Introduction to Metadata *Pathways to Digital Information*

http://www.getty.edu/research/institute/standards/intrometadata/

Metadata and the World Wide Web

Tony Gill

Introduction

Few people would argue with the assertion that catalogs are useful tools for managing collections of items, and that their usefulness increases proportionately with the size of the collection being managed. A catalog of concise, well-structured descriptions of the items in a collection should always be easier to manage than the colle ction itself, since it should provide both a distillation of the collection in terms of volume and a consistent, easily-understood structure. However, perhaps fewer people appreciate that the act of cataloging a collection is actually a process of *knowledge representation*.

Designing a catalog for a collection is ultimately a philosophical problem-solving exercise; it is an attempt to determine the most significant attributes or properties of the items in the collection, so that the *essence* of the items can be captured as concise descriptions. These concise descriptions then represent the items in the catalog, and provide a route back to the items themselves. The catalog should be much easier to search, sort and browse than the collection itself, provided that sufficient consistency in the structure and content of the descriptions is achieved, because it contains only the most essential information characterizing the items in the collection.

Computers are innately well-suited for managing catalogs; in fact, it could be argued that storing and manipulating large collections of structured data is a core component of their *raison d'être*. Database management systems have been used to store every conceivable type of catalog, from mailing lists to stock inventories to museum collections to library holdings, since they were first developed.

Computers have always employed catalogs internally as well, to keep track of different discrete data objects. In order to function correctly, they must keep an accurate record of the identity and location of every item of data stored in the various memories. For example, the operating system of a computer uses a catalog called the File Allocation Table to store the names of files and their physical position on a disk.

This type of data catalog is itself stored by the computer as data, a recursive relationship that has resulted in it being referred to as "metadata."

Many introductory articles about metadata begin by defining it simply and economically as *"data about data,"* in an attempt to demystify a term that is used considerably more often than it is fully understood. This concise and accurate definition is often then incorrectly generalized, either implicitly by the reader or explicitly by the author, to mean *"information about information."*

The unhelpful result of this undoubtedly well-intentioned semantic lenience is that the term "metadata" is now increasingly used in contexts where the term "data' would have sufficed just a few short years ago (for example, descriptions of people, objects and events), often resulting in confusion and misunderstanding.

This variety of interpretations of the term "metadata" is not altogether surprising — it is formed from two root terms that have both been adopted and re-purposed by practitioners of diverse disciplines over several millennia, ranging from epistemology and metaphysics to chemistry and computer science.

The usage of the term "metadata" in the context of this essay will borrow and synthesize meaning from the disciplines of both computer science and philosophy. Computer science provides a useful constraint for the concept of "data", by limiting it to the realm of discrete identifiable pieces of digital "computer data" — certainly still a fairly abstract concept, but considerably less so than the more general interpretation of data as facts or assertions used for analysis and inference. Philosophy, specifically metaphysics, provides the example usage of "meta" as a prefix to denote an alternate or second-order kind of relationship between two similar types of entities, and the underlying notion of the essential attributes that make up a metadata description.[1]

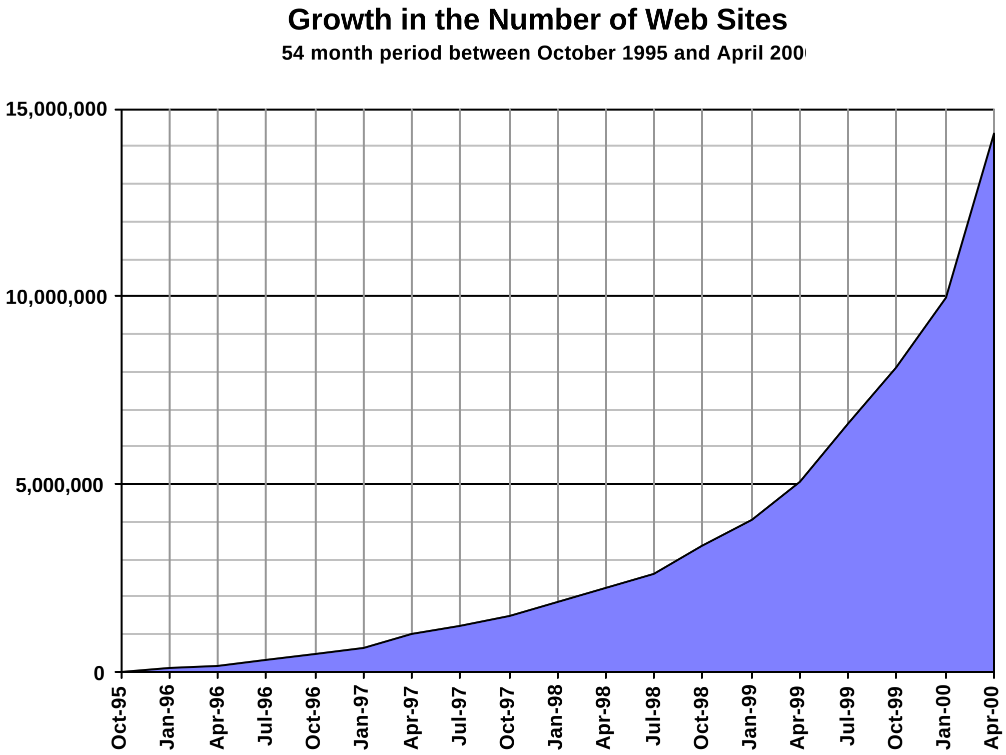
So, moving from the abstract realm to the practical, the term "metadata" in the context of this essay refers to structured descriptions, stored as computer data, that attempt to describe the essential properties of other discrete computer data objects—specifically, the data objects that make up the information on the World Wide Web, the world's largest and fastest-growing collection of data.

