Working Paper Series
ISSN 1170-487X

The Niupepa Collection: Opening the Blinds on a Window to the Past

by Te Taka Keegan, Sally Jo Cunningham and Mark Apperley

Working Paper 99/16
December 1999

Department of Computer Science
The University of Waikato
Private Bag 3105
Hamilton, New Zealand

The Niupepa Collection: Opening the Blinds on a Window to the Past

Te Taka Keegan, Sally Jo Cunningham and Mark Apperley

Department of Computer Science,
University of Waikato,
New Zealand

Abstract

This paper describes the building of a digital library collection of historic newspapers. The newspapers (*Niupepa* in Māori), which were published in New Zealand during the period 1842 to 1933, form a unique historical record of the Māori language, and of events from an historical perspective. Images of these newspapers have been converted to digital form, electronic text extracted from these, and the collection is now being made available over the Internet as a part of the New Zealand Digital Library (NZDL) project at the University of Waikato.

1. Introduction

The importance of print media was realized early in New Zealand's history, initially by missionaries who, in 1835, began the local publication of religious material. In 1840 the British Colony was established, and in 1842 the first Government periodical, *Te Karere o Niu Tireni* (The Messenger of New Zealand), was printed. In 1859 two young Māori chiefs from Ngāti Apakura, in the Waikato, returned from Poland with a printing press and began *Te Hokioi* (The War Bird), the first publication written from a Māori perspective. These developments represent the three distinctive perspectives of early publications: religious, government, and Māori perspectives.

The latter half of the 19th century saw a rise in the number of periodicals published, with more than forty separate titles aimed at a Māori audience appearing between 1842 and 1930. Many of these were published for only a few years, although some titles continued for much longer periods. A sharp decline was noted in the number of Māori publications in the 1900-1960 period (Williams, 1990). This decline has been closely associated with the 1871 amendment to the Education Act which prohibited the use of the Māori language in schooling throughout New Zealand (McRae, 1983).

In 1988 the Alexander Turnbull Library, in conjunction with the National Library's Microfilm Production Unit, and with the cooperation of libraries throughout New

Zealand, undertook a major project to microfilm these newspapers. This project concentrated on those newspapers published in Māori, or for a Māori readership, and formed a collection called *Niupepa 1842-1933*. In 1996 this collection became available in microfiche form (ATL, 1996).

The microfiche version of the Niupepa Collection is contained in 407 fiche pages, which cover 40 titles and some 19,000 individual pages. Approximately 70% of this collection represents newspapers written entirely in Māori language. Of the remainder, 27% contains parallel Māori and English text, and just under 3 % is written in English alone. Consequently the collection preserves a valuable historical collection, and a much-needed text source for scholars, teachers and students of the Māori language, showing not only a Māori perspective of New Zealand's formative history, but also the variation of written Māori over time, variation in translation over time, tribal variations in language usage, and the creation of new terms. However, while the microfiche-stored media remains secure, it is difficult to read, and a collection of this size and variation is not easily browsed, let alone searched.

2. The New Zealand Digital Library

The New Zealand Digital Library project (http://www.nzdl.org) has developed an architecture to support heterogeneous, multilingual, distributed digital libraries. We currently support about 20 collections, which range in size from a few documents up to 10 million documents; may vary in the language used in the documents, or in the language displayed in the search interface; use different browsing and indexing structures; can contain text, images, and audio; may be accessible over the WWW, or stored and searched on CD-ROM; and may physically store documents in a single site, or distribute them across hundreds of sites worldwide.

This complexity is managed by a novel, flexible macro language interface to support the creation and maintenance of digital library collections (McNab et al, 1998). Rather than viewing a digital library as a single, monolithic group of documents, our design is based on *collections*—sets of like documents—that may require radically different search, storage, and indexing strategies. For example, the Arabic Library uses storage and search mechanisms that handle non-ASCII alphabets; a collection containing French documents requires a French stemmer to support truncation of search terms; the Local Oral History collection manages audio files and image files linked to the searchable text; and the Music Library directly requires a radically different searching and indexing structure, as it permits direct searching of audio.

