Browsing a Digital Library: A New Approach for the New Zealand Digital Library

Dana McKay and Sally Jo Cunningham

Department of Computer Science
University of Waikato
Private Bag 3105, Hamilton, New Zealand
{dana,sallyjo}@cs.waikato.ac.nz

Abstract
Browsing is part of the information seeking process, used when information needs are ill-defined or unspecific. Browsing and searching are often interleaved during information seeking to accommodate changing awareness of information needs. Digital Libraries often support full-text search, but are not so helpful in supporting browsing. Described here is a novel browsing system created for the Greenstone software used by the New Zealand Digital Library that supports users in a more natural approach to the information seeking process.

1 Introduction

Browsing is a vital part of the information seeking process, allowing information seekers to meet ill-defined information needs and find new information [16,19]. Despite the importance of browsing in information seeking, however, it is relatively unsupported in many information systems [13]. The Greenstone digital library software [1] created by the New Zealand Digital Library research group [2] is an example of software that does support browsing, though to a limited extent. Greenstone is used by numerous organizations worldwide to manage and present collections of documents.

The aim of the work presented here was to create a new browsing system within Greenstone that had the flexibility to allow users to follow a more natural information seeking process. The resulting system allows users to specify parameters such as the metadata by which they wish to browse and the maximum number of documents on a page. It also allows the user to combine searching and browsing activities.

Section 2 of this paper discusses the information seeking process: examining what browsing is, why it is important and how information systems can best support it. Section 3 describes Greenstone's current browsing capabilities and considers their weak points. Section 4 presents an overview of a new browsing system and Section 5 describes an evaluation of the new system. Section 6 draws some conclusions about this work.

2 Human Information Seeking

Human information seeking behaviour is more than full text search, or wandering among the shelves in a library. The information seeking process begins with the conception of a need for information, and (if successful) ends with the satisfaction of the information seeker that they have the information they require [16]. Section 2.1 will examine more closely the information seeking process and the role that browsing plays; Section 2.2 will examine the implications of this process for information systems and their users.

2.1 The Information Seeking Process

Traditional information retrieval models have tended to marginalise human behaviour and focus almost entirely on querying using structured languages. It is on these methods that the standard (recall and precision) measures of a system's success are based, and there is a large research literature about making searching more effective [17,27].

Searching may be most of information *retrieval*, but it is not all of information *seeking* [10]. Information seeking is a process that has been studied and broken down numerous ways (see [7,16,19] for examples of this work). These models have their differences, but there are also some striking similarities. All the models begin with the user perceiving their need for information (though they may not know how to express what it is they need [23]). The models, representative of an ideal world, all end with the satisfaction of this need (though in reality users often "satisfice" [3], or simply give up [23]). All the models also describe a stage ideally suited to browsing, where the information seeker knows they have a need for information but may not know how to fulfil that need. This stage is variously called source selection [18], exploration [16] and browsing [7].

While browsing is one part of the information seeking process, it is not all of it. Moreover, the process is not necessarily linear; users generally seek information in an iterative manner switching back and forth between stages [5,18,27], particularly searching and browsing.

2.2 Information Seeking Interfaces

To be as useful as possible, information seeking interfaces should support the activities of information seekers as naturally as possible. This means supporting all stages of the information seeking process, not just searching. Despite research having long emphasized that browsing is a fundamental information seeking activity (Bates described this is 1989 [6]), many systems still do not support it [13,25].

Browsing can be supported by many different facilities, including semantic browsing using such tools as self organising maps [9,26] or phrases [24], metadata based browsing like the Greenstone classifier system [4], and subject categorisation (for example the Library of Congress classification scheme). Browsing may be within a document (for example leafing through a book) or between documents (for example wandering the library shelves). Browsing may occur for a number of reasons, including evaluation of an information source, information discovery, and clarification of an information problem [8,18,25]. A common definition of browsing is an exploratory information seeking strategy relying heavily on serendipity and being used to meet an ill-defined information need [6,8,22]

One way that conventional libraries support browsing is through subject classification of documents. However, physical libraries cannot rearrange the shelves at whim to meet the needs of the user (say, if they wanted to browse by author and they changed suddenly to title). Electronic information systems (such as digital libraries) have the opportunity to "rearrange the shelves".

For a system to support browsing effectively, and add something to conventional physical libraries, it must be flexible, to allow the user to modify their information need and information seeking strategy at will. It should support browsing for any number of reasons, including those mentioned above. For optimum information seeking effectiveness interleaving of browsing and searching should ideally be simple [11,13].

