Metadata and the Web

Tony Gill

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 before being co-opted and partially redefined by the library, archive, and museum information communities at the end of the twentieth century. Today, nearly a decade later, a Google search on "metadata" yields about 58 million results (see Web Search Engines sidebar). Metadata has quietly hit the big time; it is now 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) contain embedded Extensible Metadata Platform (XMP)[2] metadata.

As the term *metadata* has been increasingly adopted and co-opted by more diverse audiences, the definition of what constitutes metadata has grown in scope to include almost anything that describes anything else. The standard concise definition of metadata is "data about data," a relationship that is frequently illustrated using the metaphor of a library card catalog. The first few lines of the following Wikipedia entry for *metadata* are typical:

> **Metadata** (Greek: meta- + Latin: data "information"), literally "data about data," are information about another set of data. A common example is a library catalog card, which contains data about the contents and location of a book: They are data about the data in the book referred to by the card.[3]

The library catalog card metaphor is pedagogically useful because it is nonthreatening. Most people are familiar with the concept of a card catalog as a simple tool to help readers find the books they are looking for and to help librarians manage a library's collection as a whole. However, the example is problematic from an ontological perspective, because

---

[1] See http://www.exif.org/.

[2] See http://www.adobe.com/products/xmp/.

neither catalog cards nor books are, in fact, data. They are *containers* or *carriers* of data. This distinction between information and its carrier is increasingly being recognized; for example, the CIDOC Conceptual Reference Model (CRM),[4] a domain ontology for the semantic interchange of museum, library, and archive information, models the relationship between information objects—identifiable conceptual entities such as a text, an image, an algorithm, or a musical composition—and their physical carrier as follows:

> E73 Information Object *P128 is carried by* E24 Physical Man-Made Stuff

The IFLA Functional Requirements for Bibliographic Records (FRBR)[5] model makes a similar four-tier distinction between Works, Representations, Manifestations, and Items: the first three entities are conceptual entities, and only Items are actual physical instances represented by bibliographic entities.

Of course, most library catalogs are now stored as *0*s and *1*s in computer databases, and the "items" representing the "works" that they

Web Search Engines

Web search engines such as Google are automated information retrieval systems that continuously traverse the Web, visiting Web sites and saving copies of the pages and their locations as they go in order to build up a huge catalog of fully indexed Web pages. They typically provide simple yet powerful keyword searching facilities and extremely large result sets that are relevance ranked using closely guarded proprietary algorithms in an effort to provide the most useful results. The most well known Web search engines are available at no cost to the end-user and are primarily supported by advertising revenue. Web 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 browsers), the actual words on the Web page (unstructured data), and referring links (indicating the popularity of the Web resource).

---

[3] http://en.wikipedia.org/wiki/Metadata.

[4] Nick Crofts, Martin Doerr, Tony Gill, Stephen Stead, and Matthew Stiff, eds., *Definition of the CIDOC Conceptual Reference Model*, version 4.2, June 2005. Available at http://cidoc.ics.forth.gr/docs/cidoc_crm_version_4.2.1.pdf. See also Tony Gill, "Building Semantic Bridges between Museums, Libraries and Archives: The CIDOC Conceptual Reference Model," *First Monday* 9, no. 5 (May 3, 2004). Available at http://www.firstmonday.org/issues/issue9_5/gill/index.html.

[5] *Functional Requirements for Bibliographic Records* (IFLA, 1998). http://www.ifla.org/VII/s13/frbr/frbr.htm.

describe (to use the nomenclature of the FRBR model) are increasingly likely to be digital objects on a Web server, as opposed to ink, paper, and cardboard objects on shelves (this is even more true now in light of large-scale bibliographic digitization initiatives such as the Google Book Search Library Project, the Million Books Project, and the Open Content Alliance, about which more later).

So if we use the term *metadata* in a strict sense, to refer only to *data about data*, we end up in the strange predicament whereby a record in a library catalog can be called metadata if it describes an electronic resource but cannot be called metadata if it describes a physical object such as a book. This is clearly preposterous and illustrates the shortcomings of the standard concise definition.

Another property of metadata that is not addressed adequately by the standard concise definition is that metadata is normally structured to model the most important attributes of the type of object that it describes. Returning to the library catalog example, each component of a standard MARC bibliographic record is clearly delineated by field labels that identify the meaning of each atomic piece of information, for example, author, title, subject.

The structured nature of metadata is important. By accurately modeling the most essential attributes of the class of information objects being described, metadata in aggregate can serve as a catalog—a distillation of the essential attributes of the collection of information objects—thereby becoming a useful tool for using and managing that collection. In the context of this chapter, then, *metadata* can be defined as *a structured description of the essential attributes of an information object*.

