

Nobel Symposium (NS 120)

“Virtual Museums and Public Understanding of Science and Culture”

May 26-29, 2002, Stockholm, Sweden

Issues in Structuring Knowledge and Services for Universal Access to Online Science and Culture

by David Bearman and Jennifer Trant, Partners, Archives & Museum Informatics, USA
dbear@archimuse.com

Abstract

A vast quantity of information is available on-line that could dramatically raise the level of awareness of science and culture worldwide, as the means to access online information become truly ubiquitous over the next generation. Unfortunately, as the Web is now configured, it is unlikely that the average person will benefit much from the vastly increased depth of resources, or that even experts will be able to successfully tap much of the knowledge that is, in principle, available. In the most general sense, their problems are first, that the knowledge available online does not reveal itself in the terms that they, as potential users, know and secondly, that many resources are not available in a way that declares their meaning in usable metadata. The challenge to those seeking to make cultural and scientific information more readily available is to construct knowledge models that provide open architectures for digital library contributors with many different perspectives to build growing, yet archivally sound, libraries of knowledge. Developing the methods for constructing knowledge models that are sufficiently forgiving to permit useful aggregation of content, structured by a number of disciplines, yet sufficiently architecturally sound to enable useful computing across resource domains, is critical if the individual silos of content that we have constructed in the print – and now online – world are to truly become the digital museums and libraries of the future. Some tentative steps have been taken in the definition of the “semantic web”, the construction of simple ontologies, and the articulation of low level protocols for resource negotiation such as OIA. This paper will discuss the underlying issues, explore some progress to date, and present a model of how these and future contributions to user aware content navigation might be designed.

Issues in Structuring Knowledge and Services for Universal Access to Online Science and Culture

by David Bearman and Jennifer Trant, Partners, Archives & Museum Informatics, USA
dbear@archimuse.com

Introduction

Although the World Wide Web is less than a decade old, it is clearly the first implemented information technology that will fundamentally transform cultural institutions. The almost incredible, widespread availability of WWW access, which provides a bridge to vast new publics, provides the imperative for archives, libraries and museums to manifest themselves on the web. The ease with which this is done – through an extremely simple markup language (html), free browser software (with its minimal functionality), and the low start-up costs – add to our motivations for putting content on the web. Each year since the release of multiple platform versions of the Mosaic “browser” in the fall of 1993 (see <http://www.w3.org/history.html>), more museums have created a web presence and those already on the web have added more content to their web sites, until today it is safe to say without exaggeration, that all museums must have a presence on the web, and all are under tremendous pressure to document their collections in a public web space.

Yet, the success of the web has been limited to providing access to information about visiting the physical institutions, whose presence a web site announces, and to documenting the holdings of an institution for those who either know the institution exists, or find information on the site by character string searching on a search engine. The body of material in cultural institutions that has been made public on the web has not become, collectively, a cultural resource, and cannot become seamlessly available simply by the further adoption of existing methods. Indeed, as institutions make more of their holdings available on the web, they are increasingly using databases as back ends rather than by placing more simple html pages on their sites, thus limiting, rather than increasing, access to their holdings through the existing generation of search engines that rely on “crawling” web content, rather than harvesting web metadata to build indexes. The next steps in developing the web as a space for presenting cultural knowledge will require greater technological expertise to take advantage of more sophisticated protocols and adoption of standards for digital knowledge

representations. In particular, it requires that museums pay active attention to emerging approaches for metadata declaration and utilization.ⁱ

Cultural repositories, knowledge and metadata

Many different institutions collect documentation of cultural heritage, each with its own tradition of documentation and discrete audiencesⁱⁱ. These institutions are further fragmented by housing scholars from different disciplines, whose distinctive views of reality are their stock in tradeⁱⁱⁱ. In addition, the information which these institutions assemble about the objects in their care is constructed in work processes that reflect different departmental methods, and often (perhaps even usually) result in construction of data that is incompatible with that needed elsewhere in the institution, as well as incomplete from the perspective of other in-house users.^{iv} End users of cultural information themselves, come from every possible background, have every plausible interest level, and bring with them all types of prior knowledge and ignorance, and differing levels of sophistication. We have described these issues in numerous papers in the past, suggesting methods for negotiation between users' views, scholarly languages, and repository control methods in, and proposing an approach to, metadata dialogues. Even if all these recommendations were fully implemented, network wide interoperability of cultural heritage information would depend on adoption of a range of architectural standards. These fundamental constructs for interoperable cultural metadata are the subject of this paper.