Browsing is easily shown to be a vital part of the information seeking process, and very effective when combined with searching. Information systems need to recognise this importance and support browsing in ways that will allow users to become effective information seekers. The system described in this paper is a metadata-based system that allows users to configure their browsing structures and combine them with searching, thus giving the flexibility recommended in the literature.

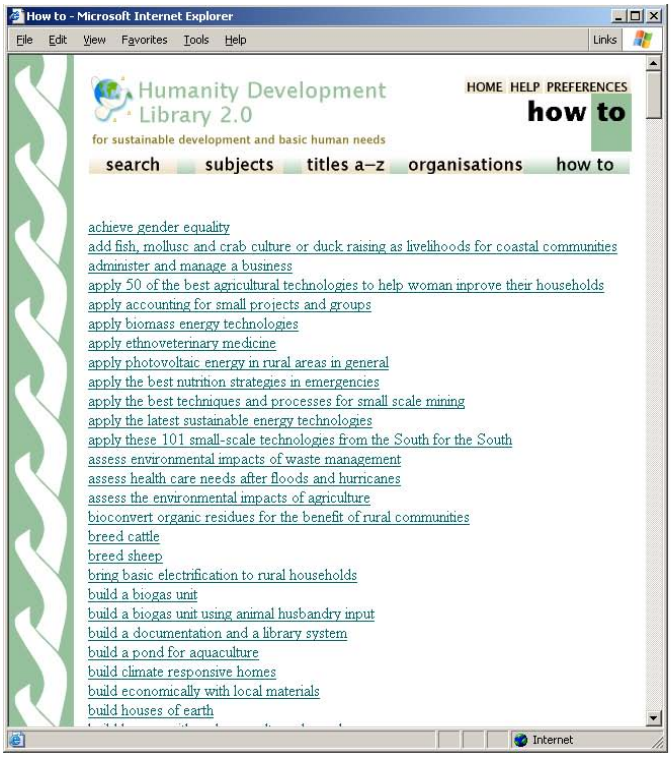
3 Greenstone and Browsing

Greenstone is a complete digital library management system, handling everything from collection building to collection presentation via a web browser. It facilitates full-text and metadata searching, and various kinds of browsing [1,4]. Greenstone is designed to allow collections to be built fully automatically (that is, to not require the manual processing of source documents) and served by inexpensive machines over a slow internet connection [28]. Greenstone is largely stateless, not keeping information about what users do from one action to the next, so as to help reduce server load. Section 3.1 discusses the current browsing facilities available in Greenstone and Section 3.2 explains why these facilities inadequately meet information seekers' needs.

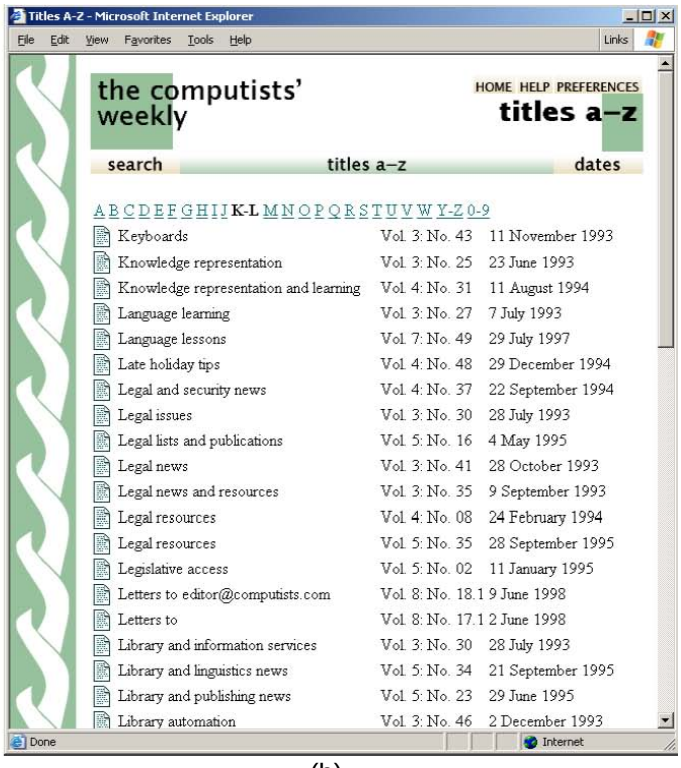
3.1 Greenstone's Current Browsing System

Greenstone's current browsing system is known as the "classifier" system. This is because documents are classified at collection build time according to their metadata, and browsing structures are pre-built ready for loading. Greenstone supports a number of different types of classifier, each suited to a specific kind of metadata. Each classifier displays information in its own way.

