Electronic Publishing '01 –
2001 in the Digital Publishing Odyssey
A. Hübler et al. (Eds.)
IOS Press, 2001

Interaction with electronic documents using the library metaphor

Guadalupe Muñoz-Martín, Ignacio Aedo, Paloma Díaz

Computer Science Department (University Carlos III of Madrid)
Avda de la Universidad, 30 28911-Leganés, Madrid, Spain
+34 91 624 9499 - lupe@inf.uc3m.es

The information available in the Internet could be more useful if in addition to its accessibility, it would be organized as users require. Internet users can feel lost if they can not find information management and retrieval services like those provided by traditional libraries. In the belief that the use of the traditional library metaphor for documents over the Internet will improve their management, we have developed a model for digital libraries called the VILMA model, and a prototype that implements it. VILMA's user interface utilizes a spatial metaphor where the typical elements (such as books, shelves, etc.) are presented in a three-dimensional world. Moreover, VILMA is composed of the three main elements of a traditional library: information entities, metadata, and processes. In VILMA, the information entities that form the library's collection, are documents in the Internet, and the metadata are those required by the Dublin Core Metadata Set. There are two types of processes in VILMA: public, which are all those related to the user, and technical, that have to do with traditional librarian's tasks. Public processes are subscription, identification, searching (analytical, expert, accidental) and customising, while technical processes are selection of documents, acquisition, classification and cataloguing, indexing, maintenance and notification. In this paper, we will present the VILMA model, and its prototype.

0. INTRODUCTION

One of the main concerns of human beings along their whole history has been the preservation and dissemination of the information to keep trace of their culture and to share knowledge. Some of the people in charge of the storage and management of information are the librarians. Librarians have years of experience organizing large quantities of diverse information. They have examined the problem of information retrieval and created regulated rules and standards to organize information, like the Dewey Decimal System [DDC, 96], the Universal Decimal Classification (UDC) scheme [McIlwaine, 95], or the classification scheme devised by the Library of Congress (LCC) [Chan, 86].

The entity in which librarians perform such activities is the library. Although libraries exist since long time ago, they have suffered an evolution mainly due to the change of the material that it organizes, and the advance in the tools that librarians use to perform their tasks. Librarians have achieved their goals making use of the means that were available at every moment in such a manner that the evolution of these means has driven the evolution of the library itself. This evolution has produced three different types of library, each one based on the previous. Before we could talk about digital libraries we had the so-called automated libraries, that were similar to traditional libraries in many ways, but with some use of technology for their management. Traditional libraries have been managing information stored in physical means such as paper or videotapes, and now, there is a new kind of medium, the electronic one. Current traditional libraries (automated libraries) are combinations of old libraries and the use of new technologies. As the technology advances, libraries are evolving into digital libraries. They utilize the new electronic medium in a bigger percentage than the automated ones. A digital library has been described as a federated structure that provides humans both intellectual and physical access to the huge and growing world-wide networks of information encoded in multimedia digital formats [Birmingham et al., 94]. A digital library must provide all of the services found currently in traditional libraries, based on principles of selection, acquisition, access, management and preservation, related to a specific client community, as well as new services made possible by the digital media and computer networks [Gladney et al., 94].

Nowadays, the availability of information has experienced an explosion mainly due to the advance of the technology and the growth of the Internet. But this explosion is tied down by chaos: there is no order or consistency among all the information available. Besides, it gets worse when whoever might produce whatever. Although some people talk about the Internet as if it was a digital library, the reality is that Internet is like a big repository of chaotic information, but not a library [Lynch, 97]. The information available in the Internet is unorganized because there is no single entity indexing it in the way a library catalog does. Moreover, the anarchic Internet does not contain standards for information management, and therefore, search mechanisms are forced to look for keywords in unstructured data. For these reasons, according to [Ferguson and Durfee, 98], users of the Internet feel lost when they can not find information management and retrieval services like those provided by traditional libraries where you make use of the librarians' knowledge, the cataloging and classification mechanisms, the notification support, or the personalized search tool among others.