Information architecture standards apply at numerous levels in information systems; here we will confine ourselves largely to a discussion of knowledge representation architectures, imagining for these purposes that data interchange and communication protocols are fully interoperating (as they temporarily are close to doing in web-based application environments). The one exception to focusing entirely on knowledge models will be a brief side journey into the world of metadata harvesting which is the current technology for building the indexes on which some Web retrieval is based. This is analogous to saying that we are concerned with standards for design of hospitals, which will focus on the nature of operating rooms, patient rooms, and nursing stations but will in no case discuss stresses on the buildings, materials used in their construction, or the installation of electrical, plumbing and other systems that will support the activities that take place in any modern building. Rather, we are examining only those elements of the architecture that are: 1) specific to cultural heritage data and

2) for which it is necessary for our community to adopt standard approaches, in order to reap the benefits of linking cultural content held by, documented by, and made available from different sources.

Dimensions of culture

Cultural heritage, as distinguished from natural heritage, consists of objects created by, or given meaning by, human activity. A collection of stones, whether made for scientific purposes, to worship or build with, or arranged by Richard Long as a work of art, are part of the cultural heritage of mankind; while stones in a rockslide, however aesthetic when perceived, are not part of culture (though a photograph depicting that aesthetic appeal will be). In other words, things become part of cultural heritage by virtue of a human action, situated not just in time and place but, indeed, in a particular human cultural context.

Cultural heritage is also distinguished from natural heritage by virtue of embodying an idea. A collection of stones, or a landscaped garden, each consists of natural objects brought together in a way that is fashioned by a human mind. We need not necessarily understand (to say nothing of appreciate or approve of) the idea in order to recognize its presence. But its presence is fundamental to its character as part of the cultural heritage, since it is, in some sense, an expression of the concrete culture in which, and by which, it was created. Finally, cultural heritage, having been made by a human agent, is the product of directed energy – someone took a specific action, which could have been a logical action of imagining the unity of several things, expressing that concept and thereby “bringing” them virtually together, or might be a physical action. To understand cultural activity, we must understand who (individually and collectively) created the cultural object at hand; when, where, and if possible, why. Cultural objects have a purpose or a use, though it is not always possible for us to know what that was for a found object, especially one from the distant past.

Cultural information retrieval, therefore, consists in being able to answer questions of who, where, why, how, when; and what was created, collected, discovered, described, published, and exhibited; and to explore the ideas this reflected or the intellectual assertions made (explicitly or implicitly). As in any information seeking, cultural information seeking begins with limited knowledge (one or several facts), and attempts to discover additional information in a (not necessarily explicitly specified) discourse model, with only some slots occupied. Information seeking that depends entirely on

human intermediation, may simply locate sources with more information about the facts we already know, and leave it to that human to read and ‘fill out their picture’. Information seeking that uses machine intermediaries to organize its findings proceeds along known informational axis, to elaborate the discourse model and populate its unknown parts and organize its results.

Documentation of cultural objects and information retrieval

The framework for this activity was articulated over fifty years ago, at the birth of the modern computer. The period immediately following World War II was an intensely creative time for information theory and technologies. The period spawned the electronic computer and operations research. It gave birth to Vannevar Bush’s famous essay *As We May Think*^v which described linking mechanisms of the kind realized in the WWW and Ranganathan’s not so famous, but equally important, high level theories of knowledge organization^{vi}. This complex analysis of the foundations of human classification, which identified six cardinal categories – time, space, personality, matter, energy and idea – has inspired efforts to classify and index human knowledge ever since.

