federal Register* 77, no. 204 (Monday, October 22, 2012): 64555; http://www.copyright.gov/fedreg/2012/77fr64555.pdf. ↩

6. See the Copyright Act of 1976, 17 U.S.C. §101. ↩

7. There is increasing discussion about embedding rights metadata into the same file as the digital surrogate of the image, thus avoiding the problem of two digital files that can and do get separated during transmission. To date, embedding rights data has been done only under limited circumstances, and the software necessary to embed the data and provide users with access to it using a free, downloadable reader is not yet widely available. ↩

8. The Copyright Act includes a number of limitations on (rights holders') exclusive rights. The most well-known of these limitations is fair use (section 107), whereby use of copyrighted works without permission of the rights holder is permitted if the use meets the statutory four-factor test. Another important exception applies to libraries and archives (section 108). Under this exception, libraries and archives are permitted to make copies of works in their collections under certain circumstances without permission of the rights holder, including, replacement copies of published works, preservation and security copies for unpublished works, and copies for users provided the copy becomes the property of the user and it is for private study, scholarship, or research. ↩

9. Examples include assumptions based on US copyright law; examples and assumptions for non-US jurisdictions are not provided herein. ↩

10. Available at http://copyright.cornell.edu/resources/publicdomain.cfm and http://copyright.cornell.edu/resources/docs/copyrightterm.pdf. ↩

11. From section 107 of the Copyright Act of 1976:

    Limitations on exclusive rights: Fair use. Notwithstanding the provisions of sections 106 and 106A, the fair use of a copyrighted work, including such use by reproduction in copies or phonorecords or by any other means specified by that section, for purposes such as criticism, comment, news reporting, teaching (including multiple copies for classroom use), scholarship, or research, is not an infringement of copyright. In determining whether the use made of a work in any particular case is a fair use the factors to be considered shall include—

    - (1) the purpose and character of the use, including whether such use is of a commercial nature or is for nonprofit educational purposes;
    - (2) the nature of the copyrighted work;
    - (3) the amount and substantiality of the portion used in relation to the copyrighted work as a whole; and
    - (4) the effect of the use upon the potential market for or value of the copyrighted work.

    The fact that a work is unpublished shall not itself bar a finding of fair use if such finding is made upon consideration of all the above factors. (emphasis added)

    Prior to passage of the Copyright Act of 1976, fair use was based on court decisions. In 1985 the US Supreme Court, in Harper & Row Publishers Inc. v. Nation Enterprises (471 U.S. 539), ruled on the applicability of the fair-use defense to unpublished works, noting

that the "author's right to control the first public appearance of his undisseminated expression will outweigh a claim of fair use" (p. 555). In order to clarify how the unpublished nature of a work was to be evaluated under the four-factor fair-use test set forth above and to reverse a growing presumption that fair use was not available as a defense against an infringement claim for all unpublished works, Congress passed an amendment to the law in 1992, and the last sentence of this section was added—the one bolded above. Notwithstanding this amendment, there is general legal consensus that courts will give greater weight to the unpublished nature of the work in fair-use cases than would be given if the work had already been published. ↩

12. All terms of original copyright run through the end of the twenty-eighth calendar year following publication, making the period for renewal registration in the above example December 31, 1973, to December 31, 1974. When checking the US Copyright Office renewal records, it is advisable to look at the years immediately preceding and following the calculated year for copyright-term expiration. This will ensure the work was not renewed properly in a different year. ↩

13. William M. Landes and Richard A. Posner, "Indefinitely Renewable Copyright" (John M. Olin Program in Law and Economics Working Paper No. 154, University of Chicago Law School, 2002), http://ssrn.com/abstract=319321 or DOI: 10.2139/ssrn.319321. ↩

14. *Report on Orphan Works: A Report of the Register of Copyrights* (Washington, DC: US Copyright Office, 2006), http://www.copyright.gov/orphan/orphan-report.pdf. ↩

15. Drafting the assumptions to be applied locally should not be used as an excuse to delay capturing rights metadata. If necessary, start with the rights information that is known and agree on the assumptions over time. ↩

16. Institutions may have zero risk tolerance or may have collections comprised primarily of works of living artists. In either case, local policy may be to seek permission. Others may feel that a good-faith judgment based on reasonable assumptions applied to the law and the facts is sufficient to allow use and defend in cases of infringement claims. ↩

Practical Principles for Metadata Creation and Maintenance

1. **Metadata creation is one of the core activities of collecting institutions and memory institutions.** Quality metadata creation is just as important as the care, preservation, display, and dissemination of collections; adequate planning and resources must be devoted to this ongoing, mission-critical activity.
2. **Metadata creation is an incremental process and should be a shared responsibility.** A metadata record may begin its life cycle as a "place holder" consisting of core data and then be enriched as it moves through the various stages of its use within an institution. By the same token, metadata creation and management distributed in a practical, reasonable way throughout the appropriate units of an institution, including but not limited to staff in acquisitions, cataloging and processing units, the registrar's office, digital asset management units, digitizing units, and conservation and curatorial departments. Ad hoc user-created metadata may be generated from work done by visiting researchers and scholars as well as other users, including non-expert users.
3. **Metadata rules and processes must be enforced in all appropriate units of an institution.** Inefficiencies, gaps in mission-critical metadata, poor-quality metadata, and negative "downstream" effects on metadata creation and workflow can be avoided by establishing and enforcing processes and procedures in all the participating units throughout an institution.
4. **Adequate, carefully thought-out staffing levels and appropriate skill sets are essential for the successful implementation of a cohesive, comprehensive metadata strategy.** An adequate number of appropriately trained staff with a variety of expertises and skills (e.g., subject expertise, cataloging experience, knowledge of controlled vocabularies, technical knowledge, research skills, knowledge of rights issues) is necessary for implementation of a successful, institution-wide metadata strategy.
5. **Institutions must build heritability of metadata into core information systems.** To avoid redundant data entry and lack of synchronization of metadata in core enterprise systems and to ensure

sharing of reliable, mission-critical information among the relevant units throughout the institution, interoperability for the automated transfer and validation of metadata from one core system to another must be achieved.

