The New Zealand Digital Library: Collections and Experience

Ian H. Witten and Rodger McNab
Department of Computer Science, University of Waikato,
Hamilton, New Zealand.

{ihw, rjmcnab}@cs.waikato.ac.nz

Phone (+64) 7 838–4246, fax (+64) 7 838–4155

Abstract*—The New Zealand Digital Library project aims to develop the underlying technology for digital libraries and make it available for others to use to create their own collections. We have built a large number of demonstration collections. Because our policy is to avoid manual processing of material, full-text indexing and—to a lesser degree—automatically created browsing structures provide the primary point of access to the material. As well as conventional textual collections, we are experimenting with collections of musical and audio material. This article describes the library structure and present and planned collections, and summarises our experiences in the project.*

The migration of information from paper to computers promises to change the whole nature of research, and in particular the methods by which people locate information. The New Zealand Digital Library project is exploring ways to impose structure on fundamentally anarchic, uncatalogued, distributed repositories of information, thereby providing information consumers with effective tools to locate what they need and peruse it conveniently and comfortably. Our goal is to produce an easy-to-use digital library system that runs on inexpensive computers at information providers' own sites and offers a public information service that information providers themselves maintain.

The aim of our project is not to set up new libraries, but to develop the underlying technology for digital libraries and make it available publicly so that others can use it to create their own collections. Not surprisingly, the technology required varies greatly depending on the kind of collection and the source of the information. Consequently, we are making several different substantial collections of public-domain text available on an experimental basis as test cases. These allow us to investigate the technical problems of gathering and indexing the material, to assess the useability of our interfaces, and to collect information on external usage so that we can improve the facilities offered. Perhaps more important, the provision of a large array of test collections allows us to explore the varying requirements of diverse collections of information and to develop a software framework that permits collections of different kinds.

Our project rests on five basic planks. First, we avoid manual processing of the material that comprises the library, and make a minimum of assumptions about conventions adopted by document repositories from which the source material is collected. For example, we do not assume the presence of any bibliographic metadata at all, nor any other information provided explicitly for organisational or indexing purposes. Second, we provide as the primary access mechanism a full-text index of the entire contents of each document, whereas other schemes index on user-supplied document descriptions, abstracts, or other document surrogates. Third, close attention is paid to the interface and to the real needs of library users. Fourth, our work directly addresses the problem of building the library in a geographically remote location with high Internet costs—an environment in which the benefits of networked library technology are especially striking. Finally, we aim to produce a library scheme that is economical in the resources it requires and can operate on a small, inexpensive, server.

This article relates our experience so far with the New Zealand Digital Library project. Because the core of any library is the collections it contains we focus on the collections we have created and the collections that we are developing. The project started in 1995 in an effort to create a digital library for computer science research (Witten *et al*., 1995), and received a seed grant shortly thereafter from the New Zealand Lotteries Grants Board to run a small pilot service. In July 1996 we received major funding from the New Zealand Foundation for Research, Science and Technology to pursue a program of research and development on "Digital Libraries for New Zealand."

We begin by describing the facilities offered by the New Zealand Digital Library, using as an illustration the first—and still the largest—collection, containing Computer Science technical reports. We then go on to review the other collections presently offered by the Digital Library, and the new ones that are being developed. A unique feature of our project is its emphasis on music and other audio material, and we describe this next. Then we discuss the issues of collection synthesis, or how the library can add value to the individual collections by relating them to others. Finally we summarise our experience in usage and collection creation.

THE COMPUTER SCIENCE TECHNICAL REPORT COLLECTION

The New Zealand Digital Library is a publicly-accessible system on the World-Wide Web[1] that provides full-text indexes to several substantial collections of information. Our flagship collection is a library of Computer Science technical reports, which currently provides a full-text index to 40,000 technical reports—one million pages, containing 400 million words—collected from over 300 sites around the world. It involves nearly 2.5 Gb of text, extracted automatically from 35 Gb of source information by a system that we have developed (Nevill-Manning *et al*., 1997). A subcollection has been set up containing technical reports gathered from sites in Germany, and our software has been transferred to the University of Bonn, where this collection is offered on an experimental basis as part of the German digital libraries program.

The current interface to the Library allows documents to be located by full-text retrieval, with the option of either Boolean or ranked search. The latter performs an "or" query on the specified terms and ranks the documents retrieved by relevance according to the well-known cosine rule (Salton and McGill, 1983). In either case queries are answered quickly, and the plain text of each matching document is instantly made available for browsing.