In this paper, we present the model for digital libraries called VILMA (VIrtual Libraries with a Multi-layer Architecture) based on the traditional library metaphor for the management of information available in the Internet by means of a multi-layer architecture. Digital libraries using this metaphor will be able to manage information as traditional libraries do, but they will also take advantage of the new electronic medium. Also, to take the metaphor further, we propose the use of a user interface based on virtual reality in the belief that it will improve the usability of the digital library.

1. THE ELEMENTS OF THE LIBRARY AND THEIR EVOLUTION

As pointed out by [Nürnberg et al., 95], there are three main elements in a traditional library: information entities, metadata, and processes. Information entities are every item in the library (single or multiple as long as the whole set is a unit of information) that provides complete information about something. Metadata are information about the information entities in the library and how it is accessed and they should identify unequivocally every document. Processes are all the actions taken over the information entities or the metadata in the library. These elements have been changing as an adaptation to the needs of the library evolution. This section presents these changes.

1.1 Information entities

They are physical documents such as books, journals, videotapes or CD-ROM's in traditional and in automated libraries, but in digital libraries we have both physical and digital documents. Digital documents are files stored in computers and could be just digitized copies of physical documents, or proper digital documents such as a collection of electronic hyperdocuments. The whole set of information entities in a library, traditional or digital, is called a collection.

1.2 Metadata

Most of the metadata necessary to identify the information entities does not depend on the medium in which these entities are stored, but some do. Some medium-independent are title, author or subject, and others medium–dependent are the size or the number of pages when the medium is paper, or the number of bytes when the medium is electronic. Another important issue related to metadata is how to access it, because they are also information entities that have to be stored somehow. Another important element related to metadata in a library is a thesaurus, that is a list of subject headings or descriptors usually with a cross-reference system for use in the organization of a collection of documents for reference and retrieval. The whole set of metadata information, the thesaurus and the indexes necessary to access it is called a catalog. As information entities do not vary from traditional to automated libraries, metadata is almost the same for both of them.

There have been several rules, called cataloguing rules to specify the metadata, like the ISBD (International Standard Bibliographic Description) since 1974. But traditional and automated libraries differ in how they store the metadata. Traditional libraries do it in paper cards that can be accessed browsing them in ordered catalogues, but automated ones store the metadata in electronic records that conform to standards such as MARC (MAchine Readable Catalog). User's interface to this metadata are OPAC's (Online Public Access Catalog). However, as digital libraries contain both physical and digital entities, which metadata are necessary for these entities change when we think of electronic documents. Electronic documents have other attributes related to the electronic medium that need to be stored such as locations, document versions or date of last modification [Heery, 96]. Metadata in this case could be stored as electronic records or could be included in the electronic document to facilitate its extraction. Currently there are several efforts to achieve a standard for digital document metadata. Some kinds of available metadata sets are: HTML META-tags, IAFA (Internet Anonymous FTP Archive) templates, Dublin Core metadata sets or the USMARC bibliographic records.

1.3 Processes

These are different when we talk about traditional, automated or digital libraries. For the first and the second one, these activities are the same, and only differ in that automated libraries use automation systems to perform some of them. These processes have been divided by [Arms, 90] in two groups: Public, such as circulation, reference, and assistance; and Technical, such as acquisition, cataloging, or shelves ordering. For digital libraries it has been described different processes such as Collection Administration, Acquisition, Cataloguing and Indexing, Borrowing, and Readers Services [Landoni and Catenazzi, 93]. A more detailed division based on the previous one distinguishes two levels of services for digital libraries [Muñoz et al., 98]: Reader-oriented, such as subscribing, walking through the library to access other services, browsing through the library shelves and through a hierarchical structure, searching through the library catalogues, consulting or asking for advice from the librarian, customizing or creating a personal user library; and Librarian-oriented, such as information acquisition, storage and maintenance of the information, cataloguing, and indexing.

Although some of the processes could sound similar among the three types of libraries, the use of digital technology to perform them incorporates substantial differences between namesake ones. For traditional and automated libraries public services that could be analogous to the reader-oriented in digital ones force readers to go physically to the library to be able to access them while digital libraries allow remote access. Remote access involves seveıal issues such as that the library could be distributed to make a better use of the resources, readers might use different languages, several users could use concurrently the same services, and the access could be at any time, among others. Also, technical/librarian-oriented services have to be performed manually by the librarian in traditional and automated libraries but they could also be carried out automatically in digital ones. This fact implies that human expertise that is crucial for these tasks has to be simulated by computer programs. Also, resource sharing could be done using the advantage of remote networked access.