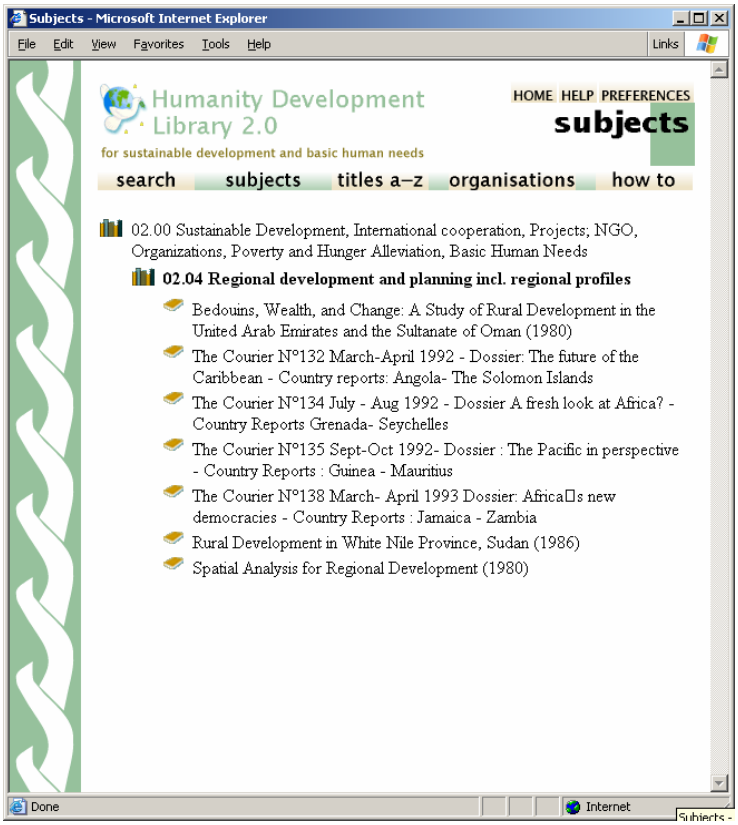
There are five main types of classifier currently implemented in Greenstone: the list, the alphabetic classifier, the hierarchic classifier, the date classifier [4] and a phrase-based classifier called "Phind" [24]



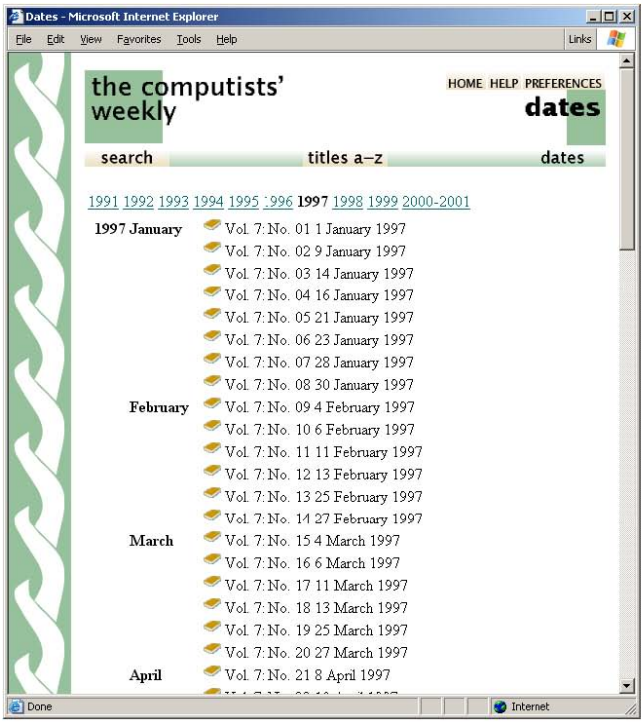
(a)



