# VISUALIZING ELECTRONIC DOCUMENT REPOSITORIES: DRAWING BOOKS AND PAPERS IN A DIGITAL LIBRARY

Andreas Rauber, Harald Bina
*Department of Software Technology, Vienna University of Technology*
*Favoritenstr. 9-11 / 188, A - 1040 Vienna, Austria*
http://www.ifs.tuwien.ac.at/ifs
{andi, harry}@ifs.tuwien.ac.at

**Abstract** While methods for retrieving documents from large information repositories have improved a lot, presentation of the retrieved documents still leaves a lot to be desired. Important information on documents is usually presented as a textual listing of available metadata attributes such as document size, author information, date of creation, and so on. This requires the user to read and abstract from the presented meta-information.

In this paper we present our *libViewer* system, a Java-Applet interfacing with a number of servers to provide an intuitive metaphor-graphics based representation of document repositories. Contrary to most other multidimensional data visualization approaches we rely on intuitive real-world metaphors to provide a visualization for untrained users rather than experts in special interfaces. We introduce a set of metaphors and present two prototype systems interfacing with Dublin Core metadata based repositories as well as the AltaVista search engine. We further provide a first usability evaluation based on comments obtained from users.

**Keywords:** Information Visualization, Metaphor Graphics, Digital Libraries, Metadata, Document Databases, User Interfaces, Usability

## 1. INTRODUCTION

Electronic document repositories have come a long way from the first file-listing based archives to modern database systems allowing complex query processing. Research in Information Retrieval further provided us with additional means to access and extract knowledge from these vast sources of information. However, with these resources opening up to

the so-called 'general public' by providing access via the Internet, more efforts have to be and are invested in providing access methods that are usable by untrained users rather than optimized for experts only.

One of the shortcomings of the current representation of document databases lies with the fact, that documents are commonly represented as sorted lists, providing additional information, such as date of creation, document size, author etc. as rather long textual descriptions. In order to decide whether a document is relevant or not, the user has to read these descriptions or use additional filtering and sorting criteria to extract the documents he or she considers most interesting. While this may be feasible for expert users, it proves rather cumbersome for non experts, who usually do not have a concept of which criteria to filter for. This is simply because filtering for document types such as 'journal' or 'hardcover books' or retrieving 'only documents with more than 100 pages' or, even worse, '200 KB' are not natural selection criteria in their known environment. Even analyzing the provided metadata to find out which concepts are available and which ranges they cover (time period, sizes, available document types) involves a lot of reading and interpretation.

On the other hand, taking a look at conventional libraries, we find a wealth of information to be conveyed by the physical representation of both the library and the books. On entering a library we find books sorted by content rather than by a specific relevance ranking criteria, allowing users to quickly identify both the topics covered by the library, the amount to which they are covered (based on the area assigned to specific topics) as well as to locate their section of interest if appropriate library maps and descriptions are provided. Within a shelve, by scanning the books sorted there, it is usually easy to tell the age of a book, the number of times it has been used before, 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 condition 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. What we want in the context of a document database is an intuitive graphical representation of the metadata usually provided only in textual form, which allows us to get an overview of the available information at one glance.

With the *libViewer* we present a tool for visualizing the documents contained in a document database based on the metadata provided by

the library system, to allow the user to gain an intuitive overview of the type and amount of information available. Contrary to other document repository visualization approaches we do not rely on abstract and dynamic mappings of concept spaces to abstract multidimensional visualization attributes. We rather favor the use of known concepts such as spatial location, physical representation, signs of intensive usage or dustiness, all well known from conventional library settings. Our main focus is to allow non-expert users to understand and feel familiar with a digital library system. While the resulting rather artistic representation may not satisfy expert users who, if trained on a system, usually prefer powerful computational features to fancy representations, non computer experts respond quite enthusiastic to the *libViewer* representation and consider it to be very helpful. The capabilities of our system are demonstrated on two prototypical applications, followed by an analysis of the feedback obtained from users.

The remainder of this paper is organized as follows: Section 2 presents an overview of metadata standards for digital document collections forming the basis for visualization. We next present the metaphors implemented in the *libViewer* system in Section 3, followed by a brief description of the *libViewer*'s client–server architecture in Section 4. Examples of the *libViewer* interfacing with two servers are presented in Section 5. Based on these experiments we present some usability evaluations as well as lessons learned for future modifications in Section 6. A comparison of our *libViewer* visualization with other approaches to document space representations is provided in Section 7, followed by some conclusions and an outlook on future work in Section 8.

