THE NEW ZEALAND DIGITAL LIBRARY
The University of Waikato

PRESS RELEASE

Computer scientists at the University of Waikato are developing a pioneering digital library system that promises to change the way people find information on the Internet.

The migration of information from paper to computers is already altering the way people track down things they want to know. Those who use the internet regularly find what they want by using searching systems, called 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. Invariably located in the US, 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.

The research group at Waikato is pursuing a rather different track by developing focused collections of high-quality information in particular areas of research and scholarship. In a pilot project designed to explore the potential of network-based digital libraries, the group has set up a Digital Library for Computer Science. This provides access to over 20,000 research reports from around 250 sites worldwide -- over half a million pages of information. It is available nationally, to universities, CRIs, government departments, and even schools. (You will find it on the internet at `http://www.cs.waikato.ac.nz/~nzdl`). A unique facility, it is attracting international attention; a German version is already operating at a site in Bonn.

To provide access to this collection, there were challenging technical problems to solve. Computer Science reports are generally stored in a format called PostScript that can't be read by ordinary internet search programs. Researchers at Waikato devised software to extract plain text from PostScript, quickly and reliably, so that each page and document could be indexed under all the words it contained.

Although the full reports remain at their original sites overseas, the plain text of each is transferred to Waikato and stored there. Researchers can search and browse through the collection without having to use the slow, and relatively expensive, trans-Pacific computer link. As well as the plain text, a facsimile image of the first couple of pages of each report is also stored locally, along with a pointer to the remote computer and directory where the original resides. In some cases all graphical figures in the original document are also extracted and transferred to New Zealand.

People can access particular documents by specifying any combination of words in them -- so-called "full-text retrieval". Search terms can be combined with logical operators like "and", "or" and "not" to create a more specific query. Alternatively a general query can be issued and the top few documents -- the ones matching most closely -- are returned. In either case queries are answered very quickly, and for each matching document all the information stored in New Zealand is instantly made available for browsing, along with a pointer that allows the original formatted document to be downloaded from the overseas site.

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, the Waikato project is also planning to construct other large collections of publicly-available information. The researchers are interested in identifying particular areas and providing access to as many documents as possible within that area, without restricting the pool of source documents by requiring them to be in any special format. The underlying philosophy is to avoid any manual processing of the information going into the library. Searches by author, date, keyword, etc., cannot be made because the research group do not want to force information providers to do anything special to make their documents available.

Two small example collections of English literature (500 or so books) have been made available. There is also a collection of "frequently asked questions" -- questions, with answers, that people have accumulated on thousands of topics, from classical guitar playing to sea kayaking.

This innovative project shows one way of dealing with the new realities of Internet publishing, where information is provided in a widely distributed manner and it is up to the consumer to locate what is needed. As a small, a 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 and the Waikato project is poised to become a national -- and international -- centre of expertise in the area.

Ian H. Witten (ihw@cs.waikato.ac.nz)
Last modified: Fri Aug 23 17:02:32 NZST 1996