# The New Zealand Digital Library Project

Ian H. Witten, Sally Jo Cunningham, and Mark D. Apperley
Department of Computer Science, University of Waikato,
Hamilton, New Zealand.

{ihw, sallyjo, m.apperley}@cs.waikato.ac.nz

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

***Abstract****—The New Zealand Digital Library project is a research programme in the Computer Science Department at Waikato University, whose aim 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. We are concerned with large collections of electronic, predominantly textual, documents, physically dispersed on computers the world over, and aim to make them accessible through a uniform interface that allows information to be located and accessed. This article describes the NZDL project, illustrated by a large example collection of Computer Science Technical Reports. The system is freely available on the World-Wide Web.*

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 goal of the New Zealand Digital Library project is to explore the potential of Internet-based digital libraries, by which we mean large collections of electronic, predominantly textual, documents, physically dispersed on computers the world over, which are accessible through a uniform interface that allows information to be located and accessed. Our vision is to develop systems that automatically impose structure on fundamentally anarchic, uncatalogued, distributed repositories of information, thereby providing users with effective tools to locate the information they need and peruse it conveniently and comfortably. As a geographically isolated but technologically advanced nation, New Zealand stands to gain markedly from effective deployment of information resources that are freely available on international computer networks.

The New Zealand Digital Library project is a research programme in the Computer Science Department at Waikato University, funded in part by the NZ Foundation for Research, Science and Technology and the NZ Lotteries Grants Board. Our aim 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 large 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. We call this testbed the "NZDL" to distinguish it from our research programme.

This article begins by presenting the facilities offered by the NZDL. We then discuss the indexing and retrieval strategies that it supports, and go on to explain how the collection is built and maintained. Finally we indicate likely directions that our project will be taking in the future.

## WHAT IS THE NZDL?

The NZDL is a publicly-accessible system on the World-Wide Web[1] that provides full-text indexes to several substantial collections of information. By far the largest is the Computer Science Technical Report (CSTR) collection, containing over 25,000 research reports from around 300 sites worldwide—over half a million pages of information. Recently, a subcollection has been set up containing technical

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

reports gathered from sites in Germany, and the NZDL software has been ported to the University of Bonn, where this collection is offered on an experimental basis as part of the German digital libraries program.

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. However, we are also constructing other collections of publicly-available information. For example, we have incorporated two small collections of English literature (totalling 550 books): the Oxford Text Archive (from the UK) and the Gutenberg collection (from the US). These are intended to show that the technology is by no means confined to the computer science literature. In addition, there is a large collection of "frequently asked questions"—questions, with answers, that people have accumulated on thousands of topics, from classical guitar playing to sea kayaking. Recently we have been requested to set up a full-text index to a US news magazine, the Computists Communique. Operating since 1991, this gives a miscellany of information on grant and funding opportunities, the Internet and World-Wide Web, online resources, research discussion lists, software offerings, development resources, and career or entrepreneurial tips—full-text retrieval makes this material interesting browsing.

The current interface to the NZDL 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. In either case queries are answered very 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 only 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 the 750,000 document TREC collection 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 CSTR is the most mature, and organisationally the most complex, collection that we offer to date. Intended as a serious research resource, it is used regularly by computer scientists worldwide. Figure 1 shows the query page and a typical response. Some of the details in the description below apply just to this collection, although the other collections are organised in a similar manner.

## INDEXING AND RETRIEVAL

We deal with material that has no cataloging information associated with it; consequently we can only approximate some of the services that are available in formally catalogued collections. If those who provide the information we use had to supply catalogue details, the scope of our collection would be greatly restricted. If we had to find the means to catalogue it ourselves, the collection would be limited by human resources. It was resolved that although the material would be gathered entirely from publicly-available repositories of text, the system should not require any effort or action on the part of

