Delivering the Maori-Language Newspapers on the Internet

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

Although any collection of historical newspapers provides a particularly rich and valuable record of events and social and political commentary, the content tends to be difficult to access and extremely time-consuming to browse or search. The advent of digital libraries has meant that for electronically stored text, full-text searching is now a tool readily available for researchers, or indeed anyone wishing to have access to specific information in a text.[1] Text in this form can be readily distributed via CD-ROM or the Internet, with a significant impact on accessibility over traditional microfiche or hard-copy distribution. For the majority of text being generated *de nouveau,* availability in electronic form is standard, and hence the increasing use of full-text search facilities. However, for legacy text available only in printed form, the provision of these electronic search tools is dependent on the prior electronic capture of digital facsimile images of the printed text, followed by the conversion of these images to electronic text through the process of optical character recognition (OCR). This article describes a project undertaken at the University of Waikato over the period 1999 to 2001 to produce a full-text searchable version of the Niupepa or Maori-language newspaper collection for delivery over the Internet.

Fundamental to this project has been the New Zealand Digital Library (NZDL) and its associated Greenstone software, de- veloped at the University of Waikato.[2] The Greenstone software architecture has been developed to support heterogeneous, multilingual, distributed digital libraries. Although it was motivated from a technological research perspective, it is a real system delivering real digital library technology, with a broad community of users internationally. The NZDL currently supports about 20

collections, which range in size from small collections of just a few documents to collections of up to 10 million documents. Greenstone collections are principally delivered via the Internet, but many are also available on self-contained CD-ROMs. Greenstone can accommodate documents in a wide variety of languages and formats, and the search interface can also be provided in different language forms. Browsing and indexing structures have been developed to accommodate different styles of collection, ranging from homogeneous collections of research papers, through structured collections such as newspapers, to diverse encyclopaedic collections containing pamphlets, books, audio clips and images. A more detailed description of the Greenstone software and its underlying philosophy is given in a later section of this article. The Niupepa project was undertaken as a challenge for Greenstone, in that the material was comprised solely of legacy printed documents, and as an attempt to make this valuable and unique collection more widely accessible and of even greater use to Maori and other readers.

Of the two principal phases of the capture task, the first, that of producing digital facsimile images of the newspaper pages, was relatively straightforward. The Niupepa collection was already available in microfiche form[3] so it was not necessary to use the original printed documents, and the handling of the material was simplified. The second phase, that of converting the images into electronic text, has been more technically challenging and time-consuming. The microfiche collection comprises 407 fiches, which cover 40 separate newspaper titles and some 19,000 individual newspaper pages. Approximately 70 per cent of the pages are written entirely in the Maori language, with the remainder mostly parallel English and Maori text and only a very small proportion (2 per cent) written solely in English. Challenges in this task included the extreme variability in the image quality of the original documents; variations in size and print density of different newspaper titles; and, most significantly, the fact that there was little or no prior experience of OCR of Maori text, and none of the commonly available proprietary OCR software packages usefully supported the Maori language.

The form of a digital newspaper collection

Facsimile images

There are two distinct phases to the digital capture of legacy printed material, although in some cases only the first will ever be carried out.[4] This first phase is that of producing a digital facsimile image - rather like a digital photograph - of the original material, made up of many individual dots or pixels. This process is usually referred to as 'scanning'. The quality of the image, or its fidelity to the original print material, is dependent both on the density of these dots (typically expressed as dots per inch, or dpi), and the accuracy or sensitivity with which the values of the dots are measured and/or recorded (typically expressed as bits per pixel). For example, if an image contains only black and white detail, then it might be quite accurately represented using just one bit per pixel, showing that a dot is either black or white. If an image contains a range of greys (or colours), it might require, say, eight bits per pixel (256 different shades of grey) or as many as 24 (three colours at 8 bits per colour) for accurate digital representation and reproduction. Both the dot density and the number of bits per pixel have a direct bearing on the amount of computer storage required to hold the image. For example, an A4 page scanned at 72 dpi and represented as just a black and white image requires approximately 63 kilobytes of storage, whereas the same page scanned at 300 dpi and with 256 grey levels occupies almost 9 megabytes.

However, regardless of the level of detail, the digital facsimile image obtained in this way is only a picture of the original document. Although it is stored digitally, this digital information relates to the picture, where it is dark and where it is light, and contains no specific information about the text characters that compose the picture. In fact, without very detailed and complex analysis, it is difficult to distinguish between a scan of a photograph and a scan of some text. In this form the image can be delivered from a digital library collection and viewed and read on a screen or printed on paper, but because the actual text is not part of the information, it is not possible to provide a full-text search facility. Image documents such as those held in digital library collections require the separate entry of metadata to catalogue them and to provide information by which they may be retrieved.

Optical character recognition and electronic text The second phase of the digital capture is optical character recognition (OCR). In this phase, the digital facsimile image described above is processed by sophisticated OCR software which identifies shapes in the image corresponding to letters or symbols from a standard character set. (Note that some software packages merge the two phases together and refer to the combined process as 'scanning' the image.) The output from this process is electronic text - a record of the characters themselves rather than their appearance. The recognition accuracy depends on a number of factors, including the sophistication of the software, the quality of the digital facsimile image, the visual quality of the original document, and the font used in the document. Typical recognition accuracies are in the range 70 to 99.99 per cent at the character level, depending on the quality of the original image. Most OCR software uses some knowledge of the language of the text (for example, a vocabulary) to assist with the recognition process. Normally a manual check is carried out as a part of the OCR phase, comparing the recognised text with the facsimile image, with corrections made as necessary.

Storage requirements for electronic text are relatively low, say, 4 kilobytes for a single space typed A4 page, because it is only the character information which is stored, and not a picture. However, as characters and words can now be identified by the computer, with the text available in this form, content-based searching (for example, full-text search) can be provided.

