Kim H. Veltman

# Towards a Global Vision of Meta-Data: A Digital Reference 'Room'[1]

Keynote at 2nd International Conference in 1997.
Published: *Proceedings of the 2nd International Conference. Cultural Heritage Networks Hypermedia*, Milan: Politecnico di Milano (Arti Grafiche Stefano Pinelli), 1999, pp. 199-209.

**Abstract**[1]

G8 pilot project 5 is devoted to *Multimedia Access to World Cultural Heritage*. The European Commission's Memorandum of Understanding concerns *Multimedia Access to Europe's Cultural Heritage*. These two projects began quite separately because they were very different in scope, one global, the other regional. In the interests of efficiency, discussions in the past year have turned to greater co-operation between the global efforts of G8 and those of the European Commission. At the first Milan Congress on Cultural Heritage (September 1996), the author outlined some subjunctive possibilities for such a co-operative framework between G8 and the Commission.[1]

More recently the European Commission hosted a meeting in Brussels (June 1997) to discuss the future of its Memorandum of the Understanding. Several speakers noted the desirability of closer co-operation between G8 and the MOU. Among them was the author, who outlined nine needs and challenges which could further this goal: 1) an open distributed processing environment (such as TINA); 2) open demo rooms serving as prototype service centres; 3) toolboxes; 4) strategies for digitizing museum content; 5) library connections ; 6) applications to education; 7) a meta-data reference "room"; 8) search, access and navigation interfaces (such as SUMS and SUMMA); and 9) a self-learning environment.[1]

This paper summarises some recent developments leading to closer collaboration between the G8 projects and those of the European Commission. Rather than attempting to explore in detail all the challenges entailed it focusses on one specific need (number 7 in the above list): namely, a meta-data reference room. It begins by examining competing models of centralised versus distributed contents. Some interim measures and recent developments in meta-data are reviewed. The limitations of number crunching as an alternative are addressed and the advantages of a centralised meta-data approach are outlined.

## 0) Background

G8 pilot project 5 is devoted to *Multimedia Access to World Cultural Heritage*. It began with a narrow focus on the leading industrial countries, which was then greatly expanded through the Information Society and Developing Countries (ISAD) conference in Midrand (May 1996) by including 42 countries from all over the world. The European Commission's Memorandum of Understanding (MOU) concerns *Multimedia Access to Europe's Cultural Heritage*. These two projects began quite separately because they were very different in scope, one global, the other regional. In the interests of efficiency, discussions in the past year have turned to greater co-operation between the global efforts of G8 and those of the European Commission. At the first Milan Congress on Cultural Heritage (September 1996), the author outlined some subjunctive possibilities for such a co-operative framework between G8 and the Commission.[2]

More recently the European Commission hosted a meeting in Brussels (June 1997) to discuss the future of its Memorandum of the Understanding. Several speakers noted the desirability of closer co-operation between G8 and the MOU. Among them was the author, who outlined nine needs and challenges which could further this goal: 1) an open distributed processing environment (such as TINA); 2) open demo rooms serving as prototype service centres; 3) toolboxes; 4) strategies for digitizing museum content; 5) library connections ; 6) applications to education; 7) a meta-data reference "room"; 8) search, access and navigation interfaces (such as SUMS and SUMMA); and 9) a self-learning environment.[3]

A subsequent smaller meeting, sponsored by the Commission, under the auspices of Arenotech and the French Ministry of Culture in Paris (July 1997), led to a draft statement concerning Common Goals of G7 Pilot Project 5 and MOU which was submitted to the steering committee for approval on 23 September 1997 and approved. This draft calls for " scenarios for use of new technologies in public sector areas such as education and commercial applications" and foresees two steps: 1) increased co-ordination with other R&D projects of the European Commission and 2) open demo rooms – as outlined in need two above-- which can serve as information centres and prototype service centres for museums. It is proposed that: In the first instance these pilot centres will be based in representative cities of the G8 and connected at an ATM level using emerging global standards. These pilot centres will then be connected with practical trials being initiated by the telephone companies to demonstrate applicability with real users. A next phase will connect these centres directly with museums and other public institutions.

