A Metaphor Graphics Based Representation of Digital Libraries on the World Wide Web: Using the libViewer to Make Metadata Visible

Andreas Rauber and Harald Bina
Institut für Softwaretechnik, Technische Universität Wien
{andi, harry}@ifs.tuwien.ac.at

Abstract

While methods for searching large digital libraries have experienced tremendous improvements recently, interfaces to such collections still have a far way to go. Most interfaces to digital libraries present themselves as various forms of sorted lists, providing metadata information on the documents in textual form. This prohibits intuitive understanding of document archives or web search results. In this paper we present the libViewer*, a Java-based user interface to digital libraries using metaphor graphics to display information on the elements in a digital library in an intuitively understandable way. Metadata on digital libraries based on the Dublin Core Initiative is mapped onto a set of metaphors to allow instant recognition and orientation in an unknown document collection, facilitating interactive retrieval and exploration.*

Keywords*: Information Visualization, Web-based Library Representation, Information Space Metaphors*

1. Introduction

When entering a library or large-scale book store, in spite of the overwhelming amount of information present in such locations users usually manage to orient themselves and find the way to their section of interest quite easily. Without being able to read the title of books from the far distance, not knowing actually where to find a book by a specific author or even without knowing a title or an author of a book, most people are able to locate the corresponding library section when looking for a dictionary, a poem collection or a story book for children. Once they found the corresponding shelve, by scanning the books sorted there, it is usually easy for them to tell the age of a book, the number of times it has been used before (at least in a public library rather than in a bookstore), as well as the amount and type of information to be expected in the books simply by looking at them. The cover of the book, the title, type of binding, the shape of the binding (brand new versus well-used and almost torn apart), the size of the book, color and other properties of an item on the shelve contain a wealth of information that most people are accustomed to and able to interpret intuitively. Thus it is easy for us to gain an intuitive overview of the contents of a library and the type of information present.

Taking a look at the emerging field of digital libraries, we are faced with a somewhat different situation. While there are collections of digital data of almost overwhelming size, representation of those collections fails to transport even part of the information in an intuitively understandable way. Research has focused on searching and retrieval of information from such collections. However, the representation of these search results is usually limited to ranked lists, giving the title, the author, date of last modification and whatever meta-information there is available, leaving the user with a long list of text to read and to interpret.

In this paper we argue in favor of the utilization of metaphor graphics for the representation of digital information to allow the user to get an instant overview of the information available. Instead of listing the meta-data as a textual description of the entries in a digital collection, we use metaphor graphics to visualize this information in an intuitively interpretable way. We present the libViewer, a Java-based interface for a digital library. It allows WWW-based access to document collections based on the Dublin Core Metadata Initiative for digital libraries. Metadata attributes are mapped onto a set of metaphors to allow instant recognition and orientation in an unknown document collection. Different types of documents, ranging from books, journals and technical reports to audio and video items are depicted as different types of books, folders and boxes, with the size of the document being encoded, for example, in the thickness of the book, or language as the color of the binding. Using these and a set of further metaphors allows us to view, understand and interpret a digital collection the way we interact with conventional collections, incorporating the wealth of information present in the physical appearance of documents.

The remainder of this paper is organized as follows. In Section 2 we give a brief review of work on library representation and interaction with digital document collections. The set of metadata attributes for digital libraries as defined by the Dublin Core Metadata Initiative and some related metadata projects are presented in Section 3. Next, Section 4 describes a set of metaphors encountered in conventional libraries, which can be used as graphical representations for the metadata of digital libraries. The libViewer interface is presented in Section 5, demonstrating the benefits of visualizing metadata using graphical metaphors, followed by some conclusions in Section 6.

2. Library Representation

With the massive increase of the amount of information available in digital form, sophisticated methods for dealing and interacting with electronic information repositories were developed. Research in Information Retrieval (IR) has produced a number of systems allowing, apart from mere database searches for titles and authors' names, full text scanning of large text corpora, retrieving documents on specific topics or describing special concepts [3, 4, 15].

While most of these methods allow the selection of a subset of entries of a digital library, we are still left with the problem of (a) identifying those items of interest from the sometimes still huge subset of items returned by search engines and (b) locating relevant information when no (more detailed) query can be identified as such. This problem can be described as document archive browsing or archive exploration as opposed to document retrieval. In order to be able to browse a collection of documents we need it to be visualized in a way that allows us to get an instant overview of the information present. This necessity of an enhanced library representation has been addressed in a number of projects, trying to provide convenient access to digital document collections.