A collection developer first determines the focus for a prospective collection and selects the collection's documents (or document sources—some collections are constructed by "harvesting" items from existing WWW sites). Building a collection often requires significant effort to make documents suitable for display or indexing; for example, the Māori newspaper collection discussed in this article required the digitalization of microfilm images, and custom software was developed to extract indexable text from PostScript for the Computer Science Technical Reports collection (Nevill-Manning et al, 1998). Indexes can be constructed to search document metadata (title, author, publication details, etc.) if available, or to search the document content at the desired levels of granularity (complete document, individual pages,

paragraphs, sections, etc.). For text documents, we use MG to construct the indexes and store documents (Witten et al., 1994). MG typically compresses text to about 25% of its original size, and compresses indexes to about 7% of the original text's size—making the total storage requirement about one-third of the original text's size. Other index types can be slotted into the digital library architecture; for example, the Music Library uses MR to index and search audio files (McNab et al, 1996).

The searching and browsing facilities provided for a particular collection are, of course, dependent on the types of indexes specified for the collection. For text indexes, the digital library architecture supports the common search engine options: stemming, truncation, phrase searching, and searching at different index granularities. Structured documents can be browsed by document section, consecutively by ordered documents in a series, and so forth. Transaction logs of user queries can be automatically maintained, and the logs can be semi-automatically analyzed to identify problems with the search interface and to provide insight into preferred user interaction patterns (Jones et al, 1998).

3. The Conversion Process

To provide an NZDL-based version of the Niupepa collection required two distinct sets of data. First, to facilitate the full-text search capability of the NZDL, the newspaper content needed to be available in electronic text form. Second, in order to preserve the form and integrity of the original newspapers, it was decided that a digital image of each page should be held as the preferred deliverable content to the user. Other data sources were also considered; for example, a comprehensive bibliography of the collection had previously been generated (Dallimore, 1990), and another research group in currently producing abstracts of the collection. Within the NZDL it is possible to integrate such resources into a single collection.

The first stage in acquiring the two principal data forms was to have digital images produced for all 19,000 pages of the collection. The most convenient form in which the images were available was on 35mm film. These images varied greatly in quality and in information density. Some of the original newspapers were crisp black and white, others were on discoloured paper with mould and ink spots almost obliterating parts of the text. Figure 1 shows an example of a poor quality image, with misalignment between the two pages of the opening. Figure 5 shows an example of one of the better quality images. The original pages varied from booklet size (210x140) to tabloid form (450x320), leading to significant variations in information density. Each 35mmm frame typically contained one opening (two pages) of a newspaper. For reasonably reliable OCR (optical character recognition) from the digital images, it was determined that scanning densities needed to correspond to approximately 300dpi on the original newspaper page. For one of the larger format newspapers, this meant an image of approximately 20×10^6 pixels.

Because of the set-up costs for digitising each of the 35mm films, both bitonal (b&w) image and grey-scale images were generated at the same time. These were produced as compressed .tif files and written to CD-ROM. The bitonal images each occupied approximately 200-300Kbytes, and the entire collection in this form required 8 CD-ROMs. The grey-scale images, however, were much larger (5-10Mbytes each), and the entire collection in grey-scale form spread over 90 CD-ROMs. Both forms of

image were captured because it was considered that grey-scale images would offer more scope for parameter adjustment during the OCR phase, but that the bitonal images would provide a more compact (and perfectly readable) form for the collection itself, and a faster delivery of content to the user.