To make the connectivity between the demo rooms a reality requires co-operation between major telephone companies which, in other contexts, may be in competition with one another. Fortunately a few projects within the ACTS programme have laid the foundations for the connectivity required. For example, the MUSIST project links Italy's Telecom Italia and Italtel with Germany's Deutsche Telekom. The VISEUM project is working at links between Germany and Canada: i.e. Deutsche Telekom, Teleglobe and Bell Canada. These connections thus provide stepping stones for a proposed initial link

between Rome and Toronto and/or Ottawa. As the museums section of the Trans European Networks (TEN) programme, MOSAIC is an obvious choice for co-ordinating such efforts.

As a first concrete step it was suggested that there might be two demo rooms linking Italy (Rome) and Canada by an Asynchronous Transfer Mode (ATM) connection. These rooms would be linked to Asynchronous Digital Subscriber Line (ADSL) trials of two telephone companies, namely, Telecom Italia and Bell (Medialinx). A preliminary meeting on 17 September, 1997, hosted by the Canadian Embassy, focussed specifically on issues of connectivity in linking Italy and Canada and on commercial dimensions.
The Ministero dei Beni Culturali is exploring these possibilities and is examining related issues of cultural content, a more formal governmental framework and an infrastructure for handling of rights management (copyright, smart cards etc.).

A second phase will add three more centres in Berlin, Paris and London. To this end, projects such AQUARELLE and VASARI offer obvious starting points, with others such as MENHIR and the Canadian project AMUSE as further partners. A third phase will expand the range of the centres to include the other G8 cities, namely, Washington, Tokyo and Moscow. As these initiatives unfold it is foreseen that both the Ministero dei Beni Culturali and the Commission will include a number of other projects within this joint framework, those with clearly international implications still under the aegis of G8, while others remain under the aegis of the MOU. While the co-operation will bring a sharing of resources and goals, the two organisations will, nonetheless, continue in an independent and inter-dependent fashion.

As was noted in the Brussels meeting, such proto-type service centres represent but one of nine challenges which need to be addressed as the Memorandum of Understanding and the Commission in general move into a next phase. Rather than attempt to explore all of these, this paper focusses specifically on the seventh of the above challenges, namely, the need for a meta-data reference room. This idea is one of the most wide reaching and will require considerable co-ordination among cultural institutions of many kinds. Taken together with the other challenges it offers a long term goal for the Commission's MOU which is consonant with the broader aims of G8 and at the same time answers a recent call for a coherent cultural policy for Europe by the Council of Ministers of Culture (30 June 1997). Only a carefully planned long term solution will protect us from the fate of doomsayers who predict that we will drown in excess information as in a second flood, rather than benefiting from the positive visions of an information society.

## 1) Introduction

Models for knowledge organization have tended towards two extremes of a spectrum. At one extreme, traditionally, there has been a vision of centralized contents. At another extreme, more recently, the rise of the Internet has favoured a model whereby everything is distributed. A number of recent developments in meta-data reflect such a model. This paper calls for a new intermediary model which links centralized meta-data with distributed contents.

## 2) Centralised versus Distributed Contents

At least since the time of the library at Alexandria there has been a dream of collecting the whole of human knowledge within a single enormous library. Panizzi revived this idea in the nineteenth century with the creation of the British Museum, which soon became a model for national libraries throughout the world. While such libraries had many advantages as major repositories of knowledge, they suffer from one major flaw. The rate of new books increases more rapidly than the spaces needed to house them. Recent experiences with the new versions of the British Library and the Bibliothèque de la France confirm this. Both buildings will be too small to house all aspects of their collections even before they are fully operational. Hence while the quest to have everything under a single roof is noble, it is simply not practical.

As an interim measure major libraries moved their extra books to nearby buildings or elsewhere. In the case of the British Library by the 1970's some of these depots were so far away that it took as long as a week for a book to be moved from the remote location to a reader's desk in the main library.