## 2. THE INFORMATION: LIBRARY METADATA

As for all types of information repositories, information about the pieces of information stored in them is provided in terms of metadata. In conventional libraries we usually find library catalogues for the various metadata attributes, listing book titles, authors, printing date 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 designed by Oren Patashnik to create bibliographies in conjunction with the LaTeX document preparation system (Lamport, 1994). 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 and so on. 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.

There further exists a whole number of different standards for digital library metadata designed for special application arenas like CSDGM (Content Standard for Digital Geospatial Metadata), CIMI (Computer Interchange Format of Museum Information), CDWA (Categories for the Description of Works of Art) and many more.

One of the most promising newer 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 such as title, creator, subject, publisher, date of creation, etc. 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.

To switch to a completely different arena of metadata representation, we might consider the page descriptions returned from Internet search engines as another type of metadata specification. Although they very much differ in the style and extent they are provided by various search engines, we still can identify a number of attributes describing the various pages, such as title, author of a page, location (URL), date of creation, relevance towards a query, the size of the page etc.

Similar metadata is provided with document collections which come in the form of book stores on the Internet such as Amazon (`http://www.amazon.com`), which also provide meta-information about the books on sale. This usually includes, apart from the standard description of a book, store-specific data such as recommendations or prices, special offers and so on.

For any of the types of metadata an intuitive visual representation would provide the user with a possibility to obtain a better overview of the documents presented to her or him without being forced to actually *read* the metadata. The goal of the *libViewer* library visualization is to provide a graphical representation for the available types of metadata in a way that is instantly recognized and interpreted by users without

requiring special training or understanding of the concept of the underlying metadata.

## 3. THE VISUALIZATION: METAPHORS

In order to support this intuitive visualization, the *libViewer* provides a number of metaphors, which are easily identifiable and relate to properties known from the real world (Cole and Stewart, 1993). These metaphors, in accordance with (Tufte, 1990), are used to (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. 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. This set can be extended to cover new types of resources as the application domain requires.

- Color: Being a very dominant feature, color can be used to represent a variety of attributes in a very distinguishing way, such as language, publication series, genre, topical classification etc. However, the fact that it is an abstract rather than a metaphorical mapping has to be kept in mind.

- Size: The amount of information available in a book or magazine is intuitively judged from the size of the physical object by its spine width, e.g. the number of pages, thus measuring the amount of information available from a specific resource.

- 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. Thus, the format can be used in a variety of ways to further distinguish between a huge number of document types and genres by imitating reality.

- Logo: When browsing a library, one recognizes the logos of well-known publishers, associating them with special types of publica-

tions. Thus, while making the library representation look more realistic and rather decorating the books, a lot of information can be conveyed by having a company logo printed on the spine.

- Text: Although the amount of text found on the spine usually is limited to a few words, a wealth of information is provided by both the text, such as title or author listing, as well as the type of text representation, like different fonts or font colors.
- Highlighting Glares: 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.
- 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.
- Well-thumbed Bindings: 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.
- Spine Alignment: 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 by promoting easier picking.
- Location: We support the concept of an array of bookshelves in order to have, similar to conventional libraries, resources on identical topics located next to each other. This information can either be provided in terms of a classification attribute, or can be dynamically created based on content analysis as e.g. with the *SOMLib* digital library system. (Rauber and Merkl, 1999)

Based on these metaphors we can define a mapping of metadata attributes to be visualized, allowing the easy understanding of documents, similar to Chernoff faces for multidimensional space representation (Chernoff, 1973). 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 (Tufte, 1983). These mappings depend on the metadata available in the respective information repository and are taken care of by the respective servers.