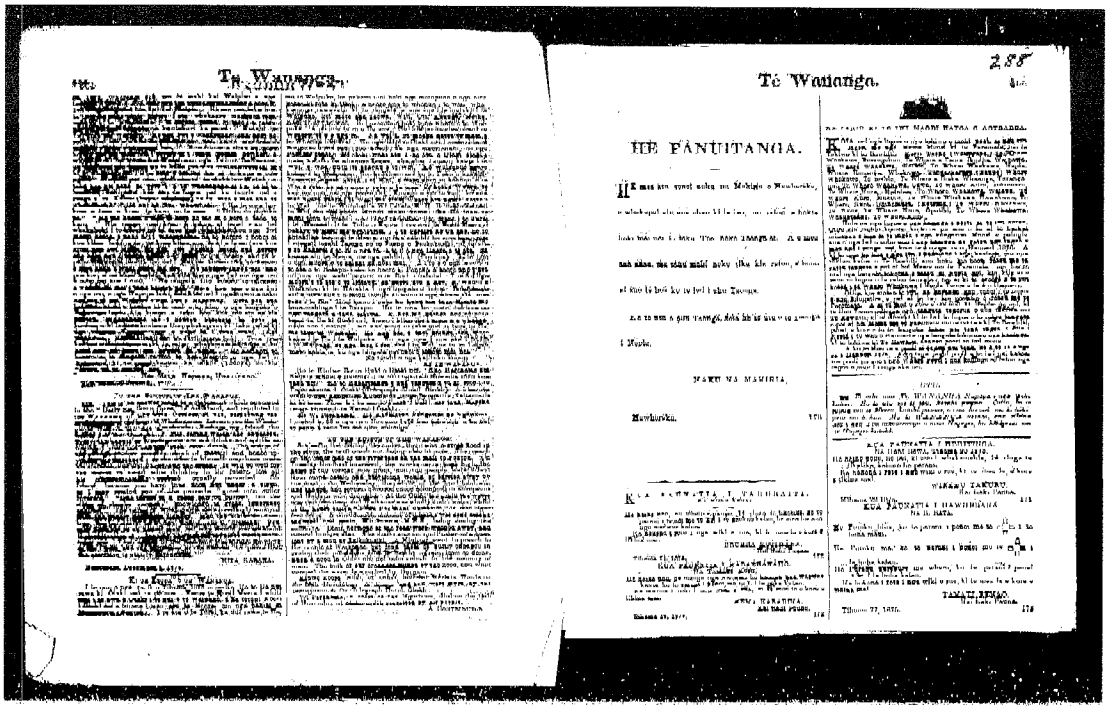

Figure 1: An example of a poor quality image from the Niupepa collection on film.

The second stage of the conversion process was to generate electronic text from the digital images. The principal technique has been to use the OmniPage™ software to automatically extract the text from the digital image of each page. Although the recognition accuracy for this process at the character level is reasonably high (75%), errors are widely distributed within a page, leading to few words recognised without any error. A post-processing stage was introduced following the OmniPage™ scan. This uses a PPM language model, a dynamic programming algorithm which investigates alternative substitutions to maximise the compressibility of the text (Teahan et al, 1998). This post-processing improved the accuracy of the electronic text to more than 90%. Finally, this text was checked against the original images by typists literate in the Māori language, and any residual errors corrected. It is worth noting that as the text is to be used only for indexing, with the digital images the principle form for delivering content to users, the electronic text does not need to be 100% accurate. However, it was a requirement that key search terms, such as people, places and dates, did need to be spelt correctly; in checking the accuracy of the text for these, in general every word has been verified.

An alternative, but much more labour intensive, approach to generating the electronic text has been to manually key in the text from the digital images. Although excellent accuracy has been achieved with this method, it is seen as neither practicable nor desirable in the long term for capturing collections of this size.

The Niupepa collection, in its present form, uses a page-level index. The text for each page in the collection is held in a separate file. These text files, plus the corresponding bitonal image files, together with files containing commentaries and bibliographic information on the titles (Dallimore, 1990), form the NZDL collection, which has been constructed using the facilities described Section 2 (McNab et al, 1998). Searching for a particular term or phrase returns a list of those pages in which the term appears. From this list, hyperlinks provide direct access to the text itself, where the search term(s) appear highlighted. From the text display, in turn, hyperlinks provide direct access to the corresponding image pages. More detail of the user interface is provided in the next section.

Approximately 12% of the total collection has so far been captured as text, and is freely accessible at the NZDL site (http://www.nzdl.org/npepa).

4. The User Interface

The web browser interface to the Niupepa collection, as with almost all the collections on the NZDL, is available in either Māori or English versions. The home page provides three facilities for accessing the collection.