The search engine for the library is the public-domain system MG (Witten *et al*., 1994). Tailored for highly efficient storage of full-text databases, MG can pack an index to a large collection of text into just 5% of the size of the original text. This is impressive because an index can easily occupy the same amount of space as the original text, or more. Further, MG responds rapidly to queries: experiments with a standard 750,000-document collection (the TREC collection; Harman, 1992) produce ranked output for queries of forty to fifty terms within three to five seconds. In fact, tests have shown that the use of compressed indexes can actually improve response time because less data needs to be retrieved from disk!

The Computer Science Technical Report collection is the most mature, and organisationally the most complex, collection offered by the New Zealand Digital Library to date. Intended as a serious research resource, it is used regularly by computer scientists worldwide. Figure 1 shows the query page on the left and a typical response on the right. The searching facilities, which can be entered on the query page, include:

- ranked and Boolean queries;
- ability to stem terms, and/or make them case-insensitive, both on a whole-query basis and a term-by-term basis;
- phrase searching;
- searching at document or page level;
- searching first pages only.

In order to bound the resources consumed by each query, a maximum of 50 documents are returned. The "query response" page on the right of Figure 1 contains information about ten of them, including the first few words of the document and four or five icons that serve as buttons. From these the user can quickly retrieve:

- the URL of the original document, with its size, creation date, and download date;
- the original document itself, downloaded from that URL;
- a facsimile image of the first page or two of the original document;
- the text extracted from the original document, crudely formatted;
- the figures extracted from it (if they are in encapsulated PostScript).

[1] at http://www.cs.waikato.ac.nz/~nzdl

Other collections are organised in a similar manner, although since they are not usually extracted from PostScript, they are somewhat simpler: the text is the original document and there is no need to provide access to it separately, nor to facsimile images of the first pages.

OTHER COLLECTIONS

Computer science is unique in that a vast amount of high-quality information already exists in digital form and is freely accessible on the Internet in the form of technical reports. In order to demonstrate that digital libraries can benefit diverse groups of users, we are also constructing other collections of publicly-available information, apart from the Computer Science Technical Reports. Presently, we have made the following collections publicly available:

- *The Computists' Communique*. An on-line AI research news magazine, operating since 1991, this includes grant and funding opportunities, industry news, Internet and Web news, online resources, research discussion lists, news and software offers, software development resources, and career or entrepreneurial tips.

- *FAQ Archive*. "Frequently Asked Questions" are an Internet phenomenon that arose as a way of reducing the number of newcomers asking the same questions in USENET discussion groups. They have developed into a corpus of questions and answers on a huge variety of topics. This collection can be searched for terms appearing within the same Frequently Asked Questions list, under the same subject heading, or within the same paragraph.

- *The HCI Bibliography*. This is a free-access online bibliographic database on Human-Computer Interaction, created in a project whose goal is to put an electronic bibliography for most of HCI on the screens of all researchers, developers, educators and students in the field. With this collection, either reference entries alone, or reference entries and abstracts, can be searched.

- *Indigenous Peoples*. This collection contains information on indigenous people found around the world. The material emanates from the Fourth World Documentation Project, which was started in 1992 to collect and distribute information of importance for indigenous people. The documents include essays, position papers, resolutions, organisational information, treaties, UN documents, speeches and declarations on social, political, strategic, economic and human rights issues. This project has become a primary information resource for universities, state and federal agencies, Indian and tribal councilmen.