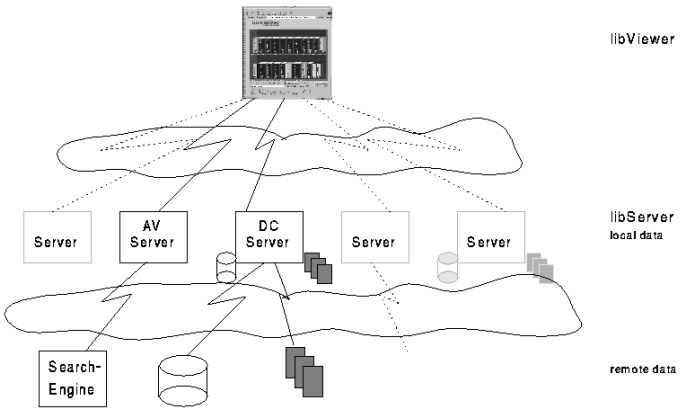


*Figure 1* *libViewer* Architecture: Applet connecting to a number of servers

## 4. LIBVIEWER ARCHITECTURE

The *libViewer* is a Java-Applet interfacing with a number of servers providing the data to be visualized. Its main task is to provide a library representation that is intuitively understandable by the untrained user by relying on concepts and metaphors taken from conventional, real world libraries. It relies on a server to provide the appropriate mapping from the metadata attributes available in a specific document repository onto (a subset of) the supported metaphors.

A conceptual view of this architecture is depicted in Figure 1. The *libViewer* applet contacts one out of a number of available servers. These servers provide access to various information repositories, which are available locally to them in terms of databases or library files. However, they can also rely on other servers to provide the information they need, actually serving as meta-servers. They retrieve the requested information from the appropriate source and provide a mapping of the available metadata onto a number of metaphors supported by the *libViewer*. This description is returned to the *libViewer* via a simple protocol. The *libViewer* in turn receives the metaphor-based description of the documents and uses it to create a graphical representation. This concept allows the *libViewer* to serve as a general interface to document repositories of all kinds, with the servers being responsible to provide appropriate mappings.

## 5. LIBVIEWER AT WORK

We currently have implemented two prototype servers for the *libViewer* system representing two different application domains. A preliminary version of the system is available online at `http://www.ifs.tuwien.ac.at/ifs/research/ir/somlib/libviewer.html` for interactive exploration.

The *DCServer* provides a mapping for a modified Dublin-Core (DC) based metadata set. With the DC set being designed specifically for digital document collections, it covers a broad range of types of metadata typically found in document databases. We furthermore extend the basic attribute set with a few attributes typically collected during library operations such as usage statistics for documents. This allows us to demonstrate the full capabilities of the *libviewer* representation.

The second server implemented so far provides a mapping of the metadata returned by the AltaVista search engine, allowing the libViewer to be used as an alternative and more intuitive interface. While being less extensive than the DC metadata set in terms of the available attributes, this setting, apart from making a prominent application, allows a straightforward comparison of conventional forms of representation with the enhanced visualization of the meta information.

### 5.1. DCSERVER: VISUALIZING THE DUBLIN CORE

The *DCServer* provides a mapping of Dublin-Core based metadata onto the *libViewer* metaphors. While we are currently concentrating on the document-oriented subset of the Dublin Core system, we have extended the basic Dublin-Core metadata set by additional attributes that are typically collected during library operation. These attributes include the number of times a specific document has been referenced, or the date when it has last been referenced. It allows us to demonstrate most of the features available with the *libViewer* within one application.

Figure 2 depicts a representation of a digital library file as created by the *DCServer*. A number of different document types such as hardcover books, paperbacks, technical reports and papers can be easily identified as their corresponding physical representations, such as 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 con-

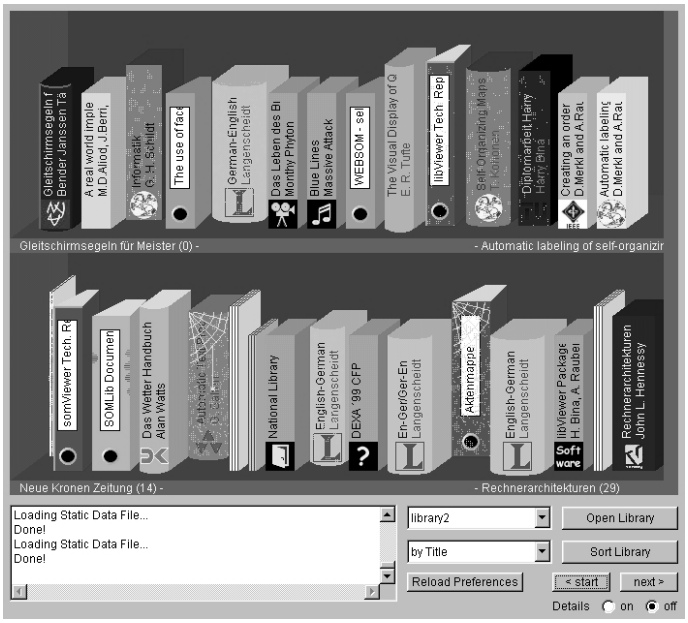


*Figure 2* *libViewer*: Visualizing DC metadata of documents in a digital library