Clearly a major issue in building any digital library collection of legacy printed documents such as the Maori-language newspapers, is whether or not to perform OCR. OCR is a time-consuming and consequently expensive process, so the trade-off between the volume of material that can be converted to digital form for a given budget and whether or not the material will be made available as electronic text is a significant one. Other similar projects have opted to provide only digital facsimiles and not electronic text in order to save costs and/or to deliver more content for the same budget.[5] These collections are unable to provide content-based search, and depend on the manual entry of metadata by which to index and retrieve each item in the collection.

Because a basic tenet of the Greenstone software is its full-text search capability, and because of the immense value added to a newspaper collection through the provision of full-text search, this was not really an issue for this project. The digital Niupepa collection *would* provide full-text search and consequently it *would* be necessary to extract electronic text using OCR techniques. However, a survey of potential users indicated a strong preference for the actual delivery to be in the form of the digital facsimiles, to preserve the form and integrity of the original newspapers, rather than the relatively featureless electronic text. This suggested that the electronic text would be used only for indexing and retrieval purposes, and would seldom be displayed or read. Consequently it was suggested that electronic text of less than 100 per cent accuracy might be acceptable. There was also an implication that both the digital facsimile images and the electronic text versions of each page would need to be stored as a part of the collection.

Accessing the collection

One of the questions in the development of any digital library newspaper collection is determining the unit of delivery. For browsing, most readers work at the level of the individual issue of a paper. However, because of the relatively unstructured and heterogeneous nature of newspaper content, to respond to a search for a word or a phrase, a finer grained unit of delivery is appropriate; the reader does not wish to go through an entire issue to locate the phrase that has been found. Given that the unit of the facsimile image (the preferred form, as we noted above) is the newspaper page, the page is the obvious choice of unit for delivery. This has implications for the indexing level used in the Greenstone software - the outcome of a search should be to pinpoint the page or pages on which the target word(s) appear. These implications are discussed later.

It was anticipated that a range of different forms of access to the collection would need to be provided. In response to a search, a reasonable outcome would be a list of the pages where the target word or phrase occurred, from which the user could then retrieve each of those pages in turn. There may be occasions when a search would be restricted to certain date ranges or to certain newspaper

titles only. For browsing, the reader may wish to access the collection by newspaper title (series) then by individual issue and page, but an alternative means of access may well be by date - 'Give me access to any newspapers published in this particular month of this particular year.' These facilities have all been included in the digital Niupepa collection, and are discussed in more detail later, together with other novel forms of access which are being considered.

There is one other aspect of access to the collection which should be mentioned at this stage. There are two associated bodies of text which it was seen could readily be added to the collection to enhance its value. The first of these was a comprehensive bibliography of the collection that had previously been compiled, and which relates to each newspaper title.[6] The second is a set of English language abstracts currently being developed for the collection, which relate to individual issues, even down to page level.[7] It was desirable that these texts should be able to be integrated into the collection, and that access for browsing and searching them, as well as cross-referencing, should be provided.

The digital capture of the newspapers

To provide a digital library-based version of the Niupepa collection required the two distinct sets of data described above; for delivery of the content in facsimile form, digital images of each page were required, and to facilitate full-text search capability, the newspaper content needed to be available in electronic text form. The first of these provides the means by which the second is acquired.

Digital image capture

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 35mm film - the original photographs from which the microfiche collection had been generated. The photographs themselves were of good quality but the original print material which they captured varied greatly in quality and in information density. Some of the original newspapers were crisp black on white text, others were

