

Niupepa: An historical newspaper collection

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

Dept of Computer Science
University of Waikato, New Zealand
{m.apperley, sallyjo, tetaka, ihw}@cs.waikato.ac.nz

The Collection

Niupepa is a collection of forty-two newspaper titles published in New Zealand during the period 1842 to 1933, comprising a total of 21,000 pages in 1750 issues. These form a unique historical record of the language of the indigenous Māori people, the evolution of the written form of this language, and of events and developments during the formative colonial history of our country. Using the Greenstone software from the New Zealand Digital Library¹, this collection is being made publicly available with full-text search capability.

Data capture

The *Niupepa* material had earlier been gathered on microfiche² from original material in libraries scattered throughout the country. The delivery of a digital library version required two distinct forms of the original data. To facilitate full-text search, the newspaper content was first converted to electronic text using optical character recognition (OCR); to maintain the form and integrity of the original newspapers, a digital facsimile of the original page was preferred for viewing.

Data capture has involved scanning 21,000 images from 35mm photographic negatives. These images vary considerably in quality—some are clean, others are badly stained—and in information density. For reliable OCR, scanning densities corresponding to approximately 300dpi on the original newspaper page were needed. OCR has been performed using FineReaderTM, utilising a dictionary of Māori words to aid recognition. Nevertheless, it has been essential for fluent Māori speakers to check the text against the original images to correct remaining recognition errors.

Delivery

The *Niupepa* collection incorporates a page-level index, with text for each page held in a separate file. These text files, together with the digital facsimiles of the original pages, files containing commentaries and bibliographic information, and English-language abstracts of individual issues, form the digital library collection. Searching for a particular term or phrase returns a list of those pages in which it appears. From this list, hyperlinks provide direct access to the text itself, where the search term(s) appear highlighted, or to the corresponding image page.

Both Māori and English language versions of the interface are provided, and in addition to the full text search capability, the collection can be browsed by series, issue or date.

Capturing this invaluable resource on microfiche secured its preservation. Creating a digital library collection opens up access for cultural, sociological, linguistic, historic and even geographical research to laypeople, schoolchildren, and scholars at home and abroad.

References

1. Witten *et al.*, “Greenstone: Open Source DL Software” in this issue.
2. “Niupepa 1842-1933”, Microfiche set, Alexander Turnbull Library, Wellington, New Zealand (1996).