The advent of the Internet seemed to promise a solution to such problems. In theory one could digitise titles and contents of books on any site and connect them via a network, thus leading to a completely distributed system. Some factors combined to cloud this picture. First, in terms of content providers, in addition to libraries and professional institutions, many individuals without training in information management placed their materials on the Internet. Second, on the user side, many of those searching for information had no clear ideas about how to ask questions. As a result the distributed model typically produced enormous amounts of general responses but seldom precise answers.

## 3) Interim Measures

To deal with the chaotic state of information distributed throughout the net, a number of initiatives are underway. These include: i) domain names; ii) mime types, iii) site mapping, iv) content mapping, v) abstracts, and vi) rating systems and agents.

### i) Domain Names, URL, URN, and URI

Present search tools typically rely on the domain names or the Uniform Resource Locators (URLs) to find things. The Internet Society has formed a consortium, which will greatly expand the number of high-level domain names (e.g. com, gov, edu) such that these can be linked with country codes to provide search strategies by topic and region. Meanwhile, the W3 Consortium is working on Universal Resource Names (URN) and Universal Resource Identifiers (URI) which will complement existing URLs and provide more subtle versions of the above. They are also working on meta-data tags to be added to the next generation of Hypertext Markup Language (HTML), called dynamic HTML,

and a new subset of Standardized Graphic Markup Language (SGML), called Extensible Markup Language (XML).

ii) Mime types

Those at the forefront (e.g. Larry Masinter, Xerox) of the next generation of Hypertext Transfer Protocol (HTTP), are working on tags for different multimedia (MIME) types, which will allow one to identify whether the message contains audio, video, text etc. This will add a further parameter to one's search criteria such that one can discover which URLs (and URNs) contain video before scanning through all the contents of a site.

iii) Site Mapping

A new technique invented at Georgia State and now being developed at Xerox PARC allows one to visualize the structure of a web site, i.e. see how many layers the site has, which pages are cross-referenced to which others such that one can recognize crucial points in the structure. A similar idea is evident in Apple's *Hot Sauce*. Microsoft is also working on a similar feature.

iv) Content Mapping

Major research labs such as Lucent (the former Bell Labs now linked with Philips), are working on defining the parameters of databases in terms of basic questions. This is being done on an inductive basis using sources with databases which are considered reliable. A combination of manual techniques and agent technologies are used to determine the parameters of the contents, such that one can know, for instance, whether the database contains video, and if so, which artists and from which year to which year.

v) Abstracts

Microsoft Windows among its tools has an *Autosummarize* function. Companies such as Apple are creating software (Vespa), which allows for automatic summaries of a page, a paragraph or a single sentence. Hence one will able to check these summaries first instead of having to search through the entire documents at the outset.

vi) Rating Systems

One of the problems with the Internet at present is that it is often very difficult to establish the quality or reliability of a given site. The W3 Consortium is developing a Protocol for Internet Content Selection (PICS) which will allow rating of sites. The are also developing a concept of digital signatures which will introduce the equivalent of a peer rating initiative for web sites.

vii) Content Negotiation

A number of models are being developed for content negotiation such as that of the TINA Consortium. This includes rights management, licensing fees and secure transactions. Others (e.g. IBM, Fraunhofer) are developing visible and invisible watermarking methods for copyright protection. These will increase the precision with which materials on the web can be handled.

viii) Agents

A great deal of research is being done on agents. Recently, Leonardo Chiariglione (CSELT), one of the key individuals responsible for the MPEG 4 and MPEG 7 standards, has initiated the Foundation for Intelligent Physical Agents (FIPA), which promises to be an international meeting ground for developments in this field. Many thinkers (e.g. Negroponte, Laurel) assume that agents will serve primarily as electronic butlers producing, as it were, tailor made selections of newspapers and other sources in keeping with our particular interests.

**4) Recent Developments in Meta-Data**

More recently there has been increasing attention to the term, *meta-data*, which is often used as if it were a panacea, frequently by persons who have little idea precisely what the term means. In its simplest form, meta-data is data about data, a way of describing the containers or the general headings of the contents rather than a list of all the contents as such. Some of the interim measures listed above could be seen as efforts in this direction.