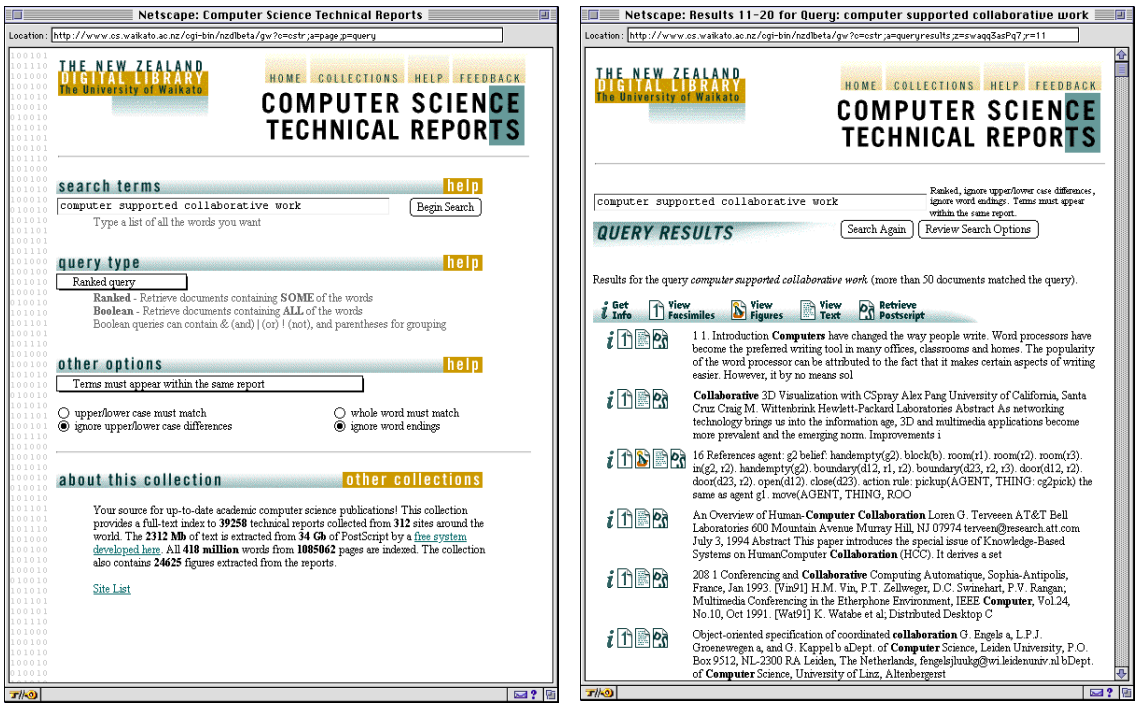

Figure 1 The Computer Science Technical Report query page and a typical response

- *Oxford Text Archive*. The public-domain texts in the Oxford Text Archive comprise 53 literary works ranging from *Wuthering Heights* to *The Red Badge of Courage*. This archive is provided by Oxford University Computing Services, which aims to serve the interests of the academic community by providing archival and dissemination facilities for electronic texts at low cost.

- *Project Gutenberg Collection*. This consists of five hundred public-domain books: *The King James Bible*; classic literary works such as *Moby Dick*, *Alice in Wonderland*, and *Paradise Lost*; reference books like *Webster's Dictionary*, the periodic table, and *The Hacker's Dictionary*. They are collected as part of the Gutenberg Project whose goal is to encourage the creation and distribution of electronic text. The texts are entered by volunteers who are encouraged to choose whatever books they like.

- *TidBITS* is a weekly electronic publication that covers news and views relating to the Macintosh computer, with a focus on Internet-related topics. News items are independent of each other, and so it is possible to search individual items, individual paragraphs, or titles of news items.

The major difference between these collections is the source and format of information, the updating policy, the granularity with which searching can be done (document, page, paragraph, etc.), the different kinds of index that need to be provided (titles, references, abstracts, etc.), and the structure and format in which the output is to be displayed.

For example, Figure 2 shows, on the left, the query response page for the query *Perlman* made to the HCI Bibliography collection. Complete references are provided, and where more information is available—as in all but the third entry—a link is created to a fuller record. At the right of Figure 2 are two sample records, which contain abstracts along with the reference themselves. The *detach* button at the top is used to detach the page so that it is retained while a second document is shown.

This is a typical example of how the structure of the information displayed needs to be tailored specifically for that kind of collection. Another example, which occurs in some of the other collections, is the provision of a summary of the contents of the collection—for example, a clickable list of titles—or even a browsing structure to allow non-searching access to the collection.