6. **There is no "one-size-fits-all" metadata schema or controlled vocabulary or data content (cataloging) standard.** Institutions must carefully choose the appropriate suite of metadata schemas and controlled vocabularies (including collection-specific thesauri and local pick lists), along with the most appropriate cataloging standards (including local cataloging guidelines based on published standards) to best describe and provide access to their collections and other resources.

7. **Institutions must streamline metadata production and replace manual methods of metadata creation with "industrial" production methods wherever possible and appropriate.** Time- and labor-intensive procedures for metadata creation should be evaluated and streamlined wherever possible (e.g., creation of core records rather than exhaustive records; metadata work and vocabulary control focused on a very few core elements or access points; elimination of redundant and outdated work flows). Automated tools (e.g., use of templates, pick lists, built-in thesauri, automated metadata generation or metadata mining) should be carefully researched and implemented as appropriate.

8. **Institutions should make the creation of shareable, repurposable metadata a routine part of their work flow.** Creation of consistent, standards-based, continuously refreshed and updated metadata enables institutions to publish information about their collections and other resources and activities in a timely, efficient manner and to more broadly disseminate that information through union catalogs and other "federated" resources via protocols such as the Open Archives Initiative's Protocol for Metadata Harvesting (OAI-PMH) and formats such as linked open data.

9. **Research and documentation of rights metadata must be an integral part of an institution's metadata work flow.** This metadata should be captured and managed in an appropriate information system that is available to the all of the individuals in the organization who need to contribute to it as well as those who need to use it.
(See "Rights Metadata Made Simple.")

10. **A high-level understanding of the importance of metadata and buy in from upper management are essential for the successful implementation of a metadata strategy.** Without a general understanding of principles 1–9 above on the part of the decision makers of an institution, it will be difficult if not impossible to consistently create adequate, appropriate metadata to enable access and use by core constituents (including internal users, the general public, and expert researchers).

Glossary

- **algorithm**A formula or procedure for solving a problem or carrying out a task. An algorithm is a set of steps in a very specific order, such as a mathematical formula or the instructions in a computer program. *See also* computer program.
- **Anglo-American Cataloguing Rules (AACR)**A data content standard for describing bibliographic materials. http://www.aacr2.org/
- **application**A software program designed to accomplish a task for an end user (e.g., word processing or project management), as distinguished from the operating system program that runs the computer itself.
- **application profile**A set of metadata elements, policies, and guidelines defined for a particular application or community. The elements may be from one or more element sets, thus allowing a given application to meet its functional requirements by using metadata from several element sets, including locally defined elements.
- **application programming interface (API)**A set of standardized requests defined by one computer program that allows another program to make requests and receive responses.
- **ASCII (American Standard Code for Information Interchange)**A seven-bit character code defining 128 characters used for information interchange, data processing, and communications systems.
- **asymmetric relationship**In the context of a thesaurus, a reciprocal relationship that is different in one direction than it is in the reverse—for example, BT/NT (for broader term/narrower term).
- **authentication**A human or machine process that verifies that an individual, computer, or information object is who or what it purports to be.
- **authority file**A file, typically electronic, that serves as a source of standardized forms of names, terms, titles, etc. Authority files should include references or links from variant forms to preferred forms. For example, in the Library of Congress Name Authority File, “Schiavone, Andrea” is the preferred name form for a Dalmatian artist active in Italy during the sixteenth century, while “Medulic´, Andrija,” “Lo Schiavone,” and several other forms are listed as variant names.

Authority files regulate usage but also provide additional access points, thus increasing both the precision and recall of many searches.

- **authority heading**A preferred, authorized heading used in a vocabulary, particularly in a bibliographic authority file that typically includes a string of names or terms, with additional information as necessary to allow disambiguation between identical headings (e.g., *United States—History—Civil War, 1861–1865—Battlefields and United States—History—Civil War, 1861–1865—Campaigns*). The types of authority headings used by the Library of Congress are the following: subject, name, title, name/title, and keyword.
- **automatic indexing**In the context of online retrieval, indexing by the analysis of text or other content using computer algorithms. The focus is on automatic, behind-the-scenes methods involving little or no input from individual searchers, with the exception of relevance feedback.
- **back-end database**A database that contains and manages data for an information system, distinct from the presentation or interface components of that system.
- **batch load**In the context of populating or contributing to databases, moving or manipulating a group of records as a single unit for the purpose of data processing, typically accomplished by the computer without user interaction, in contrast to entering records manually, one at a time. *See also* load and processing.
- **BIBFRAME (Bibliographic Framework)**A data model for bibliographic description designed to replace the MARC standards and to use the principles of linked data to make bibliographic data more useful within the library community as well as in the broader universe of information. http://www.loc.gov/bibframe/
- **Boolean operators**Logical operators used as modifiers to refine the relationship between terms in a search. The four most commonly used Boolean operators are AND, OR, NOT, and ADJ (adjacent). They may be used with parentheses and other punctuation to form logical groupings of criteria in queries—e.g., *(Castillo OR Rancho) AND Diego*.
- **browsing**The process whereby a user of a system or web site visually scans and maneuvers through navigation lists, results lists, hierarchical displays, or other content in order to make a selection, as contrasted to the user entering a search term in a search box. *See also* searching.

- **cataloger**In the context of this book, the person who enters information in records for works. *See also* end user.
- **cataloging**In the context of this book, the process of describing and indexing a work or image, particularly in a collections management system or other automated system. Cataloging involves the use of prescribed categories of information and rules—e.g., the rules described in AACR2, RDA, CCO, and CDWA.
- **Cataloging Cultural Objects (CCO)**A data content standard for describing works of art, architecture, and material culture. http://cco.vrafoundation.org/
- **CDWA (Categories for the Description of Works of Art) Lite**An XML schema for core records for art, architecture, and material culture designed to work with the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH); the elements are based on a subset of the full element set of Categories for the Description of Works of Art. http://www.getty.edu/research/publications/electronic_publications/cdwa/cdwalite.html
- **CGI script**A computer program, most frequently written in C, Perl, or a shell script, that uses the Common Gateway Interface (CGI) standard and provides an interface between a user or an external computer application and a web server. CGI scripts are most commonly used to develop forms that allow users to submit information to a web server.
- **CIDOC CRM (CIDOC Conceptual Reference Model)**An object-oriented model for the publication and interchange of cultural heritage information. http://www.cidoc-crm.org
- **classification**In the context of this book, the process of arranging works or other content objects systematically in groups or categories of shared similarity according to established criteria and using terms to identify the classes.
- **client**An application or piece of hardware that retrieves and/or renders resources or resource manifestations. Often used to denote a computer or other kinds of devices connected to a network equipped with software that enables users to access resources available on another computer connected to the same network, called a server.