In the 1950s, Ranganathan’s theories of classification influenced cataloging at the British Museum. Teams of young intellectuals applied individual terms to book cataloging in PRECIS, a “faceted” classification systems, constructed without “pre-coordination” among terms. The absence of “pre-coordinated” subject headings made possible retrieval in complex categories that were not pre-assigned by the indexers. Individual terms, such as “Germany,” might be the value assigned to the place of birth of the author, the place of creation of a work, the place of discovery of a found artifact, the subject of a book, etc., etc. But without information retrieval technologies powered by computers, the use of faceted classification gave way to easier to use “subject headings” with pre-coordinated terms such as “Workers rights, Germany, 1919-1936”. By the 1990s, when relational databases became the norm for recording documentation, though, the RDBMS represented a retrieval environment well suited to using faceted classification.

Two exceptionally simplistic mechanisms for finding information on the WWW limited retrieval through the late-1990s. The basic method was a “pointer” or “link” made by an author. It simply said: “this word or digital object is related (in some unspecified way) to this other word or digital object”. The second mechanism was equally primitive – it was character string searching from search engines that “crawled” web sites and built word

indexes. Neither subject headings with pre-coordinated concepts, nor facets with meaning attached to words, were used in search engines. Links, when written by hand, were limited in number by the energy of the author, and even when made by databases, did not carry declaration of term meaning, as do even the simplest of database field labels.

The functionality supported by web browsers that only recognized HTML and search engines that used character strings out of context are too limited to support a viable information environment. Advocates of the use of XML as a mark-up language were able, by 2000, to convince those making browser software to support XML. Almost immediately, many of these authoring web pages began to write XML, and the benefits of being able to use a semantically meaningful markup were readily apparent to those with databases being published to the web, as it permits the naming of elements of information by their meaning and function.

But because individual institutions do not construct their relational databases in using shared conventions, let alone in fifth normal form, and internal data structures are influenced by business process. Rather than being fully normalized, data is stored in forms that pragmatically relate to their functional ends. Even if all table structures were the same, naming conventions are not uniform, and there needs to be a way, even if data is normalized to relate the meaning of structured data in one database to that in other databases. The functional answer in the World Wide Web is to declare the schema of one's databases, along with definitions of elements, in a "namespace" on the Web.^{vii} Machines visiting that namespace will discover the meaning defined for given elements within the disciplinary domain that owns the namespace. If a network of namespaces with declared schema exist, and systems are developed to search them for meaning in conjunction with reporting data values found by character string searches, then intelligent retrieval could take place across the WWW.

Semantic understanding, the web and knowledge spaces

The WWW in 2002 is an unsatisfactory information environment for making meaning, because its overly simplistic system for presenting information and linking related pieces of information fails to characterize either the nature of the relation between two things or the nature of the things thus linked. Axiomatically, its present methods only allow pointing to explicit items already known to the author of one document, hence making no connection

to documents that might be created later, or be unknown to the original author.

A more robust information space was initially imagined by Tim Bernier-Lee and by the many hypertext/hypermedia innovators who came before him. It requires that the objects on the web be able to declare their nature and, that these declarations permit the generation of links that are explicitly typed, and hence can be meaningfully sorted. Instead of retrieving all matches on the character string “Edgar Degas” (including every person with that name who has ever been referenced on the web), and then all links known to authors of other works in which Degas is referenced, I should be able to find just the “Edgar Degas” I want (the 19th and early 20th century banker and artist), and limit my retrieval to references that reflect his role as an art collector. Further, I should be able to do this not only from references which are self-consciously made to that role, but by inference from the provenance of works of art or from records of art auction sales; and I should be able to make connections that were not known to the authors of the data objects I find. If we assume that soon most content on the WWW will be expressed in XML rather than HTML, and that therefore it will be stating what kind of data it is, rather than what size type font it should be displayed in (with display rules as a separate set of instructions), we have the beginning of a system of this sort.

However, for information systems (including the web) to be able to navigate to relevant documents by “knowing what they are about”, they need not just the use of a content-based markup language (XML), but also a means to relate the XML labels created by authors. These labels derive their meaning through referencing (consciously and unconsciously) different knowledge models and vocabularies used in different intellectual communities. In addition to numerous discipline-based knowledge structures (biological taxonomies, thesauri of artistic terminology, etc.), museums have created semantic models of the information they manage, such as the *Categories for Description of Works of Art* (CDWA) which focus on the relations in the life-cycle of collected objects from a scholarly perspective, SPECTRUM which focuses on the museum object and museum business processes, and the CIDOC-CRM which emphasizes the historical contextualization of objects.^{viii} In a networked environment, the value of these knowledge models will be determined more by their ability to connect to other knowledge representations by other groups, than by their ability to represent all subtle aspects of terms used for indexing aspects of cultural heritage.