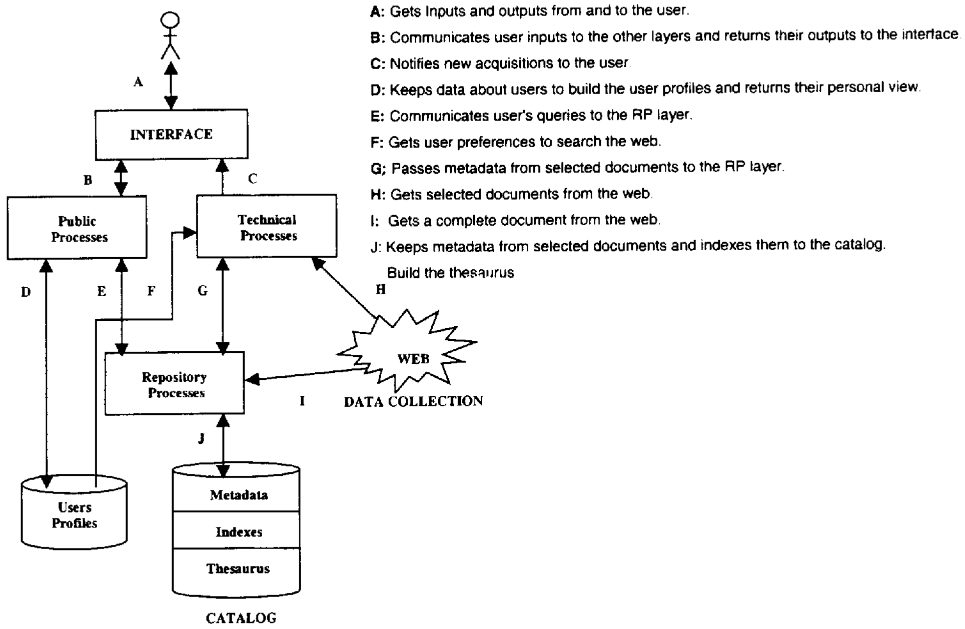
2. VILMA model

VILMA (VIrtual Library with a Multi-layer Architecture) is a model of digital library developed in order to assess the traditional library metaphor including the new features derived from the electronic media. This model provides a personalized acquisition and cataloguing system, and a virtual reality user interface that may allow a more simple access to the system. On this basis, after a first schema for the architecture, the research has evolved into the implementation of a prototype. The objective of the development of such prototype is not just the achievement of a final product, but also to verify its applicability, that is, the capability of being implemented as a real and efficient product. The next subsections present both the model and the prototype.

2.1 VILMA model

The model being developed tries to handle the three elements that a digital library should be composed of (data, metadata and processes) without loosing modularity, flexibility and scalability. Figure 1 shows a general view of this model, which is explained briefly next. This model has three main layers, called "Public Processes" (PP), composed of those processes that interact with the user, such as subscribing, identification, browsing, searching, customizing, consulting and walking processes, "Technical Processes" (TP), composed of those processes that manage the collection, such as the selection of documents, acquisition, classification, indexing and maintenance processes, and "Repository Processes" (RP), in charge of the communication between the PP layer and the catàlog. These layers communicate among them to provide the library functionality as follows.

In this model, the data collection, that is, the information entities, is obtained from the web, and it remains located in the web. The TP layer will select documents from the web based on the classification of the library and on the user preferences (arrow F) and it will acquire them to include them in the library (arrow H). When a document is acquired, it is classified to obtain its metadata, and it is indexed in the library catalog through the RP layer (arrows G and J). This layer is also in charge of the catalog maintenance, that is, to keep updated metadata of the data collection (arrows G and J), and of notifying new acquisitions to the users (arrow C). The PP layer handles all the processes related to the user. It will get, through the interface (arrows A and B), all data about every user to keep their profiles and to allow them to customize their own view of the library and it will show this personal view getting every user's information through arrows D, B and A. It will also get the users search queries and will pass them (arrow E) to the RP layer to find the required information from the catalog (arrow J). Through arrow I, the RP layer will be able to get a complete document when a user request it.