ference 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. While theoretically a mapping for the whole hierarchy of the Dublin-Core specified document types can be created that way, evaluation has shown, that a rather high-level mapping of metaphors suffices and even enhances intuitivity, since many of the subdivisions available in the metadata are rather unimportant to users looking for specific information.

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, ieee), or having the thickness of the binding represent the size of the underlying resource as e.g. 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 a spider-web covering half of a book that has not been referenced for a long time, as it is the case for

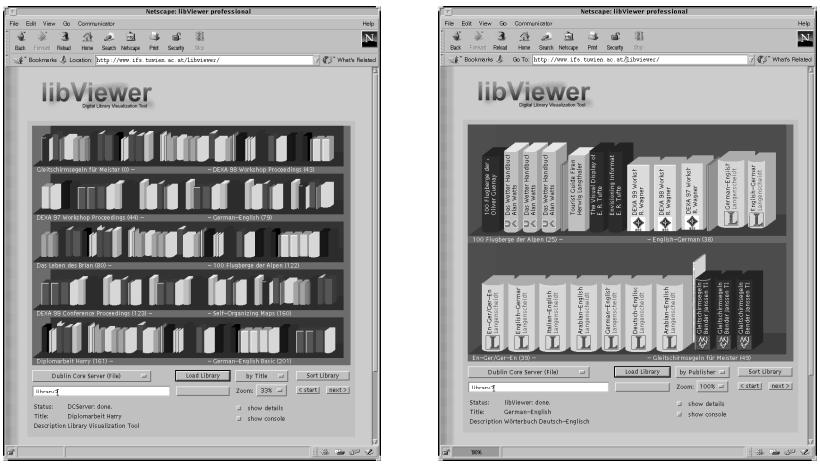


*Figure 3* libViewer: (a) Overview Representation (b) Sorted by Publisher

the fifth book in the lower shelve. On the other hand, the third book in the lower shelve is clearly identified as being frequently referenced due to its rather distorted, well-thumbed binding indicating its frequent use. Albeit hardly noticeable in the printed representation, we find a highlighting glare in the first book in the upper shelve, indicating – similar to shiny new books in libraries – the fact that it was added to the library only recently.

Furthermore, some books like the first ones in the upper shelve as well as most binders are not aligned with the backs of all the other books, making them stand out and thus promoting easier picking. 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. The alignment can thus be used to indicate some kind of relevance recommendation or, with respect to electronic book stores, indicate promotions.

In order to obtain an overview of the documents present in the library, a look from the distance is provided in Figure 3a. At this level, only the most dominant attributes are displayed, such as document type, thickness of the binding and the color, whereas the more detailed representations such as dustiness, text on the binding etc. are – similar to conventional libraries – only visible in the close-up representation.

The documents in the library can be sorted based on the metadata available. Figure 3b provides a view of the same library, again at the detailed representation level, sorted by publisher. Additional information is provided by the label in the shelve, which in this case gives the title of the first and last document in the row. Again, this is basically decided