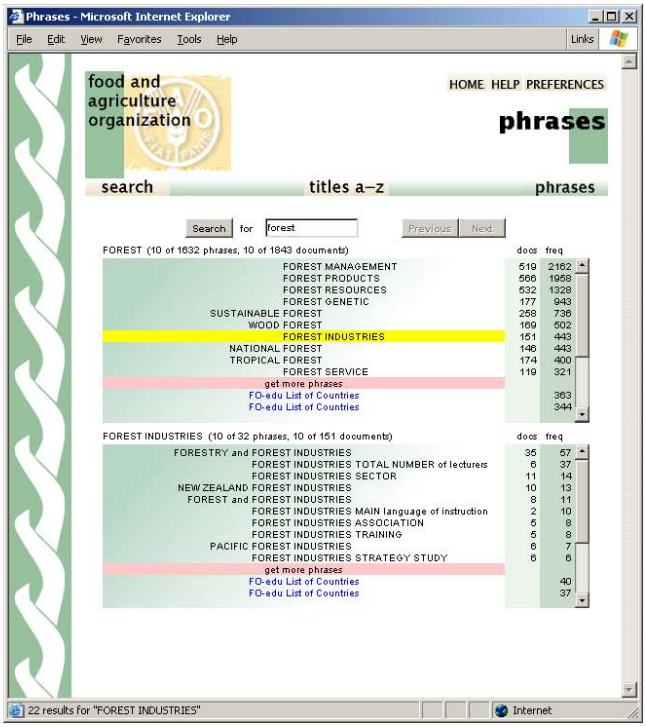
(b)



(c)



(d)



(e)

Fig. 1. The classifiers in Greenstone.
(a) Shows a list classifier of "How to" metadata.
(b) Shows an alphabetic classifier, viewed by title. The section being viewed is "K-L".
(c) Shows a hierarchic classifier. The classification being viewed "02.04" is two levels deep.
(d) A date classifier. Note the months down the side of the page.
(e) The "Phind" classifier. The word "forest" is being browsed.

The list classifier is the simplest of the classifiers; it merely sorts metadata alphabetically, and presents documents in a single long list (see Figure 1a).

The alphabetic classifier also sorts documents alphabetically, but the document list is then divided up into preset classes according to initial letter, and the classes are displayed across the top of the page (see Figure 1b). If the classes are smaller than a pre-set size the classifier will merge them (for example, 'K-L' in Figure 1b). There is no limit on the number of documents in a class.

The hierarchic classifier deals with numerical hierarchies—documents are assigned a number indicating their position in the hierarchy (much like the Dewey decimal system), and the user views the hierarchies by progressive drill-down clicking (see Figure 1c).

The date classifier is very much like the alphabetic classifier, though it uses the year as a basic unit, (as opposed to initial letter) and it displays month information down the side of the hierarchy (see Figure 1d).

The Phind classifier is not based on traditional metadata. Instead, it creates an index of phrases when the collection is built, and allows the user to browse by entering a single word or phrase and drilling down through phrases to documents (see Figure 1e).

3.2 Problems with Browsing Using the Classifiers

The classifier system in Greenstone does not support users as well as it might. The failings are in two major areas—the fact that uses cannot combine searching and browsing, and in the rigidity of the system.

As discussed in Section 2, users locate information most effectively when they can switch easily between searching and browsing. Search results in Greenstone are currently always displayed in lists, and if a search is not ranked, these lists are unsorted. Classifiers present all the documents in a collection that have the classification metadata; there is no way to search a classifier. Thus the cognitive cost of switching between searching and browsing is high, reducing information seeking effectiveness.

The rigidity of the classifier system is built-in—each classifier uses static, pre-built browsing structures, thus allowing collections to be presented only in one predetermined manner without input from the user. To illustrate how this can become a problem, imagine a collection with a number of distinct documents with the same title and different authors. The user cannot specify that they would also like to see the author metadata when browsing, much less insist upon the documents being sorted by author. Another example of how the rigidity of the classification system is detrimental to the information seeker's experience is the size of the groups displayed. Users are better able to navigate and evaluate options if they do not have to scroll [21]; yet in medium-sized collections (say 1,000 documents) users may have to scroll through three screens on a single classification, and there is no way for users to specify the largest number of documents they wish to see on a page.

The Phind classifier solves the rigidity problem, but allows browsing of only a single kind of metadata (phrases), and still does not allow collections to be filtered by search terms—and thus does not entirely solve the browsing problem.

Greenstone supports browsing in a limited way: non-searchable, static metadata classifiers. While this approach goes some way towards supporting browsing, it hinders users in their information seeking by not allowing them the flexibility necessary for truly effective information seeking. Moreover, Greenstone has strong goals relating to

usability, utility and simplicity of collection creation. The work describes here is an attempt to overcome the failings in Greenstone while still taking its goals of simple collection provision on inexpensive hardware into account.