- **clustering**In the context of automated data, clustering usually refers to the process of grouping or classifying items or data through automatic or algorithmic means rather than by incorporating human judgment.
- **collection management system**A type of database system that allows an institution to manage various aspects of its collections, including description (artist, title, measurements, media, style, subject, etc.) as well as administrative information regarding acquisitions, loans, and conservation information.
- **computer program***Also* called a program. A specific set of instructions for ordered operations that result in the completion of a task by the computer; a computer program consists of computer code. While the program is technically a type of data, computer programs are generally considered as separate from the data to which they refer (e.g., data would be the terms, scope notes, etc., in a vocabulary record). An *interactive* program acts when prompted by an action or information supplied by a user; a *batch* program automatically runs at a certain time or under certain conditions and then stops after the task is completed. A program is written in a programming language. *See also* processing.
- **computer system***See* system
- **conceptual data model**An abstract model or representation of data for a particular domain, business enterprise, or field of study, independent of any specific software or information system. Usually expressed in terms of entities and relationships. *See also* logical data model.
- **content object**In the context of a database, any entity that contains data. A content object can itself be made up of content objects. For example, a journal is a content object made up of individual journal articles, which are themselves content objects. *See also* information object.
- **controlled vocabulary**An organized arrangement of words and phrases used to index content and/or to retrieve content through browsing or searching. A controlled vocabulary typically includes preferred and variant terms and has a limited scope or describes a specific domain.
- **core elements**In the context of this book, the set of metadata elements representing the fundamental or most important information required for a minimal record. *See also* required fields.

- **cross-database searching***See* federated searching
- **crosswalk***Also* called field mapping. A chart or table (visual or virtual) that represents the semantic or technical mapping of fields or data elements from one data standard to fields or data elements in another standard that has a similar function or meaning. Crosswalks make it possible to convert data between databases that use different metadata schemes and enable heterogeneous databases to be searched simultaneously with a single query as if they were a single database (semantic interoperability). *See also* metadata mapping.
- **data**In common usage in computer science, this term is used as a singular noun to refer to information that exists in a form that may be used by a computer, excluding the program code. In other uses, *datum* is the singular and *data* is the plural, referring to facts or numbers in a general sense.
- **database**A structured set of data held in computer storage, especially one that incorporates software to make it accessible in a variety of ways. A database is used to store, query, and retrieve information. It typically comprises a logical collection of interrelated information that is managed as a unit, stored in machine-readable form, and organized and structured as records that are presented in a standardized format in order to allow rapid search and retrieval by a computer. *See also* system.
- **database field***Also* called a data field. A placeholder for a unit of information in a database that forms one of the searchable items in that database. A database field is a portion of a structured machine-readable record containing a particular category of information (e.g., *term* and *scope note* would be fields included in a vocabulary record).
- **database index***Also* called a data index. A particular type of data structure that improves the speed of operations in a table by allowing the quick location of particular records based on key column values. Indexes are essential for good database performance. The concept is distinguished from human indexing (application of keywords and other data values to a descriptive record) and automatic indexing.
- **database record***See* record

- **data content standard**Rules that determine the vocabulary, syntax, or format of content entered into data fields or metadata elements—e.g., RDA, ISO 8601 (rules for recording date and time), DACS, CCO.
- **data preprocessing***See* preprocessing
- **data processing***See* processing
- **data provider**In Open Archives Initiative nomenclature, an organization that exposes metadata records in one or more repositories (specially configured servers) for harvesting by service providers.
- **data structure**A given organization of data, particularly data elements, logical relationships between metadata elements, and storage allocations for the data.
- **data table / database table**Sets of related data elements that are organized in a grid or matrix comprising rows and columns in a database.
- **data values**The terms, words, or numbers used to populate fields in a record.
- **deep web***See* hidden web
- **default values**Values that are assumed or supplied automatically (for example, by a computer system) if a value is not specified.
- **Describing Archives: a Content Standard (DACS)**A data content standard for describing archival collections. http://files.archivists.org/pubs/DACS2E-2013_v0315.pdf
- **diacritics***Also* called diacritical marks. Signs or accent marks found over, under, or through alphabetic letters in many languages (e.g., the umlaut in German, *München*), used to indicate emphasis or pronunciation, often to distinguish different sounds or values of the same letter or character without the diacritical mark.
- **digital asset management system**A type of system for organizing digital media assets, such as digital images or video clips, for storage and retrieval. Digital asset management systems sometimes incorporate a descriptive data cataloging component, but they tend to focus on managing workflow for creating digital assets and for managing asset rights, requests, and permissions.

- **digital signatures**A form of electronic authentication of a digital document. Digital signatures are created and verified using public key cryptography and serve to tie the document being signed to the signer.
- **digital surrogate**A digital "copy" of an original work or item (e.g., a JPEG or TIFF image of a painting or sculpture, or a PDF file of an article or book). In Open Archives Initiative nomenclature, digital surrogates are often referred to as "resources."
- **document**In the context of search and retrieval, the combination of a defined, primarily self-contained, machine-readable text and the format in which it is expressed.
- **domain name**The address that identifies an Internet or other network site. On the Internet, domain names act as mnemonic aliases for IP addresses, a hierarchical numeric addressing system that enables Internet hosts to be uniquely identified. The hierarchical nature of the Domain Name System means that the authority for issuing subdomain names is delegated down the hierarchy; for example, once the Getty Trust has registered the domain name "getty.edu," it is responsible for any subdomain names such as "www.getty.edu," "shiva.getty.edu," etc.
- **Dublin Core Metadata Element Set**A set of fifteen metadata elements optimized for resource discovery on the web that can be assigned to information resources. Also often used as a "lowest common denominator" in metadata mapping. http://dublincore.org/documents/dces/
- **dynamically generated**Refers to a web page, metadata record, or other information object that is generated on demand, typically from content stored in a database and usually either in response to a user's input or from dynamic data sources that are refreshed periodically. The expression "on the fly" is often used in relation to dynamically generated content.
- **Encoded Archival Description (EAD)**A data structure standard for encoding archival finding aids in SGML or XML according to the EAD document type definition (DTD) or XML schema that makes it possible for the semantic contents of a finding aid to be machine processed. http://www.loc.gov/ead/
- **encryption**An encoding mechanism used to prevent unauthorized users from reading digital information and also for user and document

authentication. Only designated users or recipients have the capability to decode encrypted materials.