A: Gets inputs and outputs from and to the user.
B: Communicates user inputs to the other layers and returns their outputs to the interface
C: Notifies new acquisitions to the user
D: Keeps data about users to build the user profiles and returns their personal view.
E: Communicates user's queries to the RP layer.
F: Gets user preferences to search the web.
G; Passes metadata from selected documents to the RP layer.
H: Gets selected documents from the web.
I: Gets a complete document from the web.
J: Keeps metadata from selected documents and indexes them to the catalog.
Build the thesaurus

Figure 1. VILMA model

2.1.1 Public Processes Layer

This layer is in charge of all the communications with the user, and therefore will consists of a user interface that handles this communication. For the design of the user interface of an interactive system as a digital library is, we need to specify several points. According to [Shneiderman, 98], and his Object-Action Interface (OAI) model, we need: to identify which objects will be placed in the interface, to perform a task analysis, and to define the interaction styles. Task analysis refers to discovering what actions will be taken by users related to every particular object of the interface, and interaction styles implies to decide how users will accomplish the tasks and how the objects will be represented. The objects as well as their representation proposed by this model are based on the traditional library metaphor, which are rooms, books, shelves, catalogues, and OPAC. The representation of the objects of the system will imitate the real ones. The interaction styles that this model suggests are supported by the virtual reality, as the means for users to accomplish the tasks. The task analysis has been led by the knowledge about the public processes found in a library. The movement of a user in a library is directed to achieve the realization of any of the public services provided by the library. Thus, when a user comes to a library she will subscribe giving personal information and personal preferences. Once a user is in the library records, she will be able to enter the library after identifying herself. After that, she can browse the documents, read them, search the library catalogues, consult the librarian to get help, or customize her personal view of the library.

The proposal of virtual reality as the interaction style, as well as the use of the traditional library metaphor for the representation of the objects in the interface, is the conclusion of a careful examination of the final purpose of the user interface. Interfaces using virtual reality achieve a closer correspondence between the user's mental model of the computer and the image on the screen because three-dimensional images make a better use of the human visual perception and simulate the same actions taken in real life.

2.1.2 Technical Processes Layer

This layer will handle all the technical processes. These are primarily: Selection of documents, Acquisition, Classification, Indexing, Maintenance and Notification. Documents that belong to a library are chosen among the entire available documents. The modules in charge of this task decide which of those should be acquired in accordance with the library's classification and the users' preferences. This means that when a suitable document is found, a simple analysis of its content must be accomplished to decide whether the document can be chosen or not. Once the document's suitability is clear, this is acquired to be included in the library. The acquisition of a document in this model is not a "real" acquisition in the sense of getting a copy of it. To take advantage of the electronic medium and the web accessibility, the document remains on its original location, and what it is "acquired" is a pointer to that location, and a temporary copy of the document to get its metadata. Before this document can be part of the library, it must suffer a classification and cataloguing process that will extract the necessary metadata in order to identify the document. When this process is completed the document can be indexed to the library. After they are included in the library, as documents are "alive entities" that might change along the time, they require a maintenance procedure to keep them updated. Also, when a document is included in the library, those users who were interested in that type of document receive a notification.

2.1.3 Repository Processes Layer

This layer will manage the users inputs provided by the PP layer to search the catalog, and the outputs that answer those queries. It will serve as the access mechanism to the data collection. This layer as well as the TP one might be distributed and it may implement different approaches; for example, the repositories could use inverted files, relational database management, or others. The choice of one of them does not affect the model and it just depends on the implementation restrictions.

2.2 Implementation

After a first design of the model, a prototype that implements it has been built. This prototype is called VILMA, and this section will show details about the implementation of every layer of the model.

2.2.1 Public Processes Layer

In VILMA's interface, every library's room is graphically represented as a VR-model to have a closer approach to the actual physical library. The whole library is represented as a building and the inner side is divided in three floors and all of them in two rooms, one of them is a hall, and the other is the main room. Every floor is conceptually associated to a concrete public process. Thus, the ground floor provides the subscription, identification and the consulting processes.

