Breaking the Browsing Barrier for Historic Searching of Newspaper Texts.

Abstract

Traditionally, to find information in a newspaper collection it has been necessary to painstakingly browse through the individual issues hoping that relevant words will catch your eye. Researchers of New Zealand's encounter history received a boost when, in 1996, the Alexander Turnbull Library produced a collection on microfiche totaling almost 18,000 pages and covering a printing period from 1842 to 1933. While having all the newspapers in a single collection was a large step forward, browsing or searching for information in this microfiche collection was still time consuming.

By incorporating the collection into a digital library with an Internet interface, and by enabling full-text search, we have broken this browsing barrier. Making the information available this way is even more significant as the majority of the collection is written in the Māori language; we have created a Māori language resource that is sorely needed by education and research institutes, and one that provides quick and accurate access to the previously obscure source. The process involved in developing this unique digital library collection, the advantages of traditional newspaper stored in this medium, and the possibilities that we intend to investigate in the future, will be discussed in this paper.

The Māori Newspapers Collection

In 1988 the Alexander Turnbull Library[1] began scouring the libraries of New Zealand for historic newspapers that were either published in Māori or for a Māori audience. The purpose was to assist in the preservation of those rapidly decaying newspapers and to make them more readily available. Over 35 periodicals were identified, totaling over 17,700 individual pages. The Alexander Turnbull Library referred to this collection as the 'Māori Newspapers', or 'Niupepa 1842-1933'.

The Māori Newspapers form a unique and valuable resource for historians, sociologists, theologises and linguists. They contain a large amount of information on historical events and opinions embodying New Zealand's early and turbulent antiquity. The newspapers provide distinctive perspectives from government sources, from religious factions, and from indigenous sectors. Approximately 70% of the newspapers were written only in Māori and 27% were written bilingually, in Maori and English. Thus the collection also provides an immense and desperately needed text source for scholars, teachers, and students of the Māori language.

The Māori Newspapers was initially captured on microfilm, and in 1996 it was made available on microfiche and distributed to libraries throughout New Zealand. This has been an important breakthrough for researchers as it has meant that they can now access all the Māori Newspapers from the one library. However searching the collection is very time consuming. Researchers must analyse the bibliographic material to try and determine which newspapers could be pertinent. Then they must manually browse those newspaper hoping that the relevant texts will catch their eye.

[1] an entity within the National Library of New Zealand specialising in rare heritage documents

The New Zealand Digital Library

The New Zealand Digital Library (NZDL) is a research project based in the Computer Science Department at The University of Waikato. Its aim is to develop the underlying technology for digital libraries and make it available publicly so that others can use it to create their own collections. It has developed a suite of digital library software called Greenstone which will support heterogeneous, multilingual, distributed digital libraries.

The NZDL currently supports about 30 collections which cover quite a diverse range. The variations in collections include size and number of documents, language of the collection and of the interface, the types of search interfaces, the browsing and indexing structures, and the formats of the documents including texts, web pages, and multimedia. These collections are available over the WWW (http://www.nzdl.org) and can also be easily stored and made available on CD-ROM.

Motivation

Our motivation for building a collection in the NZDL of the Māori Newspapers was two-fold. First we wanted to make the collection readily available on the WWW to all researchers and students of the Māori language for no cost. In particular this would allow Māori medium education, where there is a paucity of Māori medium materials, access to a significant Māori language resource. Second, we wanted to cast off the barriers of browsing on a microfiche reader and advance the collection into an environment which incorporated a powerful full text search capability.

Building The Niupepa Collection

We created a NZDL version of the Māori Newspapers Collection and called it simply the Niupepa Collection. To do this we required two main sets of documents. First a digital facsimile of each newspaper page as the preferred deliverable to the user. Second, an electronic text version of each newspaper page that would enable the Greenstone Software to perform the full text search function on the collection.

1) Digital Facsimile

We sourced the 17,700 images from the 35mm microfilm version[2] of the Newspaper Collection, as opposed to the microfiche version, because it was the original version, and was also easier to automate. The different periodicals varied greatly in information density from booklet size of (in mm) 210x140 to tabloid form of 605x445. The individual pages also varied a great deal in image quality from crisp clean images to images that had been tarnished by mould, ink spots, or poor focus, or faded beyond recognition.

When digitizing the microfilm we generated both bitonal (b&w) and grey-scale images that were compressed as .tif files and stored on CD-ROM. Testing showed that 300dpi was the most suitable resolution for the OCR[3] process. The bitonal images were 200-300Kbytes each and were stored on 8 CD-ROMs. The grey-scale images were 5-10Mbytes each and required 90 CD-ROMs. Both forms of the image were captured because we suspected that the grey-scale images would be more

[2] with gratitude to the Alexander Turnbull Library
[3] Optical Character Recognition

suitable for the user display and more accurate for the OCR process. In practice we found the bitonal images suitable for both display and OCR, and due to the smaller file size, were a lot easer to manage.

Further image enhancement was required which included cropping and deskewing, and as the collection used a page level index, periodicals that were captured as double pages required splitting into single pages. These enhanced images were then made available to the OCR process.

