

Multilingual support; interface languages; Niupepa project



Course material prepared by

Greenstone Digital Library Project
University of Waikato, New Zealand

and

National Centre for Science Information,
Indian Institute of Science, Bangalore



Agenda

- Multilingual support
  - Multilingual content
- New language interfaces
  - Creating/updating language interfaces
- Niupepa project
  - The Niupepa Collection
  - Image Capture
  - Text Capture
  - User Interface

Multilingual support in Greenstone

Issues:

- Operating system multilingual support
- Browser multilingual support
- Charsets, encodings, fonts
- Unicode
- Html

Charsets, encoding and fonts

- Charset
  - a set of characters, in the way a human would understand them.
    E.g. A, B, C so on are characters of English (and other Latin languages)
- Encoding
  - a way of storing chars on a computer (as "bits")
    E.g. ASCII, Extended ASCII, Unicode
- Font
  - an image (or "glyph") for a particular character or character set

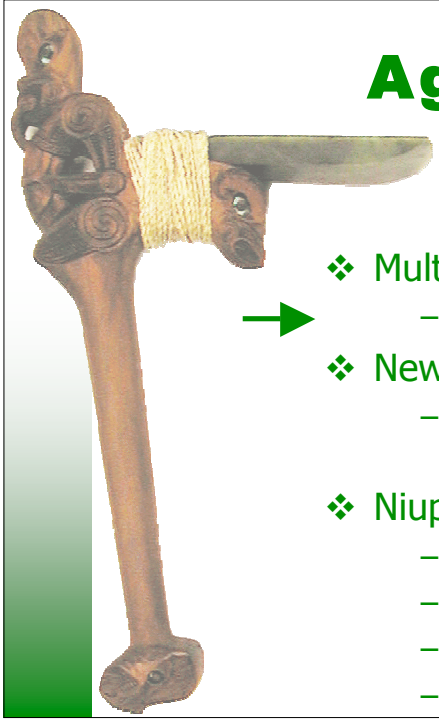
Encoding of charsets

- Widely used charsets
  - ISO 8859 series, windows series, ISO 10646 (UTF-8, UTF-16, UTF-32), ISCII, GB, Big-5, …
- Unicode
  - Emerging standard
  - Unique codes for 65,000++ characters
    Ex.. U+0041 is hexadecimal number of "A"
  - UTF-8 is a way of representing Unicode characters in a sequence of 8-bit bytes
  - UTF-16, UTF-32 are alternative representations
  - ASCII is a subset of UTF-8

Unicode in Greenstone

- Unicode is used internally
- All major charsets are automatically converted into UTF-8 through mappings
  - E.g. these charsets are mapped *from* and *to* Unicode:
    ISO 8859 series, windows series, gbk, big5, ShiftJis, uhc, ISCII, kio8_r, kio8_u, euc_jp, dos866

Mapping files are stored in gsdl/mappings folder



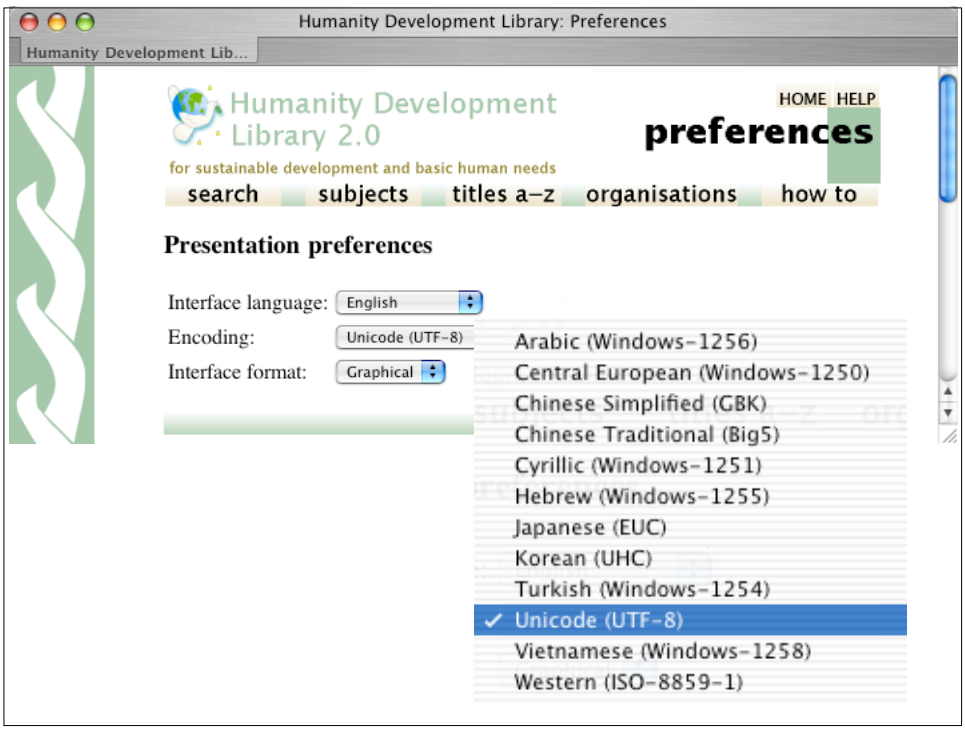
Agenda

- Multilingual support
  - Multilingual content
- New language interfaces
  - Creating/updating language interfaces
- Niupepa project
  - The Niupepa Collection
  - Image Capture
  - Text Capture
  - User Interface