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, some informative metrics are available. The July 2007 Netcraft survey of Web hosts received responses to HTTP (HyperText Transfer Protocol, the data transmission language of the Web) requests for server names from 125,626,329 "sites."[6] A site in this case represents a unique hostname such as http://www.hostname.com. The same survey in January 1996 received responses from just 77,128 Web servers; the number of Web servers connected to the Internet has grown exponentially over the past decade or so. (Fig. 1.)

---

[6] *Netcraft Web Server Survey*, July 2007. http://news.netcraft.com/archives/2007/07/09/july_2007_web_server_survey.html.

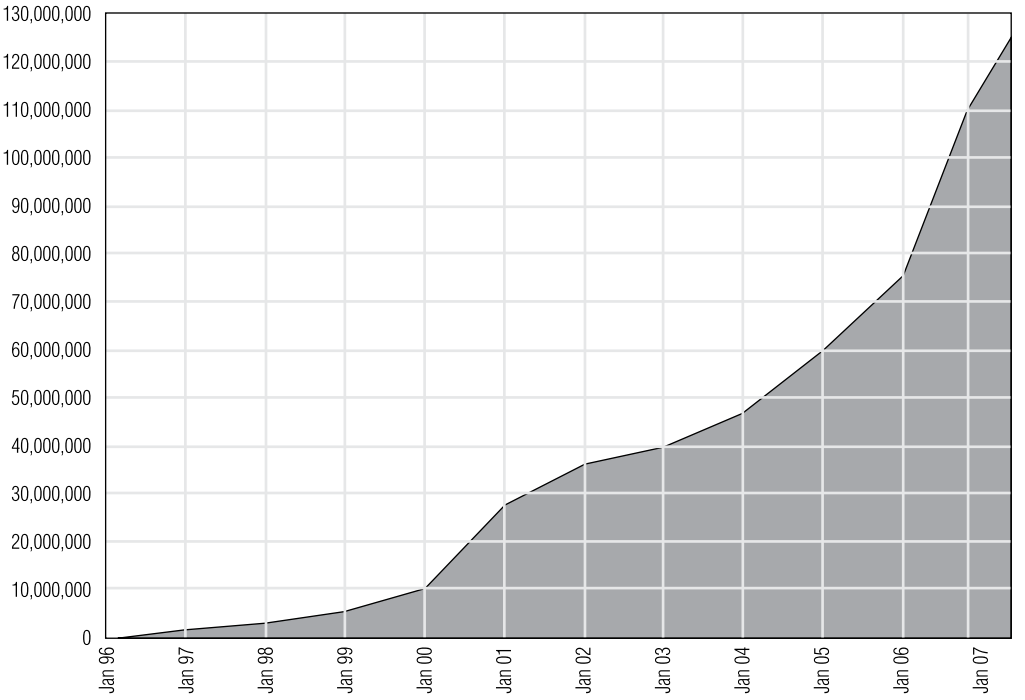

Figure 1. **Growth in the Number of Web Hosts, January 1996–July 2007.** (Source: Netcraft Survey. http://www.netcraft.com/survey/)

Although the Netcraft Web hosts 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 Web sites are hosted on each server or how many accessible pages are contained in each site.

The Visible Web versus the Hidden Web

Accurate figures for the number of pages available on the Web are much more difficult to find; two computer scientists estimated that the indexable Web comprised more than 11.5 billion pages at the end of January 2005,[7] although given the rapid increase in the amount of information on the Web, that figure is now hopelessly out of date.

The problem of determining how many pages are available on the Web is exacerbated by the fact that a large and increasing amount of the Web's content is served dynamically from databases in response to a user's input, or is in a non-Web format, or requires some kind of user authentica-

[7] Antonio Gulli and Alessio Signorini, "The Indexable Web Is More than 11.5 Billion Pages." http://www.cs.uiowa.edu/~asignori/web-size/.

tion 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 index 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, so all of this content is effectively invisible to the search engines and is not indexed.

Collectively, this content beyond the reach of search engine Web 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 than measuring the public or Visible Web. A survey published in 2001 claimed that the Deep Web was five hundred times larger than the Visible/Indexable Web,[8] although very little meaningful information can be inferred from this today; in terms of the evolution of the Internet, five years is the equivalent of a geologic era.