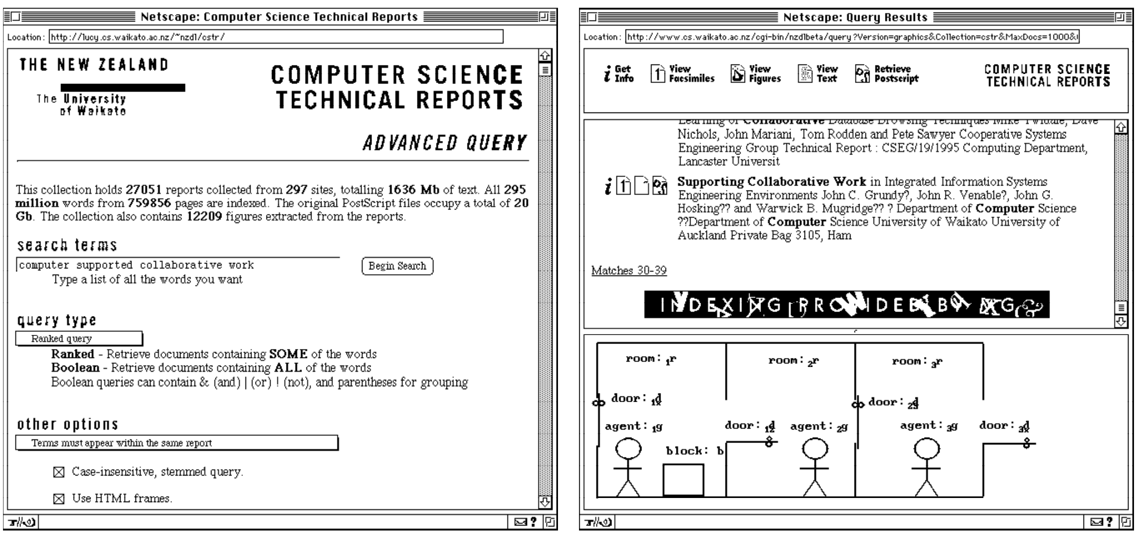


Figure 1 The CSTR query page and a typical response

participating repositories. No special software, archive organisations, or file formats are required of the providers. The only information used for cataloguing is derived from the documents themselves.

### Full-text index

These considerations led to the use of a full-text index instead of a library catalogue as the main retrieval mechanism. The index makes the entire text of all documents available for retrieval, rather than some much more restricted list of keywords as is the case in many computer-based text retrieval systems. From the point of view of building the system, this has the advantage that nothing more is needed to construct the index than the plain text of the documents in the library: there is no requirement that traditional bibliographic database information such as author, title, publisher, and so on, be manually provided. From the point of view of the user, full-text retrieval provides a powerful tool for searching for information. We consider below the extent to which it can approximate the more traditional forms of library access—by author, title, date, subject, and so on.

### Types of search

In the CSTR collection, several different kinds of search can be undertaken.

*Author/title*. In the vast majority of reports, the first page gives bibliographic information such as title and author. By limiting search term matching to this page, the user can approximate a search based on such information. For example, an initial page search for documents authored by *Knuth* will not retrieve documents that merely cite his work.

*Publication date*. Most reports also include publication date on the initial page, and the same technique serves for publication date searches. Alternatively, this facility could be simulated by permitting the user to search on the date in which a technical report was entered into its repository, although such a strategy is likely to produce uneven results because many reports are placed in repositories long after they were originally produced.

*Page searching*. The digital library stores the full text of technical reports, and supports searching over the complete document text. This is very useful in performing very general searches with high recall. However, a large number of irrelevant documents (false drops) can be expected as well. To achieve higher precision, a limited kind of proximity searching is supported by requiring the query terms to appear on the same physical page of the document. Phrase searching is implemented within MG by post-processing query results: a string search for phrases can be performed on documents returned by any query.

*Case folding/truncation/exact match*. A choice is offered of how query terms should be matched in the document returned. The user can specify whether terms can be folded to lower case, or whether a crude stemming algorithm is to be used to reduce them to root form. This allows one to search for names, or for grammatical variants of a root word.

### Indexes

To support the kinds of search identified above, a number of separate indexes to the collection are constructed that make it possible to search on a first-page, same-page, and same-document basis, and to specify whether search terms should be stemmed and case-folded or not. As far as the user is concerned, however, all that is required is to specify the level of the search and whether stemming and case-folding should be in force.