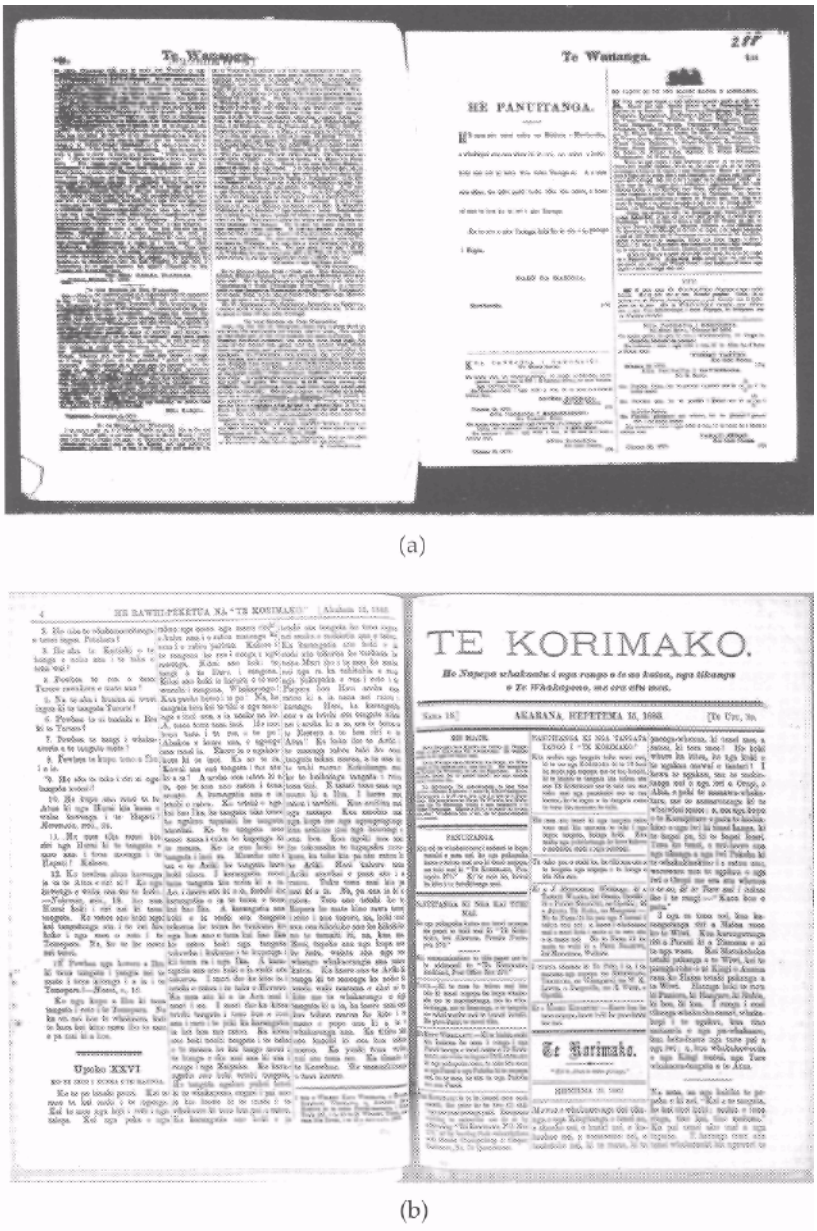

Figure 1: Example images on film, (a) is a poor quality image, with blurring and misalignment, while (b) shows one of the better images.

poorly formed text on browning paper, while still others had ink-stains, mould and pen annotations, sometimes obliterating parts of the text. Figure 1 shows two examples of images: (a) is a poor quality image, with misalignment between the two pages of the opening, while (b) shows one of the better quality images. The original pages varied from booklet size (210mm x 140mm) to tabloid form (450mm x 320mm), leading to significant variations in information density within the photographs. Each 35mm frame typically contained one opening (two pages) of a newspaper. For reasonably reliable OCR from the digital images, tests were conducted which determined that scanning densities needed to correspond to approximately 300 dpi on the original newspaper page; higher resolutions produced no noticeable improvement in recognition accuracy. For one of the larger format newspapers, this meant an image of approximately 20 million pixels.

With some 19,000 images to capture, the availability of the material on 35mm film was attractive; film, lends itself to (semi) automatic processing, not requiring manual intervention and setup for each individual page. This process was handled by an external organisation (New Zealand Micrographics), had access to scanning equipment capable of automatically handling 35mm film. Because of the set-up costs for digitising each of the 35mm films, both bilevel (black and white - 1 bit per pixel) images and grey-scale images were captured at the same time. These were produced in compressed tiff format and written to CD-ROMs.[8] The bilevel images each occupied approximately 200-300 kilobytes, and the entire collection in this form required eight CD-ROMs. The grey-scale images, however, were much larger (typically 5-10 megabytes 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 bilevel images would provide a more compact (and perfectly readable) form for the collection itself and, because of the smaller file size, a faster delivery of content to the user over the Internet.

The extraction of electronic text (OCR)

The second stage of the conversion process was to generate

electronic text from the digital images. Three different techniques were used over the lifespan of the project: manual data entry; OCR using the OmniPage™ software with post-processing software, and OCR using the FineReader™ software. It is the third of these techniques (FineReader™) which was used for the bulk of the conversion.

Manual data entry of text from, digital images is an alternative to OCR for generating electronic text but is much more labour intensive. Electronic text for a small part of the Niupepa collection was captured by this method because, at the time, the necessary labour resources were freely available. Although excellent accuracy was achieved with this method, it was seen as neither practicable nor desirable in the long term for capturing a collection of this size.

In the early stages of the work the OmniPage™ software was used to extract automatically the text from the digital image of each page. The recognition accuracy for this process at the character level was only 75 per cent and, as errors tended to be widely distributed within a page, few words were recognised without any error at all. Manual checking of the output text is always a requirement but with recognition accuracy as low as that, the correction process was very slow and arduous. One of the principal reasons for the relatively low recognition rate with OmniPage™ was the fact that it was not possible for the software to utilise any knowledge of the Maori language (alphabet, letter sequences, vocabulary, grammar), nor for it learn by example. Even worse, OmniPage™ would attempt to 'correct' Maori text into an English form. This is a problem, with the majority of proprietary OCR packages; they are packaged with dictionaries for a number of widely used languages (English, French, German, etc.) but are unable to be modified to work with user-defined dictionaries for minority languages, or to learn from, experience.

In an attempt to improve the recognition rates, a post-processing stage was introduced following the OmniPage™ scan. This used a PPM language model, a dynamic programming algorithm which investigates alternative substitutions to maximise the compressibility of the text.[9] This post-processing did improve the accuracy of the electronic text to more than 90 per cent.

However, for the latter 75 per cent of the OCR work, the project

used the FineReader™ OCR software which does allow for user-defined dictionaries, and which appears to be gaining wide acceptance as a suitable tool for this type of work.[10] With the Niupepa collection, FineReader™ produced recognition accuracies of 81 per cent at the word level and 95 per cent at the character level. This text was checked against the original images by typists literate in the Maori language, and any residual errors corrected. It is worth noting again that as the text is used mainly for indexing, with the digital images the principal form, for delivering content to users, the electronic text does not need to be 100 per cent accurate. However, it is a requirement that key search terms, such as people's names, place-names and dates, are spelt accurately if searching is to be reliable and, and by checking the accuracy of the text for these, in general every word has been verified,

A further alternative for the OCR work was considered, that of outsourcing the work. There are a number of organisations which offer an OCR service in this way. However, as it was evident that it would always be necessary to follow the OCR with a manual check, and as the manual check is time-consuming, little was to be gained by working this way, and the cost was not inconsiderable.

As indicated in the previous section, the Niupepa collection uses a page-level index. This means that the electronic text for each page in the collection is held in a separate file. These text files, plus the corresponding bilevel image files, together with files containing commentaries on the titles and the English-language abstracts, form the digital library Maori-language newspapers collection, which has been constructed using the Greenstone software described in the next section.

