Glossary

AACR (*Anglo-American Cataloguing Rules*)
A data content standard for describing bibliographic materials. http://www.aacr2.org/.

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.

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.

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, and so on. Authority files should include references or links from variant forms to preferred forms. For example, in the Library of Congress Name Authority File (LCNAF), "Schiavone, Andrea" is the preferred name form for a Dalmatian artist active in Italy during the sixteenth century, while "Medulić, 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 the recall of many searches.

back-end database
A database that contains and manages data for an information system, distinct from the presentation or interface components of that system.

CCO (Cataloging Cultural Objects)
A data content standard for describing works of art, architecture, and material culture. http://www.vraweb.org/ccoweb/cco/index.html.

CDWA (*Categories for the Description of Works of Art*)
A set of metadata categories and recommendations that may be used to design information systems and to do cataloging for art, architecture, objects of material culture, and archaeological and archival materials. http://www.getty.edu/research/conducting_research/standards/cdwa/.

CDWA Lite
An XML schema for core records for art, architecture, and material culture designed to work with the OAI-PMH; the elements are based on a subset of the full element set of *Categories for the Description of Works of Art* (CDWA). http://www.getty.edu/research/conducting_research/standards/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 interactive interface between a user or an external computer application and a World Wide 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 ontology for the mediation and interchange of heterogeneous cultural heritage information. http://cidoc.ics.forth.gr/.

client
An application 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. *See also* **server**.

conceptual data model
An abstract model or representation of data for a particular domain, business enterprise, field of study, etc., independent of any specific software or information system. Usually expressed in terms of entities and relationships. *See also* **logical data model**.

crosswalk
A chart or table (visual or virtual) that represents the semantic mapping of fields or data elements in 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). Also known as field mapping. *See also* **metadata mapping**.

DACS (Describing Archives: A Content Standard)
A data content standard for describing archival collections. http://www.archivists.org/catalog/pubDetail.asp?objectID=1279.

data content standard
Rules that determine the vocabulary, syntax, or format of content entered into data fields or metadata elements, for

Many thanks to Marcia Lei Zeng of the School of Library and Information Science at Kent State University, who reviewed the glossary and provided extremely valuable input.

example, *Anglo-American Cataloguing Rules* (AACR), ISO 8601 (rules for recording date and time), *Describing Archives: A Content Standard* (DACS), *Cataloging Cultural Objects* (CCO).

data provider (OAI nomenclature)
An organization that exposes metadata records in one or more repositories (specially configured servers) for harvesting by service providers.

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.

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, for example, a JPEG or TIFF image of a painting or sculpture or a PDF file of an article or book. In OAI nomenclature, digital surrogates are often referred to as "resources."

DTD (Document Type Definition)
A collection of markup declarations that define the structure, elements, and attributes that can be used in encoding certain type of documents in SGML or, more commonly, in XML. Examples of DTDs include the EAD DTD, the HTML DTD, and the TEI DTD. XML DTDs are gradually being replaced by the newer **XML schemas**.

Dublin Core Metadata Element Set (DCMES)
A set of 15 metadata elements that can be assigned to information resources, optimized for resource discovery on the World Wide Web. 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.

EAD (Encoded Archival Description)
A data structure standard for encoding archival finding aids in SGML or XML according to the EAD DTD or EAD XML schema, making it possible for the semantic contents of a hierarchically structured finding aid to be machine processed. http://www.loc.gov/ead/.

encryption
An encoding mechanism used to prevent nonauthorized users from reading digital information and also for user and document authentication. Only designated users or recipients have the capability to decode encrypted materials.

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 represent information objects and their relationships visually. 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.exif.org/.

field mapping
See **crosswalk**.

FTP (File Transfer Protocol)
A TCP/IP protocol that allows data files to be copied directly from one computer to another over the Internet.

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; EAD is a structured way of expressing finding aids as machine-readable data.

FRBR (Functional Requirements for Bibliographic Records)
A set of requirements and a conceptual entity-relationship model developed by the International Federation of Library Associations and Institutions (IFLA) to support bibliographic access and control. http://www.ifla.org/VII/s13/frbr/frbr.htm.

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.

folksonomy
An assemblage of concepts, represented by terms and names (called "tags"), the result of social tagging. Note that a folksonomy is not a true taxonomy. *See also* **social tagging**, **taxonomy**.