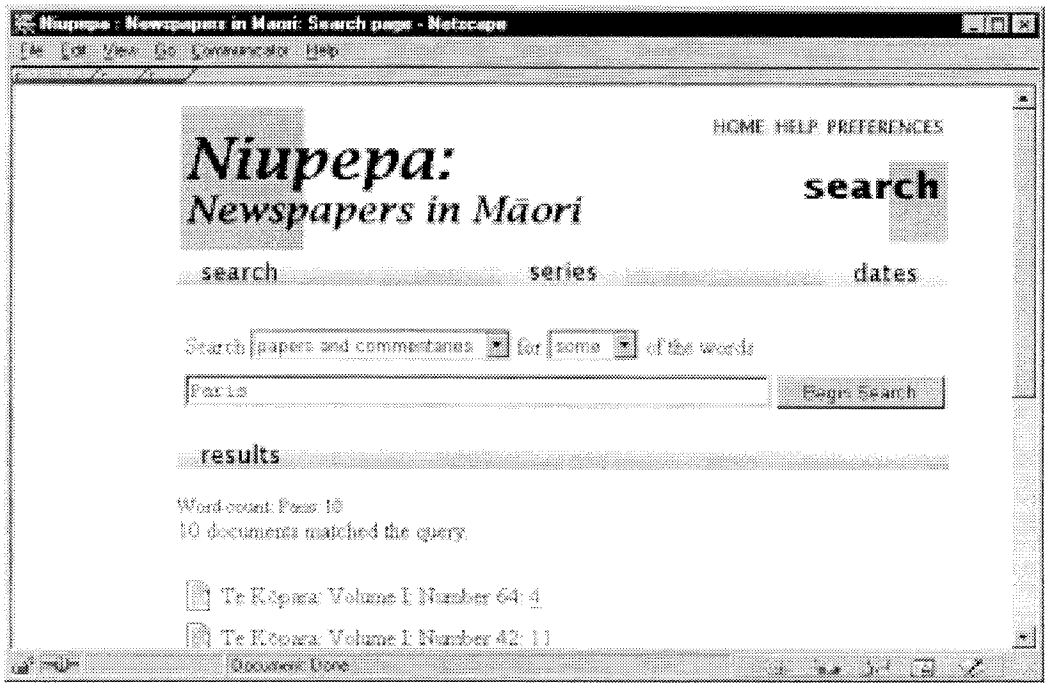

Figure 2: An example of the response to a search query.

The *Rapu*, or *Search* facility provides the standard NZDL full-text search capability, utilising the electronic text extracted from the images as described in the previous section. The collection is divided into two sub-collections, the newspapers themselves, and the bibliographic commentaries. A search can be confined to either of these sub-collections, or can be extended to cover both simultaneously. Other search options include case sensitivity, word stemming, and searching for some or all of the words in the search phrase. An example of the result of a search is shown in Figure 2.

The *Whakarārangi Taitara* or *Series* facility provides the user with the ability to browse through an index of titles and, at a lower level, a chronological index of the individual issues within a title. From this latter index the user can directly access both the text and image pages of a specific issue. Part of the lower-level index for the title *Te Karere o Nui Tireni* is shown in Figure 3.

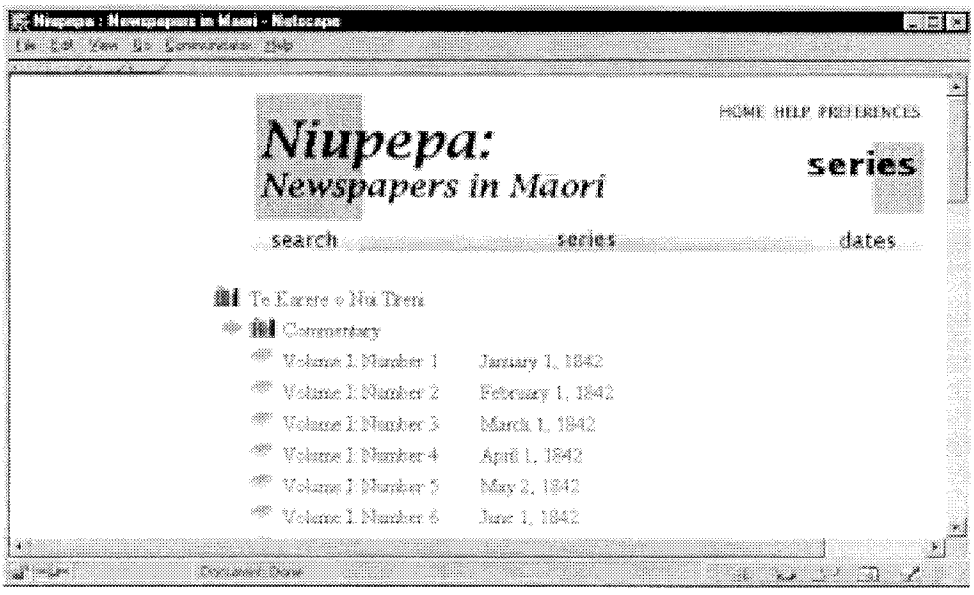

Figure 3: The **Series** option provides access to an index of titles and issues.

The *Wātaka* or *Date* facility allows the user to brows through the entire collection chronologically. Given the short life-span and sporadic publication record of some of the titles, this is a very useful index when seeking reports on historical events. This index, as shown in Figure 4, provides access to individual issues in the collection by year and month.