- **end user**In the context of systems design, the term refers to any client for whom a database system is designed and operated; from that perspective, it could include the editors or catalogers for whom an editorial or cataloging system has been designed.
- **entity relationship model**A type of conceptual data model that represents structured data in terms of entities and relationships. An entity relationship diagram can be used to visually represent information objects and their relationships. Because the constructs used in the entity relationship model can easily be transformed into relational tables, this type of model is often used in database design.
- **Exif (Exchangeable Image File Format)**A specification for an image file format for digital cameras that provides the ability to attach image metadata to JPEG, TIFF, and RIFF images. As of this writing, Exif is not maintained by any industry or standards organization but is widely used by camera manufacturers. http://www.cipa.jp/std/documents/e/DC-008-2012_E.pdf
- **false hit**In search and retrieval, an entry in a list of results that does not comply with the user's intended results. *Also* called a false drop.
- **federated searching***Also* called, cross-database searching, metasearching, and parallel searching. Performing queries simultaneously across resources residing in different domains and created by different communities. Federated searching may involve searching across multiple databases, different platforms, and varying protocols, thus requiring the application of interoperability between resources and vocabularies.
- **field mapping***See* crosswalk
- **finding aid**A descriptive tool widely used in archives. Finding aids typically take the form of hierarchical narrative descriptions of cohesive groups of archival records or collections of manuscript materials. Finding aids traditionally were paper documents; Encoded Archival Description (EAD) is a structured way of expressing finding aids as machine-readable data.

- **FOAF (Friend of a Friend)**A machine-readable ontology that models data for persons, their activities, and their relationships to other people and objects. http://www.foaf-project.org/
- **folksonomy**An assemblage of concepts, represented by terms and names (called “tags”), that result from social tagging. A folksonomy differs from a taxonomy in that it is not structured hierarchically. The authors of the folksonomy are typically the casual users of the content rather than professional indexers following standard protocols and using standardized controlled vocabularies.
- **FRBRoo**A joint initiative of the International Federation of Library Associations and Institutions (IFLA) and the International Council of Museums–International Documentation Committee (ICOM-CIDOC) to create an object-oriented ontology that both captures the semantics of bibliographic information and harmonizes those concepts in common with the CIDOC CRM, thus facilitating information interchange between the museum and library communities. http://cidoc.ics.forth.gr/frbr_inro.html
- **free-text field**A field that may contain data entered without any system-defined structure. It may be used to express ambiguity, uncertainty, and nuance in a note.
- **FTP (File Transfer Protocol)**A TCP/IP protocol that allows data files to be copied directly from one computer to another over the Internet.
- **Functional Requirements for Bibliographic Records (FRBR)**A set of requirements and a conceptual entity relationship model developed by the International Federation of Library Associations and Institutions to support bibliographic access and control. http://www.ifla.org/publications/functional-requirements-for-bibliographic-records
- **Google Sitemap**Metadata about the content of a web site that assists the Googlebot web crawler to index a site more efficiently and comprehensively.
- **granular, granularity**The level of detail at which an information object or resource is viewed or described.
- **harvester**In Open Archives Initiative nomenclature, a computer system that sends Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) requests to data providers’ repositories and harvests metadata records from them.

- **heading***Also* called a label. A string of words comprising a term combined with other information that serves to modify, disambiguate, amplify, or create a context for the main term in displays. *See also* authority heading.
- **hidden web***Also* called deep web and invisible web. The sum of the web pages that are not accessible to web crawlers, usually because they are either dynamically generated by a user querying a database or are password protected or subscription based.
- **hostname**An identifier for a specific machine on the Internet. The hostname identifies not only the machine, but also its subnet and domain—for example, "www.getty.edu." *See also* domain name.
- **HTML (HyperText Markup Language)**An SGML-derived markup language used to create documents for World Wide Web applications. HTML has evolved to emphasize design and appearance rather than the representation of document structure and metadata elements.
- **HTTP (HyperText Transfer Protocol)**The standard protocol that enables users with web browsers to access HTML documents and related media.
- **hyperlink**An abbreviated reference to a "hypertext link," a method of creating nonlinear pathways between related digital documents or to link to related objects such as image or audio files.
- **information object**A digital item or group of items referred to as a unit, regardless of type or format, that a computer can address or manipulate as a single discrete object. *See also* content object.
- **International Organization for Standardization (ISO)**A worldwide voluntary network of national standards institutes from approximately 160 countries. The standards bodies work in partnership with international organizations, governments, industries, businesses, and consumer representatives to reach consensus, set standards, and promote their use with the goal of facilitating trade and meeting the broader needs of society.
- **Internet**A global collection of computer networks that exchange information by the TCP/IP suite of networking protocols.
- **Internet directory**A thematically organized list of descriptive links to Internet sites, often created by humans who have classified sites by

their content. Best of the Web (http://botw.org) provides such directories.

- **interoperability**The ability of different information systems to work together, particularly in the correct interpretation of data semantics and functionality. *See also* semantic interoperability.
- **invisible web***See* hidden web
- **item**In the context of cataloging art, an individual object or work.
- **jargon**A characteristic terminology of a particular group or discipline that is typically not understood by a more general audience.
- **keyword**Any significant word or phrase in the title, subject headings, or text associated with an information object.
- **keyword in context (KWIC)**A type of automatic indexing in which each word in a text, title, subject heading, string of words, or term becomes an entry word in the index, with the exception of words in stop lists. Variations include KWOCs (keyword out of context) and KWACs (keyword alongside context).
- **keyword index**An index based on individual keywords found in a controlled vocabulary, text, or other content object.
- **language model**A type of automatic indexing based on term weighting and relevance prediction that attempts to predict probable query search terms based on term frequencies within documents and the inverse document frequency of terms across the target data. It is similar to the probabilistic model.
- **legacy system**An information system that has been developed and modified over a period of time and has become outdated and difficult and costly to maintain but that holds information that is important and involves processes that are deeply ingrained in an organization. Legacy systems usually are eventually replaced by new hardware and software configurations.
- **LIDO (Lightweight Information Describing Objects)**A simple XML schema for describing and interchanging core information about museum objects. http://network.icom.museum/cidoc/working-groups/lido/