4 A New Browsing System for Greenstone

This project focussed on a between-documents metadata-based browsing system. This approach was chosen because Greenstone is about presenting collections of documents (rather than single documents) and Greenstone already has an effective semantic browsing system in Phind [24]. The user capabilities the new system was to support were defined with the failings of the current system and the research on human information seeking behaviour in mind. These capabilities are as follows: users must be able to combine searching and browsing, users must be able to choose the metadata by which they browse, users should be able to browse by more than one kind of metadata at a time, and users should be able to restrict the amount of information on any one screen. The guiding principle is to give the user the richest possible browsing experience.

The new browsing system is designed to provide a rich browsing experience without being too taxing on the user—allowing users flexibility in specifying how they wish to browse, while providing a simple interface with sensible defaults. This also involved creating the system to handle alphanumeric and date metadata.

One of the major advantages of this system over the existing classifier system is that searching and browsing can be easily combined. The search offered by this interface is functionally identical to the ordinary Greenstone search, but results are presented in a browsing structure defined by the user. To avoid the loss of the useful ranking information provided by Greenstone's underlying search technology MG [29], where the search is an "any words" search, rank information is displayed next to the document metadata, similar to search engine results (see Figure 2b). If the user does not enter any search terms, then the user can browse the whole collection (see Figure 2a).

The mechanism for specifying how documents are to be browsed must allow great flexibility, but it also must to be simple enough to use without training. To that end, the user is presented with familiar web-based controls and simple language to determine their browsing preferences. They may browse one or two kinds of metadata at a time (the lists of metadata available for browsing are requested from the collection, and inserted into the interface when the browse page is displayed), and they may specify how many documents they wish to see on a page.

Because Greenstone is stateless and combining all this information to form a browsing structure is computationally expensive, the browsing structures are created only once, and the classes that are not being currently viewed are hidden in the page using dynamic HTML and JavaScript. This means that the user will get an instantaneous response when switching between the classes in a browsing structure.

Browsing more than one kind of metadata at a time allows the user to view distinguishing metadata where the primary browsing metadata values occur more than once (for example many books with the same author but different titles, when browsing first by author). It also allows the user to sort the duplicates by the second piece of metadata, (so sorting the books by author, and then sorting the books with the same author by title). Both pieces of metadata are displayed for each document, even where documents may only have one of the two pieces (see Figure 2a). Documents without the first piece of classification metadata are slotted into the browsing structure under the label "no metadata available".

The system accommodates the “number-of-documents-per-page” option by creating a two-level hierarchy, with basic divisions (for example initial letter in the first level of the hierarchy), and the second level of the hierarchy divided up so as to provide the required number of documents in each class. The classes at this level are labelled such that they are distinct from neighbouring classes (for example, if the first document in one class is called “teak chests” and the first document in the next class is called “teas of the world”, the classes will be labelled ‘teak-’ and ‘teas-’ respectively). See Figure 2b for an example of such a browsing situation.