Multilingual content: Unicode

- Operating system and browser must support your language
- Input filenames and extensions should be in ASCII (English)
- Select respective language interface from Preferences page
- Select UTF-8 encoding from preference page
- In browser, enable auto select (or select UTF-8 encoding)

  i.e. View → Encoding → Unicode (UTF-8)



Searching, browsing

- Must be able to type stuff into search box!
- Install Microsoft's IME (Input Method Editor) for your language
- Browser should support passing the query string
- Can search for words or phrases
- Advance searching is bit complicated
  - Boolean and proximity operators must be typed in Latin character set
  - Stemming: English only!
- Browsing: alphabetic sorting (e.g. AZLists) does not sort properly

Non-Unicode content

- etc/main.cfg file shows the encodings supported (lots of them)
- Make sure yours is enabled
- Greenstone maps characters into UTF-8 internally, builds the index, …
- … and converts back to native encoding upon display
- Select your language interface from Preferences page
- Select native encoding from Preferences page
- Enable Auto-select or select native encoding in the browser

E.g. Your collection is in windows-1256,
select Arabic (windows-1256) in preference page,
Arabic(windows) encoding in browser



Searching, browsing

- Note: collection is stored internally in Unicode
- Need to install IME (Unicode supported) of your language
- Browser should support passing the query string
- Otherwise the same as for Unicode content

Unmapped non-Unicode content

- Documents use encoding that Greenstone can't detect
- Greenstone will index as it is (no mapping)
- Need to have the right fonts in the OS's fonts folder
- Enable auto-select (or "User Defined") option in browser
  i.e. View → Encoding → More → User Defined
- For searching, enter query string in basic encoding format

Limitations

- Alphabetical sorting?
- Stemming?
- Boolean and proximity search operators must be entered in ASCII form
- Supports html, Word files

Agenda



- Multilingual support
  - Multilingual content
- New language interfaces
  - Creating/updating language interfaces
- Niupepa project
  - The Niupepa Collection
  - Image Capture
  - Text Capture
  - User Interface

Different interface languages

- Create a macro file for your language
  - Make a copy of english.dm & rename
  - Translate each macro value
    - If it is Unicode encoding, select Unicode format for saving

      Ex... `_textimagehome_ {____ ____ }`

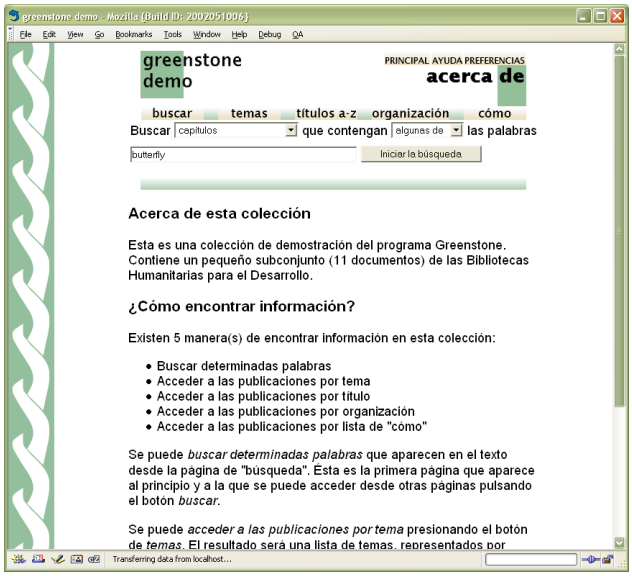
    - If it is ASCII encoding, select text document format for saving

      Ex.. `_textimagehome_ {Home page}`

Search page (English)



Search page (Spanish)



Macro file

```
package Globals
_header_  { The New Zealand Digital Library Project }
_content_ { Oops. If you are reading this then an error
            has occurred in the runtime system. }
_footer_  { Powered by <a href="www.greenstone.org">Greenstone</a>. }

package query

_content_ { _If_(_cgiargqb_ eq "large",
           _largequerybox_,_normalquerybox_) ... }

# ... the macro descriptions for _largequerybox_, _normalquerybox_,
# and other nested macros are omitted
_header_ [l=en] {Begin search }
_header_ [l=fr] {Démarrer la recherche
_header_ [l=es] {Iniciar la búusqueda
# ... and so on

# Images containing language-dependent
## "HELP" ## top_nav_button ## chelp #
_httpiconchelp_ [l=en,v=1] {_httpimg_
_httpiconchelp_ [l=en,v=0] {HELP}

# ... and so on
```