Although much of the content on the Deep Web is deliberately kept out of the public sphere, either because it is private or because some kind of fee or subscription must be paid to access it, there is a vast amount of information that is inadvertently inaccessible to Web search engines simply because it is contained in Web sites that were not designed to be accessible to the search engines' Web crawlers. This is an especially common problem for sites that generate pages dynamically in response to user input using content stored in databases. Because Web search engines often account for the vast majority of a Web site's traffic, building sites that are not accessible to Web crawlers can seriously limit the accessibility and use of the information they contain. Institutions seeking to make dynamically generated information as widely accessible as possible should design "crawler-friendly" Web 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 museum, library, and archive sectors is to contribute otherwise Deep Web collections information to union catalogs or other aggregated resources that are indexed by the commercial search engines.

Search engine providers are now also providing 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 correctly. (Both union catalogs and tools to expose Deep Web content to search engines are discussed in more detail later in this chapter.)

---

[8] Michael K. Bergman, "The Deep Web: Surfacing Hidden Value," *Journal of Electronic Publishing* 7, no. 1 (August 2001). http://www.press.umich.edu/jep/07-01/bergman.html.

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 an unimaginable amount of information, of widely variable quality, for the estimated 1 billion people who now have access to it[9]—although it is worth remembering that this is still less than one person in six globally (the myth of nearly universal access to the Web remains just that—a myth).

Unfortunately, however, finding relevant, high-quality information on the Web is not always a straightforward proposition. There is no overarching logical structure to the Web, and the core Web protocols do not offer any support for information search and retrieval beyond the basic mechanisms provided by the HTTP for requesting and retrieving pages from a specific Web address.

The disappointment of the hypertext community with the World Wide Web is clearly evident in a comment by Ted Nelson (who first coined the term *hypertext* in 1965) in a speech delivered at the HyperText 97 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."[10]

Not surprisingly, tools designed to address the resource location problem and help make sense of the Web's vast 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,[11] a distributed directory of Web sites maintained by human editors, shortly after inventing the Web itself, and search engines such as Yahoo![12] Lycos,[13] and Webcrawler[14] were launched in 1994.

The clear market leader in Web search today is Google. According to a Nielsen//NetRatings press release issued on March 30, 2006, "Google accounts for nearly half of all Web searches, while approximately one-third are conducted on Yahoo! and MSN combined."[15] According to its Web

---

[9] "Worldwide Internet Users Top 1 Billion in 2005," Computer Industry Almanac Inc., January 4, 2006. http://www.c-i-a.com/pr0106.htm.

[10] Ted Nelson, speaking at HyperText 97, Eighth ACM International Hypertext Conference, Southampton, April 6–11, 1997. Quoted in Nick Gibbins, "The Eighth ACM International Hypertext Conference," Ariadne, no. 9 (May 1997). http://www.ariadne.ac.uk/issue9/hypertext/.

[11] WWW Virtual Library: http://vlib.org/.

[12] http://www.yahoo.com/.

[13] http://www.lycos.com/.

[14] http://www.Webcrawler.com/.

[15] Press Release: "Google Accounts for Nearly Half of All Web Searches, While Approximately One-Third Are Conducted on Yahoo! and MSN Combined, According to Nielsen//Netratings, Nielson//NetRatings," March 30, 2006. http://www.nielsen-netratings.com/pr/pr_060330.pdf.

site, "Google's mission is to organize the world's information and make it universally accessible and useful."[16] In the relatively short time since the company's launch in 1998 in a garage in Menlo Park, California, it has grown to become one of the Internet's giants: it employs almost six thousand people, operates one of the five most popular Web sites on the Internet, and has a current market valuation of over $115 billion, making it the second-largest technology company in the world after Microsoft. 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 and the number of people using it continue to grow. First, it must maintain an index of the public Web that is both sufficiently current and sufficiently comprehensive to remain competitive. Currency is important because, as the Google Zeitgeist demonstrates,[17] many of the most popular searches are related to current affairs and popular culture. Any search engine that fails to maintain a sufficiently current index will not be able to deliver relevant results to queries about current events and 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 otherwise it may fail to deliver relevant results that a competitor with a more comprehensive index could provide. A study by Gulli and Signorini estimated that as of January 2005 Google had indexed about 76 percent of the 11.5 billion pages on the Visible Web.[18] Index size has traditionally been one of the key metrics on which search engines compete, so in August 2005 Yahoo! issued a press release claiming to have indexed 19 billion Web pages.[19] If the Gulli and Signorini estimate of the size of the Web is to be believed, the Yahoo! claim would imply that the Web had doubled in size in just seven months, and consequently some commentators have conducted further research, which casts doubt on the veracity of the Yahoo! figures.[20]

Third, in addition to maintaining a current and comprehensive index of the rapidly expanding Web, a search engine must be able to search the index that it has compiled by crawling the Web, ranking the search results according to relevance, and presenting the results to the user as quickly as possible—ideally in less than half a second. Much of Google's

---

16 Google Company Overview: http://www.google.com/corporate/index.html.

17 Google Zeitgeist: http://www.google.com/press/zeitgeist.html.

18 Gulli and Signorini, "The Indexable Web Is More than 11.5 Billion Pages."

19 Tim Mayer, "Our Blog Is Growing Up—And So Has Our Index," Yahoo! Search Blog, August 8, 2005. http://www.ysearchblog.com/archives/000172.html.