First floor is dedicated to search and browsing processes: the user will be able to search the library catalog, and browse the documents in the shelves represented as books. Third floor gives the personal view of the user, that is, it provides the customizing process. Floors are communicated through a lift. The virtual reality allows the user to move from one point to another in this interface to access any of the services. Figure 2 shows several snapshots of this interface. This interface has been modeled using the VRML97 International Standard ISO/IEC 14772-1:1997, and the External Authoring Interface (EAI) to add external behaviour produced by Java programs.

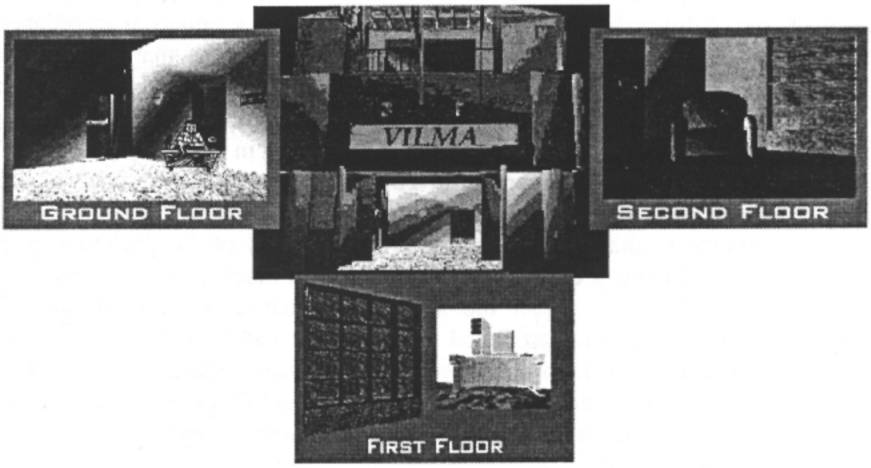


Figure 2. VILMA user interface

2.2.2 Technical Processes Layer

Selection and Acquisition: All documents belonging to the VILMA prototype must fulfill three requirements based on the nature of this digital library. 1) The "market" of the VILMA prototype is the World Wide Web, so they have to be found in the web. 2) The Dublin Core is employed as the Metadata Scheme because it is the most widely used metadata scheme from all those available for web documents, due to its ease of use and interpretability [Miller, 96], thus, they must contain this kind of information. 3) The VILMA prototype will keep computing documents, and the 1998 ACM Classification System has been chosen as the library classification (permission of use given by ACM). Documents must belong to any of the library classification subjects.

Therefore, all documents suffer a pre-classification process. That is why a web crawler written in Java to make it reusable and platform-independent, has been developed to search the web to find suitable documents for it.

Cataloguing and Classification: a "Documents Classification Module" has been written as a Java package. This module uses a simple version of the Extended Boolean Model for the Information Retrieval. In the vector space model of information retrieval [Salton, 71], documents are modeled as vectors in a high-dimensional space of many thousands of terms. The terms are derived from words and phrases in the document and are weighted by their importance within the document and within the corpus of documents. Each document's vector seeks to represent the document in a "vector space", allowing comparison with vectors derived from other sources, for example, queries or other documents. In VILMA we use a set of "Training Documents" which are documents manually classified and then presented to the system as examples of documents that belong to each classification. The system then builds class representatives each of which consists of common terms occurring in the documents known to belong to a particular classification group. When the system subsequently encounters new documents it measures the similarity between the document and the class representatives. Each time a new document is classified it is used to modify the class representative to include its most commonly occurring keywords. This model has been used as the basis of successful algorithms for document ranking, document filtering, document clustering, and relevance feedback [Jones and Willet, 97], [Baeza-Yates et al., 99]. After a document is classified, a "Metadata Extraction module" gets all the metadata elements from the document that are needed to follow the Dublin Core Element Set, version 1.0, which are: Title, Author, Subject, Description, Publisher, Contributor, Date, Resource Type, Format, Resource Identifier, Source, Language, Relation, Coverage and Rights Management, and catalog it.

