

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

**Te Taka Keegan, Mark Apperley,
Sally Jo Cunningham and Ian H Witten**

*Department of Computer Science,
University of Waikato,
New Zealand*

Abstract

This paper describes the construction and initial usage of a digital library collection of historical newspapers written in the Maori language. The newspapers (Niupepa in Maori) total over 17000 individual pages from 35 separate periodicals, and were published in New Zealand during the period 1842 to 1933. They not only form a large and desperately needed source of Maori language text, but also provide a unique historical record of the Maori Language, and of New Zealand's early encounter history as seen from a Maori perspective.

Images of these newspapers have been digitised and the text extracted from these documents using OCR techniques. The two formats (document image and text) have been linked together as the Niupepa Collection in the New Zealand Digital Library (NZDL) at the University of Waikato. The collection, previously constrained to the browsing limitations of a microfiche reader, is now freely available over the Internet with a full-text search capability.

1. Introduction

Printing was introduced to New Zealand in 1835 by missionaries, to permit the local publication of religious material. In 1840 the British Colony was established, and in 1842 the first Maori language government periodical, *Te Karere o Niu Tireni* (The Messenger of New Zealand), was printed. In 1859 two young Maori chiefs from Ngati Apakura, in the Waikato, returned from Poland with a printing press and began *Te Hokioi* (The War Bird), the first publication written from a Maori perspective (Garlick, 1995). These developments represent the three distinctive viewpoints of early publications: religious, government, and Maori.

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 Maori 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 Maori 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 Maori language in schooling throughout New Zealand (McRae, 1983).

As the Maori newspapers slipped out of production, copies of them became increasingly rare—and, of course, the older newspapers were physically deteriorating. To preserve these heritage documents and make them more widely available, 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 Maori, or for a Maori 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, covering 40 titles and some 17,700 individual pages. Approximately 70% of the pages are written entirely in the Maori language, about 27% contains parallel Maori and English text, and just under 3% is written in English alone. As well as being a unique set of historical documents, the collection is also much-needed source of Maori text for scholars, teachers and students. Its pages preserve the colonial and Maori perspectives of New Zealand's formative history, the variation of written Maori over time, different approaches to translation between English and Maori, tribal variations in language usage, and the genesis of new Maori terms.

However, to date the Niupepa Collection has not been as widely used in New Zealand as its importance to scholars and interest to the general public might suggest. A major stumbling block has been its storage on microfiche and microfilm; these media require special readers to view the documents, and with either format a collection of this size cannot be easily browsed, let alone effectively searched.

We are making the Niupepa Collection available over the WWW, implemented as a collection of the New Zealand Digital Library (NZDL) project (the NZDL software is described in Section 2). Section 3 describes the digitization of the microfiche images and the extraction of the newspaper text for indexing. Since Maori is not very well supported by OCR (optical character recognition) packages, we had to try several methods, packages and configurations to obtain an appropriate accuracy in the text recognition process. This technique is not specific to the Maori language, and can be used to improve OCR results for any language that does not have full software support. The interface to the Niupepa Collection is presented in Section 4, and Section 5 concludes with a discussion of the advantages of making these documents accessible across the WWW.

2. The New Zealand Digital Library

The Niupepa Collection is implemented using the Greenstone digital library software, developed by the New Zealand Digital Library project (<http://www.nzdl.org>). The Greenstone architecture supports heterogeneous, multilingual, distributed digital libraries. We currently support about 30 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 Maori 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

Creating 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 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 is currently producing abstracts of the collection. Within the NZDL it is possible to integrate such resources into a single collection, and one planned extension to the Niupepa Collection is the addition of these sources as they become available.

The first stage in acquiring the two principal data forms was to have digital images produced for all 17,000 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, but 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 35mm frame typically (but not invariably) 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 20x106 pixels.