One of the first applications of metaphors in the digital library arena is reported in the Bookhouse project [10], where a document database is represented as a storehouse consisting of different rooms. A number of search strategies can be followed, which in turn are indicated by various images, such as a clock to search by the time dimension, a globe for search by geographic location of books etc.

A set of various visualization techniques for information retrieval and information representation purposes was developed at Xerox PARC as part of the Information Visualization Project [14]. Information is depicted in a 3-dimensional space with the focus being on the amount of information being visible at one time and an easily understandable way of moving through large information spaces.

To overcome the basic limitations of the one-dimensional ranked-list representations of most search engines, we developed the SOMLib system [9, 13], a 2-dimensional map display which automatically organizes a set of documents by their contents. The self-organizing map (SOM) [6], a popular unsupervised neural network model, is used to produce a content-based document clustering. This approach has been used in a number of other projects for document classification so far [5, 7, 8]. A web-based interface allows the interactive exploration of documents, with the spatial organization of the collection allowing documents on similar topics to be found close to each other. This capability makes the *SOMLib* representation particularly useful in digital library exploration.

At the CNAM library, a virtual reality system is being designed for the visualization of the antiquarian Sartiaux Collection [2], where the binding of each book is being scanned and mapped into a virtual 3-dimensional library to allow the user to experience the library as realistically as possible.

While all of these methods address one or the other aspect of document, library and information space visualization, none of these provides the wealth of information presented by a physical object in a library, be it a hardcover book, a paperback or a video tape, with all the information that can be intuitively told from its very looks.

What we would like to have is a way to represent the meta-information that is available for documents in a digital library and that is important for understanding the characteristics of the documents in way that is intuitively graspable. Combining a WWW-based graphical representation of a document collection with the automatic spatial organization of the *SOMLib* system's document clustering capabilities allows a library to be perceived similar to a conventional real-world library via any standard web browser.

3. Metadata for Digital Libraries

In order to be able to provide additional information about the entries in a library, metadata is used to describe the various items. In conventional libraries we often find separate library catalogues for these various metadata attributes, listing books by title, author, topic and so on. In the field of digital libraries, a huge number of initiatives deals with the development of metadata standards for digital collections.

As one of the older examples of such metadata definitions for documents we might consider the BibTeX system. 14 different types of documents are described by a set of 24 attributes, providing a wide range of metadata in a rather flexible way.

One of the most extensive metadata formats is MARC (Machine Readable Catalogue Format), which originated in the 1960's as means of exchanging library catalogue records. It has evolved into a number of derivative standards like the USMARC in the United States, UNIMARC for international library data exchange etc. It provides a highly complex set of attributes for describing documents and it is highly developed for bibliographic and bibliographic-like data. Albeit, due to its complexity, the creation of correct MARC records requires trained specialists, limiting its application to professional library organizations.

One of the most promising standards for digital libraries is the metadata set developed by the Dublin Core (DC) Metadata Initiative (http://purl.oclc.org/dc/). It consists of a set of 15 basic attributes used to describe digital documents. While the exact specification of some attributes is not yet defined, the attributes as such have been agreed upon and are now being used in a number of projects to describe anything from webpages to digital archives. Due to space considerations we here only provide the most relevant DC attributes:

- Title: The name of a resource.
- Author or Creator: The person or organization primarily responsible for creating the intellectual content of the

resource.
- Publisher: The entity responsible for making the resource available in its present form.
- Date: The date the resource was made available.
- Resource Type: The category of the resource, such as home page, novel, poem, working paper, technical report, essay, dictionary.
- Format: The data format of the resource, used to identify the software and possibly hardware that might be needed to display or operate the resource.
- Resource Identifier: String or number used to uniquely identify the resource. Examples for networked resources include URLs, URNs or ISBN
- Language: Language of the content of the resource.

Due to its simplicity combined with its richness, the Dublin Core attribute set seems to be a promising metadata definition for digital libraries.

To provide further important information to the user of a digital library, allowing her to judge the relevance of a document by its mere graphical representation, we added a number of additional attributes to the DC metadata following the DC recommendations for attribute set extensions. This allows to accommodate information accumulated during the operation of a digital library. The additional attributes are

- Times Referenced: The number of times a resource was being referenced, i.e. being looked at and used.
- Last Time Referenced: The date when the resource was referenced for the last time.
- Location: Location of the resource in a grid-like bookshelve based on e.g. the spatial clustering of documents provided by the *SOMLib* system.
- Size: Size in terms of pages or bytes storage space.
- Bookmark: A flag indicating whether and which type of bookmark (color) is attached to this resource.