Whatever method of access is used, once a particular page has been found it can be displayed in one of two formats; a text display of the electronic text extracted from the original images, or a facsimile image display generated directly from the stored .tif image file. Usually the text version is the most appropriate initial display following a search query; search terms are highlighted within the displayed text, and are thus relatively easy to find. However, most users then prefer to switch to the

facsimile image display to capture the original context of the item. Figure 5 shows the NZDL display of a facsimile image.

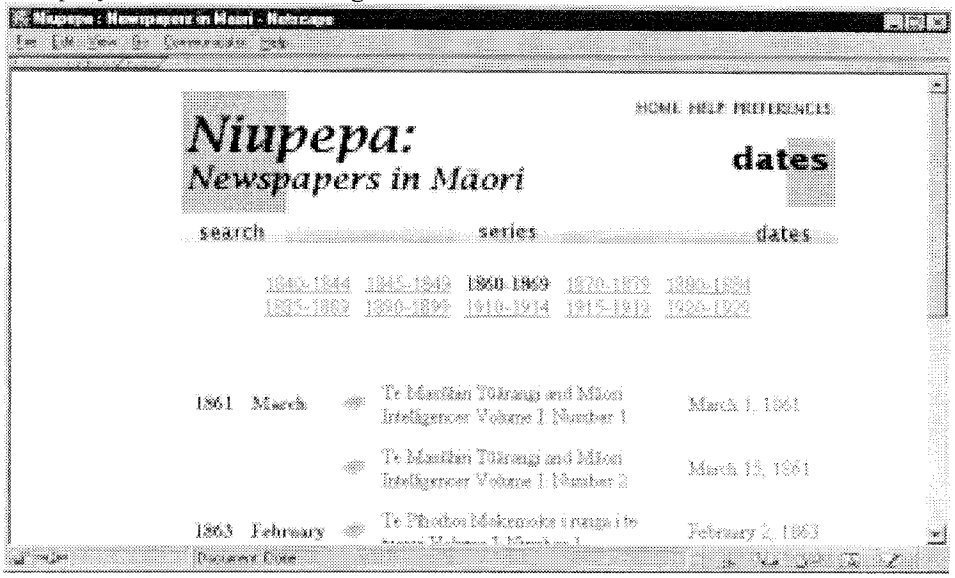

Figure 4: The collection can also be browsed chronologically using the *Dates* option.

Figure 5: An example NZDL display of a facsimile Niupepa page.

Conclusion

The Niupepa Collection provides a rich source of information for a wide variety of researchers including historians, sociologists, theologians and linguists. It provides a unique window to New Zealand's encounter history, the Māori culture and the Māori language. However, in their original newspaper form, or in the medium of microfilm and microfiche, opening the blinds to this window has been a tedious and very time consuming task. By making the Niupepa Collection available in this digital form, and over the Internet, we have not only removed the blinds, but we have also replicated the window that traditionally resided in libraries, into the many homes, schools and businesses that have an Internet connection.

It should be obvious too that major additional benefits arise from the full-text search and browsing facilities provided by the NZDL. The full-text search capability adds enormously to the value of the collection, significantly expediting the task of finding a person, a place name, an event, or an unusual word or. This facility opens up research possibilities based on the Niupepa that were never before imaginable.

Acknowledgements

This work has been carried out as a part of the New Zealand Digital Library project, and with the support of the Alexander Turnbull Library, who provided the original newspaper images on 35mm film. Image capture was carried out by New Zealand Micrographic Services Ltd.

References

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

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

Frean, Nicola. (1990) "Niupepa in the Turnbull" in *Turnbull Library Record,* **23**(1) 19-21, May.

Garlick, Jennifer. (1995) *Mäori Language Publishing – Some Issues.* Wellington, Huia Publishers.

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

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.

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.

McRae, Jane. (1983) "Māori Manuscripts – Who's Responsibility?". In *Archifacts* **4**, 2-6.

Nevill-Manning, C.G., Reed, T., and Witten, I.H. (1998) "Extracting text from PostScript" Software-Practice and Experience, **28**(5) 481-491, April.

Teahan, W.J., Inglis, S., Cleary, J.G. and Holmes, G. (1998) "Correcting English text using PPM models." Proc Data Compression Conference, edited by J.A. Storer and M. Cohn, 289-298. IEEE Press, Los Alamitos, CA.

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

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