A Public Digital Library based on Full-Text Retrieval: Collections and Experience

Ian H. Witten[1], Craig Nevill-Manning[2], Rodger McNab[1] and Sally Jo Cunningham[1]

[1] Department of Computer Science, University of Waikato,
Hamilton, New Zealand.

[2] Department of Biochemistry, Stanford University.

{ihw,rjmcnab,sallyjo}@cs.waikato.ac.nz, cnevill@stanford.edu

The NZDL aims to impose structure on anarchic and uncataloged repositories of information, providing information consumers with effective tools to locate and peruse what they need. Our goal is to produce an easy-to-use digital library system that runs on inexpensive computers at information providers' own sites and offers an information service that information providers themselves maintain. New Zealand's geographical isolation magnifies the benefits of networked digital libraries, in terms of both cost and timeliness of access to information. We are collaborating with the Medoc project in Germany to provide local indexes to German language technical-reports, and with the *J Biological Chemistry* in the United States to field-test novel browsing techniques.

The project rests on five basic planks. First, we avoid manual processing of source material, and avoid making assumptions about the document repositories from which it is collected—e.g. we do not require bibliographic metadata. Second, access is via a full-text index of the entire contents of each document, rather than document surrogates. Third, we are concerned with user interface aspects and the real needs of library users. Fourth, our systems must operate in geographically remote locations 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 operates on small, inexpensive, servers.

Full-text indexes are provided to several substantial collections of information. These collections serve as case studies. They drive our research by providing technical challenges for indexing, and human interface challenges for retrieval. This article focuses on the collections: technical details of mechanism [10], protocols [5] and novel prototype interfaces [2, 8] are available elsewhere.

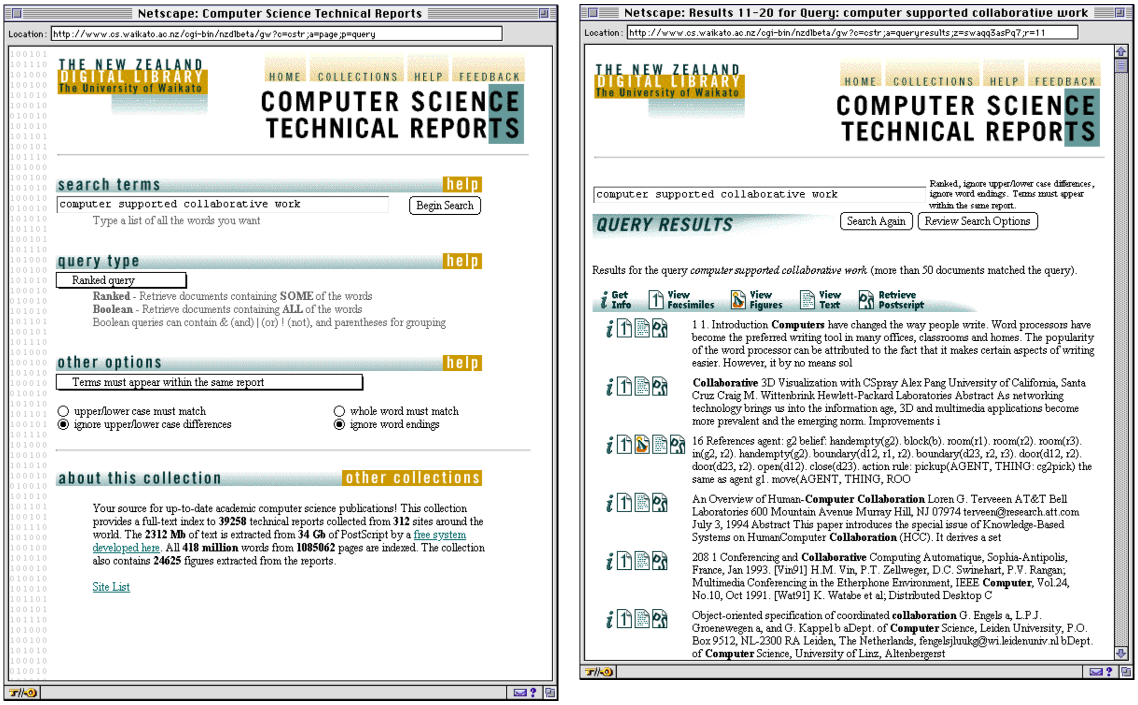
THE COMPUTER SCIENCE TECHNICAL REPORT COLLECTION

We have indexed 40,000 Computer Science technical reports—one million pages, 400 million words—from 300 sites internationally. This represents 2.3 Gb of text, extracted automatically from 34 Gb of PostScript/PDF source [7].