4. Information Space Metaphors

In order to be able to visualize the metadata present in any digital library metadata attribute set we need to identify a set of metaphors that are well understood by being easily identifyable and relating to properties known from the real world. Metaphors should (a) label the resources so that they become intuitively graspable, (b) measure them, i.e. provide quantitative information, (c) represent or imitate reality and (d) enliven or decorate the library representation [17]. Based on these premises we identified a set of metaphors to visualize the various metadata attributes in a library setting, where the mapping of attributes to metaphors needs to be flexible enough to allow personalization of the resulting visualization. Among the metaphors identified we find:

- Representation Type: Each piece of work in a digital library needs a physical representation. A set of templates is defined to represent e.g. hardcover books, paperbacks, binders, manuscripts, boxes for audio, video and software components or links to other libraries to provide a realistic visualization of library resources.
- Color: Color, being a very dominant feature, can be used to represent a variety of attributes in a very distinguishing way, such as language, publication series, genre, topical classification etc.
- Size:The amount of information available in a book or magazine is intuitively judged from the size of the physical object, e.g. the number of pages.
- Format: Format conveys, next to the type of a document, a lot of information on the genre of a document, considering, for example, oversize format books such as an atlas or art collection books vs. small paperbacks.
- Logo: When browsing a library, one recognizes the logos of well-known publishers, associating them with special types of publications. Thus, while making the library representation look more realistic and rather decorating te books, a lot of information can be conveyed using a company logo.
- Text: Although the amount of text found on the spine usually is limited to a few words, a walth of information is provided by both the text, such as title or authors listing, as well as the type of text representation, like different fonts or font colors.
- New Book: Books and other items that have been added to a collection only recently usually can be identified at large distance by their somewhat shinier color. Thus, glare effects and reflections can be used to highlight certain entries in a collection.
- Used Books: Contrary to recently added items, books that have been in a collection for a long time and which are being consulted frequently show some signs of intensive usage by crippled, well-thumbed bindings etc.
- Dust: Whereas items in a library that are frequently consulted tend to remain rather 'clean', dust usually settles on books that have not been referenced for a long time.
- Bookmarks: Similar to conventional books we can use bookmarks of different colors to mark books.
- Shelve Position: When taking a look at bookshelves we find, that books that are being used frequently, usually are not neatly aligned with all the other books nearby, but rather tend to stand out. In terms of query processing, this metaphor may be used to indicate the relevance of a resource with respect to a specific query.
- Location: Similar to conventional libraries, resources on identical topics should be located next to each other.

Based on these metaphors we can define a set of mappings of metadata attributes to be visualized, allowing the easy understanding of documents, similar to Chernoff faces for multidimensional space representation [1]. However, great care must be taken in the selection and definition of these multi-functional elements, so that the encodings can be broken by every user, avoiding the creation of graphical puzzles [16].

5. libViewer: Visualizing Metadata

The libViewer[1] is a User Interface to a digital library using Dublin Core based metadata descriptions. It is implemented as a Java-Applet allowing the simple representation of and interaction with document archives via the World Wide Web. In its current version the metaphors identified above are implemented to allow a flexible mapping of metadata attributes to graphical representations in order to best suit the requirements of the user as well as the resources present in the library. Thus, a number of mappings can be defined to optimize the representation for the requirements of a digital library, ranging from a rather realistic representation of the items in the library to a more abstract one designed for special exploration purposes. A set of 2 shelves is used to display the books in a library setting. By default the items in the library are put on the shelve in the order they are provided by the library system, listing the titles of the books as shelve labels. Used in connection with the *SOMLib* document clustering system [12], labels automatically extracted using the *LabelSOM* method [11] can be used to describe the contents of the various documents in a specific document cluster, i.e. shelve.

Figure 1 provides a sample representation of a digital library containing a number of books, technical reports, papers and multimedia resources as well as hypertext links. The various document types can be easily identified, like, e.g. the libViewer and somViewer technical reports in green binders, the 4 different Langenscheidt dictionaries as yellow hardcover books or various paperback books published by e.g. Springer. They are created by assigning each resource type a corresponding document type representation. In the given example, both journal papers as well as conference papers are mapped onto the paper representation metaphor. The difference between conference and journal papers is indicated by their color with the latter appearing in a darker color than the white conference papers. Technical reports as well as documentations are mapped to the binder metaphor with their subdivision in this particular mapping being indicated by different vertical sizes of the binders. Thus, the hierarchy of document types defined in the DC metadata can be mapped onto a hierarchy of metaphorical representations.