The Greenstone digital library system

As mentioned earlier, the Greenstone software of the New Zealand Digital Library provides a flexible and highly usable mechanism for generating and delivering real digital library collections. It can accommodate a range of document styles and forms, collection sizes, document languages, interface languages, browsing and searching structures, and storage and delivery mechanisms. This complexity is managed by a novel, flexible macro-language

interface to support the creation and maintenance of digital library collections.[11] Rather than viewing a digital library as a single, monolithic group of documents, the Greenstone 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 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 OCR work just described for the Niupepa collection. Custom software was developed to extract indexable text from PostScript for the Computer Science Technical Reports collection.[12] 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, a full-text search programme called MG is used to construct the indexes and store documents.[13] MG typically compresses text to about 25 per cent of its original size and indexes to about 7 per cent of the size of the original text, making the total storage requirement about one-third of that of the original text. Other index types can be slotted into the digital library architecture; for example, the Music Library uses a special music retrieval program to index and search files of musical notation.[14]

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 analysed to identify problems with the search interface and to provide insight into preferred user interaction patterns.[15]

The Greenstone digital library system makes its collections available over the Internet. Users of the Niupepa - or any other - collection employ an ordinary web browser, either Netscape or Internet Explorer, to access the collection at the NZDL web site. However, if the collection is likely to be used intensively in a particular organisation - for instance, libraries or universities in New Zealand using the Maori newspapers - it may be desirable to offer exactly the same collection on a local web server. This may speed access to the pages and reduce the load on the central system. Greenstone makes it very easy for other Internet servers to offer the collection - the system runs on all Windows and Linux platforms, and is freely available.

In fact, it is possible for any desktop or laptop computer to run the Greenstone software. Some collections are packaged together with the software on a CD-ROM that effectively contains a complete, self-contained copy of that collection. Greenstone is used for the dissemination of humanitarian information in developing countries - a realm where traditional publishing and distribution mechanisms have failed tragically. Access to the Internet is virtually unknown in such countries, although computing technology is often available through international programmes or donation schemes. Several Greenstone collections of humanitarian information have been produced, typically by international organisations such as UN agencies, on individual CD-ROMs, which are then distributed widely in places that the Internet does not reach-Delivering the Maori-language newspapers in digital form

The Niupepa digital library collection is provided over the Internet as one of the standard collections from the NZDL site. The web browser interface to the Niupepa collection, as with almost all the collections on the NZDL, is available in both Maori and English

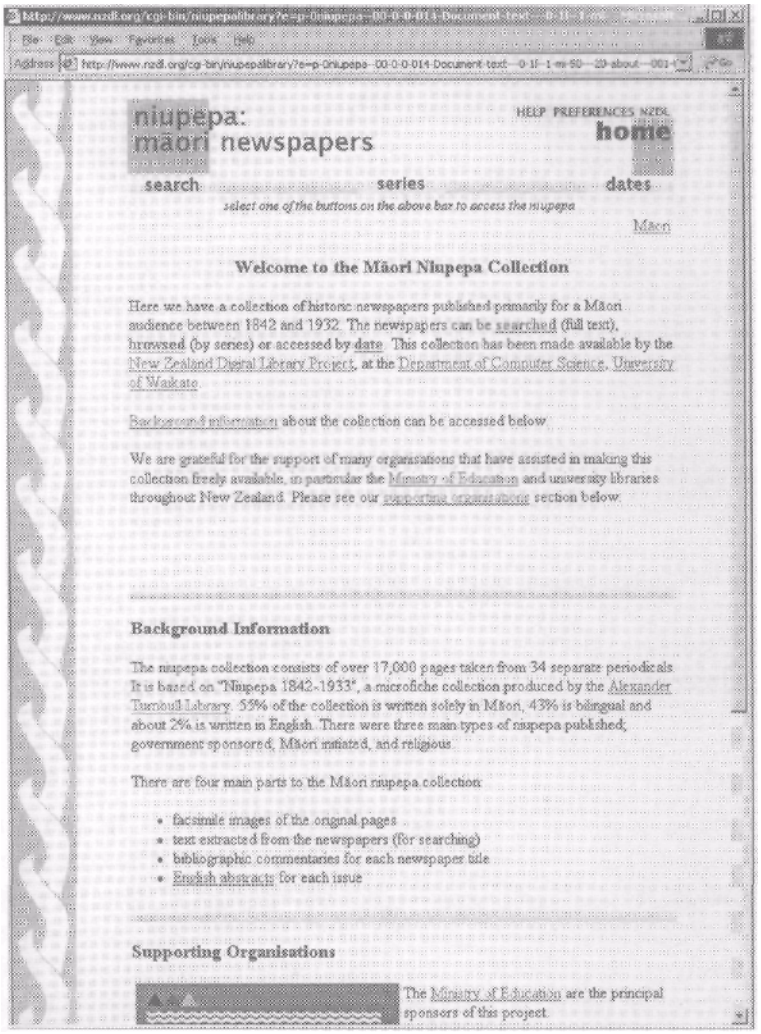

Figure 2: The Niupepa collection home page on the Internet.

language versions. The home page, shown in Figure 2, provides three facilities for accessing the collection: search, browse by series, and browse by date. The page also provides some background and explanatory information, the ability to switch from a Maori to an English interface, and from this page it is possible to initiate a search. Figure 2, and all subsequent screen shots in this chapter,