For a high recall topic search, the user specifies a document-level search. This will match to search terms appearing in any portion of the document text (including the references section), and the terms can be widely scattered. For a more focused query, the user specifies a page-level search; the search terms must then appear on the same page of the documents, although they still need not be adjacent. With both page- and document-level searching, the user can either choose exact term matching, or apply case folding and stemming to the query terms. The ability to turn case folding off is particularly useful when searching computing documents, because many software systems use acronyms that are also common English words (for example, the SMART retrieval system).

To search by author, title, or date, the user specifies that only the first page of documents should be searched, with no stemming or case folding. This type of query is not as precise as a fielded search, of course, but in practice is an effective work-around for a collection, such as ours, that is not formally catalogued. To find cited authors or papers, the user specifies a full document search with no stemming or case folding. This locates authors or titles within the body of a document, which generally includes references. Again, the search will be relatively imprecise, but anecdotal evidence from users suggests that this type of query is useful—particularly since commercial citation indexes such as SciCite provide relatively weak support for computer science.

## BUILDING THE CSTR COLLECTION

In the world of electronic information, the PostScript format provides the closest analog to paper as a document storage medium. PostScript is a page description language which is widely used on the Internet for storing technical reports and other more "serious" information (Adobe, 1985). Unlike the HTML format in which Web pages are written, PostScript cannot be read by ordinary Internet search programs and is therefore invisible to them. In order to build the CSTR index, it is necessary to be able to extract plain, unformatted, text automatically from the documents.

In fact, our system design accommodates not just PostScript but any format from which ASCII text can be extracted. For example, the Oxford Text Archives are stored in SGML, a relative of HTML; and the Gutenberg collection is stored in unadorned ASCII text. Even scanned images of textual document could be accommodated by OCR-ing them for indexing purposes: the inevitable recognition errors would reduce the quality of the index, but this can be ameliorated by using ranked queries containing redundant terms. However, PostScript files are almost universal in computer science technical report archives, and for this reason the CSTR collection currently includes only PostScript documents.

Archives of technical reports can be located through several lists maintained on the Internet, and by recursively descending the directory hierarchy looking for (possibly compressed) PostScript files. Each file is downloaded, along with its size and date, and the appropriate information is extracted.