Further attributes are mapped in a similar fashion, e.g. having the logo identify the publisher of a book if a corresponding logo is available (e.g. Springer, Langenscheidt, Vieweg), or having the thickness of the binding represent the size of the underlying resource as for the different Langenscheidt Dictionaries. Another straight-forward mapping is provided by the degree to which dust has accumulated on the back of the books, ranging from a few dust particles to

[1] A prototype of the libViewer is available at http://www.ifs.tuwien.ac.at/ifs/research/ir/libViewer

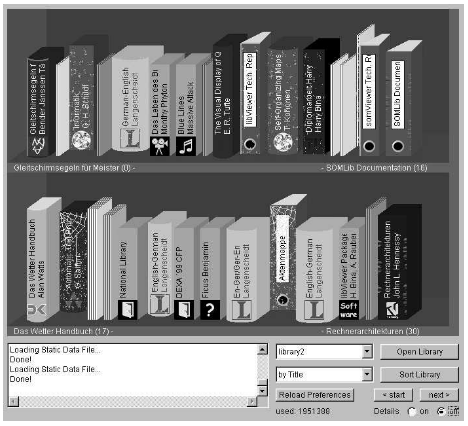

Figure 1. libViewer: Visualizing metadata of documents in a digital library

a spider-web covering half of a book that has not been referenced for a long time, as it is the case for the second book in the lower shelve. On the other hand, the last book in the lower shelve is clearly identified as being frequently referenced due to its rather distorted, well-thumbed binding indicating its frequent use. Furthermore, some books like the first one in the upper and lower shelve as well as most binders are not aligned with the backs of all the other books, indicating that they have been recently used and not been put back in place appropriately, making them stand out from all the other books and thus promoting easier ‘re-usage’. Contrary to that, some books like the third in the upper shelve or the second in the lower shelve have been pushed far into the back of the shelves due to not having been referenced recently. Albeit hardly noticable in the printed representation, we find a highlighting glare in the first book in the upper shelve, indicating the fact that it was added to the library only recently.

In order to produce a realistic visualization, a more sophisticated mapping of metadata attributes to metaphors was specified for some attributes in this library representation. Apart from the simple one-to-one mappings like resource type to representation type, resource size to representation size and publisher to logo, we chose a combined mapping of document type and publisher to e.g. define the color of some books in the library, such as the Langenscheidt Dictionaries, which have a yellow binding in the real world. Another example, in which case the resource type defines both the document type as well as the color, is represented by the mastertheses, which, apart from being hardcover books, usually can be easily identified by their black color and the logo of the University of Technology, being listed as the publisher.

Please note, that, although possible, it is not the goal of this system to represent a library as realistically as possible in terms of making all books look like their real world counterparts. Rather, we want to create a metaphorical representation which is optimized for exploration and intuitive understanding of document collections or search results. These mappings can differ for the specific information and explo-

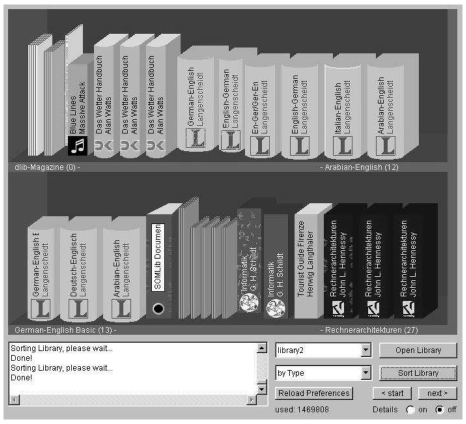

Figure 2. libViewer: Digital library representation sorted by resource type.

ration needs as well as for different information repositories. Thus, the mapping described in this example is just one out of many that are possible. For a different collection we might want to map the language of the documents to the color in the representation to clearly identify foreign language books. Another possibility would be to assign the colors of books based on their year of publication, making the various entries in e.g. a journal collection or news magazine archive intuitively visible even when they are not sorted by date. The alignment of books may be used to indicate the relevance of a certain item in the collection towards a query for the representation fo search results. However, care must be taken with the mapping definition in order to allow the graphical representations to actually serve as metaphors in the literal sense.

Apart from simply representing a static library, books can be sorted and arranged following different criteria. Figure 2 represents the same library sorted by publisher. Again, we can clearly identify the features discussed before, allowing us to judge which book contains how much information, when and how often it was used etc.