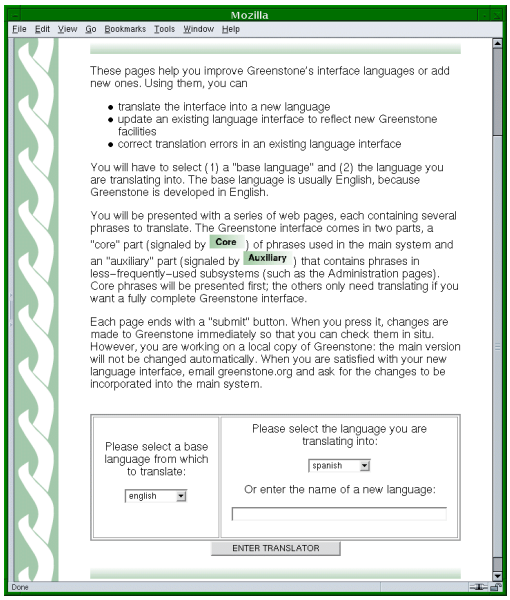
- Language specific text fragments actually gathered up in one file:
  english.dm, french.dm etc.
- Language specific fragments further divided:
  english.dm, english2.dm
  french.dm, french2.dm

Maintenance issues

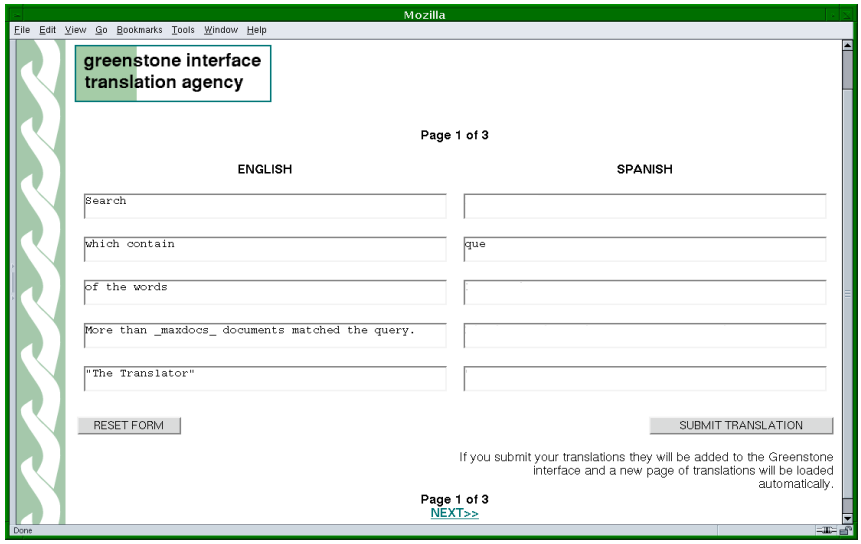


- Over 600 linguistic fragments per language in user interface
- No one person knows all languages
- No overlap between translators and developers
- Software developers and translators geographically dispersed
- Translations must keep pace with software developments