The Rise and Rise of the World Wide Web

It is impossible to determine the exact size of the World Wide Web; it has grown so large, so fast, and is so impenetrable to practical survey methodologies that it has effectively transcended our ability to measure it with any degree of precision.

However, although the actual numerical quantities will never be entirely accurate and are instantaneously out of date, carefully-designed surveys carried out at regular intervals can at least provide some insight into the trends in Web growth and usage over time.

The most recent Netcraft survey,[2] carried out on 1 April 2000, received responses to HTTP requests for server names from 14,322,950 “sites,” where a site in this case represents a unique hostname such as *http://www.hostname.com* or *http://www.hostname.org*. This is an arbitrary but simple approximation for the total number of Web sites that counts different hostnames on the same IP address as separate, but does not count separate distinct Web sites that share the same hostname: For example, *http://www.hostname.com/myWeb site/* and *http://www.hostname.com/yourWeb site/* would not be counted separately.



Source: Netcraft Survey
http://www.netcraft.com/survey/

To put this number into context, a similar type of survey conducted by Matthew Gray of the Massachusetts Institute of Technology found just 130 Web hosts in June of 1993; the Web grew by nearly *eight million percent* in less than seven years.[3]

The number of hosts is only one metric for determining the size of the Web, however; there have also been a number of attempts to count the number of individual pages available. The most recent attempt at the time of writing is the Inktomi WebMap,[4] a joint survey by the search engine company Inktomi and the NEC Research Institute, which announced in a press release dated 18 January 2000 that the Web contained in excess of one billion unique, indexable documents.[5] This does not include duplicate documents on mirror servers or documents that are "hidden" from Web crawlers, such as documents that are dynamically generated by querying underlying databases or that require some kind of user log-on.

A Selection of Web Facts

- The Tenth GVU Web Survey,[6] conducted in October 1998, found that 85% of respondents used search engines to find information on the Web, making it the second most common way of finding content (the most common method, used by 88%, is to follow hyperlinks from other pages).
- The survey by Lawrence & Giles found that, of the 15 terabytes of data that made up the estimated 800 million pages of the publicly indexable Web in February 1999, only 6 terabytes (40%) contained useful text after removing HTML tags, comments, and white space.[7]
- The same 1999 survey found that the mean number of Web pages per server was 289 and that search engines were more likely to index pages that were accessed via links from other pages.
- According to the results of a survey by Alexa Internet at the end of 1999, 80% of Web traffic is directed at just 0.5% of sites, with the top 5 sites (Yahoo, Microsoft, Excite, eBay, and AltaVista), Disney (Go.com), and AOL accounting for one click in five.[8]
- As of April 2000, new domains were being registered at a rate of one per second.[9]

Finding Needles in a Global Haystack

In view of the huge size and explosive rate of growth of the World Wide Web, it is clear that catalogs of some kind would be invaluable in helping users discover relevant information resources. Unfortunately, neither the Internet nor the World Wide Web were originally designed with the cataloging of their contents in mind; the TCP/IP suite of network protocols that enables the basic infrastructure of the Internet to function is solely a transport layer, concerned with getting packets of data from one point to another as quickly and reliably as possible, whereas the Hyper Text Transfer Protocol (or HTTP) only deals with the delivery of hyperlinked World Wide Web information.

This means that the existing network protocols do not provide any dedicated support for locating specific information resources available on the network. This sorry state of affairs falls very short of the vision of the *Memex,* a comprehensive and affordable personal reference and research tool originally proposed way back in 1945 by Vannevar Bush, believed by many to be the precursor of hypertext.[10]

The disappointment of the hypertext community with the World Wide Web is clearly illustrated by this quote from Ted Nelson (the man who first coined the term *"hypertext"* in 1965), 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.*[11]

Unsurprisingly, tools designed to address the resource location problem and help make sense of the Internet's vast information resources started to appear soon after the launch of the first Web browsers in the early 1990's; for example, Tim Berners-Lee founded the WWW Virtual Library[12] shortly after inventing the Web itself, and Yahoo![13], Lycos,[14] and Webcrawler[15] were all launched during 1994.

The tools currently available to help users find Web resources are many times larger and more powerful than their 1994 predecessors — they have to be, in order to keep up with the explosive growth in both the amount of information available and the number of users accessing it. However, there are still only two principle classes of Web resource locating tools: directories and search engines.

Directories are listings of network resources created by real people, who select, catalog and classify Web resources that they feel are appropriate for their constituency, based on factors such as accuracy, authority, and currency. Directories can either be general in scope, such as the World Wide Web Virtual Library and Yahoo!, or they can specialize in particular subject areas, such as the Art, Design, Architecture & Media Information Gateway (ADAM)[16] and the Edinburgh Engineering Virtual Library (EEVL).[17] Directories typically provide access to the resources they have cataloged both by searching and by browsing a hierarchical set of subject headings.

Search Engines, often called "spiders," "crawlers" or "robots," are automated systems that continuously traverse the Web visiting sites, saving copies of the pages and their locations as they go in order to build up a huge catalog of fully-indexed pages. They typically provide powerful searching facilities and extremely large result sets, which are relevance-ranked (using closely-guarded proprietary algorithms) in an effort to make them usable.