- **linked data**Data that is semantically linked by following a set of best practices for publishing and interlinking structured data that uses RDF syntaxes and HTTP URIs.
- **linked open data (LOD)**Linked data that is made available for use, reuse, and redistribution on the visible web.
- **link resolver**Software that uses the OpenURL standard to automatically redirect a user's request to the most appropriate copy of a networked digital object. Typically, link resolvers are used by libraries to direct their patrons from bibliographic records or abstracts to licensed subscription-based resources such as full-text electronic versions of articles, books, etc. http://www.niso.org/apps/group_public/project/details.php?project_id=115
- **load**The process of moving or transferring files or software from one disk, computer, or server to another. To *upload* means to transfer from a local computer to a remote computer; to *download* means to transfer from a remote computer to a local one.
- **logical data model**A data model that includes all entities and the relationships among them based on the structures identified in a conceptual data model and that specifies all attributes for each entity. The data is described in as much detail as possible, without regard to how it will be physically implemented in a specific database.
- **mapping**A set of correspondences between terms, fields, or element names used for translating data from one standard or vocabulary into another, or as a means of combining terms or data for search and retrieval. *See also* crosswalk.
- **MARC (Machine-Readable Cataloging) format**A set of standardized data structures for describing bibliographic materials that facilitates cooperative cataloging and data exchange in bibliographic information systems. http://www.loc.gov/marc/
- **markup language**A formal way of annotating a document or collection of digital data using embedded encoding tags to indicate the structure of the document or data file and the contents of its data elements. It also provides a computer with information about how to process and display marked-up documents. HTML, XML, and SGML are examples of standardized markup languages.

- **memory institution**A generic term used to describe an institution that has a responsibility to collect, care for, and provide access to the human record—for example, museums, libraries, and archives.
- **Metadata Encoding Transmission Schema (METS)**A standard for encoding descriptive, administrative, and structural metadata relating to objects in a digital library, expressed in XML. METS enables the "packaging" of complex digital objects that include a range of metadata as well as related digital surrogates. http://www.loc.gov/standards/mets/
- **metadata mapping**A formal identification of equivalent or nearly equivalent metadata elements or groups of metadata elements within different metadata schemas, carried out in order to facilitate semantic interoperability. *See also* mapping *and* crosswalk.
- **metadata mining**The automated extraction of metadata from electronic documents.
- **Metadata Object Description Schema (MODS)**An XML schema for bibliographic records, developed and maintained by the Library of Congress. http://www.loc.gov/standards/mods/
- **metasearch**Searching of diverse databases on diverse platforms with diverse metadata in real time via one or more protocols. The *National Information Standards Organization* MetaSearch Initiative defines metasearch as "search and retrieval to span multiple databases, sources, platforms, protocols, and vendors at once." Metasearch enables users to enter search criteria once and access several search engines simultaneously. With metasearch, fresh records are always available because searching is in real time, in a distributed environment.
- **meta tag**An HTML tag that enables metadata to be embedded invisibly on web pages (e.g., description, keywords).
- **meta tag spamming**The deliberate misuse of meta tags in order to attract traffic to a site (i.e., by boosting its ranking in search results).
- **namespace**The set of unique names used to identify objects within a well-defined domain, particularly relevant for XML, LOD, and DNS applications.
- **National Information Standards Organization (NISO)**A nonprofit association that is accredited by the American National Standards

Institute and identifies, develops, maintains, and publishes technical standards to manage information.

- **nesting**The way in which subelements may be contained within larger elements, resulting in multiple levels of metadata.
- **object-oriented programming**A programming model organized around objects rather than actions and data rather than logic, where an object is a location that has a value and is referenced by an identifier.
- **Online Public Access Catalog (OPAC)**A computerized inventory of a library's holdings.
- **ontology**In the context of this book, an ontology is a formal, machine-readable specification of a conceptual model in which concepts, properties, relationships, functions, constraints, and axioms are all explicitly defined. While an ontology is not technically a controlled vocabulary, it uses one or more controlled vocabularies for a defined domain. Identifying an existing ontology, or developing an appropriate ontology, is the first step in expressing data as linked open data (LOD).
- **Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)**A protocol used to harvest or collect metadata records from data providers. *See also* data provider, data harvester, and service provider. http://www.openarchives.org/pmh/
- **operating system**A software program that runs a computer, as distinguished from an application, which is installed into an operating system in order to enable users to perform specific tasks.
- **PageRank™**A proprietary link-analysis algorithm developed by Google founders Larry Page and Sergey Brin to assign a numerical score to each document in a set of hypertext documents based on the number of referring links. The algorithm also takes into account the rank of the referring page; thus a link from a high-ranking page counts more than a link from a low-ranking page.
- **paradigmatic relationship***Also* called a semantic relationship. A relationship between terms or concepts that is permanent and based on a known definition.
- **parsing**In processing data, a process by which data is broken or filtered into smaller, more distinct units.

- **precision**In the context of this book, a measure of search effectiveness expressed as the ratio of relevant records or documents retrieved from a database to the total number retrieved in response to the query. A high-precision search means that most of the results retrieved will be relevant; however, a high-precision search will not necessarily retrieve all relevant results. Recall and precision are inversely related (when one goes up, the other goes down).
- **preprocessing***Also* called data preprocessing. Preliminary processing or transformation of data in order to facilitate further processing, parsing, etc.
- **probabilistic model**An automatic relevance and weighting method in which terms in a text or other content object are modeled as random variables so that term frequency and distribution are used to predict the probability of relevance. *See also* language model.
- **procedure**A relatively independent portion of computer code within a larger computer program that performs a specific task in a series of steps. *Also* called a subprogram or subroutine.
- **processing***Also* called data processing. The manipulation or transformation of data through a series of operations. In batch processing, the operations are grouped together in batches and performed automatically; in >interactive processing, the operations are prompted by input from a human programmer or user. *See also* computer program.
- **program***See* computer program
- **programming language**A formal language defined by syntactic and semantic rules and used to write instructions that can be translated into machine language and then executed by a computer (e.g., PL/SQL, C++, C#, Java, Perl, Ruby, Python, BASIC).
- **protocol**A specification—often a standard—that describes how computers communicate with each other (e.g., the TCP/IP suite of communication protocols or Open Archives Initiative Protocol for Metadata Harvesting [OAI-PMH]).
- **query**In the context of retrieval, a command to look in a database and find records or other information that meet a specified set of criteria. The most precise queries are those that return the fewest false hits.