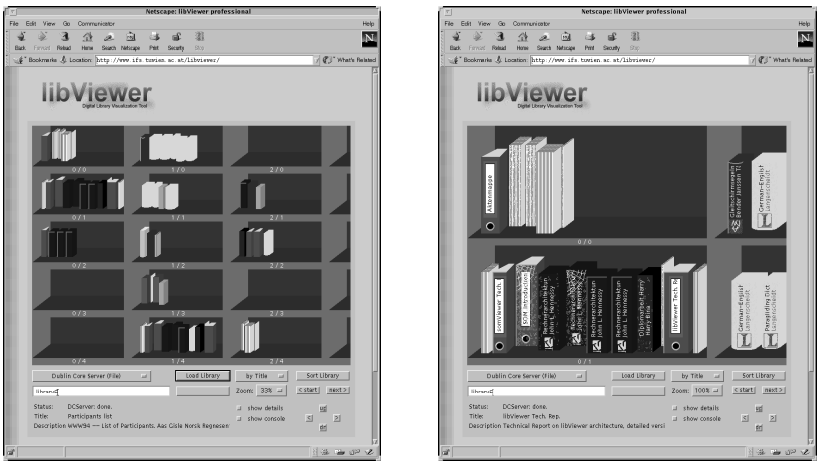


*Figure 4* libViewer: (a) Documents sorted into shelves (b) Close-up view of shelves

upon by the server, although it can be modified by the *libViewer* based on the sorting criteria.

If the server provides a classification of documents, a mored advanced shelve-representation is provided by the *libViewer* as depicted in Figure 4a. Documents are organized into different shelves, with some information on the topic being provided as shelve labels. The detailed view of these shelves again is provided in Figure 4b. In this example, we find the shelve position in terms of rows and columns printed as shelve labels, allowing the users to orient themselves. Again, the most appropriate information, such as topical labels for the shelves, should be selected by the server.

## 5.2. AVSERVER: VISUALIZING SEARCH RESULTS

The *AVServer* serves as a more intuitive visualization for the search results returned by the AltaVista search engine. With the metadata returned by search engines being both different and less detailed than the Dublin Core, it provides a good proof of the strengths of even a limited set of the *libViewer* metaphors.

A query is entered via the *libViewer* interface, passed on to the *AVServer*, which in turn forwards the query to the AltaVista search engine. It then automatically retrieves the first 10 result pages, i.e. 100 results found by AltaVista and provides a mapping of the available metadata. Based on the standard result pages from AltaVista we can extract

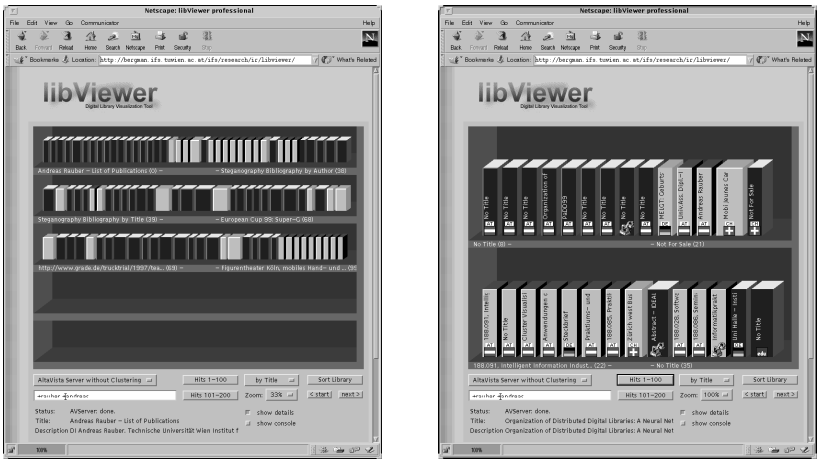


*Figure 5* AVServer Mapping: (a) 100 hits returned by AltaVista (b) Close-up view of results

a number of attributes including the title of a document, the location (URL) and thus the domain it is located in, the size of the document, the time it was created, and the language it is written in. Furthermore, a short description is provided by AltaVista based on the first few lines of the document. These attributes are now mapped onto the appropriate *libViewer* metaphors by the *AVServer*, such as the document size being mapped onto the spine width of the document, the title of the HTML-page being printed on the spine, together with a logo representing the domain of a document, the document language being mapped onto the color and so on.