In recent years some hybrid approaches have started to appear — for example, the Northern Light[18] search engine, which attempts to automatically cluster results into dynamically-generated "Custom Search Folders" according to subject, type of document, source or language, giving the kind of hierarchical organization of results more usually associated with directory services.

However, there are serious problems with both the directory and search engine approaches. Human-mediated directories generally provide good search precision at the broad subject level, and are normally considered to provide higher-quality information overall because of the human intervention in the indexing and classification process. However, this mediation is a costly, labor-intensive process that is not sufficiently scaleable to provide comprehensive up-to-date coverage of the whole Web, much of which is highly transient.

Another problem with the hand-crafted approach to cataloging Web resources is deciding upon the granularity of the resources to be described; should descriptions be created for

Web sites as a whole, or should each page be cataloged individually? Clearly, a cost-benefit tradeoff will always need to be made.

The crawler-based search engines also suffer from a number of serious problems, which affect their ability to provide an index that is both comprehensive and current, and the likelihood that users will find what they are looking for even if it has been indexed:

- Increasingly, information on the Web is being generated dynamically from databases in response to user input. This information is sometimes referred to as "the hidden Web," because it is beyond the indexing reach of the Web crawlers.
- The Web crawling components of the search engines are fully automated, which means that the indexed Web resources are selected by software algorithms rather than people, and are therefore variable in both quality and depth of indexing.
- The Web indexing playing field is not a level one: Recent research suggests that *"search engines are typically more likely to index US sites than non-US sites (AltaVista is an exception), and more likely to index commercial sites than educational ones."*[19]
- Searching large automatically-indexed databases often results in extremely large results sets, which are frequently unusable despite increasingly sophisticated information retrieval tools, relevance ranking procedures and context-aware artificial intelligence algorithms.
- As the volume of information on the Web continues to increase exponentially, the amount of network bandwidth (information-carrying capacity) required by the crawlers in order to maintain current and comprehensive indices could eventually reach unacceptable levels; ethical "codes of conduct" for Web crawlers have already existed for some years.

The search engines seem to be showing signs of strain in attempting to keep up with the explosive growth of the Web. Steve Lawrence & C. Lee Giles of the NEC Research Center conducted a scientifically rigorous survey of the search engine's coverage of Web content in February 1999.

The findings of their survey, published in the peer-reviewed journal *Nature,* suggest that the combined coverage of the 11 search engines used for the study was about 42% of the total number of unique indexable pages on the Web (i.e. not including the ever-expanding "hidden Web"), with no search engine indexing more than about 16%. In summary:

> *Our results show that the search engines are increasingly falling behind in their efforts to index the Web."*[20]

The publication of these findings subsequently seemed to prompt the search engines both to increase the size and currency of their indices, and to start quoting ever-larger numbers of the pages visited in order to generate their indices. As Danny Sullivan observes in the March 2000 issue of the Search Engine Report:

> *One of the latest trends these days is for crawlers to flaunt both how many pages they have in their index plus the larger number of pages visited to create that index. [..] Why have dual numbers returned? Because no matter how big your competition is, the Web is even bigger.*[21]

However, despite these renewed efforts by the search engines (according to Sullivan, Inktomi claimed to have an index of over 500 million pages in April 2000[22]), the outlook for their ability to keep up with the growth of the Web in the long term is not promising.

Cataloging the Web

Although initially it appears that both directories and search engines suffer from different types of problems, it seems clear that most if not all of the difficulties are the result of ambitions which are likely to prove untenable in the long term; the Web is simply too big for any single organization or service to catalog, irrespective of whether they use people or computers to generate their indices.

If there is a solution to the problem of resource discovery on the Web, it must surely be based on a distributed metadata catalog model. Ironically, the WWW Virtual Library uses just such a distributed model; however, the altruistic efforts of its volunteer curators have proved insufficient to keep pace with the growth of the Web.

The necessary technical protocols for creating distributed meshes of resource discovery databases, such as Z39.50 and WHOIS++, are already available — interoperability at a technical level is no longer a significant problem.

What is required now is the widespread adoption of standards for metadata structure, content and authentication that will allow secure interoperability on the semantic level. However, before discussing the specifics of the metadata standards currently available, it will be helpful to consider in more detail some of the specific applications that metadata can be used for, and some of the more problematic issues that arise in the description of networked resources.

Metadata Applications and Issues

Clearly, the information structure and content of Web metadata records should capture the essence of the Web resources they describes and facilitate the various tasks for which the metadata was devised.

Unfortunately, this is the point where real-world complexities start to intrude; with such a large collection of information objects to describe, spanning the breadth and depth of human knowledge and creativity, and with tens of millions of users, the number of potential applications for Web metadata is limited only by the imagination. Consequently, consensus on the most appropriate structure and content for Web metadata remains elusive, despite significant efforts worldwide; some of the more significant descriptive standards resulting from this metadata research are described below, and elsewhere on this site.

The most common application of Web metadata is generally referred to as “resource discovery,” because the metadata is intended to assist Web users discover the information they are looking for; the availability of consistent, accurate and well-structured descriptions of Web resources could enable much greater search precision and more accurate relevance ranking of the large result sets typically retrieved by search engines, for example.

Once potentially useful candidate resources and their locations have been identified, metadata can also be used to provide short descriptions or evaluations that can help the user determine the relevance of the resource, or information about any access restrictions or rights implications that may prohibit the intended use of the information. Whether or not these applications are intrinsic parts of the resource discovery process or are in fact separate applications of Web metadata remains the subject of debate.