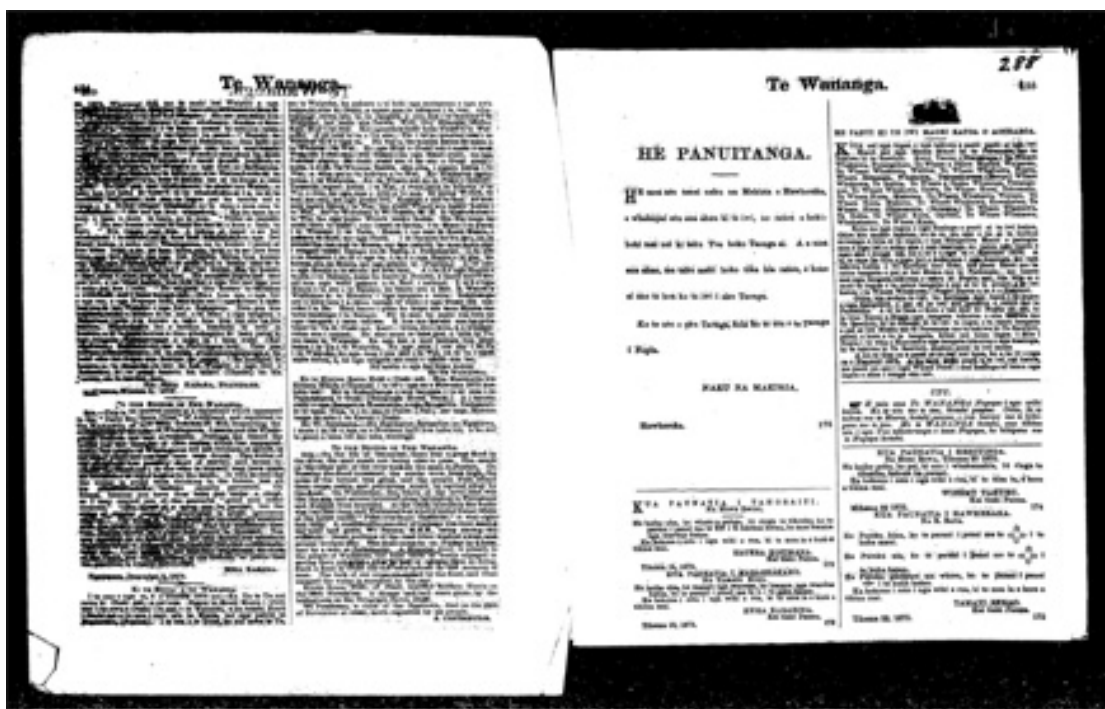


Figure 1: A poor quality image example from the Niupepa Collection on film.

Because of the set-up costs for digitizing 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.

The second stage of the conversion process was to generate electronic text from the digital images. Many methods and OCR packages were tried as there is very little software support for the Maori language. The most accurate technique that we have found has been to use FineReader® 5.0 OCR software which has to be configured to check for English-Hawaiian. The English language is selected because this enables us to use the dictionary function and we use it's additional dictionary facility to compare and update a Maori word list in the verification procedure. The Hawaiian language is selected because its preconfigured character set is more correct than the Maori language preconfigured character set.

The recognition accuracy of the OCR software is very much dependant on the quality and density of the image file. Averaging around 96% the character level accuracy will drop to low as 75% on a very dirty newspaper image. The final stage of this process is to check the text against the

original images by typists literate in the Maori language, and any residual errors corrected. This stage occurs within the FineReader® 5.0 program.

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, discussions with potential users indicate that for the Niupepa Collection the most common search terms will be personal and place names, and it is crucial for retrieval that these be spelled correctly.

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), together 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 35% 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, like nearly all the NZDL, is available in both Maori and English. The home page provides three facilities for accessing the Niupepa: full text searching, browsing by individual newspaper titles, and browsing by date of publication.

The Rapu or Search facility (Figure 2) 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 limited to one of these sub-collections, or can be extended to cover both simultaneously. The default query type is a ranked search ("search for some of the words"), and the alternative query type is an AND-ed Boolean query ("search for all of the words"). We deliberately chose to offer a simplified query interface, as previous

experience in building other collections indicated that these two query types form the vast majority of searches (Jones et al, 1998). More advanced options (stemming, case sensitivity, and number of hits displayed) are available through the "preferences" button in the upper right hand of the window.

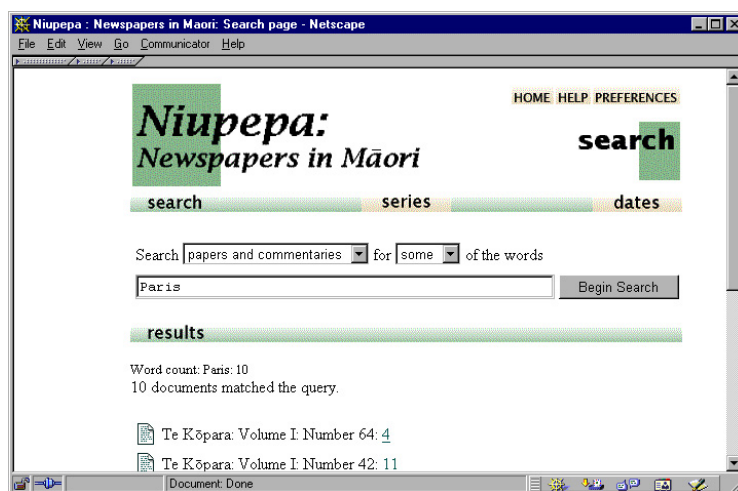


Figure 2: An example of the response to a search.