COLLECTION DEVELOPMENT

We are working on a number of other collections which we have made available on a private basis until

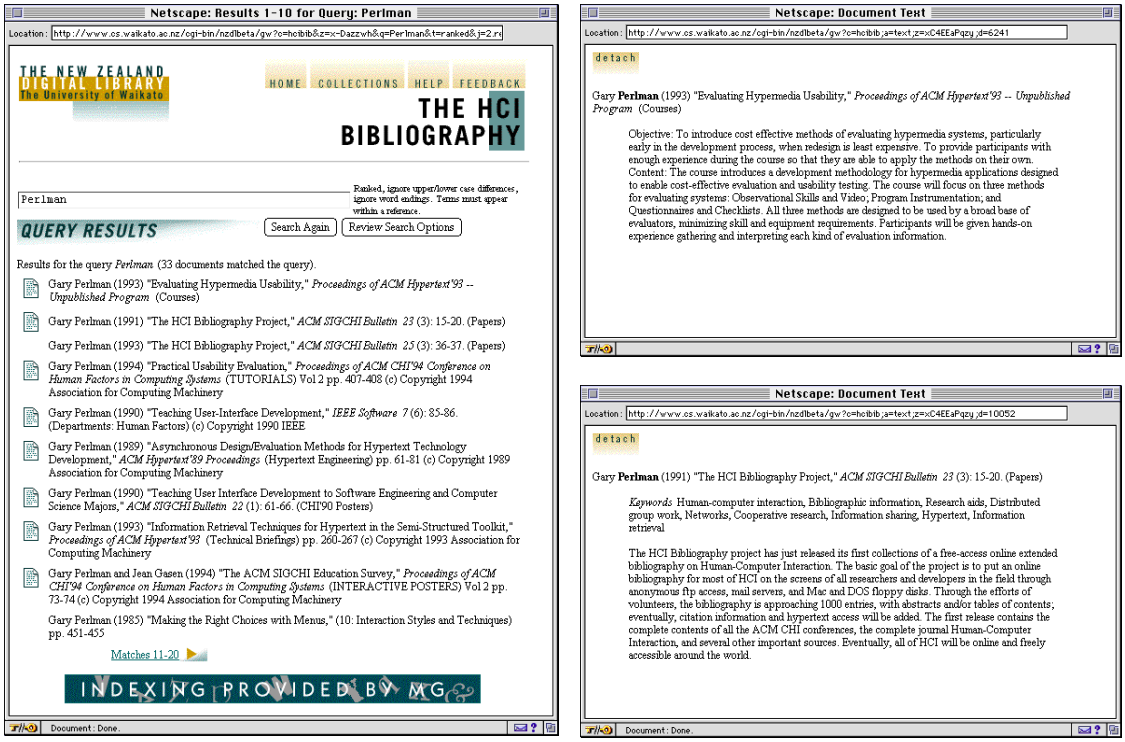

Figure 2 Results of a query for *Perlman* in the HCI Bibliography collection

they are approved for public release. Because of limited resources, we must choose carefully what to work on. We are particularly interested in novel problems for full-text retrieval which introduce new issues that have not yet been addressed in our project. Moreover, we direct our efforts in the direction of collections designed specifically to stimulate interest in the project from diverse groups of users.

Examples of collections that we are working on include:

- *The Karlsruhe Bibliographies*. This is a very large collection of bibliographies in the field of computer science, totalling about 700,000 bibliography entries. We plan to make it similar to our current *HCI Bibliography* collection, except that we are considering allowing users a choice of reference format so that they can cut and paste references into documents they are working on.

- The *Humanity Development Library*. This is a collection put together by the Global Help Project to provide ideas, experiences and solutions to workers in developing countries. The Humanity Library is the fruit of a massive humanitarian and development information transfer project to developing countries. Forty organisations are collating thousands of useful publications and resources to help solve poverty, to increase human potential, and to provide education to all. The goal is that by mid 1998, any person in a developing country with access to the Internet, or to a PC with CD-ROM drive, will have direct low-cost access to a complete library of 3,000 books with essential ideas, solutions, and know-how to help solve poverty and to meet basic human needs.

- *City Council Minutes*. In an effort to increase public awareness and participation in local government, a local city council is planning to make available on the Web records of council meetings, along with related discussion papers. Having these in searchable form will greatly increase their accessibility and usefulness. For example, one can search just motions, or paragraphs, or whole sets of minutes, to find information relevant to particular decisions, people, or places.

- *Library Catalogue*. We are converting a large library catalogue, containing about 10 million records, into a full-text-searchable collection. While conventional fielded search is appropriate when seeking an item that one knows is there and has a good deal of information about, we believe that full-text search of library catalogue entries is probably more convenient for non-professional users, in situations where one is not sure what one is looking for, and for casual browsing. We are considering adding subsidiary indexes for author, title, and keyword fields, so that if necessary the user can disambiguate query terms by restricting them to one of those fields. We are planning to conduct human factor experiments to determine the relative efficiencies of field-based and full-text searching for various kinds of information retrieval tasks.