Metadata is also often used in the management and administration of digital networked resources; this type of “administrative metadata” is essential for ensuring that Web resources are kept up to date, for example, or are free of rights restrictions that may prohibit their distribution over the Internet.

One of the more interesting consequences of the metadata research taking place around the globe is that effective cataloging, historically perceived as an arcane art practised only by librarians, museum curators and archivists, is now becoming an issue for a much wider community.

Acceptance of the importance of controlled vocabularies and formal classification schemes is becoming increasingly widespread — a fact that most experienced catalogers have taken for granted for decades (notwithstanding the fact that the sheer diversity of information on the Web is highlighting the shortcomings of the existing taxonomies for organizing the sum of human learning!).

However, the sheer scale of the Web as an information space will require new applications of the old tools and skills, such as the use of thesauri by software for automatically expanding users’ queries to include synonyms or even translations of the query terms into alternate languages, or mappings between different classification schemes and terminology authorities.

Similarly, the fact that a diverse range of vocabularies and classification schemes will need to coexist in the same vast information space means that computers must be able to identify the source authority for terms or classmark; consequently, *schema registries* will be required in order to define *namespaces* and thereby ensure that the labels used to identify the various authorities are unique and unambiguous.

While there are undoubtedly many lessons that can and should be learned from the traditional custodians of information, there are also a number of new challenges unique to the pandisciplinary, transglobal, multilingual and multicultural networked environment of the Web that will require fresh approaches and new solutions.

For example, deciding upon the most appropriate *granularity* for the resource descriptions is another issue that the would-be Web cataloger must address: How much detail about a Web resource should a catalog record contain? How many catalog records should be created for a given Web resource? Increasing user expectations regarding retrieval capabilities, combined with the flexibility and diversity of the hypertext information environment, jointly conspire to render the analogy between Web cataloging and bibliographic cataloging only partially valid. No longer content with the traditional "author, title, keyword" searches offered by library catalogs, users now expect to be able to search for words or phrases appearing within the body text of Web resources. A hybrid approach that incorporates hand-crafted site-level descriptions produced by skilled catalogers and augments them with automated full-text indexes, could provide the most effective solution providing the results are relevance-ranked accordingly.

Another significant conceptual difficulty arises from the need to describe the relationships between networked resources and other objects: What exactly should metadata describe? Strictly speaking, metadata should describe the properties of an object which is itself data, for example a Web page, a digital image or a database — which is analogous to the librarian's practice of cataloging "the thing in hand." For networked resources, however, these properties are often not very interesting or useful for the purposes of discovery; for example, if a researcher is interested in discovering images of famous artworks on the Web, they would generally search using the properties of the original artworks (e.g. CREATOR = Picasso, DATE = 1937), not the properties of the digital copies or "surrogates" of them (e.g. CREATOR = Scan-U-Like Imaging Labs Inc., DATE = 2000-02-29).

Both the "granularity" and "surrogacy" problems have at their root the need to describe the *relationships* between different objects (not all of which will exist on the Web) in the various records describing those objects; for example, a record describing a Web page within a site should indicate its membership of the site, and a scanned image of a Picasso painting on the Web should identify the painting from which it was derived.

Of course, none of the problems described above are new–the traditional guides to information resources, such as librarians, museum curators and archivists, have been wrestling with the seemingly-impossible task of "Modeling the World" in order to describe information resources for decades. But the urgent need to catalog the Web has made these fundamentally epistemological issues significant for a new and much larger community.

Tools for Web Cataloging

Over the last few years, a plethora of tools for cataloging the Web have appeared — most of them Web-accessible themselves.

Some tools simply provide basic metadata creation and editing features, allowing syntactically-correct metadata records to be created and edited manually without the need to understand the complexities of the various encoding syntaxes. Other tools provide more sophisticated features, such as the ability to convert between different metadata formats or automatically extract embedded metadata from Web pages; some tools even attempt to generate metadata automatically by making inferences from the contents of documents. A detailed list of metadata-related tools is maintained on the Dublin Core Web site.[23]

In addition to hosting the Dublin Core Web site, OCLC also operates a developing service called CORC[24] (Cooperative Online Resource Catalog), that provides an insight into how the various Web cataloging tools can be provided in an integrated system to support the creation and maintenance of a collaborative Web resource catalog. CORC provides a suite of Web cataloging tools that can be used by participating librarians to add to the central shared CORC database, which at the time of writing contained 229,075 resource descriptions.

Standards for Metadata on the Web

In order for metadata to be as useful and cost-effective as possible, it is essential that its structure, semantics and syntax conform to widely supported standards, so that it is effective for the widest possible constituency, maximizes its longevity and so that processing can be automated as far as possible.

Three metadata standards efforts are particularly pertinent in the Web context: The “keyword” and “description” meta tags as implemented by the search engines, the Dublin Core Metadata Initiative, and the Resource Description Framework. These are discussed below.

Search Engine Meta Tags

The AltaVista search engine originally popularized the use of two simple metadata elements, “keywords” and “description,” that can be embedded in Web pages by their authors using the HTML meta tag. The original intention was that the “keyword” metadata could be used to provide more effective retrieval and relevance ranking, whereas the “description” would be used in the display of search results to provide a more accurate summary of a Web resource.

With the exception of meta tags that are automatically (and somewhat pointlessly) inserted into Web pages by authoring tools such as the "generator" tag, "keywords" and "description" are now the most commonly-used meta tags on the Web. The Lawrence & Giles 1999 survey[25] ascertained that they were used in the homepages of 34% of sites, for example.