A representation of the query result for *+rauber +andreas* is presented in Figure 5a. The first 100 results returned by AltaVista are depicted in the order they were returned by the query engine. The features represented at the small-scale representation allow us to obtain a quick overview of the 100 result pages. The titles of the first and the last document in each shelve are given as shelve labels as well as the according document number. Language, as far as returned by AltaVista, is encoded in terms of color, with the blue-red-white documents indicating english-language documents, gold-black-yellow indicating german language documents. Documents for which AltaVista did not return a language identification are colored grey. We also find the different document lengths indicated by the width of the spine to be easily discernible. Additionally, even at that distance, we can easily differentiate between

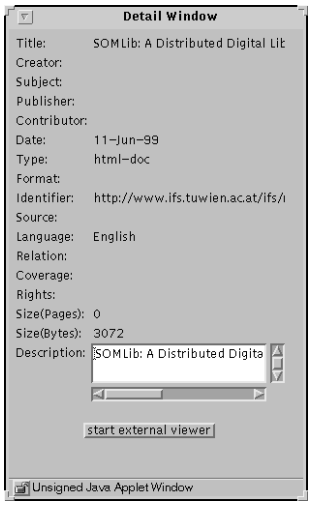


*Figure 6* Details Window: Detailed textual representation of metadata

documents that have been created recently and thus have a highlighting glare as opposed to older documents.

As the mouse pointer moves across the shelves, more information on the respective pages is depicted both in the status bar, where the title and the document description is listed, and in a details window as presented in Figure 6, where all the metadata available on the respective document is listed.

Clicking on the middle of the first shelve gets us to the close-up representation depicted in Figure 5b. Here we can now obtain more detailed information on the documents. The title of each page, if available, is printed on the spine. The domain where a document is located is indicated by a logo on the spine, with the nations flag serving as logos for national domains, such as the austrian, german or swiss flag and labels serving as logos for domains such as *com, edu, org* as e.g. for the last document in the lower shelve. If no logo is available for a given domain an unknown-domain logo is mapped onto the spine, as for the eigth document in the upper shelve, which happens to be located in Hong Kong. Clicking on a document or on the button in the details window opens the appropriate document in another browser.

## 6. USABILITY EVALUATION

Following the first prototype implementations we presented the system to a number of users, consisting of librarians, computer scientists

and non-computer-science students, to obtain some feedback on their view of its usability. Most of them were immediately fascinated by the graphical representation, and the metaphors turned out to be, for the most, rather self-explanatory.

While at the beginning we intended to provide a rather sophisticated mapping of attributes and combinations of attributes to metaphors, we found, that rather simple mappings usually suffice to provide the information necessary. One of the initial goals was to make the books in the library as much as possible resemble their real-world counterparts. This lead to — apart from sophisticatedly mapped logos of publishers — combinatorial mappings of, e.g., document type and publisher onto color. For example, dictionaries, if they were published by Langenscheidt, were colored yellow, whereas dictionaries by e.g Pons were colored green. However, while this very real-world like representation allowed users to easily recognize the books and to spot and name them correctly even in large collections, in turned out to be unnecessarily complex for more general applications.

As the most helpful features we found to be, apart from the obvious document type representations, the size of the document, which is much easier to be told from the visualization than from the textual representation. However, especially with the AltaVista Server, we found the differences in document length for small documents to be too little discernible on some queries, calling for a somewhat more adaptive mapping from document size to spine width employing a logarithmic mapping function. It turned out that this also corresponds to some extent to the real world, where for most larger documents, such as books, a thinner type of paper is used, thus not leading to a linear increase of spine width with respect to the number of pages.

The highlighting glare used to identify new documents turned out to be another very prominent feature, although some users did not notice this metaphor when they first came across it, suggesting to make it more dominant.

We obtained some mixed reports on the mapping of language onto color in the AltaVista Server, which some people found to be perfectly intuitive whereas others were somewhat irritated at the beginning. However, all of them interpreted that metaphor perfectly alright after a short time, and it turned out to be one of the most helpful features for telling relevant documents apart.

The country flag as logo helped a lot to locate the documents people were looking for, especially if the query was not formulated very precise, and when people were looking for information they knew was available in a specific country. Complaints were of course filed when no flag for a

specific country was yet available and we keep working on improving that list. Some (computer-literate) people noted, that the country flags were too dominant as opposed to the labels for non-national domains such as specifically '.com', '.org' and '.net', which may contain pages from national sources, yet people might fail to identify them simply because they are not listed with the appropriate flag.

The classification of documents into separate shelves turned out to be very helpful in terms of segmenting a larger number of books. Although we have so far not included the automatic classification of documents returned by AltaVista as part of the SOMLIB digital library system (Rauber, 1999), simply the segmentation into smaller chunks as well as the more realistic representation of the library in the small-scale representation found promising response in terms of usability.