Each of these pose interesting problems. For example, the left part of Figure 3 shows the result of a query for *marriage* in the Humanity Development Library (the repetitions in the list are caused by repeated sections in the original document collection), and on the right is the fourth document, *Closing the Gap Between Supply and Demand: Fertility Decline in Egypt*. Because this collection is structured as a hierarchical system of documents emanating from many different sources, it is rather hard to orient oneself in the collection. Consequently a navigational aid is presented at the top of the document returned (on the right) showing the section it resides in, along with the sibling subsections, preceded by the titles of hierarchically enclosing sections. This is generated automatically from the document structure, and by clicking on the folder and document icons the user can browse the collection very conveniently.

A number of other projects are presently being discussed with clients.

- *Historic Maori Newspapers*. A collection of nineteenth century New Zealand newspapers in the Maori language has recently been made available on microfiche for scholars and others interested in our country's history. However, the utility of any corpus of newspaper material is severely limited if no index is available, and indexing such a collection manually would be a formidable task. We are investigating the feasibility of scanning the material for presentation on the Web and OCRing it to create a machine-readable approximation to be used for indexing. The machine-readable version does not have to be perfect, just good enough to create a useful index. One option is to use a semi-automated process where important items such as names are checked manually.

- *Computer Science Paper Titles and Abstracts*. Many publishers provide public services that give access to journal article titles and abstracts, and conference programs are often published too. We plan to harvest these automatically and add them to our computer science library.

MUSIC AND AUDIO COLLECTIONS

Despite their variety, the collections above are all composed of conventional textual material. In addition, we are experimenting with the collection, indexing, and retrieval of audio and musical material.

Melody index

Our "melody index" system retrieves music on the basis of a few notes that are sung, hummed, or otherwise entered (McNab *et al*., 1996). Music librarians are often asked to find a piece of music from a few sung notes. The magnitude of this task may be judged by the fact that the Library of Congress holds over six million pieces of sheet music (not including tens of thousands of operatic scores and other major works), and the National Library of France has 300,000 hours of sound recordings, 90% of it music. As digital libraries develop, these collections will increasingly be placed on-line through the use of optical music recognition technology.

With the melody index, users can literally sing a few bars and have all melodies containing that sequence of notes retrieved and displayed—a facility that is attractive to casual and professional users alike. Such systems will form an important component of the digital music library of the future. With them, researchers will analyse the music of given composers to find recurring themes or duplicated musical phrases, and musicians and casual users alike will retrieve compositions based on remembered (perhaps imperfectly remembered) musical passages.

The system transcribes melodies automatically from microphone input. It then searches a database for tunes containing melodic sequences similar to the sung pattern, and ranks the tunes that are retrieved according to the closeness of the match. We have implemented different search criteria involving melodic contour, musical intervals and rhythm; carried out tests using both exact and approximate string matching; and performed user studies on how people remember tunes. We conclude from these experiments that people need a choice of several matching procedures and should be able to explore the results interactively in their search for a particular melody. This is exactly what the system offers.

Figure 4 shows two screen displays from the melody retrieval system. In each case the input sung was the first eight notes of *Auld Lang Syne*. For the query on the left, the simple query mode was used. The transcribed input is shown at the top, and titles of tunes returned by the database search appear below. Tunes are ranked according to how closely they match the sung pattern. Any of the returned tunes may be selected for display, and the user can listen to the displayed tune. The right shows the "advanced