Querying and retrieval. Figure 1 shows a typical query and response. Searching facilities include ranked and Boolean queries; the ability to stem individual terms and/or render them case-insensitive; phrase searching; document, page, or first-page-only searches. Since the collection is not formally cataloged users cannot perform conventional author/title/date searches, but most reports include this sort of bibliographic information in the first page and fielded search can be approximated by restricting the query's scope appropriately.

The response shows the first few words of retrieved documents. Buttons allow users to retrieve the original document's URL, size, creation date, and download date; the PostScript itself; a facsimile image of the first page; the document text with query terms highlighted; and (where possible) the figures it contains. The extracted text and figures enable a quick scan to determine whether the document is worth downloading.

Collection maintenance. Other technical report servers suffer frequent maintenance problems caused by changes in the bibliography file format, and inconsistencies in the information, in the repositories that they index—which is why we do not use cataloguing information stored with the repository itself. The



computer supported collaborative work

Type a list of all the words you want

Ranked query

Ranked - Retrieve documents containing **SOME** of the words
Boolean - Retrieve documents containing **ALL** of the words
Boolean queries can contain & (and) | (or) ! (not), and parentheses for grouping

Terms must appear within the same report

○ upper/lower case must match
● ignore upper/lower case differences

○ whole word must match
● ignore word endings

Your source for up-to-date academic computer science publications! This collection provides a full-text index to **39258** technical reports collected from **312** sites around the world. The **2312 Mb** of text is extracted from **34 Gb** of PostScript by a free system developed here. All **418 million** words from **1085062** pages are indexed. The collection also contains **24625** figures extracted from the reports.

computer supported collaborative work

Ranked, ignore upper/lower case differences, ignore word endings. Terms must appear within the same report.

Results for the query *computer supported collaborative work* (more than 50 documents matched the query).

1 1. Introduction **Computers** have changed the way people write. Word processors have become the preferred writing tool in many offices, classrooms and homes. The popularity of the word processor can be attributed to the fact that it makes certain aspects of writing easier. However, it by no means sol

Collaborative 3D Visualization with CSpray Alex Pang University of California, Santa Cruz Craig M. Wittenbrink Hewlett-Packard Laboratories Abstract As networking technology brings us into the information age, 3D and multimedia applications become more prevalent and the emerging norm. Improvements i

16 References agent: g2 belief: handempty(g2). block(b). room(r1). room(r2). room(r3). in(g2, r2). handempty(g2). boundary(d12, r1, r2). boundary(d23, r2, r3). door(d12, r2). door(d23, r2). open(d12). close(d23). action rule: pickup(AGENT, THING: cg2pick) the same as agent g1. move(AGENT, THING, ROO

An Overview of Human-**Computer Collaboration** Loren G. Terveeen AT&T Bell Laboratories 600 Mountain Avenue Murray Hill, NJ 07974 terveen@research.att.com July 3, 1994 Abstract This paper introduces the special issue of Knowledge-Based Systems on HumanComputer **Collaboration** (HCC). It derives a set

208 1 Conferencing and **Collaborative** Computing Automatique, Sophia-Antipolis, France, Jan 1993. [Vin91] H.M. Vin, P.T. Zellweger, D.C. Swinehart, P.V. Rangan; Multimedia Conferencing in the Etherphone Environment, IEEE **Computer**, Vol.24, No.10, Oct 1991. [Wat91] K. Watabe et al; Distributed Desktop C

Object-oriented specification of coordinated **collaboration** G. Engels a, L.P.J. Groenewegen a, and G. Kappel b aDept. of **Computer** Science, Leiden University, P.O. Box 9512, NL-2300 RA Leiden, The Netherlands, fengelsjluukg@wi.leidenuniv.nl bDept. of **Computer** Science, University of Linz, Altenbergerst

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

collection is maintained automatically by periodically examining the technical report repositories for changes and updating accordingly. New sites are detected by various means (mostly manual) and added to the index.

Size and scalability. The public-domain MG search engine is tailored for highly efficient storage of full-text databases [9]. The text in this collection is compressed to 840 Mb; several indexes are added, totaling 700 Mb; and first-page facsimiles and extracted figures add another 680 Mb, for a total of 2.2 Gb—coincidentally, just the same size as the original text, and 6% of the total PostScript source.