As people kept interacting with the system, some new features that should be included were listed. Sorting the books in different ways already helped people a lot once they identified the feature that was most relevant to them, such as the domain. However, especially for larger numbers of books, filtering unwanted book types or domains would have been helpful.

One of the issues raised by some users concerned the orientation of books. Although the text on the spine consists of only a few words, which generally can be read without much effort and without actually turning ones head, some argued, that a pile of books might be easier to read than the vertical shelve-position with the text being given in horizontal position rather than vertical. In fact, the shelve position of books is only required in the real world where gravity prevents you from simply picking a book out of the bottom of a pile. With digital libraries, having the books piled on top of each other as several piles rather then sorted into a shelve would not provide this type of problems. Still, the effect of re-orienting the books has to be evaluated as some people opposed to it, saying they might feel uncomfortable with such a representation, as books 'hovering in space' are unnatural. This orientation–question needs to be analyzed in more detail.

With respect to the graphical representation, some users wanted a 3-dimensional view of the library allowing them to actually move through the library and pick up books. Still, as most users did not have any special 3-d plug-ins available at their systems, not to mention special 3-d viewing devices, they felt sufficiently satisfied with the 2-dimensional representation, especially as the 3-d effects in terms of book and shelve representation were considered more than realistic enough to create a 3-dimensional impression. However, it definitely merits further consideration.

## 7. RELATED WORK

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. In this Section we provide pointers to research work addressing a variety of issues in information space visualization, ranging from content-related approaches via information organization to document visualization as such.

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 (Hahn et al., 1996, Hearst, 1994, Salton et al., 1993). Apart from document retrieval, systems analyzing a set of documents to provide question answering are emerging (Aliod et al., 1998), which try to analyze the semantics of a question and try to create an answer based on the information stored in an underlying document collection.

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.

To overcome the basic limitations of the one-dimensional ranked-list representations of most search engines, we developed the *SOMLib* digital library system (Rauber and Merkl, 1998; Rauber, 1999; Rauber and Merkl, 1999), a 2-dimensional map display which automatically organizes a set of documents by their contents (available at `http://www.ifs.tuwien.ac.at/ifs/research/ir/somlib`). The self-organizing map (SOM) (Kohonen, 1995), 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 (Kaski et al., 1997; Lin et al., 1991; Merkl, 1997). 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, which is similar to real-world library organization. This capability makes the *SOMLib* represen-

tation particularly useful in digital library exploration. However, in spite of the 2-dimensional topical clustering, no further meta-information on the documents can be extracted from the standard *SOMLib* Web interface.

Another map-based representation of documents is provided by the Nemo project (Hascoët and Soinard, 1998), showing the main attributes of a set of documents as icons of different color, patterns and text, however without support for automatically organizing the documents according to their content. Still, the visualization must be viewed rather in the perspective of multidimensional information visualization, not focusing on intuitively interpretable real-world metaphors.

An approach for visualizing the contents of texts is presented in (Rohrer et al., 1998), where the main concepts of a text are used to span a multidimensional shape which is rendered to form 3-dimensional shapes, allowing the detection of documents on similar topics as documents exhibiting a similar shape.

One of the first applications of metaphors in the digital library arena is reported in the Bookhouse project (Pejtersen, 1989), 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. However, apart from metaphors for different interaction mechanisms, no visual representation for various types of documents and for available metadata are created, making the *libViewer* a prominent complementary system.

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 (Robertson et al., 1993). Information is depicted in a 3-dimensional space with the focus being on the amount of information visible at one time and an easily understandable way of moving through large information spaces, focusing on the visualization of the content rather than the metadata of documents.

At the CNAM library, a virtual reality system is being designed for the visualization of the antiquarian Sartiaux Collection (Cubaud et al., 1998) The binding of each book is being scanned and mapped into a virtual 3-dimensional library to allow the user to experience the collection as realistic as possible. Here, the purpose is not so much to provide intuitive access to information, but rather to allow the user to experience the real collection in a virtual setting.

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. Thus, they still have to rely on textual metadata to present important information on the documents. This obviously calls for a combination of the various approaches in order to obtain a digital library usable by everybody instead of experts only.