Explicitly declared ontologies, whose relations were mapped in a set of agreed primitives could provide the underlying information required for correct automatic data typing, and declaration of meanings between information from discrete disciplines, and domains of practice within cultural repositories. This would enable us to overcome the non-essential differences between field or element labels, and to verify the synonymy of content between such labeled fields and elements.

For example, in an art museum, a “Creator” is a person, or organization, or culture that makes a work of art. In libraries, an “Author” is a person or organization that writes a literary work. To Ranganathan, both are Personality with Idea bringing Energy to make Matter in Time and Space. In libraries, the “time” and “place” of record are those of another act, “publishing”, which involves another personality (the “Publisher”). The analogous act in the life of a work of art, “first exhibition or publication”, has the same legal implication for copyright, but a very different place in the definitive documentation of the work. The action of the person in each of these cases is central both to understanding the relation, and to grasping its significance within the documentation context of the type of institution that recorded it.^{ix}

While the labels are different in the case of “creator” and “author”, the formal declaration of their meaning would reveal that they are synonymous. While the label may be the same “date of publication” and “place of publication” in the two cases, and the meaning is the same in a formal sense, the significance of the fact in the two domains is very different, and in the world of art, the primary significance of “publication” (e.g. to make public) is achieved by exhibition of a work of art. It is left to the user to move ahead with her analysis, informed by this distinction.

rdf and the semantic web

The Semantic Web Initiative of World Wide Web Consortium (w3c.org) is currently trying to build the consensus and toolsets required to reform the web from a simplistic linking mechanism to a robust information space.^x The Resource Description Framework (rdf) working group is elaborating a means for formal knowledge models (terms which could then be used in XML labels) to be declared in “Name Spaces” (web resources where the web retrieval tools will need to go to find the meaning of the data they are seeking). The Resource Description Framework (rdf) is something between

a dialect of XML and an xml application that currently serves as the vehicle by which formal ontologies from one domain are related to those from other domains.^{xi} For ease of visualization, knowledge models are expressed as ‘nodes’ (object types) and ‘arcs’ (relationship types) in rdf diagrams, though in the language itself, the expression is only that of a non-procedural code (see RDF Primer).

The semantic web initiative is designed to move from rdf to its exploitation as a vehicle for linking content based on meaning. Although in its early days, the initiative has developed some tools and projects that are demonstrating the promise of a semantically aware network.

Before the full set of tools that would be required to realize the semantic web, including interoperable knowledge models are deployed, communities of interest can build web resource sets that share metadata or have maps between their ontologies and act as mini-semantic webs. These communities, which will need to harvest the metadata of their members and publish it with integrated retrieval tools, are rapidly adopting OAI-MHP, the Open Archives Initiative Metadata Harvesting Protocol, to realize their aims.^{xii} The OAI-MHP is a simple protocol enabling a metadata creating institution to place its metadata “outside the front door” in a recognized container, so that a harvesting agent serving the metadata collation service can come by on a regular or irregular interval and collect it. When the metadata is brought back to the collecting site, the rules for understanding its content have been agreed within the community being served, so full semantic web based understanding of meaning is not required.

A similar protocol, RSS or RDF Site Summary, was designed to publish small quantities of content on a continuous basis to metadata harvesters.^{xiii} It has been used for press releases and announcements that feed into news wire services, but could also be deployed to publish new content resources of cultural institutions.

The differences between the business processes these protocols were designed to support has led one to be more push oriented (RSS), while the other is more pull oriented (OAI), but each is designed to supply data to a trusted agent that is harvesting metadata from known sites.