In order to provide more information on resources in the library that cannot be visualized using the metaphors available (e.g. short abstracts, keywords), detailed metadata information can be viewed in a separate details window as the mouse pointer moves across the items on the shelves. Furthermore, clicking on any of the items opens the resource if a link to the electronically available version is provided in the resource metadata entry and a viewer for the corresponding type of resource (file viewer, audio player etc.) is defined.

While the basic document styles implemented in the current version concentrate on text document representation providing a somewhat simplified representation for all other resource types, more sophisticated graphic styles can be defined for various types of audio, video and image files. Thus the library representation can be tailored to the requirements of any digital library collection. While the current implementation allows only low-level interaction with the library collection, more sophisticated methods for querying the library as well as displaying it at a lower level of detail from larger distance, allowing the user to get an instant overview over a large collection, are under investigation. Such a representation can then serve as a sophisticated interface to a document-cluster based information archive representation like the *SOMLib* digital library system.

6. Conclusions

We have presented a Web-based visualization tool for digital libraries. Contrary to most current approaches, large document collections, be it a library as such or the result of a query to a search engine, are not depicted as ranked lists of titles, with additional information like document type, author, size, etc. being provided as textual descriptions. Rather, well-known graphical metaphors are used to produce an intuitively understandable representation of the metadata of the documents. This type of information space visualization allows intuitive and straight forward analysis of large document collection, providing an ideal setting for interactive browsing and exploration. In connection with the document clustering provided by the *SOMLib* system, the libViewer can be used to provide user-centered visualization of digital information repositories and search results via the World Wide Web.

References

[1] H. Chernoff. The use of faces to represent points in k-dimensional space graphically. *Journal of the American Statistical Association*, (68):361–368, 1973.

[2] P. Cubaud, C. Thiria, and A. Topol. Experimenting a 3d interface for the access to a digital library. In *Proc. ACM Conf. on Digital Libraries (DL98)*, Pittsburgh, PA, 1998.

[3] U. Hahn, M. Klenner, and K. Schnattinger. Automatic concept acquisition from real-world texts. In *AAAI Spring Symp. on Machine Learning in Information Access*, Stanford, USA, 1996.

[4] M. Hearst. Using categories to provide context for full-text retrieval results. In *Proceedings of RIAO, Intelligent Multimedia Information Retrieval Systems and Managemen*, New York, NY, 1994.

[5] S. Kaski, T. Honkela, K. Lagus, and T. Kohonen. WEBSOM—self-organizing maps of document collections. In *Elsevir Publ*. Elsevir Publications, 1997.

[6] T. Kohonen. *Self-Organizing Maps*. Springer Verlag, Berlin, Germany, 1995.

[7] X. Lin, D. Soergel, and G. Marchionini. A self-organizing semantic map for information retrieval. In *Proc. Int'l ACM SIGIR Conf. on R & D in Information Retrieval*, Chicago, IL, 1991.

[8] D. Merkl. Exploration of text collections with hierarchical feature maps. In *Proc. of the 20th Int'l ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 186 – 195, Philadelphia. PA, 1997.

[9] D. Merkl and A. Rauber. CIA's view of the world and what neural networks learn from it: A comparison of geographical document space representation metaphors. In *Proc. 9th International Conf. on Database and Expert Systems (DEXA98)*, Vienna, Austria, 1998.

[10] A. Pejtersen. A library system for information retrieval based on cognitive task analysis and supported by an icon-based interface. In *Proc. of the ACM SIGIR Conference on Information Retrieval (SIGIR'89)*, 1998.

[11] A. Rauber. LabelSOM: On the labeling of self-organizing maps. In *Proc. International Joint Conference on Neural Networks*, Washington, DC, 1999.

[12] A. Rauber. SOMLib:A digital library system based on neural networks. In *Proc. ACM Conference on Digital Libraries*, Berkeley, CA, 1999.

[13] A. Rauber and D. Merkl. Creating an order in distributed digital libraries by integrating independent self-organizing maps. In *Proc. Int'l Conf. on Artificial Neural Networks (ICANN'98)*, Skövde, Sweden, 1998.

[14] G. Robertson, S. Card, and J. Mackinlay. Information visualization using 3d interactive animation. *Comm. of the ACM*, 36:57 – 71, April 1993. ACM.

[15] G. Salton, J. Allan, and C. Buckley. Approaches to passage retrieval in full text information systems. In *Proc. of the Int'l. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 49 – 58, Pittsburg, USA, 1993.

[16] E. Tufte. *The Visual Display of Quantitative Information*. Graphics Press, Connecticut, 1983.

[17] E. Tufte. *Envisioning Information*. Graphics Press, Connecticut, 1990.