More specifically there are a number of serious efforts within the library world. The Library of Congress is heading work on the Z.39.50 protocol, designed to give inter-platform accessibility to library materials. This is being adopted by the Gateway to European National Libraries (GABRIEL) and the Computer Interchange of Museum Information (CIMI) group.

A number of meta-data projects are underway. For instance, the Defence Advanced Projects Agency (DARPA), in co-ordination with the National Science Foundation (NSF), NASA and Stanford University are working on meta-data in conjunction with digital library projects. DARPA itself is working on Knowledge Query Markup Language (KQML) and Knowledge Interchange Format (KIF). The Online Computer Library Centre (OCLC) has led a series of developments in library meta-data (Dublin Core, Warwick Framework). In essence these projects have chosen a core subset of the fields in library catalogues and propose to use these as meta-data headers for access to the complete records. An alternative strategy is being developed by the Institut für Terminologie und angewandte Wissensforschung (Berlin). They foresee translating the various library schemes such as the Anglo-American Cataloging Rules and the Preussusiche Regeln into templates using Standardized General Markup Language (SGML). This approach will allow interoperability among the different systems without the need for duplicate information through meta-data headers.

## 5) Number Crunching or the Limits of Brute Force

Each of the above initiatives is laudable and useful in its own right. They will all contribute to easier access to materials and to efficiencies in that users can sometimes rely on overviews, excerpts and abbreviations rather than needing to consult the whole database in the first instance. But all of these remain short term solutions in that they do not solve questions of how one determines variant names, places etc. Meanwhile some members of the computer industry continue to argue that the troubles surrounding the Internet are merely a passing phase; that although connectivity and search engines and were initially too slow, as soon as these hindrances are resolved, all will be well. While rhetorically attractive, such reassurances are not convincing for several reasons.

First, there is a simple question of efficiency. A local database may have only local names. The name for which one is searching may only exist in specialized databases. Going to a typical database does not guarantee finding the name. Going to all databases just to identify the name is highly inefficient The same problem applies to subjects, places, different chronological systems etc. It applies also to different media. If I am looking for one particular medium such as video then it makes sense to look at sites with video, but not all sites in the world. Searches to find anything, anywhere, anytime should not require searching everything, everywhere, every time. As the number of on-line materials grows apace with the number of users, the inefficiencies of this approach will become ever greater.

A second reason is more fundamental. Even if computer power were infinite and one could search everything, everywhere, every time, this would not solve the problems at hand. Names of persons and places typically have variants. If I search for only one variant the computer can only give me material on that variant. If, for example, I ask for information about the city of *Liège*, the computer can at best be expected to find all references to *Liège*. It has no way of knowing that this city is called *Luik* in Dutch, *Lüttich* in German and *Liegi* in Italian. This is theoretically merely a matter of translation. But if every place name has to be run through each of the 6,500 languages of the world each time a query is made, it would be an enormous burden to the system. And it would still not solve the problem of historical variants. For instance, *Florence* is known an *Firenze* in modern Italian but was typically written as *Fiorenza* in the Renaissance. It would be much more practical if every advanced search for a place name went through a standard list of names with all accepted variants. Such a standardised list acting as a universal gazetteer needs to be centralised.

The same basic principle applies to variant names of authors, artists etc. If I have only one standard name, the computer finds that name but it can never hope to find all the variants. Sometimes these variants will be somewhat predictable. Hence the name *Michel de France*, will sometimes be listed under *de France*, sometimes under *France, Michel de*. In other cases the variants are more mysterious. Jean Pélerin, for instance, is known as Viator, which is a Latin equivalent of his name, but other variants include Le Viateur, and Peregrinus. No simple translation nor even a fuzzy logic programme can be expected

to come up with all the actual variants of names. Needed is a central repository to ensure that these variants can be found efficiently. In the case of artists names, for instance, Thieme-Becker's *Allgemeine Künstler Lexikon* offers a useful starting point, as do the great library catalogues (e.g. National Union Catalogue, British Library and Bibliothèque Nationale). These lists need to be collated to produce one authority list with all known variants, much in the way that the Getty found it needed in the case of its (in house) Union List of Names (ULAN). The problem applies also to subjects,[4] as anyone who has tried to find things in foreign versions of *Yellow Pages*, will know. In Toronto, for example, a person wishing to know about train schedules will find nothing under *Trains*, but needs to look under *Railroads*. A person looking for a paid female companion will find nothing under *Geisha*, *Call Girl* or *Prostitute*, but will find 41 pages under the heading *Escort Service*.