Indexing: after the classification and extraction of metadata from a document, a program written in Java will index the metadata extracted from the document to the repositories in order to create the library's thesaurus and to facilitate the access to the metadata under a user's search query.

Notification: When a document is included in the library, a notification Java program is in charge of checking those users interested in documents that belongs to the same classification than that document, and send them a warning about the acquisition.

Maintenance of the information: As the documents for VILMA are obtained from the web, a robot will check first that documents still exists, and that all the metadata about every document keeps the same. It will change those metadata that differ from the original. Given that life of web documents is not fixed, a study about the timeline of web documents using Dublin Core is being carried out to provide the robot with the appropriate period for the checking out of metadata.

2.2.3 Repository Processes Layer

This layer consists of the repositories for the library catalog, which stores all the metadata elements for every document that has been indexed to the library. It also consists of the access mechanisms that allow searching this catalog under a user query.. This catalog is a PostgreSQL database running under Linux Debian accessed by the previous modules using JDBC connectivity.

3. RELATED WORKS

In order to verify the novelty and contributions of this work, some of the current digital libraries available via web have been reviewed and compared to the prototype VILMA. The contributions of this research are two: the use of the traditional library metaphor adapted to the electronic and users needs, and the use of the virtual reality for the interface for digital libraries. Thus, firstly we will talk about those that do not use any of them, and secondly we will see those that use any or both of them.

Most of the current digital libraries available via web work as enhanced search engines that use hypermedia to interact with the user. This means that they do not take advantage of the services provided by the traditional library metaphor to manage documents. Also, they keep on using hypermedia based interfaces which, according to [Callaghan and Hand, 96] are limited in function and sophistication. The study carried out by [Theng et al., 00] on the ACM Digital Library, ACMDL (http://www.acm.org), the Networked Computer Science Technical Reference Library, NCSTRL (http://www.ncstrl.org), and the New Zealand Digital Library, NZDL (http://www.nzdl.org) to investigate the purpose and usability of digital libraries throws two main conclusions. First of them is that users prefer traditional library services, and second one is that users feel lost if the digital library do not utilize an appropriate model.

Other reviewed libraries like the previous are: The Library of Congress[1], those from the Digital Libraries Initiative, phase I: University of California Berkeley[2], University of Illinois, University of Michigan[3], University of California Santa Barbara[4], Carnegie Mellon University[5], and Stanford[6], and the German Digital Library Project[7].

However there are some digital libraries that share some of the characteristics of VILMA. For example, the Networked Digital Library of Theses and Dissertations has tried a desktop VR interface as well as an inmersive one [Phanouriou et al., 99].

4. CONCLUSIONS

We have presented a model for digital libraries, and its prototype, that uses the traditional library metaphor in two ways: it implements the services provided by traditional libraries adapting them to the electronic medium in the belief of the generality of these services, and it uses the spatial metaphor for the objects in its interface. This model also provides a virtual reality interface to allow a more realistic interaction with this spatial metaphor. VILMA also provides new services that take advantage of the electronic medium. These features are supposed to improve the management of electronic documents in a network in two different aspects. Firstly it provides all the services needed for documents management. And secondly, because taking advantage of the new electronic medium, it eases the access to the documents in several senses. It makes the reader able to customize her preferences and her library view. It approaches the mental model of the reader to the conceptual model of the system by using metaphors and virtual reality,

The hypothetic value of this model for digital libraries will be checked when the development of the prototype VILMA is finished at the end of March. Then, an evaluation with experts (librarians, CHI experts, etc.) and another with real and potential users will be carried out. The results of these two types of evaluation will be analyzed and will confirm or refute the hypothesis of this research.

5. REFERENCES

[Arms, 90] Arms, C. "Campus Strategies for libraries and electronic information" Ed. Digital Equipment Corporation, 1990

[Baeza-Yates et al., 99] Baeza-Yates, R. and Ribeiro-Neto, B. "Modern Information Retrieval". Addison-Wesley/ACM Press, 1999.