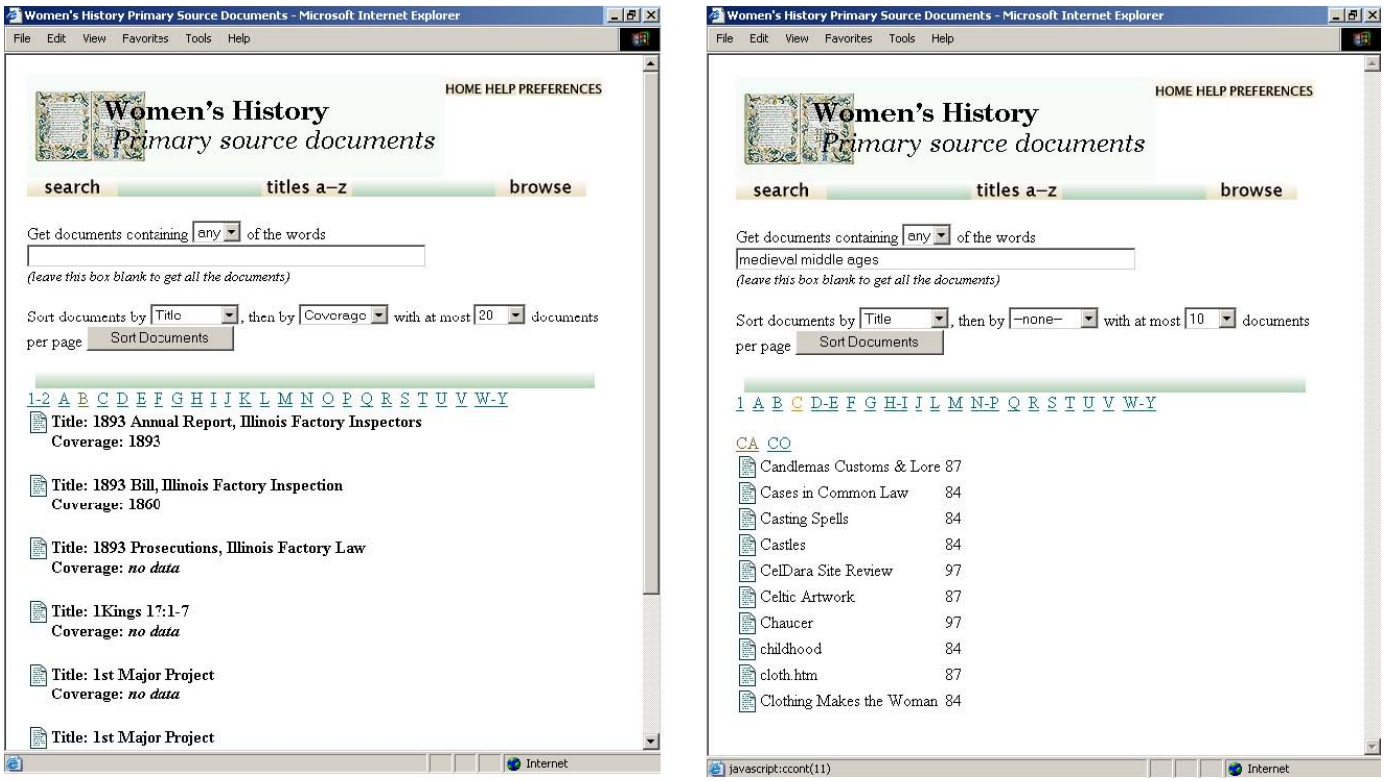


Fig. 2. The new browsing system.
(a) Shows browsing of a whole collection by two pieces of metadata.
(b) Shows browsing of ranked search results by title metadata. Note the ranking information on the right and the second level of the hierarchy

Users rarely change the defaults on information seeking interfaces; therefore while this interface offers a lot of flexibility it must also have sensible defaults [14]. By default, the entire collection is displayed for browsing by title metadata; this default is with a view to giving the user a good overview of the system. The default second piece of metadata by which to browse is determined by whatever metadata the collection has—it is hard to tell automatically what will be useful for any given collection, so the default second piece of metadata is the first detected piece of metadata that is not the title. The default number of documents per page is 20—this is approximately one screen-full, so as to avoid wasted screen real-estate, but also to lessen the need for scrolling. Searching defaults to “any” words, as this is less likely to give a “no match” result and therefore less likely to frustrate the user [14].

The new browsing system has been designed to offer flexibility in a very simple manner. The major advantages it has over Greenstone's existing classifier system are the combination of searching and browsing, and the ability to interactively change browsing structures to meet changing information needs.

5 Evaluation

There were two main components to the evaluation of this system: a technical evaluation (in Section 5.1) and a user study (in Section 5.2).

5.1 Technical Evaluation

Systems must be technically sound to be useful. There are two aspects of the new browsing system that can be meaningfully evaluated: scalability and time constraints.

The new browsing system attempts to address the scalability issues faced by the old system (i.e. browsing lists potentially growing very long in medium large collections) by introducing a second level to the browsing hierarchy. Consider a collection with 26,000 documents, with the initial words of titles evenly distributed through the alphabet; the title browsing interface of the old system would display 1,000 documents in a long list, for each letter of the alphabet. The new system would divide these classes of 1,000 documents up into smaller classes of documents, containing, say, 50 documents each (this is determined by the "maximum number of documents per page" setting on the interface). This means there will be twenty subclasses across the top of the page under the top-level classes. This is still reasonably usable. Of course, it is possible with this system to have so many documents that it becomes unusable too, but this number is much larger than in the old system.