20 Matthew Cheney and Mike Perry, "A Comparison of the Size of the Yahoo! and Google Indices, 2005." http://vburton.ncsa.uiuc.edu/oldstudy.html.

rapid rise to dominance in the search engine market can be attributed to its sophisticated and patented PageRank™ relevance ranking algorithm, which ranks the importance of relevant pages according to the number of links from other pages that point to them.[21] The PageRank™ value of each Web page and the text contained in the Title HTML tag are really the only metadata that Google uses to any meaningful extent in providing its search service—the search itself is performed on an index of the actual data content of the HTML pages. Fourth, a market-leading search engine such as Google must be able to respond to hundreds of millions of such search requests from users all around the world every day.[22]

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 small clusters of very powerful servers, Google has developed a massive parallel architecture comprising large numbers of inexpensive networked PCs, which Google claims is both more powerful and more scalable than the use of a smaller number of more powerful servers.[23]

Google's server cluster was reported to comprise more than fifteen thousand PCs in 2003; the company has provided little official information about its hardware recently, but given the explosive growth in both the amount of information on the Web and the number of Web users, coupled with a wide range of new services offered by Google (e.g., Google Print, Google Scholar, Google Images, GMail, Froogle, Blogger, Google Earth), the number of server nodes is undoubtedly much greater today. There is widespread speculation on the Web that the Google server cluster today comprises anywhere between 100,000 and 1,000,000 nodes[24] and that it could in fact be the most powerful "virtual supercomputer" in the world.

Can the Search Engines Keep Up?

Can the 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? This is a difficult question to answer; analysts have been predicting since

---

[21] "Our Search: Google Technology." http://www.google.com/technology/.

[22] Danny Sullivan, "Searches per Day," from SearchEngineWatch.com. http://searchenginewatch.com/reports/article.php/2156461.

[23] Luiz André Barroso, Jeffrey Dean, and Urs Hölzle, "Web Search for a Planet: The Google Cluster Architecture," *IEEE Micro* 23, no. 2 (April 2003). http://labs.google.com/papers/googlecluster-ieee.pdf.

[24] Brian Despain, "*Google—The Network?*" entry for September 22, 2005, on the blog *Thinking Monkey*. http://www.thinkingmonkey.com/2005/09/google-network.shtml.

before the new millennium that the Web would outgrow the search engines' abilities to index it, but so far 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 main 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."[25] However, compare this with the January 2005 study by Gulli and Signorini,[26] which estimated that Google had indexed about 76 percent of the 11.5 billion pages on the Web, and it seems that the search engines provide significantly better coverage now than they did in the Web's infancy. Clearly, the search engines in general and Google in particular have been able to scale up their technology better than most people predicted at the end of the twentieth century.

But common sense suggests that there has to be some kind of limit to this continuous and rapid expansion. Even if Google's innovative, massively networked supercomputer architecture is technically capable of indefinite expansion, perhaps other kinds of constraints will prove insurmountable at some point in the future. A recent article by one of Google's principal hardware engineers warns that unless the ratio of computer performance to electrical power consumption improves dramatically, power costs may become a larger component of the total cost of ownership (TCO) than initial hardware costs.[27] This could become a significant barrier to the continued expansion of the Google platform in the future, particularly if energy costs continue to rise. A million interconnected servers consume a tremendous amount of electrical power.

Metadata to the Rescue?

In the early days of the Web, many people, particularly 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 classical 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, such as the subject gateways funded

---

[25] Steve Lawrence and C. Lee Giles, summary of "Accessibility of Information on the Web," *Nature* 400 (July 9, 1999): 107–9.

[26] Gulli and Signorini, "The Indexable Web Is More than 11.5 Billion Pages."

[27] Luiz André Barroso, "The Price of Performance: An Economic Case for Chip Multiprocessing," *ACM Queue* 3, no. 7 (September 2005). http://acmqueue.com/modules.php?=name=Content&pa=showpage&pid=330.

by the Electronic Libraries Programme for the higher education sector in the United Kingdom.[28] 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 expensive) MARC record, complete with subject headings and controlled terminology and conforming to standardized 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; unlike books, Web sites often change, and sometimes they disappear altogether.