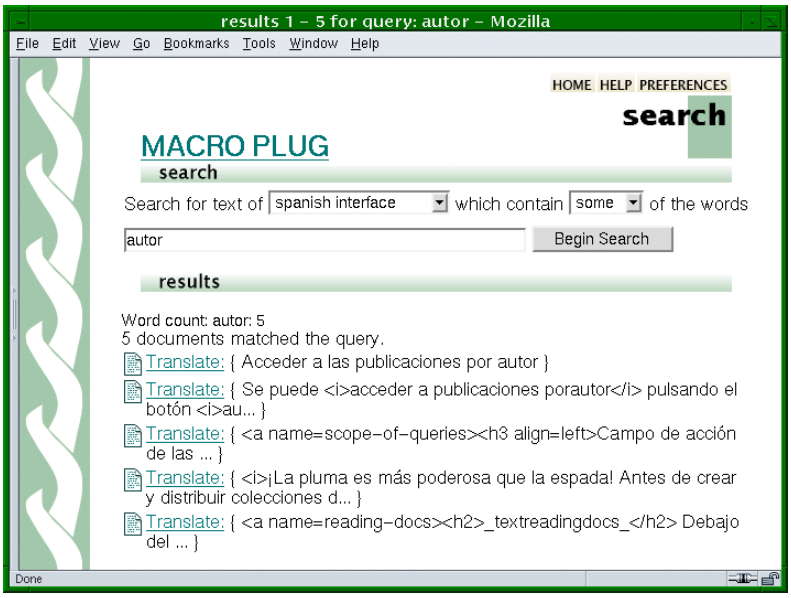
Translator's interface



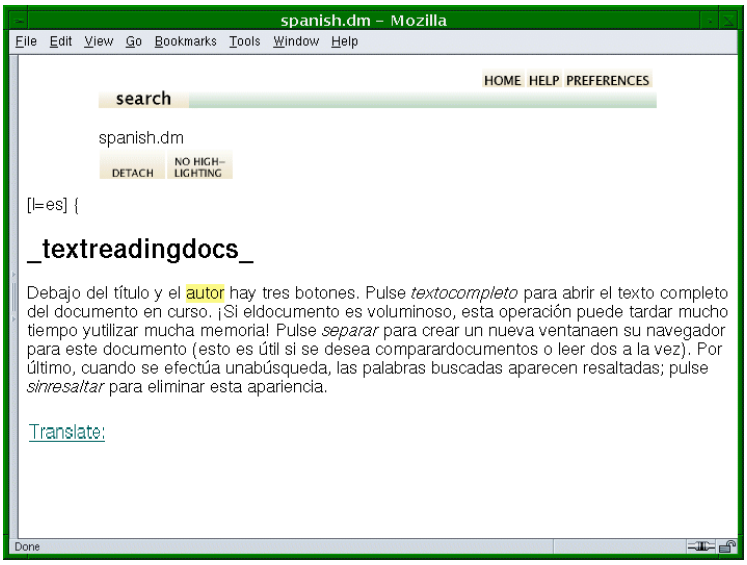
New phrases, and superseded ones



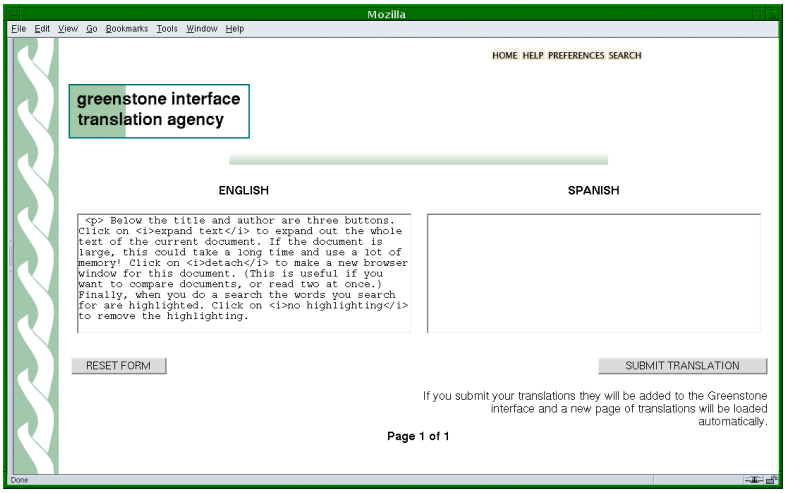
Search for an error ...



... find offending fragment ...



... and fix it up



Generating Icons

- Use image editor
- Translate the content of the tab or icon
- Retain the same file names
- Place all tabs and icons into single folder and copy the folder to gsdl/images

English



ttitlgr.gif



ttitlof.gif



ttitlon.gif

Hindi



ttitlgr.gif



ttitlof.gif



ttitlon.gif

Documentation and Help

- *Developer' Guide* (developer.pdf)
- Greenstone archives collection
- http://www.unicode.org



Agenda

- Multilingual support
  - Multilingual content
- New language interfaces
  - Creating/updating language interfaces
- Niupepa project
  - The Niupepa Collection
  - Image Capture
  - Text Capture
  - User Interface

NZ Maori Language History
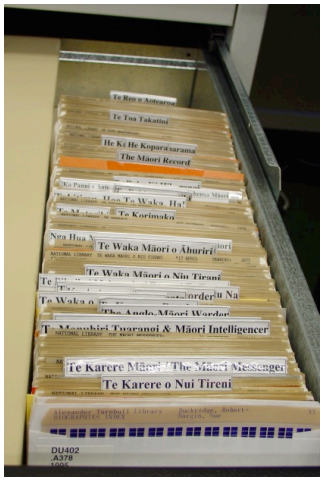
- Medium is Maori 1840s
- Maori population ~30% 1890s
- Education Acts prohibit Maori 1867 & 1871
- Urban shift after WWII
- Language Loss imminent 1979
- Revival initiatives begin 1980s

Maori Language Publishing

- Traditionally an oral language
- Missionaries printed scripture 1830s
- 50% Maori were literate 1850-1900
- Many periodicals, publications written in Maori
- Language did not recover with population in 1900s
- Publication in Maori ceased
- A desperate need arises 1980s

What is the Niupepa Collection?

- Libraries throughout NZ hold newspapers from the time when Maori language publishing was flourishing
- In 1988/1989 the Alexander Turnbull Library undertook to source the best copies of these periodicals and preserve them on microfilm
- Called the "Niupepa Collection," this is available on microfiche at most libraries.

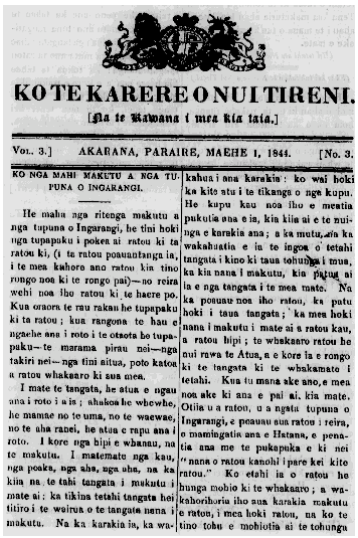




What's in the Niupepa Collection?

- Approximately 18,000 pages in 40 periodicals
- Mostly written in Maori or for Maori audience
- Written from 3 distinct perspectives: Missionaries, Government and from Maori themselves
- Written in a very clear style
- Invaluable source of historic information

Te Karere o Nui Tireni



KO TE KARERE O NUI TIRENI.