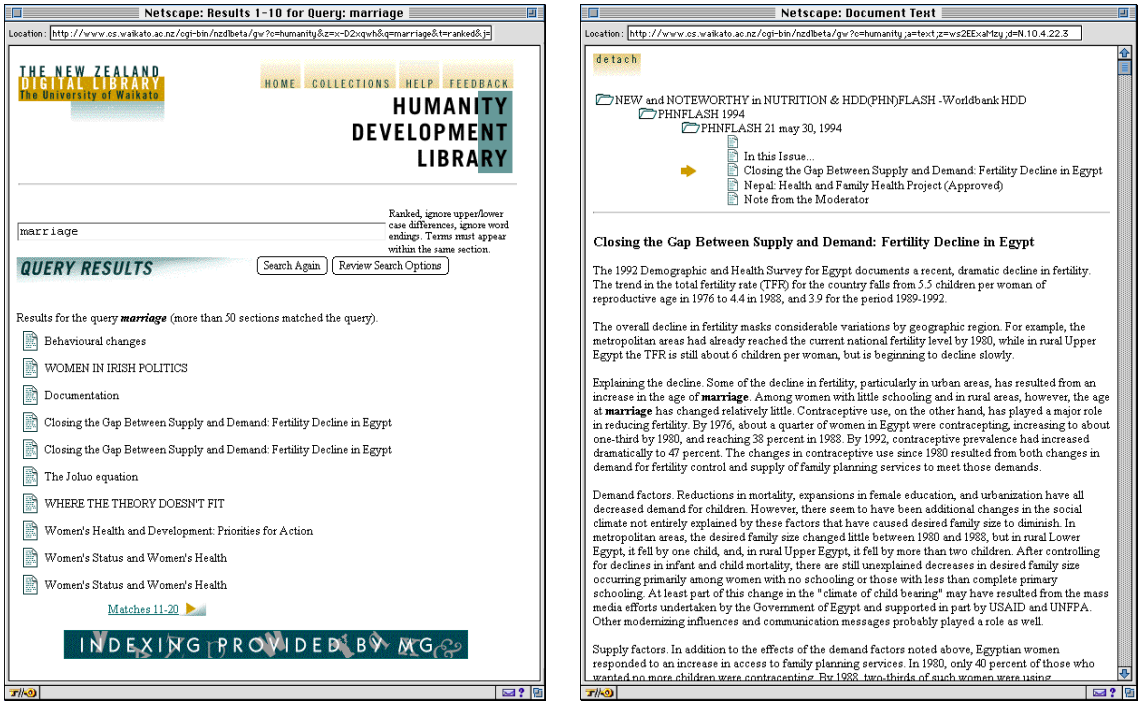

Figure 3 Results of a query for *marriage* in the Humanity Development Library

query" mode which allows the user to control the tightness of the matching by specifying pertinent parameters.

It is necessary to have a database of tunes stored in the form of music notation, and our test database is a collection of international folk tunes from various countries formed by amalgamating a collection of 1,700 tunes, most of North American origin, from the Digital Tradition folksong database (Greenhaus 1994), with the 7,700 from the Essen database of European and Chinese melodies (Schaffrath 1992). There are just over half a million notes in the database, and the average length of a melody is around fifty notes. The database is segmented into the following parts:

- North American (and British) folksongs (1650 tunes)
- German ballads and folksongs (5550 tunes)
- Chinese ethnic and provincial songs (2150 tunes)
- Irish folksongs (50 tunes).

Optical music recognition

One current limiting factor for the melody index is that few score collections are available in machine-readable form. As digital libraries develop, however, scores will be placed online through the use of optical music recognition technology. We are working in the development of optical music recognition (Bainbridge and Bell, 1996; Bainbridge, 1997) and have already built a prototype service that accepts images of music and returns an audio or MIDI file containing a performance of the music, or a file suitable for input to a music editing package. Eventually, music submitted for processing will automatically be added to the database of tunes that can be searched using the melody index.

Development of audio collections

Although the melody index uses a special search engine that is capable of applying musical knowledge to match sequences of notes, there is ample opportunity for audio collections that use conventional text search engines as the basic searching mechanisms. Currently under discussion are:

- *Oral History Collections*. Oral history is becoming available in both tape and transcript form from a number of sources. We are investigating the feasibility of using a timed transcript to provide full-text access to audio material.