Granularity, modularity, and sustainability

Cultural institutions have complex stories to tell, and are tempted to tell them in narratives, with much multimedia accompaniment and careful design. They tend to handcraft the intellectual content they publish on the web. But failure to attend to the engineering of their information content - the size and metadocumentation of informational chunks and the interfaces between them, will result in creation of unmanageable presentations. These presentations will be hard to update, impossible to migrate to new platforms, and unable to recombine in other informational context.

However special the story that the cultural institution has to tell, it will be better if it can be integrated with other data from other institutions. However perfectly it is told, there will be updates and changes, and there will be a need to move the information to new display environments and software. Huge investments will be made over the next decades in crafting information to present over the Internet; only interchangeable information parts will be sound investments.

Too often very large and complex informational objects, such a historical tours or games, are authored in the way that books are written – by individuals, as integral creations, entirely dependent on the software environments in which they will initially be manifested – only for the institution to discover that such efforts do not scale, cannot be migrated, and will not connect to other information and activities outside themselves. But it is already too late.

If functionality is to be extended over time, new functions must address predictable information objects, and interact with them in predictable ways. A high level function, such as a “quiz” must be able to use any information object in the environment as an “answer” and as the framework for a “question”. A timeline or map must be able to display any object in the information environment in a predictable way, if it contains place or date metadata. After a web site is launched, new methods may be added - such as glossary functions, tables of contents, or annotation functions – and these must behave predictably with respect to existing content.

All this means that basic principles of information engineering must be respected from the outset in the construction of cultural information utilities,

not the least because cultural knowledge bases will be built up over many years or decades.

Implications for cultural heritage

One of the attractions of large bodies of cultural heritage materials in digital form is their potential use in a wide variety of contexts. Users from many different perspectives and points of view can move objects from the presentation context of the institution and create their own meanings and narratives.^{xiv} Museums themselves can also take advantage of this power to extract and deliver the knowledge represented in their collections in a manner that meets users needs^{xv}

The promise of the web is to virtually unite and re-unite digital objects in contextual information spaces.^{xvi} However, our current web practices stand squarely in the way of achieving those goals. Flash-built, exhibition-focused web features that present the equivalent of a closed CD-ROM on the web, may have a sound pedagogical and communications goals. But as they are now implemented, the digital objects in these expensive and labor-intensive resources are rarely reusable, and rarely locatable outside their local navigation. They stand in the face of the developing perspectives of museums as sources of information for society.^{xvii}

If content from a multiplicity of cultural repositories of different types is to be made accessible through searching, we need to move beyond authored links and character string based indexing as mechanisms for finding relevant objects in a networked environment. XML, RDF, and the semantic web provide us with the facilities to do this, and in combination with some short-term metadata harvesting utilities such as OAI and RSS, we can move towards realizing the integration functions of true semantic linking, without having to wait for all the facilities of the true semantic web to be implemented. What is critical is that we begin to work together to surface the true impediments local practice has on collective knowledge construction.

Scholarly information sharing combined with tools such as are present in collaboration environments, will make the dream of the fully distributed digital library meaningful. Casual users of the WWW will be able to improve the accuracy of both their searching and linking by having self-knowledge associated with the objects at the end of each link and explicit relationship types associated with links. Instead of being the functional

equivalent of a gigantic term occurrence list (which returns thousands times the number of hits that anyone could possibly follow), the web could become more the functional equivalent of a gigantic classified index, and the information that it contains could be more accessible as a result.

And, if care has been taken in construction of the components, the cultural information spaces we construct will be able to survive over considerable time and build their value as information objects are contributed from many different sources.

While these architectural solutions are not in themselves sufficient to ensure a robust and universally accessible knowledge resource on the Web^{xviii}, we certainly will not achieve one without them.