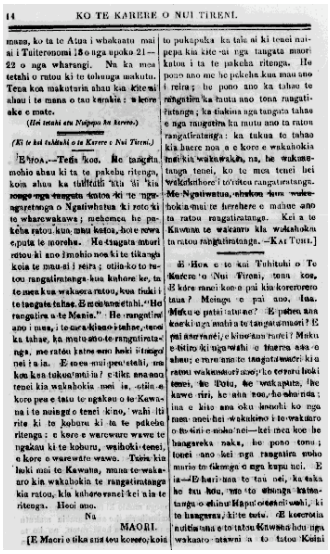
[Na te Wawana i mea kia tata.]

Vol. 3.] AKARANA, PARAIRE, MAEHE 1, 1844. [No. 3.

KO NGA MAHI MAKUTU A NGA TUPUNA O INGARANGI.

He maha nga ritenga makutu a nga tupuna o Ingarangi, he tini hoki nga tupapaku i pokea ai ratou ki te ratou ki, (i te ratou posauntanga ia, i te mea kahore ano ratou kia tino rongo noa ki te rongo pai)—no reira wehi noa iho ratou ki te hacre po. Kua onaora te rau rakau he tupapaku ki ta ratou; kua rangona te hau e ngaehe ana i roto i te otaota he tupapaku—te marama pirau nei—nga takiri nei—nga tini aitua, poto katoa a ratou whakaaro ki sua mea.

I mate te tangata, he atua e ngau ana i roto i a is; ahakoa he whewhe, he mamae no te uma, no te waewae, no te aha ranei, he atua e ngau ana i roto. I kore nga hipi e whanau, na te makutu. I matemate nga kau, nga poaka, nga ahe, nga aha, na ka kiia na te tahi tangata i makutu i mate ai: ka tikina tetahi tangata hei titiro i te waiura o te tangata nena i makutu. Na ka kurakia ia, ka wakahua i ana karakia: ko wai hoki ka kite atu i te tikanga o nga kupu. He kupu kau noa iho e meatia pukutia ana e ia, kia kiia ai e te nuinga e karakia ana; a ka mutu, ara ka wakahuatia e ia te ingoa o tetahi tangata i kino ki taua tohunga i mua, ka kia nana i makutu, kia patua ai e nga tangata i te mea mate. Na ka pouaua nos iho ratou, ka putu hoki i taua tangata; ka mea hoki nana i makutu i mate ai a ratou kau, a ratou hipi; te whakaaro ratou he aui rawa te Atua, e e kore ia e rongo ki te tangata ki te whakamate i tetahi. Kua tu mana ake ano, e mea non ake ki ana e pai ai. kia mate. Otira u a ratou, u a ngatu tupuna o Ingarangi, e peauau aua ratou i reira, o mamingatia ana e Hatana, e penatia ana me te pukapuka e ki nei "nana o ratou kanohi i pare kei kite ratou." Ko etahi ia o ratou he hunga mohio ki te whakaaro; a wakahoriboria iho aua karakia makutu e ratou, i mea hoki ratou, na ko te tino tohu e mohiotia ai te tohunga



14 KO TE KARERE O NUI TIRENI.

mana, ko ta te Atua i whakaatu mai ai i Tuiteronomi 18 o nga upoko 21—22 o nga wharangi. Na ka mea tetahi o ratou ki te tohunga makutu, Tena koa makuturia ahau kia kite ai ahau i te mana o tau karakia: a kore ake e mate.

(Hei tetahi atu Nupepa ka korero.)

(Ki te kai tuhituhi o te Karere o Nui Tireni.)

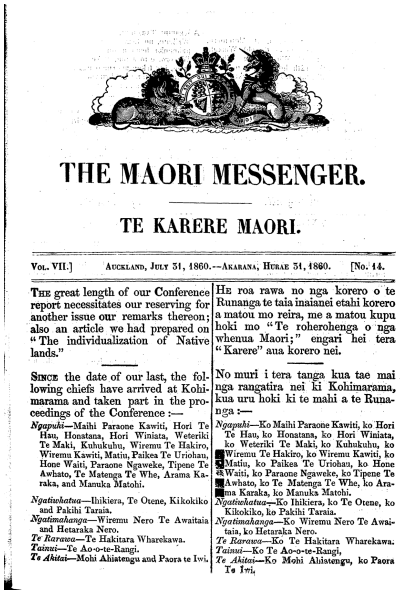
E Hoa,—Tena koe. He tangata mohio ahau ki ta te pakehu ritenga, koia ahau ka tiaki[illegible] ki te nga-ngaetanga o Ngatiwhatua ki roto ki te whare whakawa; nohoneke he pakeha ratou, kua ano kotea, hoi e rewa e puta te morehu. He tangata ataori ratou ki ano i mohio noa ki te tikanga koia te mau ai i reira; otira ko te ratou rangatiratanga kua kahore ke, na te mea kua wakaea ratou, kua tuki i te tangata tahae. E ara ano tetahi, "Ho rangatira a te Mania." He rangatira ano i nuia, i te aro kino o te hoa, tera ka tahae, ka mutu ano te rangatiratanga, me ratou katoa ano kua i tangi nei i a ia. E mea mai pea tetahi, tena koa kau tokou anti ia? e tika ana ano tenei kia wahakokia mai ia, otira e kore pea e taia te ngakau o te Kawana i te noinga o tenei kino, wahi iti rite ki te koburu ki ia te pakeha ritenga: e kore e wareware wawe te ngakau ki te koburu, waiboki tenei, e kore e warewaho wawo. Taria kia hoki mai te Kawana, mana te wakaaro kia wahakokia te rangatiratanga kia ratou, kia kahore ranei kei aia te ritenga. Hoi ano.