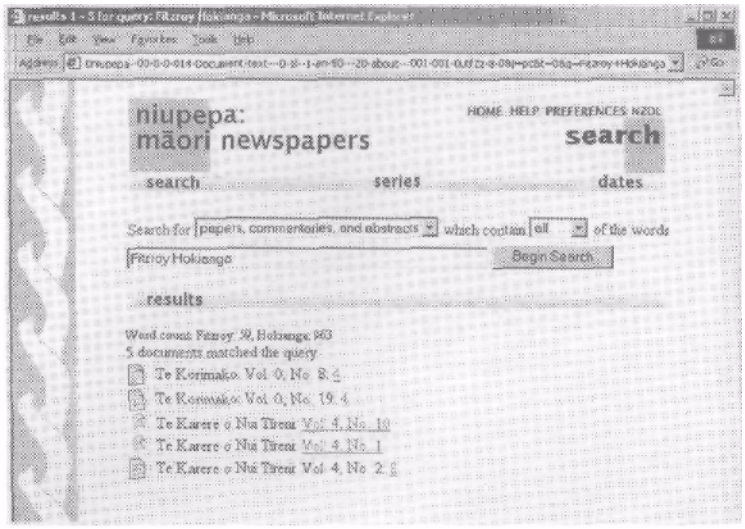

Figure 3: The result of searching the entire collection for pages containing both of the words Fitzroy' and 'Hokianga'.

show the collection as accessed over the Internet using a standard web browser (Internet Explorer).

Search

The *Rapu,* or *Search* facility (see Figure 3) provides the standard Greenstone full-text search facility, utilising the electronic text extracted from the images as described above. The Niupepa collection is divided into three sub-collections: the newspapers themselves, the bibliographic commentaries, and the English-language abstracts. A search can be confined to a single sub-collection, or it can be extended to cover all three simultaneously. Basic search options are a choice between some or all of the selected words but access to a more comprehensive set of search options, including word stemming and case sensitivity, is available through the 'Preferences' screen.

Figure 3 shows the result of a search for the two words Fitzroy' and 'Hokianga'. Five documents have been found which contain both of these words on the same page; three of these 'hits' are in the newspaper collection itself, and two occurred in the English-language abstracts. Hyperlinks in the search results provide direct

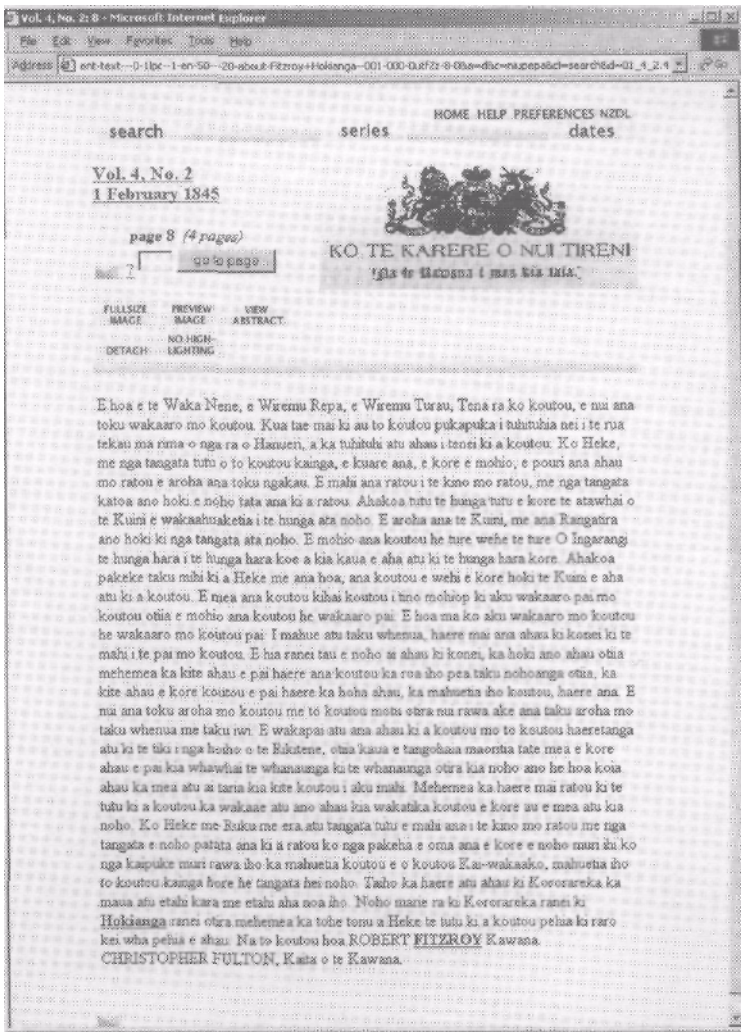

Figure 4: A page of extracted electronic text displayed by following one of the search result hyperlinks (Figure 3).

access to these pages. Selecting the last of these hyperlinks *(Te Karere o Nui Tireni,* Vol. 4, No. 2:8) leads to the display shown in Figure 4, the electronic text extracted from page 8 of volume 4, number 2, of *Te Karere o Nui Tireni.* In this text, the occurrences of the search terms are highlighted by underlining. From this page, a number of navigation options are provided. The other pages of the issue can be accessed, either by the arrow buttons at the top or