The Whakararangi 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 (Figure 3).

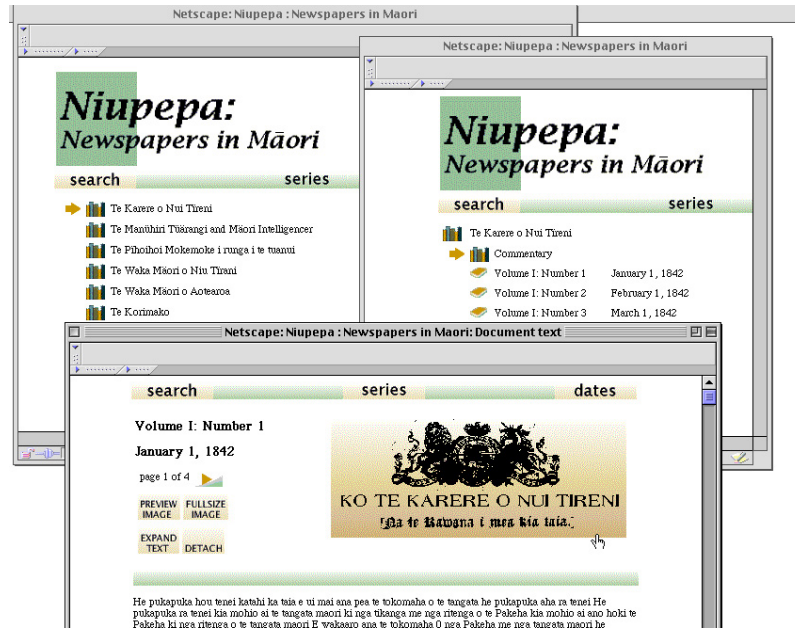


Figure 3: The Series option provides access to an index of titles and issue.

The Wataka or Date facility allows the user to browse 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.



Figure 4: The collection can be browsed chronologically in the Dates option.

Once a particular page has been found, it can be displayed in one of two formats: as text extracted from the original images, or as a facsimile image generated directly from the stored .tif 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 5. An example display of a facsimile Niupepa page.

5. Conclusion

The Niupepa Collection is a rich source of information for historians, sociologists, theologians and linguists; it provides a much-needed source of

Maori language text for New Zealand classrooms; and the collection is expected to be of interest to the general public as well. The Niupepa are a unique window to New Zealand's encounter history, the Maori culture, and the Maori language. However, in their original newspaper form, or in the media of microfilm or 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 accessible over the Internet, we have not only removed the blinds, but have also replicated the window that traditionally resided in libraries, into the many homes, schools and businesses that have an Internet connection.

An important feature of the WWW version of the Niupepa Collection is that the interface supports both browsing and full text searching. Users can easily scan issues from a given date, to identify co-occurring events; quickly isolate a particular series for sequential scanning; or pinpoint items of potential interest by searching for a personal name, place name, event, unusual turn of phrase, and so forth. These search capabilities open research possibilities on the Niupepa that were previously unimaginable. The collection is also expected to be of great interest to the New Zealand public, allowing the exploration of aspects of the country's heritage that were previously nearly inaccessible.

From the perspective of digital libraries research, this work provides an opportunity to investigate issues in information retrieval and information seeking for a language that has not previously been studied. As the

collection progresses to completion, and develops a user base, we will maintain logs of searching and browsing interactions. These transaction logs will be analyzed for insights into preferred user information seeking patterns, and the results used to tailor the search interface to the needs of the user population.

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 grateful to the New Zealand Ministry of Education for their 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., Smith, L.A., Witten, I.H., Henderson, C.L., and Cunningham, S.J. (1996) "Toward the digital music library: tune retrieval from acoustic input", Proceedings of Digital Libraries '96, Austin (Texas), 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) "Maori Manuscripts — Whose Responsibility?", 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.", Proceedings of the Data Compression Conference, edited by J.A. Storer and M. Cohn, IEEE Press, Los Alamitos, CA., 289-298.

Williams, Sheila. (1990) "The Maori Language Printed Collections",
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.