Na MAORI.

[E Maori e tika ana tou korero; koia to pukapuka ka tuia ai ki tenei nupepa kia kite ai nga tangata maori katoa i ta te pakeha ritenga. He pono ano me he pakeha kua mau ano i reira; he pono ano ka tahae te rangatira ka tauia ano tona rangatiratanga; ka tiakina nga tangata tahae e nga rangatira ka mutu ano ta ratou rangatiratanga: ka tukua te tahae kia haere noa a e kore e wahakokia mai kia wakawakia, na, he wakamatangia tenei, ko te mea tenei hei wekaehohore i to'ou rangatiratanga. Me Ngatiwatua, ahakoa kua wakahokia mai te hurehere o mahue ano ta ratou rangatiratanga. Kei a te Kawana te wananga kia wahakokia ta ratou rangatiratanga. —Kai Tohu.]

E Hoa e te kai Tuhituhi o Te Karere o Nui Tireni, tena koe. E kore ranei koe e pai kia korerorero tatou? Mauria e pai ana. Ina. Mutu o patai atu ano? E pehea ana kaerikinga mahi a te tangatamuri? E pai ake tenei, e kino ana tenei? Mutu e titiro ki nga wahi e tuerea ana e ahau; e mea ana te tangata maori kia ratou wakamatau? aue; ko te ratou loki tenei, he Roti, he wekapuke, he kawe riri, ke ana noa, he aha noa; ina e kiia ana oku loaohi ko nga mea anei hei wakakino i te wakaaro o te tini e mohu nei—kei mea koe he hangareka nake, he pono tenei tenei ano kei nga rangatira mohio marie te tikanga e nga kupu nei. E ia—E hari ana to tau nei, ka taka he tau hou, mo te oranga katoatanga o ehinu Hapori o te ao katoa, ki te hangarau, ki te tuiu. E kororotia nuitia ana e to tatou Kawana hau nga wakaaro otawai a to tatou Kaihi

A bilingual newspaper



THE MAORI MESSENGER.

TE KARERE MAORI.

Vol. VII.] Auckland, July 31, 1860.—Akarana, Hurae 31, 1860. [No. 14.

The great length of our Conference report necessitates our reserving for another issue our remarks thereon; also an article we had prepared on "The individualization of Native lands."

Since the date of our last, the following chiefs have arrived at Kohimarama and taken part in the proceedings of the Conference:—

Ngapuhi—Maihi Paraone Kawiti, Hori Te Hau, Honatana, Hori Winiata, Weteriki Te Maki, Kuhukuhu, Wiremu Te Hakiro, Wiremu Kawiti, Matiu, Paikea Te Uriohau, Hone Waiti, Paraone Ngaweke, Tipene Te Awhato, Te Matenga Te Whe, Arama Karaka, and Manuka Matohi.

Ngatiwhatua—Ihikiera, Te Otene, Kikokiko and Pakihi Taraia.

Ngatimahanga—Wiremu Nero Te Awaitaia and Hetaraka Nero.

Te Rarawa—Te Hakitara Wharekawa.

Tainui—Te Ao-o-te-Rangi.

Te Akitai—Mohi Ahistengu and Paora te Iwi.

He roa rawa no nga korero o te Runanga te taia inaianei etahi korero a matou mo reira, me a matou kupu hoki mo "Te roherohenga o nga whenua Maori;" engari hei tera "Karere" aua korero nei.

No muri i tera tanga kua tae mai nga rangatira nei ki Kohimarama, kua uru hoki ki te mahi a te Runanga:—

Ngapuhi—Ko Maihi Paraone Kawiti, ko Hori Te Hau, ko Honatana, ko Hori Winiata, ko Weteriki Te Maki, ko Kuhukuhu, ko Wiremu Te Hakiro, ko Wiremu Kawiti, ko Matiu, ko Paikea Te Uriohau, ko Hone Waiti, ko Paraone Ngaweke, ko Tipene Te Awhato, ko Te Matenga Te Whe, ko Arama Karaka, ko Manuka Matohi.

Ngatiwhatua—Ko Ihikiera, ko Te Otene, ko Kikokiko, ko Pakihi Taraia.

Ngatimahanga—Ko Wiremu Nero Te Awaitaia, ko Hetaraka Nero.

Te Rarawa—Ko Te Hakitara Wharekawa.

Tainui—Ko Te Ao-o-te-Rangi,

Te Akitai—Ko Mohi Ahistengu, ko Paora Te Iwi,

Ngatiwaru—Mohi Te Rongomawhata and Te Reweti Puata.

Te Arawa—Paraksia Tarapoa, Peraera, Te Ngahue, Hori Te Kouku, Pereuara, Matene, Hunioua, Te Manihera, Kahumu, Rawharitua, und Te Kuhe.

Ngatiwaru—Ko Mohi Te Rongomawhata, ko Te Reweti Puata.