The images also needed to be resized and redefined into a format suitable for the WWW. We created two images that we made available to the user interface; a large scale image suitable for reading the actual text of the newspaper and a smaller preview image that showed the whole page in one screen. Both images were stored in a .gif format, the larger image was 50-200kbytes and the preview images 10-20kbytes.

2) Text Extraction

To extract the text files we required an OCR package that would support the Māori language and could also support the use of an additional dictionary that could be updated in the verification process. We found the FineReader® 5.0 software package to be the most suitable. It has support for 176 different languages, including Māori, and a very accurate recognition engine. Though the recognition accuracy was very much dependant on the density and quality of the image file.

The OCR process consisted of loading a 'batch' of images, 'blocking' the images, 'training' a recognition pattern, 'reading' the blocks, 'verifying' the subsequent text that had been generated and correcting it where necessary, and then saving the text file to an appropriate format. We employed Māori literate typists for the OCR process and found that it could take anyway between 2 and 15 minutes per page. The variation again being due to the density and quality of the image file.

A consistent naming convention was used which enabled the Greenstone software to easily access adjoining pages in a periodical, or corresponding formats of an individual page.

The User Interface

The Niupepa Collection can be freely accessed on the WWW[4] where it has been implemented as a collection of the New Zealand Digital Library. On the website there are 3 principle methods of accessing newspaper pages; browse by title, browse by date and search by content. It is the latter method where the full text search potential is realised, and the browsing restrictions of the microfiche reader are cast aside.

The website's default language is Māori due to the nature of the collection, due to the funding support, and to make a statement about the Māori language i.e. an indigenous language can find purpose in a computer environment such as a digital library. It should be noted that the interface language can be easily switched to one of 10 other languages.

The modern Māori language character set uses a macron symbol over a vowel to signify that the vowel sound should be lengthened. We incorporated this into our user interface by using the Unicode macron symbols to represent the lengthened vowels[5]. However recognition of macron

[4] www.nzdl.org/niupepa/
[5] this is an official standard certified by the Taura Whiri i te Reo Māori (*NZ Māori Language Commission*)

characters has not been incorporated into the search facility as the orthography of the original newspaper texts does not indicate whether the vowels sounds are lengthened or not.

Enhancements

We have been fortunate to be able to incorporate two additional sets of material into the Niupepa Collection. The first is a 'commentary' about each periodical that was written by Gail Dallimore. Content in the commentary includes bibliographic data, background information, a listing of main subjects covered in the periodical and physical locations of specific issues of the periodical.

The second additional set on material is a result of some work being undertaken by a team headed by Jane McRae at the University of Auckland. They have been producing 'abstracts in English'- a hypertext linked summary for each issue of certain newspapers. This gives non-Māori literate readers incites into what has been written.

The commentaries and English abstracts can be accessed through the 'Browse by Title' screen or by specifying them as an option while performing a search function.

Future Developments

As the digitalization of the collection nears completion we will begin to maintain logs of searching and browsing interactions. We will use these logs to obtain insights into our user preferences and will try and tailor our interface to the needs of the user population.

Possible enhancements to the user interface could include; a graphical timeline where the time period may be selected by moving an adjustable slider along a timeline; a search by selecting a certain location in a map; high-lighting on the facsimile image areas that match the search criteria.

Conclusions

An important and unique collection of historic documents has been captured in digital form and made conveniently accessible over the Internet. There are cost considerations and extra work may be required to undertake OCR with a minority language but the rewards that having a full text search ability bring, justify the extra effort. Removing the browsing barriers opens up the resource quite significantly.

The techniques developed are equally applicable to a wide range of legacy text collections, especially those written in an indigenous language. The resultant digital resource has the potential to contribute significantly to the promotion and preservation of language and culture.

Acknowledgements

This work has been carried out with the support of the Alexander Turnbull Library, who provided the original newspaper images on 35mm film. Image capture was performed by New Zealand Micrographic Services Ltd. We are also very grateful to the New Zealand Ministry of Education for their continued financial support of this project.

Bibliography

ATL (1996) Niupepa 1842-1933, Microfiche set, Alexander Turnbull Library, Wellington, New Zealand.

Dallimore, Gail. (1990) He Arahi, He Tohu o Nga Pepa te Maori: A Bibliography of Maori Newspapers, 1840-1900.
Unpublished research report.

Garlick, Jennifer. (1995) Maori Language Publishing — Some Issues. Wellington, Huia Publishers.

Jones, S., Cunningham, S.J. and McNab, R. (1998). "An analysis of usage of a digital library", Proceedings of the European Conference on Digital Libraries ’98, Heraklion, Lecture Notes in Computer Science no. 1513, Springer, 261-277.

McNab, R.J., Witten, I.H., and Boddie, S.J. (1998) "A distributed digital library architecture incorporating different index styles", Proceedings of Advances in Digital Libraries ’98, IEEE CS Press, Los Alamitos, Calif., 36-45.

Williams, Sheila. (1990) "The Maori Language Printed Collections", Turnbull Library Record 23(1), 12-18, May.