REFERENCES

ⁱ Note that the most important standards developments change quickly. See, for example, David Bearman, "Cultural Heritage Information Standards in a Networked World", in *Prometheus: New Technologies in Culture* (Athens, Lambrakis Research Foundation) p.39-52; revised in *Archives and Museum Informatics*, vol.8#2 p.91-107 or David Bearman, "Standards for Networked Cultural Heritage", *Archives and Museum Informatics*, vol.9 #3, p.279-307

ⁱⁱ Bearman, David and Jennifer Trant, "Unifying Cultural Memory," with David Bearman, *Information Landscapes for a Learning Society, Networking and the Future of Libraries* 3, 1998. And presentation at UK Office of Library Networking Conference, July 1998. Paper available at <http://www.archimuse.com/papers/ukoln98paper/index.html> Presentation available at www.archimuse.com/papers/ukoln98/uklon98.pdf

ⁱⁱⁱ Bearman, David and Jennifer Trant, "Beyond Simple Resource Discovery: A framework for metadata declarations of disciplinary schema to support research in heterogeneous collections," *International Symposium on Information Technology in Museums - Integrated Applications*, Bonn, Germany, December 1-2, 1997. (published in German).

^{iv} David Bearman, "Data Relationships in the Documentation of Cultural Objects", *Visual Resources*, v.11, p.295-306

^v Bush, Vannevar, "As We May Think", *The Atlantic Monthly*, July, 1945, Volume 176, No. 1; pages 101-108. Available online at <http://www.theatlantic.com/unbound/flashbks/computer/bushf.htm>

^{vi} [S.R. Ranganathan's works see Ranganathan, S.R. Colon Classification, Basic Classification. 6th ed. New York: Asia Publishing House, 1963.; for PRECIS, see Austin, Derek. PRECIS: A](#)

[Manual of Concept Analysis and Subject Indexing](#). London: The British Library Bibliographic Services Division, 1984.

^{vii} see <http://www.w3c.org/2001/sw/>

^{viii} [for the CIDOC Conceptual Reference Model](#), see <http://cidoc.ics.forth.gr/>; [for the Categories for Description of Works of Art](#), see <http://www.getty.edu/research/institute/standards/cdwa/>; [for SPECTRUM](#), see <http://www.mda.org.uk/spectrum.htm>.

^{ix} to understand the central role of actions in knowledge models of culture, see Carl Lagoze, Jane Hunter and others. ABC model. Papers include from 2001 DC Tokyo conference <http://www.cs.cornell.edu/lagoze/papers/DC2001.pdf>.; A new version, which will be published through the Journal of Digital Information special issue on metadata, is available at http://metadata.net/harmony/JODI_Final.pdf.

^x Semantic Web Initiative, W3C. <http://www.w3c.org/2001/sw/>

^{xi} RDF Primer 2002. W3C Working Draft 26 April 2002 Edited by Frank Manola, and Eric Miller, This version: <http://www.w3.org/TR/2002/WD-rdf-primer-20020426/>

^{xii} Lagoze/Van de Sompel 2001. Lagoze, Carl and Herbert Van de Sompel, The Open Archives Initiative: building a low-barrier interoperability framework, ACM/IEE Joint Conference on Digital Libraries (JCDL), 2001, available at <http://www.w3c.org/2001/sw/>

^{xiii} RSS (RDF Site Summary) RSS 1.0 specification released on 2000-12-06 <http://purl.org/rss/1.0/>

^{xiv} Peter Walsh, "The Unassailable Voice", Museums and the Web 1997, Los Angeles, Archives & Museum Informatic <http://www.archimuse.com/mw91/speak/walsh.html>

^{xv} Sledge 1995. Looking for Mr. Rococo. ICHIM 1995. AAT hierarchies/faceted classification and searching

^{xvi} David Bearman, "Information Strategies and Structures for Electronic Museums", Information: The Hidden Resource, Museums and the Internet. Proceedings of the Seventh International Conference of the MDA, 1995 ed. by Anne Fahy and Dr. Wendy Sudbury (Cambridge, UK, Museum Documentation Association, 1995) p.5-22 ; also David Bearman, "Museum Strategies for Success on the Internet", Museum Collections and the Information Superhighway (London, Science Museum, 1995) p.15-27; also published in Spectra, vol.22#4 p.18-24 <http://www.nmsi.ac.uk/infosh/bearman.htm>

^{xvii} MacDonald, George F. (1991). "The museum as information utility." Museum Management and Curatorship 10: 305-311

^{xviii} David Bearman and Jennifer Trant, "Economic, Social, Technical Models for Digital Libraries of Primary Resources," New Review of Information Networking, #4, 1998, pp 71-91. Paper available at <http://www.archimuse.com/publishing/amico/>