All sizes increase linearly with the volume of text. Retrieval is independent of database size, taking two disk seeks per query term and two per document retrieved. Database inversion is a potential limiting factor, but a recently-developed algorithm can invert an estimated 5 Gb of raw text in twelve hours with only 40 Mb of main memory [6]. Collections of tens of gigabytes seem quite feasible. However, extrapolation on the basis of raw size is a gross oversimplification: other factors—such as number of documents or richness of vocabulary—can send things awry. Our actual experience extends to rich-vocabulary (library catalog) collections of 10 million short documents, and 3.5 Gb of raw text; inversion times are a few hours.

As collections grow, it will inevitably become necessary to partition the database. Partitioning may be desirable in any case—indeed the separation between collections in our current system is dictated by user needs rather than technical limitations. Scalable full-text indexes exist: experiments indicate that the results of retrieval performed in parallel on the segments of a partitioned database can be combined with little degradation in effectiveness [3]. We will rely on such techniques to provide full scalability.

OTHER COLLECTIONS

To demonstrate how digital libraries can benefit diverse groups of users we have built many other collections of publicly-available information.

- *The Computists' Communique* and *TidBITS* magazines: searchable on individual news item, its title, and individual paragraph.
- *FAQ Archive*: searchable within entire FAQ list, by subject heading, or individual paragraph.
- *The HCI Bibliography*: searchable by reference entry, with or without abstract.

- *Humanity Development Library*: humanitarian and development information collected by the Global Help Project.
- *Indigenous Peoples*: position papers, resolutions, treaties, UN documents, speeches and declarations on social, political, strategic, economic and human rights issues.
- *Oxford Text Archive* and *Project Gutenberg*: public-domain collections of English text.

Major differences between these collections include the source and format of information, updating policy, granularity of searching (document, page, paragraph, etc.), different kinds of index (titles, references, abstracts, etc.), structure and format in which output is displayed, and provision of summaries of the collection's contents—e.g. lists of titles and hierarchical, Web-accessible browsers like that shown in Figure 2. The challenge is to enable information providers to tailor the system to new document sets without programming.

MUSIC COLLECTIONS

Our "melody index" retrieves music on the basis of notes that are sung, hummed, or played [4]. Users can literally sing a few bars and have melodies containing that motif retrieved. Such systems will enable researchers to analyze music for recurring themes or duplicated phrases, and musicians and casual users alike will retrieve compositions based on remembered (even imperfectly remembered) passages.

Querying and retrieval. The system transcribes melodies automatically from microphone input, searches a database for tunes containing similar melodic sequences, and ranks matches using features such as melodic contour, musical intervals and rhythm. Figure 3 shows the response when a user sang the first eight notes of *Auld Lang Syne*. The transcribed input appears at the top; titles of similar items, ranked according to matching score, appear below. Any of the tunes may be selected for audio replay or visual display.

Collection maintenance. The database comprises 9,400 international folk tunes (half a million notes) stored in musical notation, an amalgam of the Digital Tradition database of North American folk songs with the Essen database of European and Chinese melodies. Though few machine-readable score collections are presently available, optical music recognition technology will change this. We have built a prototype service that accepts images of music and returns corresponding audio or MIDI files (based on [1]). Music submitted for processing can be automatically added to the database of indexed tunes.

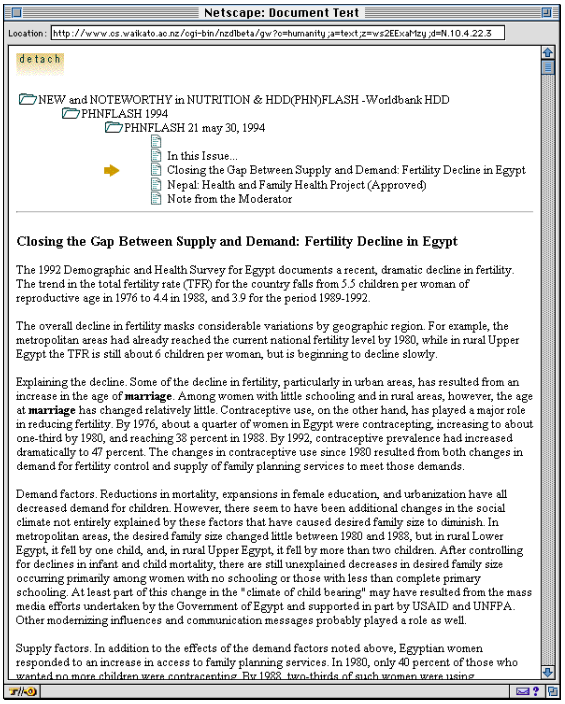
COLLECTION CREATION AND DISTRIBUTION

It takes a skilled technician a few hours to build a simple collection and integrate it into the digital library framework, excluding any time needed to gather the information. There is remarkable variety in the requirements for different collections: we have already created a dozen and experience shows that each poses new problems. Differences occur in the format of the source information, interface structure, indexing needs, and maintenance/update regime. We estimate that the number of collections must triple before we reach a state where we can be reasonably confident that new ones can be accommodated within the existing framework. When we reach that point we plan to design an end-user interface for collection building.