The time constraints on the new system are more worrying. Users hate waiting for web pages to load [18,22], so load time has a large impact on the usefulness of a page. Unfortunately because the new system produces pages that contain entire browsing structures, the pages are very large. The browsing interface itself is 6.78kB and each document is 0.04kB in the browsing structure. This means that over a 56kbps modem the interface will take 1 second to load, and each document in the browsing structure will add about 0.05 seconds—with a large collection this adds up very quickly. A collection of 1,073 documents (browsed by title and coverage) was shown to take 66 seconds to load over a 56kbps modem connection, precluding this interface from being used over a low speed connection. However, for a high speed connection or a local collection, this time drops to under 1 second, and once the page is loaded then the entire browsing hierarchy is available instantaneously. When we compare this to the classifier system, the total load time over a 56kbps would be 88 seconds for title metadata only. However, this time is in smaller chunks as the user loads each part of the hierarchy, and thus the wait time is more palatable to the user (each individual page would take about 3 second to load over a 56kbps connection).

A transaction log analysis of 42 collections in the New Zealand Digital Library [2] from June 21st to December 19th 2001 shows that approximately 9.5% of all actions are browsing with the classifier system, and that 37% of the time when a user looked at one part of a classifier (say the 'A' section of a title classifier) their next action was to look at another part of the same classifier (say the 'B' section). This has an associated time cost under the old system, but under the new system it is instantaneous.

5.2 User Study

There were two user studies performed on the concept embodied by this interface, one to determine how users actually want to browse, and one to determine the predictability of the system (i.e. to determine whether users could guess what the system would show them given the interface). Both these studies were paper based studies using index cards to represent documents. For more information on these studies see [20].

The first study asked user to arrange the paper documents into a browsing structure that they would find useful for locating information on a specific topic. Eight out of ten study participants created some metadata-based browsing scheme. Moreover, of the participants who created metadata-based browsing structures, five used more than one kind of metadata, something they couldn't do with the old system. When asked to comment on whether the organisation they had created was appropriate for an electronic information system, three users commented that they "would also like searching", and three users said that an information system should be able to present more than a single view of an information system. This indicates that users want the flexibility offered by the new system but not available in the old system.

The second study asked participants to arrange documents as they believed the interface would arrange them (being shown a picture of the interface where the documents were to be sorted first by title and second by coverage). A high level of comprehension was shown, with seven out of eight users sorting the documents properly, and the eighth user commenting that this was not how he believed the interface would sort the documents, but it was how he would like it to.

Evaluating the new browsing system both technically and with user studies shows that it has only on major flaw: the amount of time a browse page may take to load (and even that is ameliorated by the fact that it then provides better performance than the standard browsing interface 37% of the time). The new system handles large numbers of documents better than the old system, is readily comprehensible, and allows users the flexibility they want in a browsing system.

6 Conclusions

A novel browsing system was created within the Greenstone software. The system fits cleanly within the Greenstone software, and does not require any extra effort on the part of the user or the collection maintainer to use. This new system is a metadata-based between documents system, and was designed with human information seeking needs and the failures of the old system in mind.

The new system has some technical issues when it comes to page load time, but this problem can be solved by using local collections or a fast connection. Furthermore the total load time for a browsing structure is actually less in the new system than the old system.

The new system allows users to combine searching and browsing, in keeping with both the literature on information seeking behaviour and the user experiment carried out as a part of the work done for this investigation. The new system also allows the user more flexibility in determining the way in which they browse, also in keeping with experimental results and information seeking literature. In both these areas, the new browsing system is a vast improvement over the old classifier system making information seeking easier and more effective. The new browsing system can handle many more documents than the old system before the browsing structures become unusable.

In sum, a novel browsing system that allowed users dynamic interaction with information collections and fully supported the three main browsing activities of evaluating information sources, finding “new” information and clarifying information problems was implemented in Greenstone.