Unfortunately, many of the major search engines have now stopped using meta tags to improve relevance ranking, and some have even stopped indexing meta tags, because of the increase (or at least the perceived increase) in *meta tag spamming* or *spoofing.* Meta tag spamming is the term given to the deliberate misuse of meta tags in order to boost a site's ranking in search results, for example by repeating keywords hundreds of times or by using sexually-explicit keywords. The following policy statements are from the Web sites of AltaVista, Excite and Northern Lights respectively:

> *Why aren't METAtags given preference? Consider the opportunity for abuse and spamming. [..] Basically, METAtags are a band aid to help you deal with pages that don't state what they are about in clear text, right up front. Do it right to begin with, and you don't need METAtags at all. You'll get far better results in terms of search engine traffic that way.*[26]
>
> *Unfortunately, meta tag information is not always reliable. It may or may not accurately reflect the content of the site. In general, our spider does not honor metatags. This means we do not index the content of the meta tag.*[27]
>
> *While our crawler does make note of META tags, Northern Light does not assign any particular relevance to words contained in META tags, nor do we use them to control descriptions on our results list.*[28]

According to Search Engine Watch,[29] the only search engines that use the "keywords" meta tag to provide more effective relevance ranking are those based upon the Inktomi search engine (Inktomi lists America Online, Freeserve.net, Goto.com, LookSmart, HotBot, MSN and Yahoo! among its customers). Inktomi claims that their search engine can detect common spamming techniques, and "penalizes" documents it suspects of containing inappropriate metadata by ranking them lower.[30]

The "description" tag is also used by some search engines (e.g. AltaVista, Inktomi, Excite) to provide more naturalistic descriptions of sites in results displays, when compared to the automatically generated summaries from the first few lines of the document that are generally used otherwise.

Although the search engines all have different approaches with respect to metadata and relevance ranking, they appear to have one characteristic in common—they all use the contents of the HTML <TITLE> tag as the single most significant factor in the ranking of result sets.

Dublin Core

The Dublin Core Metadata Element Set[31] (a.k.a. "Dublin Core" or just "DC") is a set of 15 information elements that can be used to describe a wide variety of information resources on the Internet for the purpose of simple cross-disciplinary resource discovery. The 15 elements (described in more detail elsewhere on this site) are:

> *Contributor, Coverage, Creator, Date, Description, Format, Identifier, Language, Publisher, Relation, Rights, Source, Subject, Title, and Type.*

The 15 elements and their meanings have been developed and refined by a group of librarians, information professionals, and subject specialists through an ongoing consensus-building process that has included seven international workshops to date and an active mailing list.[32]

From the outset, the development of the Dublin Core element set has been underpinned by a number of guiding philosophies:

- The elements must be simple to understand and use, so that any creator of networked resources would be able to describe their own work without requiring extensive training.
- Every element is both optional and repeatable.
- The elements should be international and cross-disciplinary in scope and applicability.
- The element set should be extensible, to allow discipline or task-specific enhancements.
- The most important strategic application of the element set would be for embedded descriptions of Web resources, created by the resource authors, which meant a syntax that could be accommodated within HTML's <META> tag.

Early adopters of the Dublin Core soon encountered the types of problems discussed in the previous section, which have resulted in a number of additional extensions and refinements to the simple core element set:

- The Warwick Framework,[33] a conceptual container architecture for diverse heterogeneous metadata packets; prototype SGML and MIME implementations of the Warwick Framework have been developed, but perhaps the most important contribution of this work is the formalization of requirements that led to the development of the Resource Description Framework (discussed below).
- Interoperability Qualifiers[34] that can be used either to refine the semantics of the element or to provide more information about the encoding scheme used for an element's value.

- Acknowledgement of the *1:1 Principle,* which states that the most robust solution to the granularity and surrogacy issues described previously is to use separate metadata "sets" or "packets" for each discrete object (item or collection, network resource or otherwise), and to describe the relationships between them using an enumerated list of relationship types.

There are now a number of large-scale deployments of Dublin Core metadata around the globe — the official Dublin Core Web site lists15 in North America and Mexico, 27 in Europe and 12 across Asia and Australia.[35] Some of these initiatives are on a national scale, for example the Australian Government Locator Service[36] and the CCTA Government Information Service in the UK, open.gov.uk.[37]

However, although significant progress in raising awareness and increasing deployment of the Dublin Core has been made over the last few years, there is still a long way to go before it can begin to deliver on its promise of better resource discovery on the Web. The Lawrence & Giles 1999 survey,[38] for example, found that only 0.3% of Web sites contained Dublin Core metadata. This poor uptake, in global terms at least, is undoubtedly due at least in part to the reluctance of the major search engines to support Dublin Core:

> *Search engine support is crucial for success, as demonstrated by the lack of support for the existing Dublin Core meta tags. [..] Practically no one uses these tags, and the reason why is because none of the major search engines does anything with them. They don't index them, nor do they provide a way to search within the Dublin Core meta tag fields.*[39]

Another factor that has hindered the widespread adoption of Dublin Core metadata is the length of time it has taken to reach consensus on approved Interoperability Qualifiers.[40] Qualifiers for refining element semantics and identifying formal encoding schemes were originally proposed as the "Canberra Qualifiers"[41] during the fourth Dublin Core Workshop in Australia in March 1997, but the initial set of approved qualifiers was not formally accepted as part of the Dublin Core "registry" until April 2000, more than three years later.