## 8. CONCLUSIONS

We have presented the *libViewer* system as a general interface to electronic document collections. It provides an intuitive graphical representation of documents based on the metadata available. Instead of reading textual descriptions of documents, their types, sizes, age etc., metaphor graphics are used to convey this information in a self-explanatory way. As a Java Applet the *libViewer* may contact a number of servers which provide the mapping of metadata available in the repository they serve onto the metaphors supported by the *libViewer* client. We have demonstrated the capabilities of our system using two servers. The *DCServer* connects with Dublin-Core based library description files and provides mappings demonstrating the wealth of information that can be conveyed to the user in graphical form. The *AVServer* provides a mapping from the results returned by AltaVista allowing users to get an improved, intuitive visual representation of the documents found.

Both systems have shown to be very helpful in initial usability evaluations and the graphical representations were highly appreciated. Some issues raised during these evaluations, such as the orientation of books or a different calculation of the spine width to allow for better separation between small and large documents, are currently under investigation. While initially being conceived solely as a representation system, additional functionality requested by users will make the *libViewer* evolve into an application supporting more flexible query interfaces as well as additional interaction and result modification facilities.

Following the promising results of our initial evaluations with the *DCServer* and the *AVServer* we now plan to implement servers to some other widely used document repositories. These shall serve as a broader basis for more advanced usability evaluations.

## References

Aliod, M., Berri, J., and Hess, M. (1998). A real world implementation of answer extraction. In *Proc. Workshop on Natural Language and Information Systems (NLIS'98) in Conjunction with DEXA '98*, Vienna, Austria.

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

Cole, W. and Stewart, J. (1993). Metaphor graphics to support integrated decision making with respiratory data. *Intl. Journal of Clinical Monitoring and Computing*, 10:91–100.

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

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

Hascoët, M. and Soinard, X. (1998). Using maps as a user interface to a digital library. In *Proc. Int'l ACM SIGIR Conference on Research and Development in Information Retrieval*, Melbourne, Australia.

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

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

Kohonen, T. (1995). *Self-Organizing Maps*. Springer Verlag, Berlin, Germany.

Lamport, L. (1994). *LatTex*. Addison-Wesley, USA.

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

Merkl, D. (1997). Exploration of Text Collections with Hierarchical Feature Maps. In *Proc. of the ACM SIGIR Conference on Research and Developement in Information Retrieval (SIGIR'97)*, Philadelphia, PA.

Pejtersen, A. (1989). 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 Research and Developement in Information Retrieval (SIGIR'89)*.

Rauber, A. and Merkl, D. (1998). 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.

Rauber, A. (1999). SOMLib: A Digital Library System Based on Neural Networks. In *Proc. ACM Conference on Digital Libraries (ACMDL'99)*, Berkeley, CA.

Rauber, A. and Merkl, D. (1999). The SOMLib Digital Library System. In *Proc. European Conference on Digital Library Systems*, Paris, France. LNCS, Springer Verlag.

Robertson, G., Card, S., and Mackinlay, J. (1993). Information visualization using 3d interactive animation. *Communications of the ACM*, 36:57 – 71.

Rohrer, R., Ebert, D., and Sibert, J. (1998). The shape of shakespeare: Visualizing text using implicit surfaces. In *IEEE Symposium on Information Visualization (INFOVIS'98)*, North Carolina.

Salton, G., Allan, J., and Buckley, C. (1993). 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.

Tufte, E. (1983). *The Visual Display of Quantitative Information.* Graphics Press, Connecticut.

Tufte, E. (1990). *Envisioning Information.* Graphics Press, Connecticut.

## Biographies

**Andreas Rauber** is Research Assistant at the Department of Software Technology at the Vienna University of Technology. He received his MSc at the Vienna University of Technology in 1997 and is since involved in research in the fields of Data Mining, Information Visualization, and Digital Libraries.

**Harald Bina** is a Graduate Student at the Department of Software Technology at the Vienna University of Technology. His research is focused on Information Visualization and Usability Evaluation.

Andreas Rauber, Harald Bina: **Visualizing Electronic Document Repositories: Drawing Books and Papers in a Digital Library**. In: Proceedings of the 5. IFIP 2.6 Working Conference on Visual Database Systems (VDB5), May 10. - 12. 2000, Fukuoka, Japan, Advances in Visual Information Management - Visual Database Systems, pp. 95 - 114, Kluwer Academic Publishers, Norwell, MA.