Google Sitemap
Metadata about the content of a Web site that assists the Googlebot Web crawler to index a site more efficiently and comprehensively. www.google.com/webmasters/sitemaps/.

granularity
The level of detail at which an information object or resource is viewed or described.

harvester (OAI nomenclature)
A computer system that sends OAI-PMH requests to OAI data providers' repositories and harvests metadata records from them.

header metadata
Metadata embedded in the header part of a digital file.

Hidden Web (also known as Deep Web, 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 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.

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. Yahoo! provides numerous 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**.

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 important information and involves processes that are deeply ingrained in an organization. Legacy systems usually are eventually replaced by a new hardware and software configuration.

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 and books. http://www.niso.org/standards/standard_detail.cfm?std_id=783.

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.

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 datafile and the contents of its data elements. This markup 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 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.

metadata mining
The automated extraction of metadata from electronic documents.

metasearch
Searching of diverse databases on diverse platforms with diverse metadata in real time by means of one or more protocols. The NISO 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. http://www.niso.org/committees/MS_initiative.html.

META tag
An HTML tag that enables metadata to be embedded invisibly on Web pages, for example, Description, Keywords.

META tag spamming
The deliberate misuse of meta tags in order to attract traffic to a site, for example, by boosting its ranking in search results.

METS (Metadata Encoding Transmission Schema)
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/

MODS (Metadata Object Description Schema)
An XML schema for bibliographic records, developed and maintained by the Library of Congress. http://www.loc.gov/standards/mods/.

namespace
The set of unique names used to identify objects within a well-defined domain, particularly relevant for XML applications. An XML Namespace is a W3C recommendation for providing uniquely named elements and attributes in an XML instance. A namespace is declared using the reserved XML attribute xmlns, the value of which must be a URI (Uniform Resource Identifier) reference. For example, the Dublin Core Metadata Element Set, Version 1.1 (original 15 elements) has the approved DCMI namespace URI as http://purl.org/dc/elements/1.1/.

nesting
The way in which subelements may be contained within larger elements, resulting in multiple levels of metadata.

network bandwidth
Derived from the term used to describe the size or "width" of the frequencies used to carry analog communications such as television and radio. For Internet purposes, bandwidth is generally (and incorrectly) used to refer to the rate of data transfer.

OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting)
A protocol used to harvest or collect metadata records from data providers. http://www.openarchives.org/pmh/.

object-oriented
A programming or data modeling methodology that utilizes the notion of classes and their properties. Members (or instances) of a class share the same properties—for example, color or weight (however, note that although members of a class all share the same properties, the values of those properties do not need to be the same). Classes can contain subclasses, members of which inherit the properties of the parent or "superclass."

ontology
A formal, machine-readable specification of a conceptual model, in which concepts, properties, relationships, functions, constraints, and axioms are all explicitly defined.

OPAC (Online Public Access Catalog)
A computerized inventory of a library's holdings.

Open WorldCat
A subset of the WorldCat union bibliographic database made available by OCLC to certain Web search engines and online book retailers. http://www.oclc.org/worldcat/open/.

PageRank™ (Google)
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, such that a link from a high-ranking page counts more than a link from a low-ranking page. http://www.google.com/technology/.

precision
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; for example, in a database containing 100 records relevant to the topic "book history," a search retrieving 50 records, 25 of which are relevant to the topic, would have 50 percent precision (25/50). (Definition from ODLIS, *Online Dictionary for Library and Information Science*, http://lu.com/odlis/.) *See also* **recall**.

protocol
A specification—often a standard—that describes how computers communicate with each other, for example, the TCP/IP suite of communication protocols or the OAI-PMH.

RDF (Resource Description Framework)
An application of XML that enables the creation of rich, structured, machine-readable resource descriptions. http://www.w3.org/RDF/.