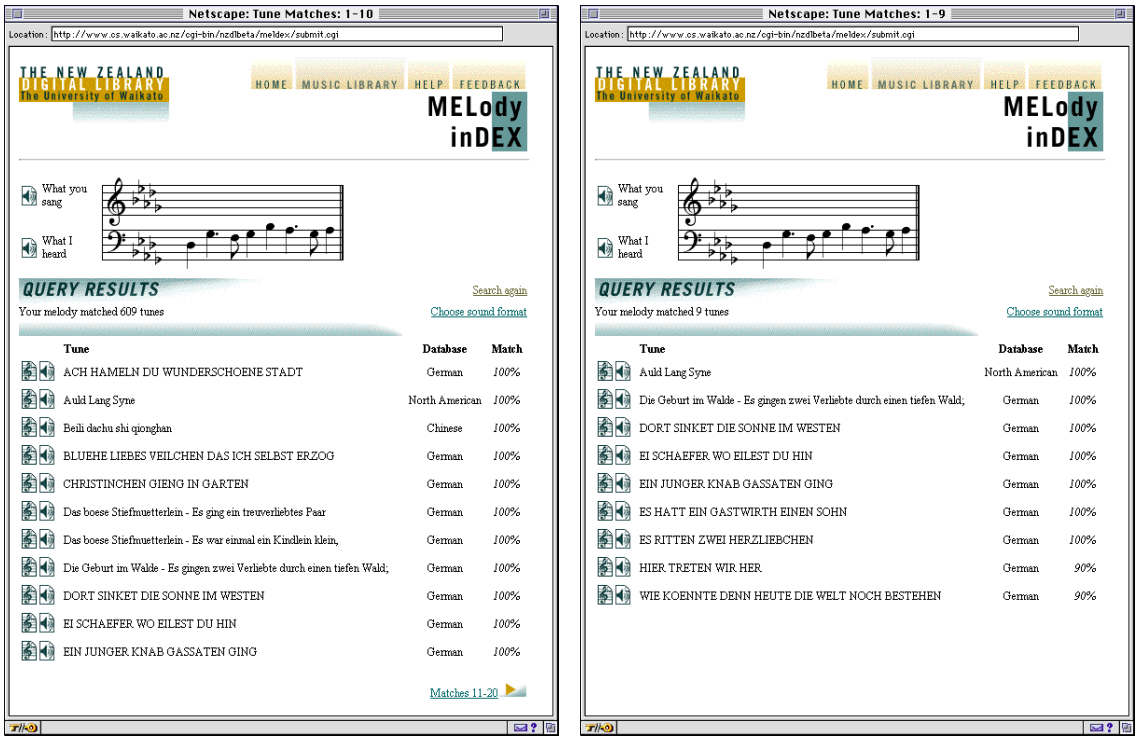

Figure 4 Using the melody index

- *Visually Impaired Readers*. The Web offers interesting distribution possibilities for the visually impaired. Currently, organisations such as the Royal New Zealand Foundation for the Blind produce audio tapes and large-print versions of books for distribution to such readers. Many potential users have access to voice-synthesis technology that enables them to read books that are available in electronic form. Such equipment can also permit access to the Web, provided that Web pages present information, and more particularly choices, textually rather than graphically. We are considering the issues involved in setting up Web-based library services designed expressly for partially-sighted and blind users.

COLLECTION SYNTHESIS

Returning from the music and audio department to the New Zealand Digital Library as a whole, it is apparent from the preceding pages that the Library is conceived as a number of independent collections, each providing a demonstration of what the underlying full-text-search technology can do in a different area. This has proved to be an excellent approach and provides several useful independent demonstrations for particular classes of user. However, intriguing possibilities are presented by combining some of the collections. For example, it would be useful to permit joint searches on the *Indigenous Peoples* text base and the *Humanity Development Library*. To do this would simply be a matter of constructing a joint index to the two collections.

More interesting would be a combination of different kinds of information, permitting, for example, cross-searching of the Computer Science Technical Reports, the computer science bibliography collections, and computer science paper titles and abstracts. As well as raising the question of the comparability of automatic relevance ranking information for different kinds of text, this will present some interesting interface issues. For example, a paragraph-level search on a document collection might be initiated if a title search fails, or a search on a name to be extended from a bibliography to a full report collection if only a few items match.

Potentially even more useful is the possibility of adding value to collections by correlating information from different sources. The full text of a large number of documents provides an enormously valuable resource that can be mined for further information, both to enhance the way in which it is presented to the user and for bibliometric research and analysis. The main tool we envisage for correlating this information is off-line (overnight) analysis whereby each item from one collection is sought in the same or another collection in an attempt to find matching documents. This methodology can add value to collections either by inserting new information or by cleaning the data already present.

Examples of such activities include

- matching bibliography collections against technical report collections to provide bibliographic information for the reports;
- correlating bibliography entries with bibliographies themselves, to identify and eliminate duplication;
- identifying reference sections in technical reports, seeking references in the collection (and in bibliographic collections), and inserting hyperlinks to the referenced document (or to its abstract);
- performing experiments on clustering within the collections to identify clusters of similar documents and present them to users.