Subsets of the collections can be written to 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 benefit from the same interface and facilities.

FUTURE WORK

The NZDL is organized as several independent collections, each showing what the underlying technology can do in a different area. Joint searches could be enabled simply by constructing a combined index. But interesting issues arise when cross-searching different kinds of information—e.g. computer science technical reports, bibliography collections, and paper titles and abstracts—because automatic relevance ranking must be compared for different kinds of text, and because user interface issues are non-trivial. For example, a paragraph-level search on a document collection might be initiated if a title search fails, or a search on a name might be extended from a bibliography to a report collection if few items match.



Netscape: Document Text
Location : http://www.cs.waikato.ac.nz/cgi-bin/nzdlbeta/gw?c=humanity;a=text;z=ws2EExaMzy;d=N.10.4.22.3
detach

NEW and NOTEWORTHY in NUTRITION & HDD(PHN)FLASH -Worldbank HDD
PHNFLASH 1994
PHNFLASH 21 may 30, 1994
In this Issue...
Closing the Gap Between Supply and Demand: Fertility Decline in Egypt
Nepal: Health and Family Health Project (Approved)
Note from the Moderator

Closing the Gap Between Supply and Demand: Fertility Decline in Egypt

The 1992 Demographic and Health Survey for Egypt documents a recent, dramatic decline in fertility. The trend in the total fertility rate (TFR) for the country falls from 5.5 children per woman of reproductive age in 1976 to 4.4 in 1988, and 3.9 for the period 1989-1992.

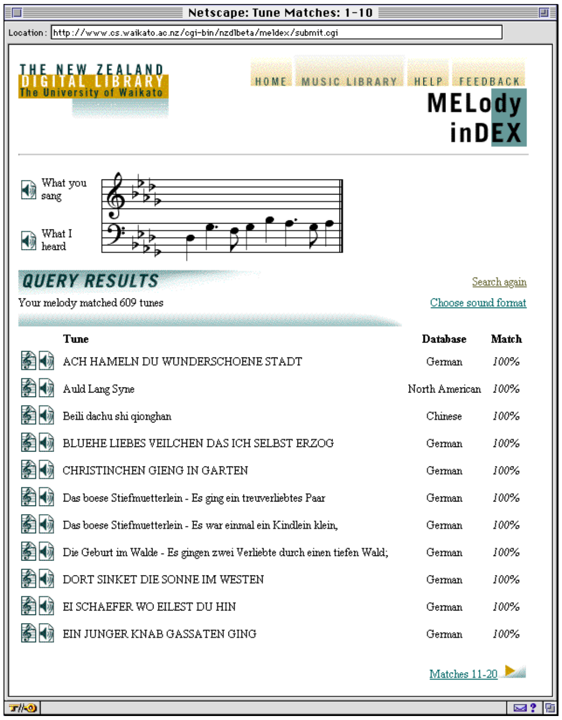
The overall decline in fertility masks considerable variations by geographic region. For example, the metropolitan areas had already reached the current national fertility level by 1980, while in rural Upper Egypt the TFR is still about 6 children per woman, but is beginning to decline slowly.

Explaining the decline. Some of the decline in fertility, particularly in urban areas, has resulted from an increase in the age of **marriage**. Among women with little schooling and in rural areas, however, the age at **marriage** has changed relatively little. Contraceptive use, on the other hand, has played a major role in reducing fertility. By 1976, about a quarter of women in Egypt were contracepting, increasing to about one-third by 1980, and reaching 38 percent in 1988. By 1992, contraceptive prevalence had increased dramatically to 47 percent. The changes in contraceptive use since 1980 resulted from both changes in demand for fertility control and supply of family planning services to meet those demands.