- **query expansion**Reformulating a query in order to return a broader or more comprehensive set of results (e.g., adding synonyms to a search term).
- **recall**A measure of a search system's effectiveness in terms of retrieving all results that are possibly relevant, expressed as the ratio of the number of relevant records or documents retrieved over all the relevant records or documents. A high recall search retrieves a comprehensive set of relevant results; however, it also increases the likelihood that marginally relevant content objects will also be retrieved. Recall and precision are inversely related (when one goes up, the other goes down).
- **record**In the context of this book, a coherent, discrete group of populated fields or metadata elements. *Also* called a logical record.
- **relational database**A database organized on a relational model that organizes data into one or more tables of rows and columns with a unique key for each row. The rows in a table can be linked to rows in other tables by storing the unique key of the row to which it should be linked.
- **relationship**In the context of this book, a link between two types of data, records, files, or any two entities of the same or different types in a system or network.
- **relevance ranking**Ranking and sorting of query results, typically estimated by an algorithm that calculates the number and weight of occurrences of the search term in the targeted data. Relevance ranking frequently does not correspond to the actual relevance of the information retrieved in a search for the user's information-seeking needs.
- **required fields**Data fields or metadata elements that are required to meet a standard or the requirements of a system's operations.
- **Resource Description and Access (RDA)**The cataloging standard for libraries that, as of this writing, has begun to replace AACR2. http://www.rdatoolkit.org/
- **Resource Description Framework (RDF)**A standard model for data interchange on the web that extends the linking structure of the web to use URIs to name relationships between things. RDF enables structured and semistructured data to be exposed and shared across different applications. http://www.w3.org/RDF/

- **resource discovery**The process of searching for specific information objects on the web.
- **retrieval**In the context of this book, the activity of using a search or other method to find records or other data in a database. *See also* query.
- **robot***See* web crawler
- **schema***Also* called scheme. The organization, structure, and rules for encoding information that supports specific communities of users. The plural forms of the word schema are "schemas" and "schemata." *See also* XML schema.
- **schema registry**An authoritative source of names, semantics, and syntaxes for one or more schemas.
- **screen scraping**A technique in which display data (usually unstructured) is automatically retrieved and extracted, for example from a web page.
- **search engine**A computer program that allows users to search electronic resources. In the context of the World Wide Web, the term usually refers to a program that searches a large index of web pages generated by an automated web crawler. *See also* web search engine.
- **searching**Operations or algorithms intended to determine if one or more data items meet defined criteria or possess a specified property.
- **semantic interoperability**The ability of different agents, services, and applications to communicate data while ensuring accuracy and preserving the meaning of the data.
- **semantic linking**A method of linking terms in a database according to the meaning of and relationships between terms.
- **Semantic Web**An evolving, collaborative effort led by the World Wide Web Consortium (W3C) whose goal is to provide a common framework that will allow *data* to be shared and reused across various applications and enterprise and community boundaries. It derives from W3C director and inventor of the World Wide Web Tim Berners-Lee's vision of the web as a universal medium for data, information, and knowledge exchange.
- **server**An application that supplies resources or resource manifestations. Often used to refer to a networked computer that acts

as a source of data and/or applications used by multiple client computers or devices. *See also* client.

- **service provider**In Open Archives Initiative nomenclature, an institution or organization that harvests metadata from data providers and uses the aggregated metadata as a basis for building value-added services.
- **Simple Knowledge Organization System (SKOS)**An endeavor of the World Wide Web Consortium that develops specifications and standards to support the use of knowledge organization systems (KOS) such as thesauri, classification schemes, subject heading lists, and taxonomies within the framework of the Semantic Web. http://www.w3.org/2004/02/skos/
- **social bookmarking**The decentralized practice and method by which individuals and groups create, classify, store, discover, and share web bookmarks or "favorites" in an online "social" environment.
- **social tagging**The decentralized practice and method by which individuals and groups create, manage, and share terms, names, etc.—called "tags"—to annotate and categorize digital resources in an online "social" environment. A folksonomy is the result of social tagging. Also referred to as collaborative tagging, social classification, social indexing, mob indexing, folk categorization. *See also* tagging.
- **sorting**The automated process of organizing a results list, data elements in a record, or other data in a particular sequence based on established criteria or attributes of the data—for example, alphabetically, by parent string, or by an associated date. There may be primary sort criteria and secondary sort criteria (e.g., an algorithm can be formulated to first sort place names in a results list alphabetically and then to sort by the parent string).
- **spamming**Used in reference to meta tags, the abuse of metadata that web page creators include in the HTML header area of their pages in order to increase the number of visitors to a web site. Keyword spamming entails repeating keywords multiple times in order to appear at the top of search engine result listings or listing keywords that are irrelevant to the site in order to attract visitors under false pretenses.
- **specifications**In the context of designing an information system, the formal, detailed description of user and technical requirements, including specific descriptions of procedures, functions, screens,

reports, materials, other features, and hardware. *See also user requirements*.

- **spider***See* web crawler
- **SQL (Structured Query Language)**A special-purpose command language used with relational databases to perform queries and other tasks.
- **SRU/SRW (Search and Retrieve via URL/Search and Retrieve Web Service)**Companion protocols for web search queries utilizing the CQL Common Query Language. http://www.loc.gov/standards/sru/
- **stop list**In the context of search and retrieval, words in a vocabulary or target data that are ignored in searching or matching because they occur too frequently or are otherwise of little value in retrieval for a given domain. Common stop lists for a text contain articles, conjunctions, and prepositions, although these words are typically not included in a stop list for a vocabulary.
- **surrogate***See* digital surrogate
- **system***Also* called a computer system. A number of interrelated hardware and software components that work together to store and convert data into information by using electronic processing. *See also* database.
- **tagging**In the context of the web, the act of associating terms (called "tags") with an information object (e.g., a web page, an image, a streaming video clip), thus describing the item and enabling keyword-based classification and retrieval. Tags—a form of user-generated metadata—from communities of users can be aggregated and analyzed, providing useful information about the collection of objects with which the tags have been associated. *See also* social tagging.
- **taxonomy**An orderly classification that explicitly expresses the relationships, usually hierarchical (e.g., genus/species, whole/part, class/instance)—between and among the things being classified. A taxonomy can be used as a controlled vocabulary. *See also* folksonomy.
- **TCP/IP (Transmission Control Protocol/ Internet Protocol)**The International Organization for Standardization (ISO) standardized suite of network protocols that enables information

systems to communicate with other information systems on the Internet regardless of their computer platforms.