References

1. The Greenstone Software. http://www.greenstone.org, accessed March 6 2003.
2. The New Zealand Digital Library. http://www.nzdl.org, accessed March 6 2003.
3. Agosto, Denise. Bounded Rationality and Satisficing in Young People’s Web-Based Decision Making. *Journal of the American Society for Information Science and Technology* 53:1 2002 16–27.
4. Bainbridge, David, Mckay, Dana and Witten, Ian. The Greenstone Developer’s Guide. Dept. of Computer Science, University of Waikato 2001. Available at http://www.greenstone.org, accessed March 6 2003.
5. Baldonado, Michelle Q. Wang. A User-Centered Interface for Information Exploration in a Heterogeneous Digital Library. *Journal of the American Society for Information Science* 51:3, 2000 297–310.
6. Bates, Marcia J. The Design of Browsing and Berrypicking Techniques for the Online Search Interface. *Online Review* 13:5 1989 407–424.
7. Beaulieu, Micheline. Interaction in Information Searching and Retrieval. *Journal of Documentation* 56:4 2000 431–439.
8. Chang, Shan-Ju and Rice, Ronald E. Browsing: A Multidimensional Framework. *Annual Review of Information Science and Technology* 28 1993 231–276.
9. Chen, Hsinchun, Houston, Angela L., Sewell, Robin R., and Schatz, Bruce R. Internet Browsing and Searching: User Evaluations of Category Map and Concept Space Techniques*. Journal of the American Society for Information Science* 49:7 1998 582–603.
10. Crabtree, Andy, Twidale, Michael B., O’Brien, Jon and Nichols, David M. Talking in the Library: Implications for the Design of Digital Libraries. *Proc. 2nd ACM International Conference on Digital Libraries*, Philadelphia, Pennsylvania 1997 221–228.
11. Cunningham, Sally-Jo, Knowles, Chris, and Reeves, Nina. An Ethnographic Study of Technical Support Workers: Why We Didn’t Build a Tech Support Digital Library. *Proc. 1st Joint ACM/IEEE-CS Conference on Digital Libraries*. Roanoke, Virginia 2001 189–198.
12. Henninger, Scott and Belkin, Nicholas. Interface Issues and Interaction Strategies for Information Retrieval Systems. *CHI Companion ’95*. Denver, Colorado 1995 401–402.
13. Jacso, Peter. Savvy Searching Starts with Browsing. *Online and CD-ROM Review* 23:3 1999 169–172.
14. Jones, Steve, Cunningham, Sally Jo and McNab, Rodger. An Analysis of Usage of a Digital Library. *Proc. 2nd European Conference on Research and Advanced Technology for Digital Libraries*, Heraklion, Greece 1998 261–277.
15. Knepshield, Pamela A. Savage and Belkin, Nicholas. Interaction in Information Retrieval: Trends Over Time. *Journal of the American Society for Information Science* 50:12 1999 1067–1082.
16. Kuhlthau, Carol. Inside the Search Process: Information Seeking from the User’s Perspective. *Journal of the American Society for Information Science* 42:5 1991 361–371.
17. Lawrence, Steve and Giles C. Lee. Context and Page Analysis for Improved Web Search, *IEEE Internet Computing* 2:4 1998 38–46.
18. Lazar, Jonathon, Bessiere, Katie, Ceaparu, Irina and Shneiderman, Ben. Help! I’m Lost: User Frustration in Web Navigation. *IT&Society* 3:1 2003 18–26.
19. Marchionini, Gary. Information Seeking in Electronic Environments. Cambridge University Press, New York. 1995
20. McKay, Dana. Browsing and Greenstone: a Study of Browsing in Digital Library

Software. Available from the Dept. of Computer Science, University of Waikato. 2002.
21. Nielsen, Jakob. The Changes in Web Usability Since 1994", *Alertbox* December 1 1997, available at http://www.useit.com/alertbox/9712a.html, accessed March 6 2003
22. Nielsen, Jakob. The Top Ten New Mistakes of Webpage Design, *Alertbox* May 30, 1999, available at http://www.useit.com/alertbox/990350.html, accessed March 6 2003.
23. Nordlie, Ragnar "User Revealment"—a Comparison of Initial Queries and Ensuing Question Development in Online Searching and Human Reference Interactions.*Proc. 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Berkley, California 1999 11–18.
24. Paynter, Gordon, Witten, Ian, Cunningham Sally Jo, and Buchanan, George. Scalable Browsing for Large Collections: a case study. *Proc. 5th ACM Conference on Digital Libraries*, San Antonio, Texas 2000 215–218.
25. Salampasis, Michail, Tait, John and Bloor, Chris. Evaluation of Information Seeking Performance in Hypermedia Digital Libraries. *Interacting with Computers* 10 1998 269–284.
26. Shukla, Preeti. *Cartography for Collections.* Masters Thesis. Available from the Dept. of Computer Science, University of Waikato. 2003.
27. Spink, Amanda, Bateman, Judy and Jansen, Bernarnd J. Searching the Web: a Survey of Excite Users. *Internet Research: Electronic Networking Applications and Policy* 9:2 (1999) 117–128.
28. Witten, Ian, and McNab, Rodger. The New Zealand Digital Library: Collections and Experience. *The Electronic Library* 15:6 (1994) 495–504.
29. Witten, Ian, Moffat, Alistair and Bell, Timothy. Managing Gigabytes: Compressing and Indexing Documents and Images. Van Norstrand Rheinhold, New York (1994).