As a result, metadata standards for describing Internet resources have appeared, such as META tags, the Dublin Core Metadata Element Set (DCMES), and the Resource Description Framework (RDF). These are discussed in more detail below (note, however, that many search engines make little or inconsistent use of embedded metadata, since it cannot always be trusted).

META Tags

The AltaVista search engine originally popularized the use of two simple metadata elements, "keywords" and "description," that can be easily and invisibly embedded in the <HEAD> section of Web pages by their authors using the HTML META tag. Here is an example:

> <META NAME="KEYWORDS" CONTENT="data standards, metadata, Web resources, World Wide Web, cultural heritage information, digital resources, Dublin Core, RDF, Semantic Web">
>
> <META NAME="DESCRIPTION" CONTENT="Version 3.0 of the site devoted to metadata: what it is, its types and uses, and how it can improve access to Web resources; includes a crosswalk.">

The original intention was that the "keyword" metadata could be used to provide more effective retrieval and relevance ranking, whereas the "description" tag would be used in the display of search results to provide an accurate, authoritative summary of the particular Web resource.

Dublin Core

The Dublin Core Metadata Element Set (DCMES)[29] is a set of fifteen information elements that can be used to describe a wide variety of

---

[28] See http://www.ukoln.ac.uk/services/elib/.

[29] *Dublin Core Metadata Element Set, Version 1.1, Reference Description*. http://www.dublincore.org/documents/dces.

resources for the purpose of simple cross-disciplinary resource discovery. Although originally intended solely as the equivalent of a quick and simple "catalog card" for networked resources, the scope of the Dublin Core gradually expanded over the past decade to encompass the description of almost anything. The fifteen elements are *Contributor, Coverage, Creator, Date, Description, Format, Identifier, Language, Publisher, Relation, Rights, Source, Subject, Title*, and *Type*.

The fifteen Dublin Core elements and their meanings have been developed and refined by an international group of librarians, information professionals, and subject specialists through an ongoing consensus-building process that has included more than a dozen international workshops to date, various working groups, and several active electronic mailing lists. The element set has been published as both a national and an international standard (NISO Z39.85-2001 and ISO 15836-2003, respectively). There are now a significant number of large-scale deployments of Dublin Core metadata around the globe.[30]

Resource Description Framework

The Resource Description Framework (RDF)[31] is a standard developed by the World Wide Web Consortium (W3C) for encoding resource descriptions (i.e., metadata) in a way that computers can "understand," share, and process in useful ways. RDF metadata is normally encoded using XML, the Extensible Markup Language.[32] However, as the name suggests, RDF only provides a *framework* for resource description; it provides the formal *syntax*, or structure, component of the resource description language but not the semantic component. The *semantics*, or meaning, must also be specified for a particular application or community in order for computers to be able to make sense of the metadata. The semantics are specified by an RDF vocabulary, 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 metadata elements in the domain. RDF vocabularies can be expressed either as RDF schemas[33] or as more expressive Web Ontology Language (OWL)[34] ontologies.

The CIDOC CRM[35] is a pertinent example of an ontology that provides the semantics for a specific application domain—the interchange of rich museum, library, and archive collection documentation. By expressing the classes and properties of the CIDOC CRM as an RDF

---

[30] *Dublin Core Projects*. http://www.dublincore.org/projects/.

[31] *Resource Description Framework*. http://www.w3.org/RDF/.

[32] *Extensible Markup Language (XML)*. http://www.w3.org/XML/.

[33] *RDF Vocabulary Description Language 1.0: RDF Schema*. http://www.w3.org/TR/rdf-schema/.

[34] OWL Web Ontology Language Guide: http://www.w3.org/TR/owl-guide/.

[35] See http://cidoc.ics.forth.gr/official_release_cidoc.html; and note 4.

schema or OWL ontology, information about cultural heritage collections can be expressed in RDF in a semantically unambiguous way, thereby facilitating information interchange of cultural heritage information across different computer systems.

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

RDF, RDF schemas, and OWL are all fundamental building blocks of the W3C's Semantic Web[36] activity. The Semantic Web is the vision of Sir Tim Berners-Lee, director of the W3C and inventor of the original World Wide Web: Berners-Lee's vision is for the Web to evolve into a seamless network of interoperable data that can be shared and reused across software, enterprise, and community boundaries.

A Bountiful Harvest

The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)[37] provides an alternative method for making Deep Web metadata more accessible. Rather than embed metadata in the actual content of Web pages, the OAI-PMH is a set of simple protocols that allows metadata records to be exposed on the Web in a predictable way so that other OAI-PMH-compatible computer systems can access and retrieve them. (Fig. 2.)