Hence a fully distributed model for housing collections may be extremely attractive because it means that museums, galleries and other cultural institutions can remain in control of the databases and information pertaining to their own collections. The disadvantage is that there are already hundreds and there will soon be tens of thousands of individual repositories and if every user around the world has to visit all of these sites for every search they do, this approach will become hopelessly inefficient.

**6) Centralized Meta-Data**

An alternative is to link this distributed model of individual collections with a centralized repository for meta-data. The basic idea behind such a repository is to use the methods established by thousand of years of library experience as a general framework for searching libraries, museums, galleries and other digitized collections. This centralized meta-database will have three basic functions:

First, it serves as a master list of all names (who?), subjects (what?), places (where?), calendars, events (when?), processes (how?) and explanations (why?). This master list contains all typical variants and versions of a name, such that a person searching for Vinci, Da Vinci or Leonardo da Vinci, will be directed to the same individual.

Second, this master list contains a high-level conceptual map of the parameters of all major databases in cultural and other institutions. Hence, in the case mentioned above of the user searching for Chinese art of the Han dynasty, the master list will identify which databases are relevant. Recent initiatives in site mapping and content mapping will aid this process.

Third, this master list of names and subjects is linked to indexes of terms (classification systems), definitions (dictionaries), explanations (encyclopaedias), titles (bibliographies), and partial contents (reviews, abstracts, and citation indexes). Thus this centralized database effectively serves as a digital meta-reference room which links to distributed contents in libraries, museums, galleries and other institutions. This process of contextualisaton of otherwise disparate information enables the centralized source to act as a service centre in negotiating among distributed content sources.

Libraries have long ago discovered the importance of authority lists of names, places and dates. Indeed, a number of international organizations have been working in this direction during the past century, including the Office Internationale de Bibliographie, Mundaneum, the International Federation on Documentation (FID[5]), the International Union of Associations (UIA[6]), branches of the International Standards Organization (e.g. ISO TC 37, along with Infoterm) as well as the joint efforts of UNESCO and the International Council of Scientific Unions (ICSU) to create a World Science Information System (UNISIST). Over 25 years ago, the UNISIST committee concluded that: "a world wide network of scientific information services working in voluntary association was feasible based on the evidence submitted to it that an increased level of cooperation is an economic necessity".[7] Our recommendation is that this world-wide network should include both cultural and scientific information.

As a first step one would combine the lists of names already available in RLIN, OCLC, BLAISE, PICA, GABRIEL, with those of the Marburg Archive, the *Allgemeine Künstler Lexikon*, *Iconclass*, the Getty holdings (ULAN, Thesaurus of Geographic Names), and the lists owned by signatories of the MOU. This will lead to a future master list which is essential for all serious attempts at a meta-data approach to cultural heritage and knowledge in general. Because such a list represents a collective public good it is important that it should be placed in safekeeping with UNESCO. Senior officials at UNESCO already support this idea. It would make sense to link this list with related bodies such as UNISIST or ICSU. A series of copies will be replicated in various centres around the world.

The basic framework for such a digital reference room might come under the combined purview of the European Commission's Memorandum of Understanding in its next phase and the G8 pilot projects 5 (Multimedia Access to World Cultural Heritage ) and 4 (Bibliotheca Universalis). A series of national projects can then add country specific information. These national projects can be organized by consortia of industry and government. By contributing lists from a given country, that country receives access to the centralized meta-data base.