Te Arawa—Ko Paraksia Tararoa, ko Peraera, ko Te Ngahue, ko Hori Te Kouku, ko Pereuara, ko Moteene, ko Hiuia, ko Te Manihera, ko Kahumu, ko Rawharitua, ko Te Kuhe.

DIED

On the 25rd of last month, at Meremere, Kerehi Tarapuni, one of the old Waikato Chiefs. His friend Wiremu Te Wheoro writes—"At the 12th hour of the morning, when near death, he arose and looked up at the clouds drifting over Waikato, and then upon the waters of the river. After gazing a long time, he bade farewell in these words, 'Farewell ye shadows of the Waikato and Taupiri! Farewell ye tribes! Hold fast to kindness!' Here his farewell ceased and he fell asleep. * * * He was beloved by all the people of Waikato. They were never injured by him. All he did was for good and the suppression of evil."

KUA MOE

A Kerehi Tarapuni, he rangatira kaumatua o Waikato, no te 25 o nga ra o Hune. Ko te pukapuka tenei a tona hoa a Wiremu Te Wheoro, e mea ana, "No tona tatanga ki te mate, i te tekau ma rua o nga haora o te ata, ka tahi ka maranga ki runga, ka tahi ka titiro ki te ao rere mai i roto i Waikato; ka tahi ka huri te titiro ki te wai o Waikato; ka roa te tirohanga, ka puta iana mihi—'Hei konei ra e te atarangi i Waikato! Hei konei ra i Waikato, hei konei ra i Taupiri! Hei konei ra e nga iwi! Kia mau ki te atawhai!' Ka mutu tana poroporoaki, ka tahi ka moe. * * * He tangata e manaakitia ana e nga iwi katoa o Waikato. Kahore hoki he kino mana e kitea ki runga ki nga iwi, ko te pai anake, ko te pehi i te kino."

Intelligence.

CANTERBURY.

On the 26th of June last, Hobepa, Nopera, and all the other chiefs of the Ngaitos tribe, wrote a letter to the Ngaituhuriri, of Canterbury, to recommend them to abide by the "four laws of our mother, Queen Victoria, who is the nourisher of both the whites and the Maories." On receipt of this letter, the natives of Canterbury met at Rapaki to consider the contents thereof; and they all came to the resolution to be like the Ngaitos, and "cleave to the laws of the Queen by which they have become men."

Afterwards they gave a token of their good feeling towards the whites "who are living on the land covered with darkness, that is Taranaki." They heard that many of them are in distress, their property having been destroyed by the spoiler—and that their friends the white people of Canterbury were collecting money to assist the aged, the women, and the children: and they thought they would do the same. Accordingly, at Kaiapoi they collected 10l. 14s., at Rapaki 4l. 14s. 6d., at Port Levy 4l. 7s.; and other villages are doing the same thing but have not yet reported the amount. But because some of the Europeans suspected them of an intention to go to Taranaki, to join those who are in arms, they were grieved and

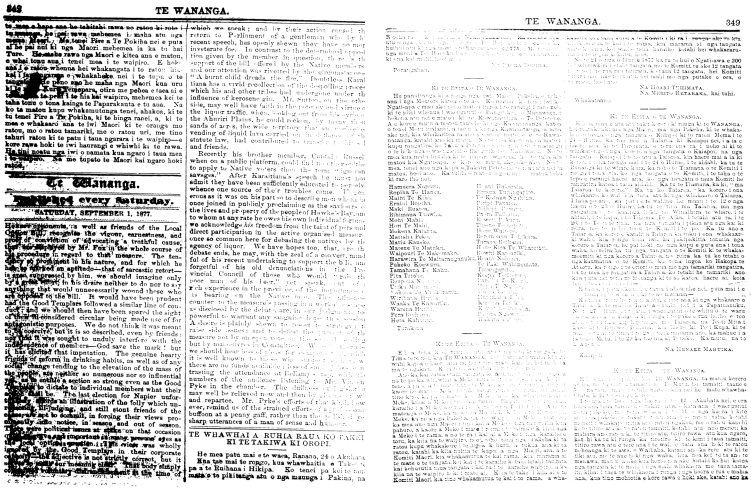
He Korero.

KATAPERE.