The lengthy delay in reaching consensus on qualifiers was certainly not caused by a lack of effort or commitment from those involved; the Dublin Core Metadata Initiative is a voluntary international standards effort, and the participants regularly donate significant time and resources to the cause of improved Web resource discovery.

Notwithstanding the effort required to reach international cross-disciplinary consensus on any topic, the intellectual difficulty in reaching agreement on qualifiers is partly the result of well-intentioned attempts to apply Dublin Core far more broadly than what it was originally designed for — simple discovery of "document-like objects" on the World Wide Web.

CIMI, the Consortium for the Computer Interchange of Museum Information, conducted a detailed two-phase investigation into the utility of Dublin Core metadata over a three-year period. Starting in 1998, Phase I looked at simple unqualified Dublin Core for museum information resource discovery, whereas Phase II extended the “testbed” to include the use of qualified Dublin Core metadata for the interchange of richer descriptions between museums.

CIMI found that the unqualified implementation of the Dublin Core Metadata Element Set could be an effective tool for the coarse-grained discovery of museum information resources in a cross-disciplinary networked environment, particularly if the recommendations in the CIMI Guide to Best Practice were followed.[42]

However, CIMI also found that Qualified Dublin Core (DCQ) could not be recommended for information interchange within the museum community, because it could not support the rich descriptions that museums need to share. This was due to a combination of constraints imposed by the underlying data model of the element set, which was originally designed for the description of text-based Web resources, and the "dumb-down" rule for the application of “semantic refinement” qualifiers, which stipulates that qualifiers can refine but not extend the semantics of any given element.

Regardless of the success or failure of the Dublin Core in its current guise to be widely adopted for resource discovery on the Web, the Herculean and ongoing effort has resulted in a deliverable that could prove even more significant in the long-term — international, cross-disciplinary consensus on the key requirements for effective resource discovery on the Web.

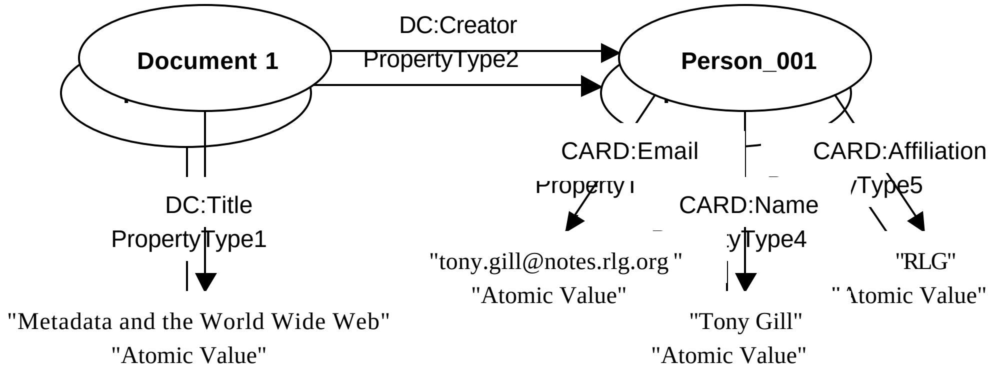
The lessons learnt in the Dublin Core Metadata Initiative have helped to build the foundations of another metadata standard: the Resource Description Framework.

Resource Description Framework

The Resource Description Framework,[43] produced as part of the World Wide Web Consortium’s Metadata Activity, is a metadata application of XML,[44] the Extensible Markup Language, the successor to HTML and the future language of the Web. Its development was informed by previous work such as PICS[45] (Platform for Internet Content Selection), the Dublin Core/Warwick Framework initiative, and the metadata activities of major software vendors such as Microsoft and Netscape.

The Resource Description Framework is built upon a simple but robust data model that allows r*esources* to be described in terms of their *properties*. The *values* of the properties can be either *atomic* in nature, such as text strings or numbers, or they can in turn be other *resources*, which can have *properties* of their own.

This data model is often depicted visually using a type of diagram called a *directed labeled graph,* also known as a *node and arc diagram.* A generalized example of an RDF description could take the following form (all of the examples and diagrams in this section are based heavily on Eric Miller's excellent examples[46]):



As the name implies, RDF is a *framework* for resource description; it has to be adapted in order to serve specific communities or applications through the use of *RDF Schemas,* which use the XML *Namespace* mechanism to unambiguously identify the particular semantics of the property types.[47]

To illustrate this by example, a description of this essay and its authorship could feasibly be described using two RDF Schemas, each based on a different metadata standard with different semantics; Dublin Core element definitions could be used for the description of the Web document, whereas the semantics of the elements in the vCard[48] scheme could be used to describe the properties of the author. In this example, the namespace mechanism is used to specify that property types prefixed with "DC" refer to Dublin Core element semantics and those prefixed with "CARD" refer to vCard semantics:

Using this highly extensible and robust logical framework, 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 specifically declared using the namespace mechanism.

In effect, RDF is a practical implementation of the Warwick Framework, in that it supports the coexistence of heterogenous "packets" of metadata, but it could in principle accomplish much more than the Warwick Framework set out to achieve — RDF could enable the Web to evolve into a global semantic network.

Metathreats and Metaopportunities

It is just two years since this essay was first published, and although much progress has been made in terms of the standards and tools to support the deployment of metadata on the World Wide Web, practical solutions to some of the underlying social, political and economic problems remain elusive.

This should not be too surprising — factors such as trust, privacy, authenticity, and authority have always been critically important in the dissemination of information, and the ease with which the Web allows information to flow exacerbates the need to address these issues in the networked environment.