RDF schema
A set of semantics within a defined namespace for use with specific applications of RDF.

recall
A measure of the effectiveness of a search expressed as the ratio of the number of relevant records or documents retrieved in response to the query to the total number of relevant records or documents in the database; for example, in a database containing 100 records relevant to the topic "book history," a search retrieving 50 records, 25 of which are relevant to the topic, would have 25 percent recall (25/100). (Definition from ODLIS, *Online Dictionary for Library and Information Science*, http://lu.com/odlis/.) *See also* **precision**.

relevance
The extent to which information retrieved in a search of a library collection or other resource, such as an online catalog or a bibliographic database, is judged by the user to be applicable to ("about") the

subject of the query. Relevance depends on the searcher's subjective perception of the degree to which the document fulfills the information need, which may or may not have been expressed fully or with precision in the search statement. Measures of the effectiveness of information retrieval, such as precision and recall, depend on the relevance of search results. (Definition from ODLIS, *Online Dictionary for Library and Information Science*, http://lu.com/odlis/.)

relevance ranking
The algorithmic process, a feature of many search software applications, by which results in a result set are sorted or ranked according to their relevance. In OPACs, for example, relevance is computed based upon the number of occurrences of the search term in the record that is retrieved, and the weight assigned to the field(s) in which the search term appears. (Definition from ODLIS, *Online Dictionary for Library and Information Science*, http://lu.com/odlis/.) Google's PageRank™ is an example of a relevance ranking algorithm.

resource discovery
The process of searching for specific information objects on the Web.

robot
See **Web crawler**.

schema
A set of rules for encoding information that supports specific communities of users. Also called "scheme." 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**.

semantic interoperability
The ability of different agents, services, and applications to communicate data while ensuring accuracy and preserving the meaning of the data (definition based on Marcia Bates and Mary Niles Maack, *Encyclopedia of Library and Information Sciences*, 3rd ed. [New York: Marcel Dekker, forthcoming]).

Semantic Web
An evolving, collaborative effort led by the W3C whose goal is to provide a common framework that will allow data to be shared and re-used across various applications as well as across enterprise and community boundaries. It derives from W3C director and inventor of the World Wide Web Sir 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 (OAI nomenclature)
An institution or organization that harvests metadata from data providers and uses the aggregated metadata as a basis for building value-added services.

SGML (Standard Generalized Markup Language)
International Standards Organization standard ISO/IEC 8879:1986; a markup language first used by the publishing industry, for defining, specifying, and creating digital documents that can be delivered, displayed, linked, and manipulated in a system-independent manner. XML and HTML are derived from SGML.

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, and so on (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* **folksonomy**, **tagging**.

spamming
Used in reference to meta tags. The abuse of metadata that creators include in the HTML header area of their Web 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.

spider
See **Web crawler**.

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/.

surrogate
See **digital surrogate**.

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.

TCP/IP (Transmission Control Protocol/ Internet Protocol)
The ISO standardized suite of network protocols that enables information systems to communicate with other information systems on the Internet, regardless of their computer platforms.

TEI (Text Encoding Initiative)
An international cooperative effort to develop guidelines for standard encoding schemes (i.e., the TEI and TEI Lite DTDs) for literary and linguistic texts. http://www.tei-c.org/.

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/conducting_research/standards/intrometadata/intro.html specifies that the hypertext transfer protocol "http" should be used to retrieve the document intro.html from the host www.getty.edu in the directory research/conducting_research/standards/intrometadata.

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*.

Visible Web
The subset of the World Wide Web that is visible to Web browsers and indexable by search engines' Web crawlers. 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 for describing works of art and architecture and their visual surrogates. http://www.vraweb.org/projects/vracore4/index.html

W3C (World Wide Web Consortium)
The main international standards organization for the World Wide Web.

Web 2.0
A phrase used loosely by the Web development community to refer to a perceived "second generation" of Web technologies and applications. Wikis, folksonomies, gaming, podcasting, blogging, and so on, are all considered Web 2.0 applications.

Web browser
A software application that enables users to view and interact with information and media files on the Web. Internet Explorer, Mozilla Firefox, and Netscape Navigator are examples of Web browsers.

Web crawler (robot, spider)
A software program that systematically traverses the Web, either for the purpose of generating a searchable index of Web content or to gather statistics.

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.

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.

wiki
A collaborative Web site 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.

XHTML (Extensible HyperText Markup Language)
A reformulation of HTML in XML.

XML (Extensible Markup Language)
A simple, flexible markup language derived from SGML. Originally designed for large-scale electronic publishing, XML is now playing an increasingly important role in the 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 W3C standard. http://www.w3.org/TR/xmlschema-0/.

XMP (Extensible Metadata Platform)
A markup language, based on 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/index.html.

Z39.50
An ISO 23950 and ANSI/NISO Z39.50 standard information retrieval protocol. Z39.50 is a client/server-based protocol for searching and retrieving information from remote databases.

Introduction to Metadata 3.0 ©2008 J. Paul Getty Trust