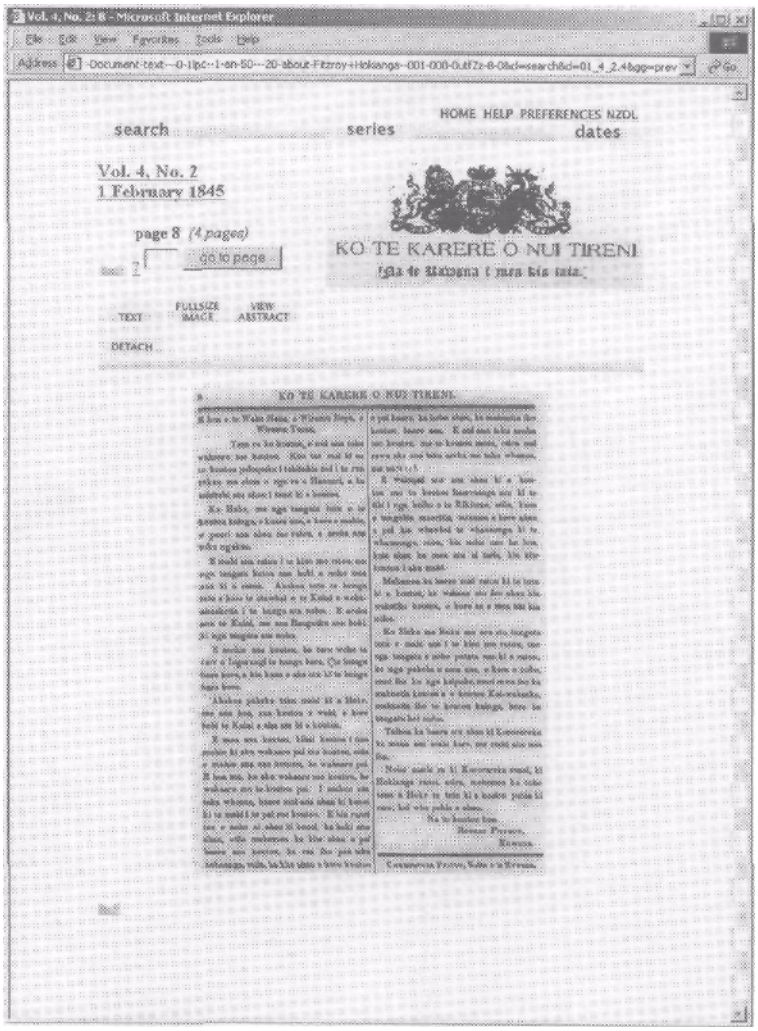

Figure 5: The preview image page corresponding to the electronic text of Figure 4.

bottom of the text (these would lead to the preceding page, page 7), by using the 'go to page' facility, or by selecting the issue information at the top left that leads directly to the front page of the issue. Other options include the ability to view the digital facsimile page image, either in full size or preview form, or the English-language abstract for this issue. Figure 5 shows the preview image page, and Figure 6, the abstract. It should be noted that

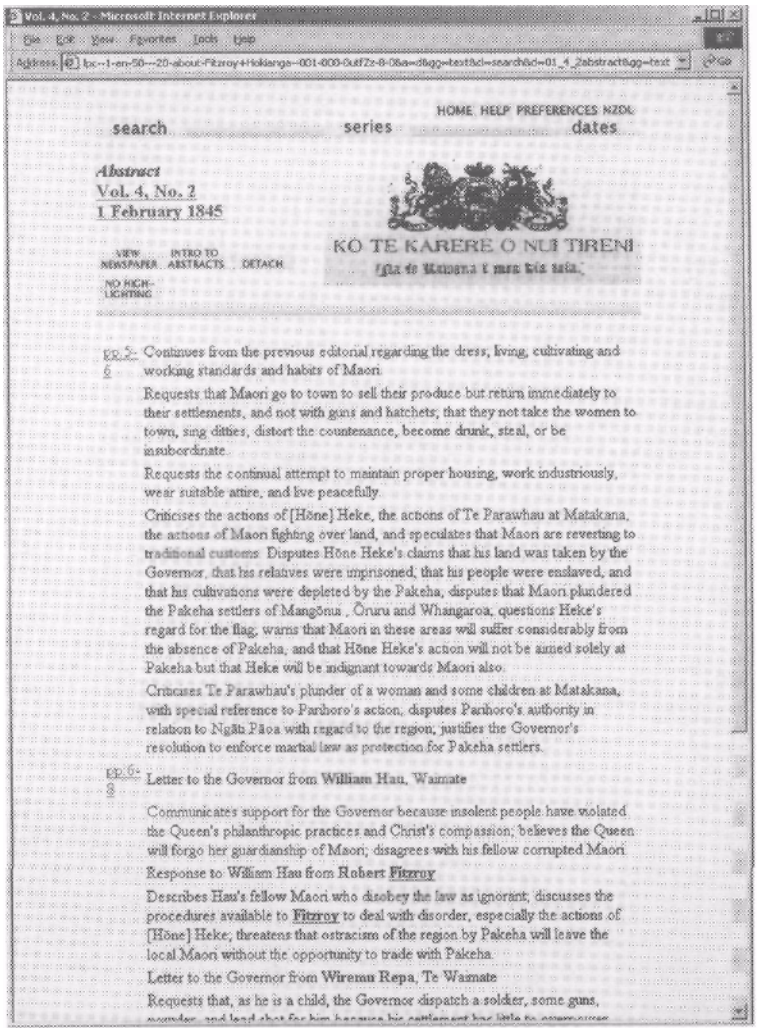

Figure 6: The abstracts corresponding to the page shown in Figures 4 and 5.

search terms are not currently highlighted on the page images, only in the electronic text.