## 7) Conclusions

Models for knowledge organization have ranged on the one hand from dreams of a single centralised source for all contents (e.g. the Library at Alexandria), to a fully distributed model on the other. We have shown that, although they may be conceptually attractive, both of these extremes are impractical. It was shown that these problems will not be resolved as a result of a) recent innovations on the Internet, b) new initiatives with respect to meta-data or c) even through the advent of nearly infinite computing power which promises to increase greatly the possibilities of number crunching using brute force. In the end, all of these solutions are piecemeal and short term.

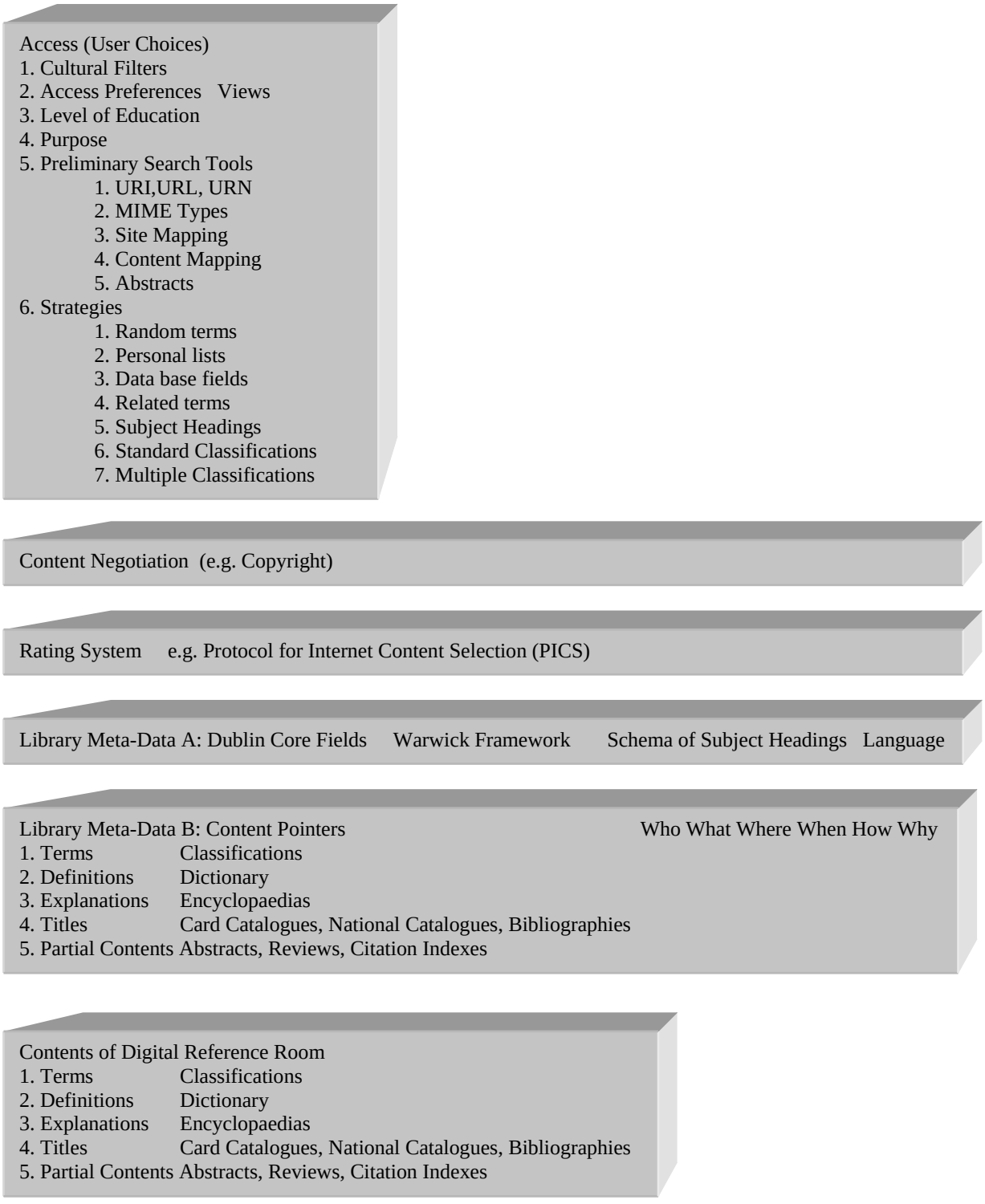
This paper outlines an alternative model, entailing centralised meta-data in the form of a digital reference room and distributed content sources. This digital reference room, will