Demand factors. Reductions in mortality, expansions in female education, and urbanization have all decreased demand for children. However, there seem to have been additional changes in the social climate not entirely explained by these factors that have caused desired family size to diminish. In metropolitan areas, the desired family size changed little between 1980 and 1988, but in rural Lower Egypt, it fell by one child, and, in rural Upper Egypt, it fell by more than two children. After controlling for declines in infant and child mortality, there are still unexplained decreases in desired family size occurring primarily among women with no schooling or those with less than complete primary schooling. At least part of this change in the "climate of child bearing" may have resulted from the mass media efforts undertaken by the Government of Egypt and supported in part by USAID and UNFPA. Other modernizing influences and communication messages probably played a role as well.

Supply factors. In addition to the effects of the demand factors noted above, Egyptian women responded to an increase in access to family planning services. In 1980, only 40 percent of those who wanted no more children were contracepting. By 1988, two-thirds of such women were using

Figure 2 Browsing the result of a query for *marriage* in the Humanity Development Library



Your melody matched 609 tunes

Tune	Database	Match
ACH HAMELN DU WUNDERSCHOENE STADT	German	100%
Auld Lang Syne	North American	100%
Beili dachu shi qionghan	Chinese	100%
BLUEHE LIEBES VEILCHEN DAS ICH SELBST ERZOG	German	100%
CHRISTINCHEN GIENG IN GARTEN	German	100%
Das boese Stiefmuetterlein - Es ging ein treuverliebtes Paar	German	100%
Das boese Stiefmuetterlein - Es war einmal ein Kindlein klein,	German	100%
Die Geburt im Walde - Es gingen zwei Verliebte durch einen tiefen Wald;	German	100%
DORT SINKET DIE SONNE IM WESTEN	German	100%
EI SCHAEFER WO EILEST DU HIN	German	100%
EIN JUNGER KNAB GASSATEN GING	German	100%

Figure 3 Using the melody index

Value can be added to collections by mining the full text and correlating the contents with different sources. Examples include matching bibliography collections against technical report collections to infer metadata; eliminating duplication by self-matching bibliography entries; identifying reference lists and inserting hyperlinks to cited documents in the collection (and in bibliographies); analyzing document sets to infer topic clusters. Phrases suitable for a browsable hierarchical index can be inferred directly from the text, and help users build intuition about the content of a collection [8].

CONCLUSION

The NZDL will enable institutional end-users to create focused information collections in particular areas—in contrast to the less selective approach taken by Internet search engines—and make them publicly available. We have paid particular attention to the specific needs of a variety of document sources, providing flexible index granularity as well as browsing and searching interfaces tailored to content.

Digital libraries can be constructed whenever repositories of suitable text exist. Extracting all search information from the documents themselves is feasible if full-text indexing is used, and eliminates the manual cataloguing involved in library creation and maintenance. Although New Zealand's geographical isolation provided the initial impetus for making access to information in electronic form simpler and more efficient, the techniques we have devised are applicable internationally.

ACKNOWLEDGMENTS

This research is supported by the New Zealand Foundation for Research, Science and Technology. Thanks to all members of the project team, and to our information providers.

REFERENCES

[1] Bainbridge, D. and Bell, T.C. (1996) "An extensible optical music recognition system." *Proc Australasian Computer Science Conference*, Melbourne, 308–317.

[2] Beale, R., McNab, R.J. and Witten, I.H. (1997) "Visualizing sequences of queries: a new tool for information retrieval." *Proc IEEE Information Visualization*, London, 57–62.

[3] de Kretser, O., Moffat, A., Shimmin, T. and Zobel, J. (1997) "Methodologies for distributed information retrieval." Submitted for publication.

[4] 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. ACM Digital Libraries*, 11–18.

[5] McNab, R. and Witten, I.H. and Boddie, S.J. (1997) "A distributed digital library architecture incorporating different index styles." Submitted to *Advances in Digital Libraries*.

[6] Moffat, A. and Bell, T.A.H. (1995) "In situ generation of compressed inverted files." *J. American Society for Information Science* 46(7): 537–550.

[7] Nevill-Manning, C.G., Reed, T., and Witten, I.H. (in press) "Extracting text from PostScript" *Software—Practice and Experience.*

[8] Nevill-Manning, C.G., Witten, I.H. & Paynter, G.W. (1997) "Browsing in digital libraries: a phrase-based approach," *Proc ACM Digital Libraries*, 230–236.

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

[10] Witten, I.H., Cunningham, S.J. and Apperley, M.D. (1996) "The New Zealand digital library project" *D-Lib magazine* <www.dlib.org>, November.