The OAI-PMH supports interoperability (which can be thought of as the ability of two 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 (see Glossary). As the names suggest, an OAI data provider is a source of metadata records, whereas the OAI harvester retrieves (or "harvests") metadata records from one or more OAI data providers. Since both an OAI data provider and an OAI data harvester must conform to the same basic information exchange protocols, metadata records can be reliably retrieved from the provider(s) by the harvester.

Although the OAI-PMH can support any metadata schema that can be expressed in XML, it 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 between different systems.

---

[36] Semantic Web. http://www.w3.org/2001/sw/.

[37] Open Archives Initiative: http://www.openarchives.org/.

Google's Sitemap, part of a suite of Webmaster tools offered by that search engine, also supports the OAI-PMH. By exposing a metadata catalog as an OAI data provider and registering it with Google's Sitemap, otherwise Deep Web content can be made accessible to Google's Web crawler, indexed, and made available to the search engine's users.

Meta-Utopia or Metagarbage?

In his oft-quoted diatribe, "Metacrap: Putting the Torch to the Seven Straw-men of the Meta-Utopia,"[38] journalist, blogger, and science fiction writer Cory Doctorow enumerates what he describes as the "seven insurmountable obstacles between the world as we know it and meta-utopia." 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 below.

- *"People lie."* Metadata on the Web cannot be trusted, because there are many unscrupulous Web content creators that publish misleading or dishonest metadata in order to draw additional traffic to their sites.
- *"People are lazy."* Most Web content publishers are not sufficiently motivated to do the labor involved in carefully cataloging the content that they publish.
- *"People are stupid."* Most Web content publishers are not smart enough to catalog effectively the content that they publish.
- *"Mission: Impossible—know thyself."* Metadata on the Web cannot be trusted, because there are many Web content creators who inadvertently publish misleading metadata.
- *"Schemas aren't neutral."*[39] Classification schemes are subjective.
- *"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.

---

[38] Cory Doctorow, "Metacrap: Putting the Torch to the Seven Straw-men of the Meta-Utopia," August 26, 2001. http://www.well.com/~doctorow/metacrap.htm.

[39] Doctorow confusingly uses "schema" here to refer to classification schemes, not the more common meaning of a metadata schema or data structure.

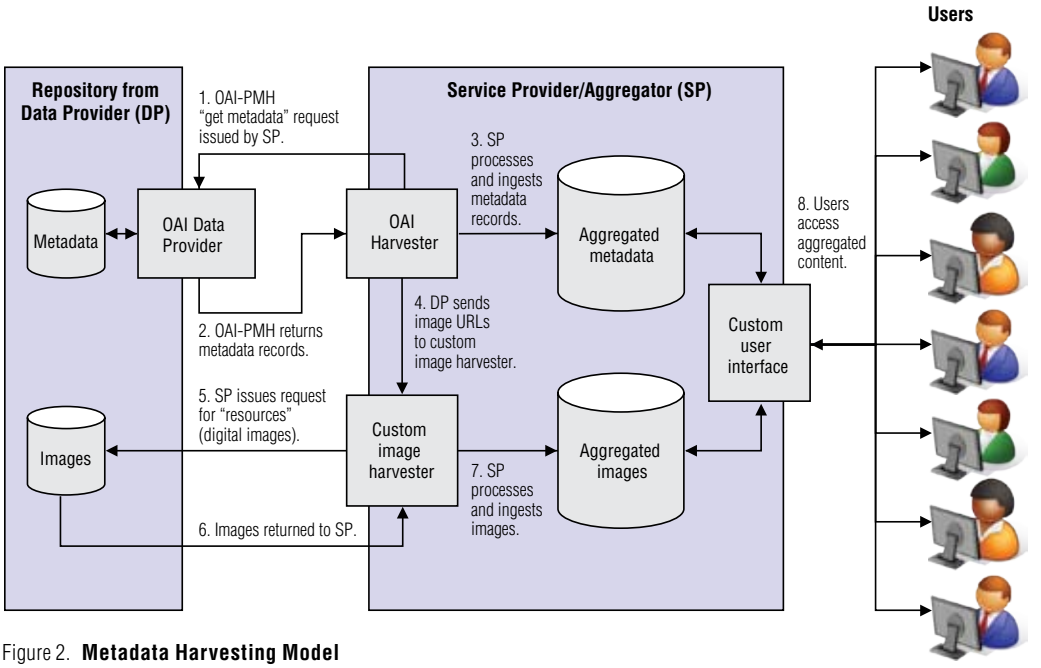

Figure 2. **Metadata Harvesting Model**

Doctorow's 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. META tag spam is designed to increase the likelihood that a Web site will appear in a search engine's search results and to improve the site's ranking in those results. There is plenty of incentive to increase a Web site's visibility and ranking with the search engines. Increased visibility and higher ranking can dramatically increase the amount of user traffic to a Web site, which results in greater profits for a commercial site's owners 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 either 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. But because most search engines do not utilize embedded metadata, there is usually no incentive for the vast majority of honest Web publishers to expend the additional 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 originally developed by AltaVista.