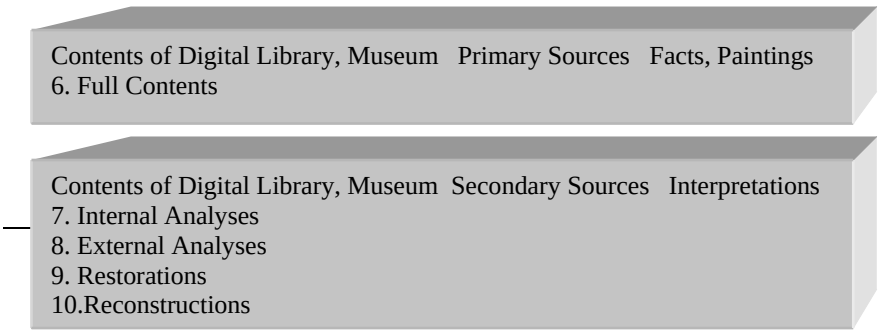
combine in virtual space the resources of famous reference collections such as the British Library, the Bibliothèque Nationale and the Vatican Library, and thus serve as an entry point for digital libraries of primary and secondary sources on a global scale.[8] It would be fitting if the European Commission working in tandem with the G8 pilot projects and UNESCO, would co-ordinate such a project in conjunction with consortia of industry and governments. Centralised meta-data in a digital reference room will present considerable challenges, but it offers a long term answer to the problems of an information age on a global scale.

**Acknowledgements**

This alternative has grown out of many years of experience in major libraries (e.g. British Library, Göttingen, Vatican) and research institutes (Wellcome Institute for the History of Medicine, Herzog August Bibliothek, Getty Center for the History of Art and the Humanities --now the Getty Research Institute-- and the McLuhan Program in Culture and Technology). It has matured this past year as a consultant to Stuart McLeod (CEO, Bell MediaLinx) and while doing research on new media for Eric Livermore (Advanced Networks, Bell Northern Research). I am very grateful to these gentlemen and to the many individuals at the above institutions for their generous support and encouragement over the years. I owe a great deal to Dr. Ingetraut Dahlberg, the founder of the Gesellschaft für Klassifikation and the International Society for Knowledge Organisation for many stimulating suggestions, ideas, and references such as bringing to my attention the work of UNISIST.

## Appendix 1 Some Key Elements of the SUMS-SUMMA Model (©1997) as a Framework for a Meta-Data Digital Reference Room






**Notes**

[1] Keynote: *2nd International Conference. Cultural Heritage Networks Hypermedia*, Milan, September 1997.

[2] "World Access to Cultural Heritage: An Integrating Strategy", Acts of Congress: *Beni Culturali. Reti Multimedialità, Milan,, September 1996*, Milan, 1997, (in press).

[3] See the author's: "The Future of the Memorandum of Understanding (MOU) for Multimedia Access to Europe's Cultural Heritage," Draft Document of the Memorandum of Understanding.

[4] For an example of this problem in the context of historical studies see the author's: Past Imprecision for Future Standards: Computers and New Roads to Knowledge", *Computers and the History of Art*, London, vol. 4.1, (1993), pp. 17-32.

[5] Based on its French name: Fédération Internationale de la Documentation

[6] Based on its French name: Union Internationale des Associations

[7] UNISIST. *Synopsis of the Feasibility Study on a World Science Information System*, Paris: UNESCO, 1971, p. xiii.

[8] This digital reference room will also be a fundamental resource for a new software which is truly universal in its scope, namely a System for Universal Multi-Media Access (SUMMA), provisionally to be developed at Maastricht.