Browse by series

The *Whakarārangi Taitara* or *Series* facility provides the user with the ability to browse through an index of all titles and, at a lower

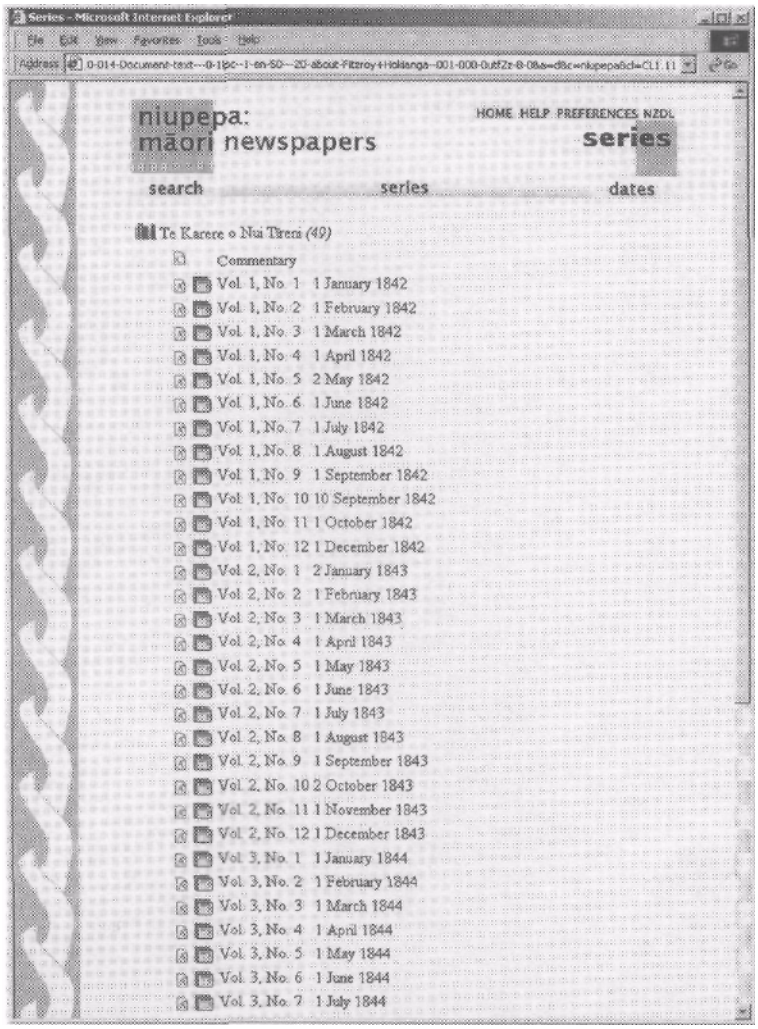

Figure 7: Browsing the collection by series, with the lower level index for *Te Karere o Nui Tireni* shown.

level, with a chronological index of the individual issues within a title. Figure 7 shows this lower-level index for *Te Karere o Nui Tireni;* the icons alongside each issue provide access to the English abstracts and the issue respectively. The first item in the list also provides access to the bibliographic commentary for this title. Selecting an issue from this index leads directly to displays of the form of Figures 4 and 5.

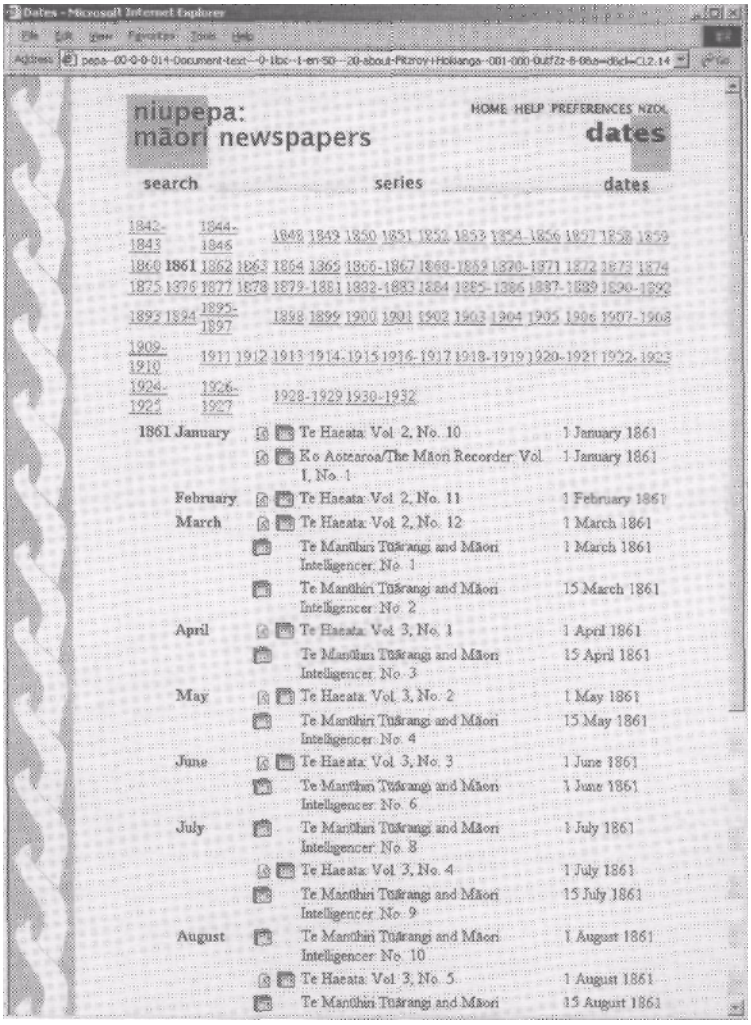

Figure 8: Browsing by date, showing three different titles appearing during the year 1861.

Browse by date

The *Wātaka* or *Date* facility allows the user to browse through the entire collection chronologically. Given the short lifespan and sporadic publication record of some of the titles, this is a very useful index when seeking reports on historical events. Figure 8 shows an example display for the year 1861. It can be seen that at least

three different titles appeared during this period: *Te Haeata, Ko Aotearoa,* and *Te Manuhiri Tuarangi.* Again, icons in this index provide access directly to the individual issues, and to the English abstracts where they are available.

Whatever method of access is used, once a particular page has been found it can be displayed in one of three formats: a text display of the electronic text extracted from the original images, or a facsimile image either as a reduced size preview, or as a full-size image which may be larger than the available display screen. The text version is the default initial display following a search query; search terms can be highlighted within the displayed text, and are thus relatively easy to find. However, most users prefer to switch to the facsimile image display to capture the original form of the item. Access to the abstracts, where available, is also provided from any of the page displays.