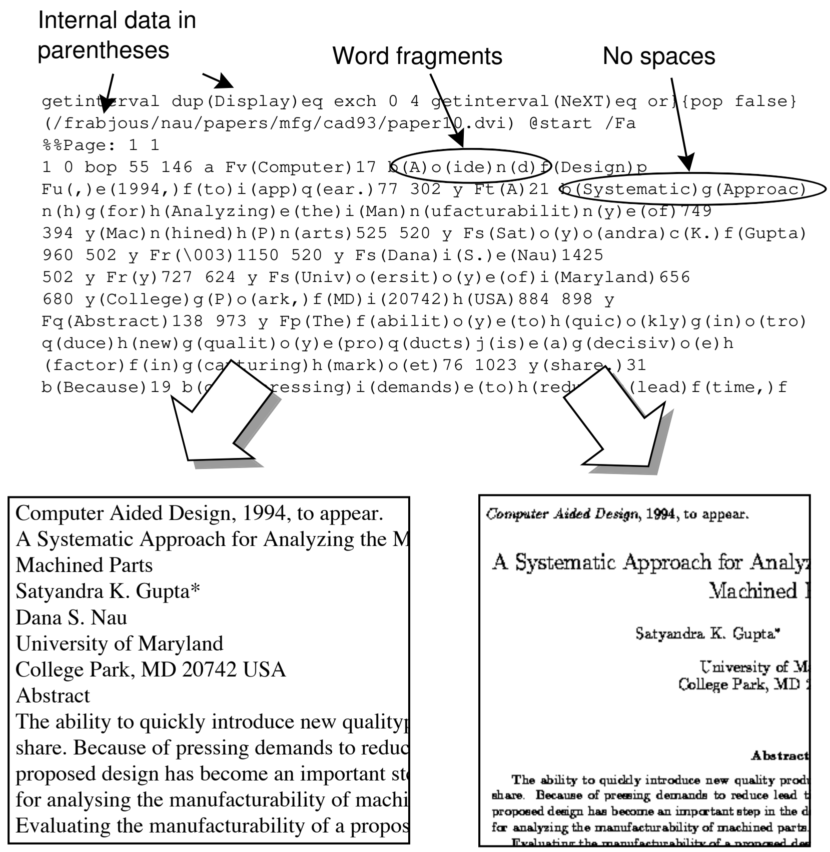


Figure 2 Conversion from PostScript: a PostScript file, the text extracted from it, and a facsimile image

### Information stored centrally

With an Internet-based digital library, a crucial question is how much information to store centrally. It was decided that the library would comprise an index and search engine, and the documents themselves would remain in their original repositories. In addition to the index, a facsimile image of each document's first page or two is retained at the NZDL site so that users can read the title and abstract, and sample the look and feel of the original. The plain, unformatted, text must be extracted from the documents to build the index, and it proves expedient to retain a full copy of this text locally as well. This is useful in its own right for browsing: users can examine the text without going to the trouble of downloading and printing PostScript. Moreover, the text of the collection provides an excellent foundation for bibliometric research. Figure 2 illustrates (at the top) an original document in PostScript form, and (at the bottom) the two files extracted from it: on the left the full ASCII text, and on the right a facsimile image of the first page.

### Text extraction

While the words of a technical report usually appear as plain text within a PostScript file, they are thoroughly intermixed with PostScript language commands and internal data. Words appear within parentheses, but so does internal information such as font names and error messages. Spaces are not explicit, but are coded implicitly in terms of the placement of words on the page. Finally, whole words are not always bracketed together: to give greater control over spacing, letters and word fragments are often placed individually. Figure 2 illustrates some of these problems. Extracting plain text from such documents presents an interesting technological challenge which we have addressed as part of our project (Reed *et al.*, in preparation).

### Gathering the information

One of the motivations for our project is efficient use of the expensive transpacific Internet link. Computer scientists can search for and preview technical reports locally before downloading the full document file, thus encouraging exploration without concern for network charges. However, to build the full-text index it is necessary to examine the contents of each report. Transmitting all of them across the Pacific would negate any cost benefits the project might offer. Consequently a distributed scheme was used to create the initial collection. In the first stage, a computer in North America downloaded each technical report, extracted the facsimile images and raw text, and sent them to New Zealand. This process is depicted in Figure 3.

### Maintaining the collection

In order to maintain the technical report collection, it is necessary is to ensure that the documents indexed continue to exist. This can be checked by periodically examining the technical report repositories for changes and updating the collection accordingly. One source of growth for the collection is when new documents in known repositories are located during a routine examination of currently indexed sites. New sites are detected by various means: monitoring standard Internet lists for new additions, manually scanning the newsgroups that announce them, and encouraging users to email