- **Text Encoding Initiative (TEI)**An international cooperative effort to develop guidelines for standard encoding schemes—i.e., the TEI and TEI Lite document type definitions (DTDs)—for literary and linguistic texts. http://www.tei-c.org/
- **transliteration**The process of rendering the letters or characters of one alphabet or writing system into the corresponding letters or characters of another alphabet or writing system, generally based on phonetic equivalencies. While a common noun will often be translated, a proper name in a non-Roman alphabet is more often transliterated. There are often multiple standards for transliterating from one writing system to another, thus producing multiple variant names.
- **truncation**In searching and matching, the action of cutting off characters in a search term in order to find all terms with a certain common string of characters; this typically involves the user employing a wildcard symbol to search for a string of characters no matter what other characters follow (or precede) that string (e.g., searching for *arch** will retrieve *arch, arches, architrave, architecture, architectural history,* etc.).
- **Unicode**A sixteen-bit character-encoding scheme and standard for representing letters, characters, and diacritical marks in most of the world's modern scripts. http://unicode.org/
- **unique identifier**A number or other string that is associated with a record or piece of data, exists only once in a database, and is used to uniquely identify and disambiguate that record or piece of data from all others in the database.
- **URI (Uniform Resource Identifier)**A short string that uniquely identifies a resource such as an HTML document, an image, a downloadable file, or a service. URLs and URNs are types of URIs.
- **URL (Uniform Resource Locator)**A type of URI consisting of an Internet address that tells users how and where to locate a specific file on the World Wide Web. A URL includes not only the name of a file, but also the name of the host computer, the directory path to get to that file, and the protocol needed in order to use it (e.g., http://www.getty.edu/research/publications/electronic_-publications/index.html specifies that the hypertext transfer protocol

"http" should be used to retrieve the document "index.html" from the host "www.getty.edu" in the directory "research/publications/-electronic_publications/index.html.")

- **URN (Uniform Resource Name)**A type of URI consisting of a unique, location-independent identifier of a file available on the Internet. The file remains accessible by its URN regardless of changes that might occur in its host and directory path. For example, urn:issn:0167-6423 is the URN for the journal *Science of Computer Programming*.
- **user interface**The portion of the design and functionality of a cataloging, editorial, search and retrieval, or other system or web site with which end users interact, including the arrangement of displays, menus, clickable text or images, pagination, etc. A user interface that is easy for users to utilize is called user friendly.
- **user requirements**In system design, the initial formal explanation of functionalities, displays, and reports expressed from the point of view of the user's needs and expectations. *See also* specifications.
- **Virtual International Authority File (VIAF)**A federated resource that provides integrated access to millions of records from authority files compiled by libraries and other memory institutions from around the world. http://www.viaf.org
- **visible web**The subset of the World Wide Web that is visible to web browsers and indexable by search engines' web crawlers. In order to be accessible to web crawlers, the pages must be accessible simply by following links (i.e., not generated dynamically in response to user input) and not protected by a password.
- **VRA Core 4.0**An XML schema developed by the Visual Resources Association (VRA) and supported by the Library of Congress, VRA Core is used for describing works of art and architecture and their visual surrogates. http://www.loc.gov/standards/vracore/schemas.html
- **web browser**A software application that enables users to view and interact with information and media files on the web. Mozilla Firefox, Google Chrome, and Apple's Safari are examples of web browsers.
- **web crawler**A software program that systematically traverses the web, either for the purpose of generating a searchable index of web content or to gather statistics. *See also* robot and spider.

- **web search engine / Internet search engine**A software program that collects data taken from the content of files available on the web and puts them in an index or database that web users can search in a variety of ways. The search results provide links back to the pages matching the user's search in their original location.
- **web server**A computer that is able to respond to HTTP requests from clients known as web browsers and return the appropriate HTTP responses—most typically serving an HTML page.
- **website**A collection of related electronic pages (*web pages*), generally formatted in HTML and found at a single address where the server computer is identified by a given host name.
- **wiki**A collaborative website that contains pages that any authorized user can edit. Wikis typically retain all former versions of each page, allowing the revision history of a page to be tracked and for unwanted revisions to be reversed.
- ***Wikipedia***A free, collaborative, volunteer-driven, web-based encyclopedia that utilizes wiki software to allow anyone to edit articles. http://en.wikipedia.org/wiki/
- **World Wide Web**A vast, distributed wide-area client-server architecture for retrieving hypermedia documents over the Internet.
- **World Wide Web Consortium (W3C)**The main international standards organization for the World Wide Web.
- **XML (Extensible Markup Language)**A relatively simple, flexible markup language used for publication and exchange of a wide variety of data on the web.
- **XML schema**A machine-readable definition of the structure, elements, and attributes allowed in a valid instance of a conforming XML document. XML schemas are expressed using the XML Schema Definition language, a World Wide Web Consortium (W3C) standard. http://www.w3.org/TR/xmlschema-0
- **XMP (Extensible Metadata Platform)**A markup language, based on the Resource Description Framework (RDF), for recording and embedding metadata about digital assets. Developed by Adobe Systems and supported across the company's range of software products and file formats. http://www.adobe.com/products/xmp.html

- **Z39.50**A client/server-based protocol for searching and retrieving information from remote databases.