It can no longer be argued that the lack of metadata on the Web is caused by a lack of standards; a range of usable metadata standards are now available, from simple search engine “keyword” and “description” tags, to a comprehensive architecture for creating interoperable knowledge representations. Nor can a lack of tools be blamed.

Creating good metadata requires time and money, but there is little incentive for content creators to expend much of either on the creation of metadata descriptions, because many search engines don’t use them. The metadata that does exist, most of which is created in good faith, is not being used by search engines because they cannot rely on it to provide accurate and faithful descriptions. The missing ingredient is *trust*, without which the Web's resource discovery cake has a bitter taste.

Traditionally, publishers who made fraudulent claims or who published misleading information would end up facing either legal action or bankruptcy, or possibly both. Most nations have extensive legal provisions for dealing with libel, theft of intellectual property, publication of offensive materials, false advertising etc. in the traditional publishing industries.

In fact, there have been a number of lawsuits over disputed uses of Web metadata[49] in recent years, most notably a series of cases involving Playboy Enterprises as both the plaintiff and the defendant. So far, the judgments in these cases appear to have been rational and just.

However, recourse to traditional legal measures is costly and time-consuming, particularly across international boundaries, and the world's judicial systems are ill-equipped to keep up with the pace of technological change in the networked environment.

Ultimately, the architects who are responsible for the ongoing development of the Web are also responsible for enabling the exchange of trust in the Web environment — governments and legal systems do not have the right skills or resources to accomplish this without resorting to restrictive, heavy-handed measures.

Fortunately, various constituencies within the Web developer community are fully aware of this responsibility, as evidenced by the current and emerging technologies to support digital signatures based on public key infrastructures such as VeriSign's DigitalID,[50] the CREN Certificate Authority Service,[51] and the W3C's XML-Signature initiative.[52]

The widespread adoption of digital signatures will ultimately enable metadata descriptions of Web resources to be digitally signed—the Resource Description Framework has been designed from the outset to support digitally-signed descriptions, for example.

Once the authority and authenticity of metadata descriptions can be easily and reliably established, search engine and portal providers will be much more willing and enthusiastic to use them to enhance the resource discovery service they provide for their users.

Some search engine and portal builders may want to produce their own metadata descriptions, since they can then exercise editorial control over the style of description, the indexing techniques and the classification or rating methods. However, if they are not familiar with cataloging, they will rapidly discover that there's a lot more to the art of description than meets the eye!

Museums, libraries and archives, however, have long been expert in the business of capturing, authenticating, and making sense of knowledge through the description of objects and collections, and have been trusted as providers of accurate, impartial information for centuries. In addition to the vast repositories of high-quality knowledge they possess, they are also rich in the less tangible currencies of trust, credibility and authority.

The availability of a robust, secure, and semantically-powerful metadata architecture will not only allow "memory institutions"[53] such as museums, libraries, and archives to more effectively meet their own institutional missions in providing access to their information treasures; it will also empower them to fulfil a role as trusted, nonpartisan guides to the best information the Web has to offer, and thereby act as guardians of our shared cultural record.

> *Presumably man's spirit should be elevated if he can better review his shady past and analyze more completely and objectively his present problems. He has built a civilization so complex that he needs to mechanize his record more fully if he is to push his experiment to its logical conclusion and not merely become bogged down part way there by overtaxing his limited memory. His excursion may be more enjoyable if he can reacquire the privilege of forgetting manifold things he does not need to have immediately to hand, with some assurance that he can find them again if they prove important.*[54]

Notes

[1] Interestingly, metadata supports both the Platonic notion of essence as being separate to an entity, and the Aristotelian notion of the essence being inherent to an entity.

[2] Netcraft Web Server Survey, April 2000: http://www.netcraft.com/survey/Reports/0004/

[3] **Matthew Gray's** Web Growth Summary, http://www.mit.edu/people/mkgray/net/Web-growth-summary.html

[4] Inktomi Webmap: http://www.inktomi.com/Webmap/

[5] *"Web Surpasses One Billion Documents"*, Inktomi Press Release dated 18 January 2000, http://www.inktomi.com/new/press/billion.html

[6] Source: GVU's Tenth WWW User Survey (conducted October 1998) http://www.gvu.gatech.edu/user_surveys/survey-1998-10/graphs/use/q52.htm

[7] **Steve Lawrence & C. Lee Giles,** summary of *"Accessibility of Information on the Web"*, Nature Vol. 400, pp 107-109, 8 July 1999. http://www.nature.com. A summary of the findings is available from http://www.wwwmetrics.com/ .

[8] *"Internet Trends Report 1999",* Issue 4Q99, Alexa Research, 1 February 2000, http://www.alexaresearch.com/top/TrendsReport_4Q99.pdf

[9] Source: Netcraft http://www.netcraft.com/survey/

[10] **Vannevar Bush,** "As We May Think", *The Atlantic Monthly,* July 1945. http://www.w3.org/History/1945/vbush/vbush-all.shtml

[11] **Ted Nelson,** speaking at HyperText 97, the Eighth ACM Conference on Hypertext, Southampton April 6-11th 1997. Source: *Nick Gibbin's Trip Report* on UK Web Focus Conference Centre at http://www.ukoln.ac.uk/web-focus/events/conferences/www6/focus/hypertext97/gibbins/report.html

[12] WWW Virtual Library: http://www.vlib.org

[13] Yahoo!: http://www.yahoo.com/

[14] Lycos: http://www.lycos.com/

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

[16] ADAM: http://adam.ac.uk/

[17] EEVL: http://www.eevl.ac.uk/

[18] Northern Light: http://www.northernlight.com/

[19] **Steve Lawrence & C. Lee Giles,** summary of "*Accessibility of Information on the Web*", Nature Vol. 400, pp 107-109, 8 July 1999. http://www.nature.com. A summary of the findings is available from http://www.wwwmetrics.com/.