Future interface enhancements

The Niupepa collection in digital library form provides enormous advantages for readers, the most significant being the greater availability and the provision of full-text search. However, the very presence of this material in electronic form, opens up a range of other possibilities for adding value to the collection. Already the bibliographic commentaries and the English-language abstracts have been incorporated into the digital collection. Browse by date and browse by series have also been added to the basic search facility. However, a date restricted search has yet to be implemented, although technically this will not be a difficult task - it is more a question of how best to provide the interface for this feature. Tests are also being carried out with the automatic identification and indexing of specific *types* of information within the collection. For example, at the word level, work has been carried out to identify place-names and names of people. An experimental interface based on an index of place-names has already been built.[16] This allows a search to be carried out from a place-name or a map reference, which returns all pages that contain references to places within a limited distance from that location. A map-based display can show all place-names referred to on a page or a set of pages,

and can allow the user to narrow a search to a particular region. On a larger scale, it has been suggested that it may be possible to identify units or particular kinds of texts of interest to students of Maori language and culture, such as *waiata* (songs), *karakia* (incantations) and *whakapapa* (genealogy). Other variations on these ideas, which will add value to the existing collection, are also being considered.

This chapter has described the development of a digital library collection of the Maori-language newspapers. Although the undertaking was significantly influenced by the availability of the Greenstone software and by the pre-existence of the Niupepa collection on microfiche, it was strongly motivated by a desire to secure and help preserve the collection, to make it more widely available and accessible, and to significantly improve its utility by the provision of content-based searching and indexing.

Bibliography

Brown, C. W. & Shepherd, B, J., 1995. *Graphic File Formats.* Greenwich, CT: Manning Press. Cartwright, J., Chantiny, M, Hori, J., & Peacock, K., 2000. The Digital Landscape: The Hawaiian Newspapers and War Records and Trust Territory Image Repository of the University of Hawaii, *First Monday,* 5(6); http://firstmonday.org/ĩssues/issue5_6/cartwright/indtⁱx.html Dallimore, Gail, 1990. He Arahi, he Tohu o nga Pepa a te Maori: A Bibliography of Maori Newspapers, 1840-1900. Unpublished research report, Alexander Turnbull Library, Wellington. Hrafnkelsson, Orn, 2000. The Icelandic Experiment: Digitising Newspapers and Magazines from the 18th and 19th Centuries. Unpublished paper: Gutenberg 2000 8th Annual Conference of SHARP, 3-8 July, Mainz, Germany. Jones, *S.,* Cunningham, S. J. & McNabb, R., 1.998. An Analysis of Usage of a Digital Library, European Conference on Digital Libraries '98, Heraklion, Crete. Berlin: Springer-Verlag (Lecture Notes in Computer Science Series, No. 1513),pp.261-277. Kenney, A. R. and Rieger, O. Y., 2000. *Moving Theory into Practice: Digital imaging for libraries and archive.* Mountain View, CA: Research Libraries Group. Lesk, M., 1997. *Practical Digital Libraries: Books, Bytes and Bucks.* San Francisco: Morgan Kaufmann. Long, Brendan, 2000. Niupepa: Interface Design. Unpublished Project Report, 0657.420. Department of Computer Science, University of Waikato.

Lossau, Norbert, 1998. The Center for Retrospective Digitisation at Gottingen University Library, *Vine,* 107:39-43. McNab, R. J., Smith, L. A., Witten, I. H., Henderson, C. L. & Cunningham, S. J., 1996. Toward the Digital Music Library: tune retrieval from acoustic input, Proc *Digital Libraries '96.* New York: ACM Press, pp.11-18 Nevill-Manning, C. G., Reed, T. & Witten, 1. H., 1998. Extracting text from Postscript, *Software-Practice and Experience,* 28(5):481-91. Teahan, W. J., Inglis, S., Cleary J. G. & Holmes, G., 1998. Correcting English text using PPM models, *Proc Data Compression Conference.* Los Alamitos: CA: IEEE Press, pp.289-98. Witten, I.H., McNab, R., Jones, S., Cunningham, S. J., Bainbridge, D. & Apperley, M., 1999. Managing complexity in a distributed digital library, IEEE *Computer,* 32(2): 74-79. Witten, I. H., Moffat, A. & T. C. Bell, 1999. *Managing Gigabytes: compressing and indexing documents and images.* Second edition. San Francisco: Morgan Kaufmann. Witten, T. H. and Bainbridge, D., forthcoming. *How to Build a Digital Library.* San Francisco: Morgan Kaufmann.

Notes

1 See Lesk (1997) or Witten and Bainbridge (forthcoming) for further information on digital libraries.
2 The New Zealand Digital Library is at http://www.nzdl.org, while the Greenstone software can be downloaded from http://www.greenstone.org
3 *Niupepa 1842-1933,* Microfiche set, Alexander Turnbull Library, Wellington, New Zealand, 1996.
4 As noted in Kenney & Rieger 2000:14-15.
5 For instance, the Hawaiian newspaper project. See Cartwright, Chantiny et al. 2000.
6 Dallimore 1990.
7 See the Introduction, p.xi regarding the English abstracts.
8 As suggested by Brown & Shepherd 1995:439-44.
9 See Teahan, Inglis et al 1998:289-98.
10 As used, for example, by Lossau 1998 and Hrafnkelsson 2000.
11 Described in Witten, McNab et al. 1999.
12 See Nevill-Manning, Reed & Witten 1998.
13 Described in Witten, Moffat & Bell 1999.
14 See McNab, Smith et al 1996.
15 See Jones, Cunningham & McNabb, 1998.
16 See Long 2000.