Doctorow's other points are less convincing, particularly if we look at the subset of Web content created by museums, libraries, and archives. Librarians, museum documentation staff, and archivists are typically diligent, well-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, EAD, CDWA Lite, and VRA Core), classification schemes, controlled vocabularies, and community-specific cataloging rules (such as AACR, DACS, and CCO) 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 of descriptive metadata created by human beings for centuries.

Playing Tag

Another recent development in the field of metadata on the Web that significantly weakens Doctorow's case are the so-called folksonomies. A folksonomy is developed collaboratively within a specific user community when many people use a shared system to label Web content, such as Web pages or online images, with descriptive terms, or tags. People are individually motivated to tag Web content because it allows them to organize and find the content at a later date; they are effectively building their own personal catalogs of Web content. With folksonomies, any terms or names can be used, without restriction—unlike taxonomic classifications, in which a fixed hierarchical list of carefully constructed descriptive terms must be used.

The folksonomy aspect comes into play when all the tags applied to a specific Web resource by multiple users are aggregated and ranked. If one person applied the term *impressionism* to a Web site, it doesn't really say very much. However, if several hundred people use this term and it is the most commonly used tag for that Web site, then it is a pretty safe bet that the Web site is about Impressionism and Impressionist art.

This is analogous to Google's PageRank™ algorithm: each time an individual user labels a Web resource with a specific descriptive tag, it counts as a "vote" for the appropriateness of that term for describing the resource. In this way, Web resources are effectively cataloged by individuals for their own benefit, but the community also benefits from the additional metadata that is statistically weighted to minimize the effects of either dishonesty or stupidity.

The two most well known examples of folksonomy/tagging sites on the Web are del.icio.us[40] and Flickr.[41] Del.icio.us enables users to create

[40] http://del.icio.us.

[41] http://flickr.com/.

tagged personal catalogs of their favorite Internet bookmarks, whereas Flickr is a digital photo sharing site that enables users to tag photos for easier retrieval. It is interesting to note that both companies were acquired by Yahoo! in 2005. Clearly, the world's second most popular search engine company sees significant value in community-generated metadata.

In Metadata We Trust (Sometimes)

Metadata is not a universal panacea for resource discovery on the Web. The underlying issues of trust, authenticity, and authority continue to impede the widespread deployment and use of metadata for Web resource discovery, and this situation is unlikely to change as long as the search engines can continue to satisfy the search needs of most users with their current methods (indexing the Title HTML tags, the actual words on Web pages, and ranking the "popularity" of pages based on the number of referring links).

However, human-created metadata still has an extremely important role within specific communities and applications, especially in the

Libraries and the Web

The Web has dramatically changed the global information landscape—a fact that is felt particularly keenly by libraries, the traditional gateways to information for the previous two millennia or so. Whereas previous generations of scholars relied almost entirely on libraries for their research needs, the current generation of students, and even of more advanced scholars, is much more likely to start (and often end) their research with a Web search.

Faced with this new reality, libraries and related service organizations have been working hard to bring information from their online public access catalogs (OPACS), traditionally resources hidden in the Deep Web beyond the reach of the search engines' Web crawlers, out into the open. For example, OCLC has collaborated with Google, Yahoo! and Amazon.com to make an abbreviated version of its WorldCat union catalog accessible as Open WorldCat. The full WorldCat catalog is available only by subscription.

But the most striking example of collaboration between libraries and a search engine company to date is undoubtedly the Google Book Search–Library Project.[1] This massive initiative, announced late in 2004, aims to make the full text of the holdings of five leading research libraries—Harvard University Library, the University of Michigan Library, the New York Public Library, Oxford University Library, and Stanford University Library—searchable on the Visible Web via Google.

By adding the full text of millions of printed volumes to its search index, the Google Book Search–Library Project will enable users to search for words in the text of the books themselves. However, the results of searches will depend on the works' copyright status. For a book that is in the public domain, Google will provide a brief

bibliographic record, links to buy it online, and the full text. For a book that is still in copyright, however, Google will provide only a brief bibliographic record, small excerpts of the text in which the search term appears (the size of the excerpts depends on whether the copyright holder is a participant in the Google Books Partner Program,[2] a companion program for publishers), and links to various online booksellers where it can be purchased.