USAGE, DISTRIBUTION, COLLECTION CREATION

Our initial digital library prototype, the New Zealand Digital Library for Computer Science, has already gained widespread acceptance both nationally and internationally. Although its existence has not been widely advertised, since it became available two years ago 2,700 queries have been made from 300 sites nationally, and 42,000 from about 80 countries internationally. (These figures exclude the 7,000-odd queries from Waikato University.) Although this level of usage is not high compared with many Internet information services, it is aimed at a rather specialised market. We plan to advertise the service more extensively once we have set up a North American mirror site.

Our goal is to produce an easy-to-use digital library system that runs at information providers' own sites. We do not yet have any such sites, except at the University of Bonn to which an initial version of the software has been ported. However, we have recently ported the system to the Linux operating system and it runs very well on a modern PC configuration that is powerful (100 MHz processor, 128 Mb memory, 4 Gb disk) but nevertheless fairly inexpensive.

Moreover, we are investigating distributing subsets of the New Zealand Digital Library collections on CD-ROM. Coupled with a local Web server, this provides a powerful means whereby users without convenient access to the Web, or for whom access is unreliable or expensive, can still benefit from just the same interface and facilities.

Currently it takes a skilled person about a day or two to build each collection and integrate it into the Digital Library framework, excluding any time that is necessary to gather the information in the first place. We aim eventually to make it easy for users to create their own collections. However, we have encountered a remarkably wide variety in the requirements that different kinds of information pose. We have already created eight collections and our experience is that each one poses new problems: the information comes in different formats, the requirements of the interface are different, the indexing needs are different, the maintenance and update regime is different. Each new collection stimulates us to solve new problems, and we estimate that the number must triple before we reach a state where we can be reasonably confident that new collections can be accommodated within the existing framework. When we reach that point we plan to work on an end-user interface for collection building.

CONCLUSION

The New Zealand Digital Library project is making it easy to create focused collections of high-quality information in particular areas on the Internet, in sharp contrast to the approach taken by most other search engines. We are developing novel software infrastructure that enables those who manage and maintain such collections to make them publicly available.

Several lessons have already been learned from the project. On the Internet it is becoming crucial to have a professional-looking shop front, even for an experimental service, and we have benefited greatly from skilled help for our Web page layout and design. Each new collection area comes with its own particular idiosyncrasies, and it is necessary for success to be sensitive to user needs and not to try to force different collections into the same Procrustean mould. With the right software infrastructure, it is however possible to respond very flexibly and rapidly to different requirements, and we believe that our software development effort will eventually converge to a point where most new collections can be accommodated very easily.

REFERENCES

Bainbridge, D. and Bell, T.C. (1996) "An extensible optical music recognition system." *Proc Australian Conference on Computer Science*. Melbourne; January.

Bainbridge, D. (1997) *Extensible optical music recognition*. PhD thesis, Department of Computer Science, University of Canterbury, New Zealand.

Greenhaus, D. (1994) *About the Digital Tradition*. http://www.deltablues.com/DigiTrad-blurb.html.

Harman, D.K.E. (1992) "Proc. TREC Text Retrieval Conference," Gaithersburg, MD: National Institute of Standards Special Publication, 500-207.

McNab, R.J., Smith, L.A., Witten, I.H., Henderson, C.L. and Cunningham, S.J. (1996) "Toward the digital music library: tune retrieval from acoustic input" *Proc Digital Libraries '96*, 11-18.

Nevill-Manning, C.G., Reed, T., and Witten, I.H. (1997) "Extracting text from PostScript" *submitted to Software—Practice and Experience.*

Salton, G. and McGill, M.J. (1983) *Introduction to modern information retrieval*. New York: McGraw Hill.

Schaffrath, H. (1992) "The ESAC databases and MAPPET software." In *Computing in Musicology*, Vol 8., edited by W. Hewlett and E. Selfridge-Field. Menlo Park, CA: Center for Computer Assisted Research in the Humanities.

Witten, I.H., Moffat, A., and Bell, T.C. (1994) *Managing Gigabytes: Compressing and indexing documents and images*. Van Nostrand Reinhold, New York.

Witten, I.H., Cunningham, S.J., Vallabh, M., and Bell, T.C. (1995) "A New Zealand digital library for computer science research" *Proc Digital Libraries '95*, 25-30, Austin, Texas, June.