Contributors

- **Murtha Baca** is head of the Digital Art History program at the Getty Research Institute in Los Angeles. She holds a PhD in art history and Italian language and literature from the University of California, Los Angeles. Baca is the author of numerous articles in the field of art documentation and controlled vocabularies and the editor of Introduction to Art Image Access (Getty Research Institute, 2002) and a coeditor of Cataloging Cultural Objects: A Guide to Describing Cultural Works and Their Images (American Library Association, 2006). She has taught workshops and seminars on metadata, visual resources cataloging, and thesaurus construction at museums, universities, and other organizations in North and South America, Europe, and Asia. She teaches graduate seminars on metadata and thesaurus construction in the Department of Information Studies at UCLA.
- **Tony Gill** is the global director of library science and information management at an advertising agency in New York. He has been an adjunct professor at New York University's Graduate School of Arts and Science, where he taught Interactive Technology in Museums as part of the university's Museum Studies Program. Gill spent thirteen years developing collaborative, standards-based solutions for the creation and delivery of digital cultural heritage information before assuming his current position. He has degrees in communication in computing (Middlesex University) and physics and philosophy (King's College, London). He is the author of a number of publications on the use of information technology in the cultural heritage sector and is a coauthor of the CIDOC Conceptual Reference Model (ISO Standard 21127:2006).
- **Anne J. Gilliland** is professor of information studies and moving image archive studies in the Department of Information Studies and director of the Center for Information as Evidence at UCLA. She holds a PhD in information and library studies, with a cognate in business information systems, from the University of Michigan. An internationally recognized expert on archives and information as evidence, Gilliland's numerous publications include *Enduring Paradigm, New Opportunities: The Value of the Archival Perspective in the Digital Environment* (Council on Library and Information Resources, 2000). She has provided expert testimony to the US House

and Senate on appropriations for electronic records research and has acted as an adviser to the National Archives and Records Administration on methods for managing and preserving digital records.

- **Maureen Whalen** is the former associate general counsel at the J. Paul Getty Trust, where her work focused on cultural property and intellectual property matters. She received her undergraduate and law degrees from the State University of New York at Buffalo and her master's degree in library and information science from UCLA. Whalen was active in the Museum Attorneys Group and was a member of the Section 108 Study Group, formed to make recommendations to the Librarian of Congress for alterations to Section 108 of the US Copyright Act to take into account current technologies. She is an adjunct faculty member in the Department of Information Studies at UCLA, teaching graduate seminars on intellectual property and intellectual freedom.
- **Mary S. Woodley** (1951–2013) received a PhD in classical archaeology as well as a master's degree in library and information science from UCLA. From 1986 to 1999 she was a librarian in the technical services department of the Getty Research Institute. In 1999 she joined the professional staff of the Oviatt Library of California State University, Northridge (CSUN), where she was collection-development coordinator and subject liaison for the departments of art, archaeology, and anthropology. Woodley was an active member of the Dublin Core Metadata Initiative, serving on its advisory board, and was an elected member of the Cataloging and Classification Executive Committee of the American Library Association's Association for Library Collections and Technical Services (ALCTS). She created a workshop on managing digital projects for the ALCTS and the Library of Congress and taught an advanced seminar on knowledge management at CSUN.

About

Metadata provides a means of indexing, accessing, preserving, and discovering digital resources. The volume of digital information available over electronic networks has created a pressing need for standards that ensure correct and proper use and interpretation of the data by its owners and users. Well-crafted metadata is needed more now than ever before and helps users to locate, retrieve, and manage information in this vast and complex universe.

The third edition of *Introduction to Metadata*, first published in 1998, provides an overview of metadata, including its types, roles, and characteristics; a discussion of metadata as it relates to web resources; and a description of methods, tools, standards, and protocols for publishing and disseminating digital collections. This revised edition is an indispensable resource in the field, addressing advances in standards such as linked open data, changes in intellectual property law, and new computing technologies, and offering an expanded glossary of essential terms.

Citation Information

Chicago

Baca, Murtha, ed. *Introduction to Metadata*. 3rd ed. Los Angeles: Getty Publications, 2016. http://www.getty.edu/publications/intrometadata.

MLA

Baca, Murtha, ed. *Introduction to Metadata*. 3rd ed. Los Angeles: Getty P, 2016. 26 Mar. 2022 <http://www.getty.edu/publications/intrometadata>.

Permanent URL

http://www.getty.edu/publications/intrometadata

Revision History

Any revisions or corrections made to this publication after the first edition date will be listed here and in the project repository at https://www.github.com/gettypubs/intrometadata, where a more detailed version history is available. The revisions branch of the project repository, when present, will also show any changes currently under consideration but not yet published here.

July 20, 2016

- First release

November 9, 2016

- Add links to e-book editions
- Update Cataloguing in Publication record

Other Formats

- Paperback
- Amazon Kindle (MOBI)
- Apple iBookstore (EPUB)
- Google Play (EPUB)
- Barnes & Noble NOOK (EPUB)

Copyright

The Getty Research Institute Publications Program
Thomas W. Gaehtgens, *Director, Getty Research Institute*
Gail Feigenbaum, *Associate Director*

First edition 1998
Third edition 2016

Published by the Getty Research Institute, Los Angeles
Getty Publications
1200 Getty Center Drive, Suite 500
Los Angeles, California 90049-1682
www.getty.edu/publications

Murtha Baca, *Series Editor*
Tom Fredrickson, *Manuscript Editor*
Marissa Clifford, *Proofreader*
Gary Hespenheide, *Designer*
Amita Molloy, *Production*

Eric Gardner, *Digital Designer and Developer*
Greg Albers, *Digital Project Manager*

Library of Congress Cataloging-in-Publication Data

- Names: Baca, Murtha, editor. | Getty Research Institute, issuing body.
- Title: Introduction to metadata / edited by Murtha Baca.
- Other titles: Introduction to metadata (2016)
- Description: Third edition. | Los Angeles : Getty Research Institute, [2016] | Includes bibliographical references.
- Identifiers: LCCN 2015046323 (print) | LCCN 2015050428 (ebook) | ISBN 9781606064795 (pbk.) | ISBN 9781606064801 (epub) | ISBN 9781606065006 (online)
- Subjects: LCSH: Metadata. | Database management. | World Wide Web. | Information organization.
- Classification: LCC QA76.9.D3 I599 2016 (print) | LCC QA76.9.D3 (ebook) | DDC 025.3--dc23
- LC record available at http://lccn.loc.gov/2015046323

Note: The editor and authors of this publication are aware that the noun "metadata" (like the noun "data") is plural and, therefore, should take a plural verb form. However, in order to avoid awkward locutions, it has been treated here throughout as singular.