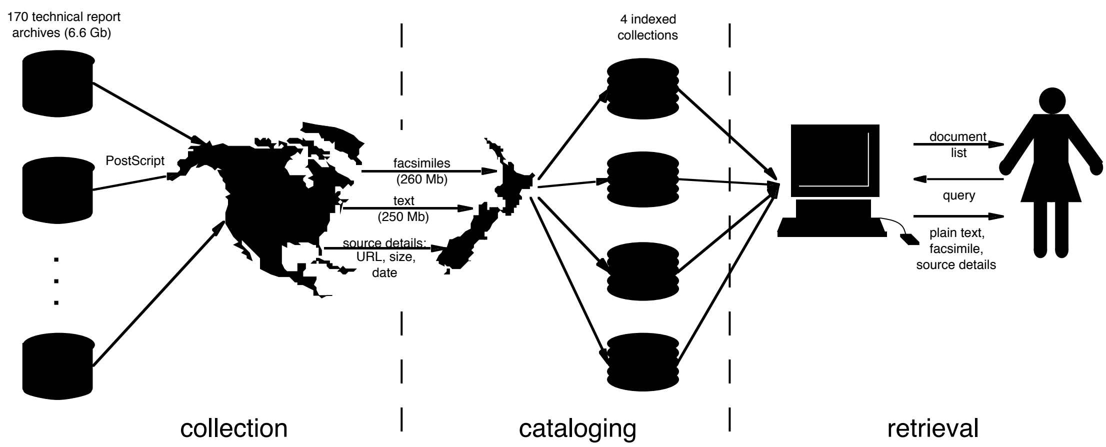


Figure 3 Gathering the information for the CSTR collection

suggestions to a central coordinator.

## FUTURE DIRECTIONS

So far our efforts have been devoted to setting up the current version of the NZDL system. It provides a unique facility for information access, and we have received enquiries from a number of organisations internationally. We are also surveying and critiquing other digital library projects, and in some cases setting up links with them. Table 1 shows digital library projects and related information around the world; further information on each can be accessed from our Internet site (see caption).

A reference librarian using the CSTR collection would immediately notice several limitations. The range of years covered is narrow, since computing departments generally have only a few years of their reports available in computer repositories. The collection does not index all computer science technical reports available over the Internet (for example, the home pages of researchers will often include a scattering of their own work). The absence of formal cataloging makes it difficult to locate a single document by title or author. The strength of the collection lies in providing access to very recent research results which are often not yet formally published. The field of computer science changes very rapidly—indeed, the charge is often laid that an article is out of date by the time it appears in a print journal—and given the even longer publication cycle for a citation index for the field, printed computing reference guides cannot provide timely access to rapidly-changing research fronts.

In pursuit of our overall goal of developing network-based technology for creating and automatically maintaining collections, we are beginning to investigate a number of issues surrounding digital libraries. Some of these are concerned with how users will interact with systems in the future. For example, we are closely monitoring NZDL usage to study library users' needs, and looking at the possibility of novel interfaces that cater to a wide spectrum of users.

Others address more technical concerns. For example, we are developing more comprehensive methods for abstracting layout and bibliographic information from document files. One serious weakness in the current NZDL is that conventional bibliographic information is lacking: we are considering the possibility of using a large collection of bibliographic references to computer science papers, and attempting to match technical reports in our collection with the appropriate reference to an actual publication. This would increase the degree of organisation in our library enormously.

Another area in which we are interested is collaborative browsing. In practice people often work collaboratively and we would like to apply techniques of "computer-supported cooperative work" to allow two people to search collaboratively. A particularly interesting and potentially productive situation corresponds to the traditional librarian-assisted search in which one participant is an expert on the collection and how to use searching techniques, while the other is seeking a particular piece of information. We see this as a vital role for librarians in the digital libraries of the future.

Finally, we are assessing potential topic areas for future experimental collections. While questions of copyright are of central concern to digital libraries, we have decided not to address them in our project, which is primarily concerned with technical issues. From a technical point of view, imposing restrictions on access to our collections is not difficult. Consequently we are most interested in public information, and in particular information of a national character which would be of widespread interest to people in this country. We would be very interested to receive ideas about what might be suitable.

## CONCLUSION

The new realities of Internet publishing are that information is provided in a widely distributed manner and it is up to the consumer to locate what is needed. Those who use the Internet regularly find what they want by employing search engines to seek out all electronic documents that contain certain combinations of words, names, or even acronyms. This is a researcher's dream—the ability to call up all relevant information at the click of a mouse. Large, high-profile, invariably US-based, these systems digest and index vast tracts of information. Voraciously and indiscriminately, they collect all text that anyone, anywhere, places on computers attached to the Internet. As a result, people have to sift through enormous volumes of trash before they find what they want. The dream is becoming a nightmare.