It is perhaps ironic that, due to the dysfunctional and anachronistic state of existing copyright legislation, this scenario is almost the exact reverse of the familiar library catalog metadata example: Rather than search metadata catalogs in order to gain access to full online texts, the Google model helps users to search full online texts in order to find metadata records!

But open access to the rich content of printed books is clearly an idea whose time has come. The Google Book Search–Library Project may be the most ambitious project of its kind to date, but it is neither the first large-scale book digitization project (e.g., the Million Book Project has already digitized over 600,000 volumes)[3] nor the last. At the same time that Google was striking deals with libraries to digitize their collections, the Internet Archive and its partner, Yahoo! were busy recruiting members for the Open Content Alliance.[4]

The Open Content Alliance is a diverse consortium that includes cultural, nonprofit, technology, and government organizations that offer both technological expertise and resources (e.g., Adobe Systems, HP Labs, Internet Archive, MSN, Yahoo!) and rich content (e.g., Columbia University, the UK's National Archives, the National Library of Australia, Smithsonian Institution Libraries, the University of California). It has a broad mission to "build a permanent archive of multilingual digitized text and multimedia content" and "to offer broad, public access to a rich panorama of world culture."[5]

The Open Content Alliance has launched the Open Library,[6] which, like Google Book Search, will make the full texts of large quantities of books accessible via Yahoo!'s search engine while simultaneously respecting copyright restrictions. However, unlike the Google initiative, the Open Library is committed to making the full text of every digitized book available free of charge on the Web.

The undeniably positive result of these various initiatives is that within the next decade or so the Web will be vastly enriched by the addition of a huge and freely accessible corpus of the world's literature. Unfortunately, however, unless the copyright situation improves dramatically (e.g., through the introduction of proposed new legislation for "orphan works"),[7] it seems that the corpus of literature soon to be freely available on the Web will not include any significant quantity of copyrighted material from the twentieth and twenty-first centuries.

---

[1] Google Books–Library Project: http://books.google.com/googlebooks/library.html.

[2] Google Books–Partner Program: http://books.google.com/googleboos/publisher.html.

[3] Million Books Project Frequently Asked Questions: http://www.library.cmu.edu/Libraries/MBP_FAQ.html.

[4] Open Content Alliance: http://www.opencontentalliance.org/.

[5] Open Content Alliance Frequently Asked Questions: http://www.opencontentalliance.org/faq.html.

[6] The Open Library: http://www.openlibrary.org/.

[7] Report on Orphan Works: A Report of the Register of Copyrights, January 2006, U.S. Copyright Office. http://www.copyright.gov/orphan/. See also "Rights Metadata Made Simple," p. 63.

museum, library, and archive communities for whom metadata is really just cataloging with a different name. All the necessary standards and technology components to facilitate intracommunity knowledge sharing are now in place:

- Descriptive data structure standards for different kinds of community resource descriptions, for example, MARC,[42] Dublin Core, MODS,[43] EAD,[44] CDWA Lite,[45] and VRA Core;[46]
- Markup languages and schemas for encoding metadata in machine-readable syntaxes, for example, XML and RDF;
- Ontologies for semantic mediation between data standards, for example, CIDOC CRM and IFLA FRBRoo;[47]
- Protocols for distributed search and metadata harvesting, for example, the Z39.50 family of information retrieval protocols (Z39.50,[48] SRU/SRW[49]), SOAP,[50] and OAI-PMH.[51]

By combining these various components in imaginative ways to provide access to the rich information content found in museums, libraries, and archives, it should be possible to build a distributed global Semantic Web of digital cultural content and the appropriate vertically integrated search tools to help users find the content they are seeking therein.

---

[42] http://www.loc.gov/marc/.

[43] http://www.loc.gov/standards/mods/.

[44] http://www.loc.gov/ead/.

[45] http://www.getty.edu/research/conducting_research/standards/cdwa/cdwalite.htm.

[46] http://www.vraweb.org/projects/vracore4/index.html.

[47] http://cidoc.ics.forth.gr/frbr_inro.html.

[48] Z39.50 Maintenance Agency: http://www.loc.gov/z3950/agency/.

[49] SRU (Search/Retrieve via URL): http://www.loc.gov/standards/sru/.

[50] Simple Object Access Protocol (SOAP): http://www.w3.org/TR/2000/NOTE-SOAP-20000508/.

[51] Open Archives Initiative Protocol for Metadata Harvesting (OAH-PMH): http://www.openarchives.org/OAI/openarchivesprotocol.html.