[1] http://www.loc.gov/, [2] http://elib.cs.berkeley.edu/, [3] http://dli.grainger.uiuc.edu/, [4] http://www.si.umich.edu/UMDL/,

[5] http://alexandria.sdc.ucsb.edu/, [6] http://www.informedia.cs.cmu.edu/, [7] http://diglib.stanford.edu/, [8] http://medoc.informatik.tu-muenchen.de/

[Birmingham et al., 94] Birmingham, W., Drabenstott, K., Frost, C., Warner, A. and Wills, K., "The University of Michigan Digital Library: This is Not Your Father's Library". In Proceedings of the Digital Library 94 Conference (DL'94). Schnase, J. L., Leggett, J. J., Furuta, R. K.. Metcalfe, T. (eds.), 19-21, June 1994, College Station, Texas, U.S.A., pp. 53-59.

[Callaghan and Hand, 96] Callaghan, M. and Hand, C. "Presentation and Representation of Implicit Knowledge in the World Wide Web. Workshop on Knowledge Media for Improving Organizational Expertise - Impacts of new methods and enabling technologies", at International Conference on Practical Aspects of Knowledge Management, Basel, Switzerland, 30-31 October 1996.

[Chan, 86] Chan, L.M., "Library of Congress Classification as an online retrieval tool: potentials and limitations". Information Technology and Libraries, 5 (3), pp. 181-192.

[DDC, 96] Dewey Decimal Classification and Relative Index. Edition 21, 1996. Forest Press, division of OCLC Online Computer Library Center, Inc.

[Ferguson and Durfee, 98] Ferguson, I.A.; Durfee, E.H. "Artificial intelligence in digital libraries :moving from chaos to (more) order". International Journal on Digital Libraries; Vol. 2, N. 1 October 1998.

[Gladney et al., 94] Gladney, H., Fox, E., Ahmed, Z., Ashany, R., Belkin, N., and Zemankova, M. "Digital Libraries: Gross Structure & Requirements". In Proceedings of the Digital Library 94 Conference (DL'94). Schnase, J. L., Leggett, J. J., Furuta, R. K., Metcalfe, T. (eds.), 19-21, June 1994, College Station, Texas, U.S.A.

[Heery, 96] Heery, R. "Review of Metadata Formats", Program, Vol. 30, No. 4, October 1996, pp. 345-373.

[Jones and Willet, 97] Jones, K. S.and Willet, P. eds. Readings in Information Retrieval. Morgan Kaufmann, 1997.

[Landoni and Catenazzi, 93] Landoni, M. y Catenazzi, N. "Hyper-books and visual-books in an electronic library" The Electronic Library, Vol. 11, n 3, june 1993

[Lynch, 97] Lynch, C.A., "Searching the Internet". Scientific American, March, 1997, pp. 52-56. Also available at: URL: http://www.sciam.com/0397issue/0397lynch.html]

[McIlwaine, 95] McIlwaine, I.C., "Preparing traditional classification for the future: Universal Decimal Classification". In: New roles for classification in libraries and information networks: reports from the Thirty-sixth Allerton Institute. Cataloging and Classification Quarterly, 21 (2), pp. 49-58.

[Miller, 96] Miller, M., "Metadata for the masses" , Ariadne, the web version Issue 5, September 1996. Also available at: URL:
http://www.ariadne.ac.uk/issue5/metadata-masses/

[Muñoz et al., 98] Muñoz, G., Aedo, I., Díaz, P. Virtual reality and agents in a digital library. In Proceedings of Second European Conference on research and advanced technology for Digital Libraries (Crete, Greece, September 1998) LNCS 1513 Springer, pp. 681-682.

[Nürnberg, et al., 95] Nürnberg, P. J., Furuta, R., Leggett, J. J., Marshall, C. C., Shipman III, F. M. "Digital libraries: issues and architectures" In Proceedings of the second annual conference on the theory and practice of digital libraries, Digital Libraries' 95. June 11-13, 1995 - Austin, Texas, USA.

[Salton, 71] Salton, G. The SMART retrieval system - experiments in automatic document processing, Prentice-Hall, Englewood Cliffs N.J. 1971.

[Shneiderman, 98] Shneiderman, B. Designing the user interface. Addison Wesley Longman. Chapter 2, pp. 61-67.