| | | |
|---|---|---|
| **Digital Library Research Projects** | Berkeley University<br>Berkeley University Library<br>British Library<br>Carnegie-Mellon University (Informedia project)<br>Carnegie-Mellon University (Universal Library project)<br>Library of Congress<br>Stanford University<br>Tufts University (Perseus project)<br>University of California at Santa Barbara<br>University of Illinois<br>University of Indiana Music Library (Variations project)<br>University of Michigan | |

| | | |
|---|---|---|
| **Computer Science Technical Report Systems** | Consortium of German universities | MEDOC |
| | Consortium of US universities | NCSTRL |
| | Karlsruhe University | Bibliography collection |
| | NASA | NTRS |
| | Queen Mary and Westfield College | Research interests collection |
| | Stanford University CS Department | |
| | University of Colorado | Harvest |
| | University of Indiana | UCSTRI |

| | |
|---|---|
| **Lists of DL Resources** | *Communications ACM* Digital Libraries special issue (April 1995)<br>Cornell University<br>Digital Library initiative<br>Johns Hopkins University<br>Tennessee Technological University<br>Texas Library Resource Sharing Program<br>Universiteit Antwerpen<br>University of Maryland Baltimore County<br>Yahoo! |

| | |
|---|---|
| **Lists of CS Technical Report Resources** | Computer Science Technical Report Archive Sites<br>Cornell Technical Report site list<br>Electronic Journals<br>NASA<br>Simon Fraser University<br>Yahoo! |

| | |
|---|---|
| **Related Projects and Information** | American Memory<br>Annotated Bibliography of Digital Library related sources<br>D-Lib magazine<br>Definition and Purposes of a Digital Library<br>De Montfort University Institute for Electronic Library Research<br>DIGLIB mailing list<br>International Federation of Library Associations and Institutions<br>Internet Public Library<br>Loughborough Library Electronic Journals Service<br>SuperJournal<br>Tennessee Library report on Digital Libraries<br>Text Encoding Initiative<br>US Navy |

Table 1 Related Digital Library projects, resources, and information
(see under "Related Work" at http://www.cs.waikato.ac.nz/~nzdl)

Our vision is rather different: the development of focused collections of high-quality information in particular areas. The New Zealand Digital Library project is developing novel software infrastructure that enables those who manage and maintain such collections to make them publicly available. As a small, geographically isolated country, we find the struggle to keep up with the explosion of printed information increasingly difficult. We suffer from diseconomies of scale in a world of exponentially growing information. We stand to benefit greatly from networked digital libraries. This new technology will help New Zealand become an international centre of expertise in the area.

## REFERENCES

Adobe Systems Incorporated (1985) *PostScript language reference manual*. Addison Wesley, Reading, Massachusetts.

Reed, T., Nevill-Manning, C.G. and Witten, I.H. (in preparation) "Extracting text from PostScript." Computer Science Department, University of Waikato, New Zealand.

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

## ABOUT THE AUTHORS

Ian H. Witten is a Professor of Computer Science at the University of Waikato, having come to New Zealand four years ago from the University of Calgary, Canada. His research interests include digital libraries, text compression, machine learning, programming by example, and interactive systems, and he has published around 150 refereed papers and six books including *Managing Gigabytes: Compressing and indexing documents and images* (Van Nostrand Reinhold, 1994, co-authored with A. Moffat and T. Bell). He directs the New Zealand Digital Library project and another large research project on machine learning and its application to agriculture.

Sally Jo Cunningham is a Lecturer in Computer Science at the University of Waikato, having come to New Zealand five years ago from Louisiana State University, USA. As well as a doctorate in Computer Science, she has a degree in the humanities (Asian Studies) and has held a summer internship at the Library of Congress. She is Research Editor of the Journal of Library and Information Science Research (LIBRES). Her research interests include digital libraries, machine learning, computer education, and computer applications in textiles, and she has published widely on an eclectic variety of topics. She is Objective Leader for *Collections and Usage Studies* in the New Zealand Digital Library project.

Mark D. Apperley is a Professor of Computer Science and Chairperson of Department at the University of Waikato. His research interests include all aspects of human-computer interaction, interface design, computer-supported cooperative work, and visual programming, and he has published widely in journals and conferences. He is Objective Leader for *User Interfaces for Readers* in the New Zealand Digital Library project, and also directs another large research project on tools for computer-supported cooperative work.