[20] **Steve Lawrence & C. Lee Giles,** summary of "*Accessibility of Information on the Web*", Nature Vol. 400, pp 107-109, 8 July 1999. http://www.nature.com. A summary of the findings is available from http://www.wwwmetrics.com/ .

[21] **Danny Sullivan,** *"Numbers, Numbers — But What Do They Mean?"* The Search Engine Report, 3 March 2000, http://www.searchenginewatch.com/sereport/00/03-numbers.html

[22] **Danny Sullivan,** *"Search Engine Sizes"* 11 April 2000, http://www.searchenginewatch.com/reports/sizes.html

[23] Metadata related tools: http://purl.org/dc/tools/index.htm

[24] Cooperative Online Resource Catalog (CORC): http://www.oclc.org/oclc/corc/

[25] **Steve Lawrence & C. Lee Giles,** summary of "*Accessibility of Information on the Web*", Nature Vol. 400, pp 107-109, 8 July 1999. http://www.nature.com. A summary of the findings is available from http://www.wwwmetrics.com/ .

[26] AltaVista meta tags policy: http://doc.altavista.com/adv_search/ast_haw_metatags.shtml

[27] Excite meta tags policy: http://www.excite.com/info/getting_listed/meta_tags/

[28] Northern Light meta tags policy: http://www.northernlight.com/docs/gen_help_faq.html#q12

[29] **Sullivan, Danny,** *"Search Engine Features for Webmasters"* Search Engine Watch, 2 February 2000, http://www.searchenginewatch.com/webmasters/features.html

[30] http://www.inktomi.com/products/portal/search/Partners/goto.html#Spoofing

[31] The official Dublin Core Web site is http://purl.org/dc/ and the official reference definition of the Dublin Core Metadata Element Set, Version 1.1 is at http://purl.org/dc/documents/rec-dces-19990702.htm

[32] Archived at http://www.mailbase.ac.uk/lists/dc-general/

[33] **Lorcan Dempsey & Stu Weibel,** *"The Warwick Metadata Workshop: A Framework for the Deployment of Resource Description"* D-Lib Magazine, July/August 1996, http://www.dlib.org/dlib/july96/07weibel.html

[34] Dublin Core Interoperability Qualifiers: http://www.mailbase.ac.uk/lists/dc-general/2000-04/0010.html

[35] Projects Using Dublin Core Metadata Organized by Geographical Region: http://purl.org/DC/projects/index.htm

[36] Australian Government Locator Service: http://www.naa.gov.au/recordkeeping/gov_online/agls/summary.html

[37] CCTA (UK) Government Information Service: http://open.gov.uk

[38] **Steve Lawrence & C. Lee Giles,** summary of "*Accessibility of Information on the Web*", Nature Vol. 400, pp 107-109, 8 July 1999. http://www.nature.com. A summary of the findings is available from http://www.wwwmetrics.com/.

[39] **Danny Sullivan,** *"The New Meta Tags Are Coming - Or Are They?"* The Search Engine Report, 4 December 1997, http://www.searchenginewatch.com/sereport/97/12-metatags.html

[40] **Stu Weibel,** *"Approval of initial Dublin Core Interoperabiity Qualifiers"* 17 April 2000, announcement posted on the public dc-general mailing list, http://www.mailbase.ac.uk/lists/dc-general/2000-04/0010.html

[41] **Stu Weibel, Renato Ianella & Warwick Cathro,** *"The 4th Dublin Core Metadata Workshop Report"* D-Lib Magazine June 1997, http://www.dlib.org/dlib/june97/metadata/06weibel.html

[42] **Consortium for the Computer Interchange of Museum Information,** *"CIMI Guide to Best Practice: Dublin Core"*, 12 August 1999, http://www.cimi.org/documents/meta_bestprac_final_ann.html

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

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

[45] Platform for Internet Content Selection (PICS): http://www.w3.org/PICS/

[46] **Eric Miller,** *"An Introduction to the Resource Description Framework"*, D-Lib Magazine May 1998, http://www.dlib.org/dlib/may98/miller/05miller.html

[47] RDF Schema Specification 1.0: http://www.w3.org/TR/2000/CR-rdf-schema-20000327/

[48] vCard: http://www.imc.org/pdi/

[49] **Danny Sullivan,** "Meta Tag Lawsuits", Search Engine Watch, http://www.searchenginewatch.com/resources/metasuits.html

[50] Verisign DigitalID: https://www.verisign.com/client/index.html

[51] CREN: http://www.cren.net/ca/index. html

[52] XML-Signature: http://www.w3.org/Signature/, or CREN Certificate Authority Service http://www.cren.net/ca/index.html

[53] **Lorcan Dempsey,** *"Scientific, Industrial, and Cultural Heritage: a shared approach",* Ariadne Issue 22, 12 January 2000, http://www.ariadne.ac.uk/issue22/dempsey/

[54] **Vannevar Bush,** "As We May Think", *The Atlantic Monthly,* July 1945. http://www.w3.org/History/1945/vbush/vbush-all.shtml