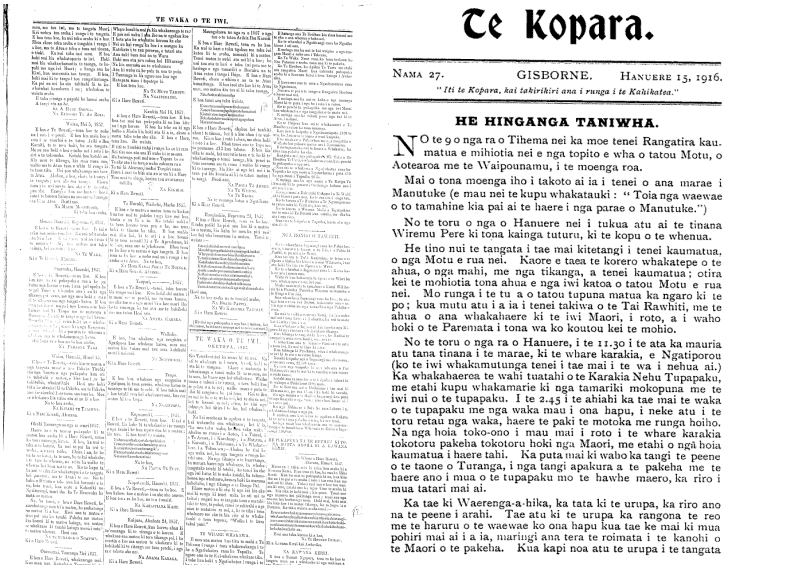
No te 26 o nga ra o Hune i tuhituhi atu ai nga rangatira o Ngaitos—na Hobepa Temuihengia, na Nopera te Ngiha, na ratou katoa hoki tetahi pukapuka ki a Ngaituhuriri o Katapere, e mea atu ana ki a ratou kia u tonu ki "nga tikanga maranga o to tatou Whaea, o Kuini Wikitoria, o te kai whangai o nga Pakeha, o nga Maori." No konei i huihui ai nga Maori o Katapere ki Rapaki kia hurihia e ratou nga whakaaro o roto o tenei pukapuka; a ka whakaae katoa ratou kia pera ano ratou me Ngaitoa. "kia piri ano ki nga ture o te Kuini i whakatupu tangata ai ratou."

Muri iho ka puta to ratou tohu aroha ki nga Pakeha "o te kainga e tauria nei e te pouritanga, a Taranaki." Ka rongo hoki ratou, ka mate nga Pakeha i te rawa kore, kua pau hoki nga taonga i te pahua: ki te whakaaro hoki o o ratou hoa Pakeha kia kawea ki Katapere etahi o nga korobeke, wahine, tamariki,—e kohikohi ana i te moni mo tera; ka mea ano nga Maori ma ratou hoki tetahi wahi o tera kohikohinga. Na, kua kohikohi a Kaiapoi 10l. 14s., a Rapaki 4l. 14s. 6d., a Poturiwi 4l. 7s. Kahore ano etahi o nga kainga i oti noa te kohikohi. No te mea hoki ka tapatoria ratou e etahi o nga Pakeha, ka mea e haere pea ki Taranaki ki

Image variation



Size and volume of text

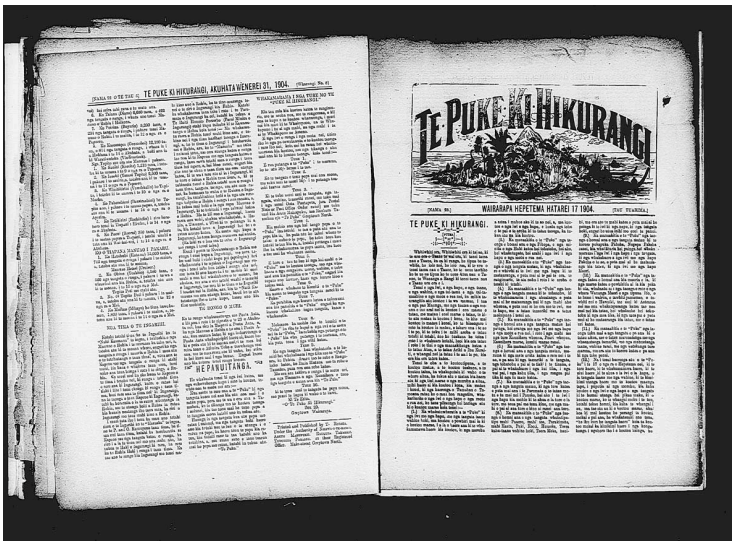


Capturing print material

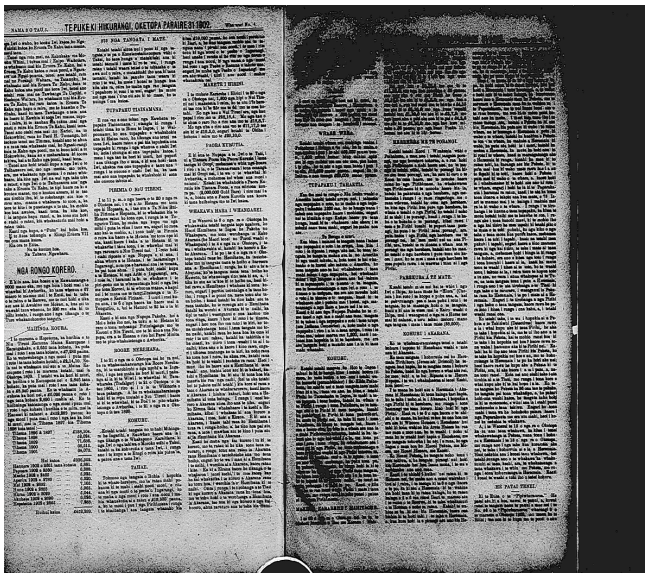
1. Capture a digital facsimile image
   - ❖ digital photograph of the original image comprising individual dots or pixels
2. Extract the text using OCR
   - ❖ need electronic text as a record of the actual characters if text is to searchable

- ❖ Images were available on 35mm film
- ❖ Good quality photographs, but original print material varied substantially in quality and information density

Examples on film



Examples on film



Examples on film



Examples on film



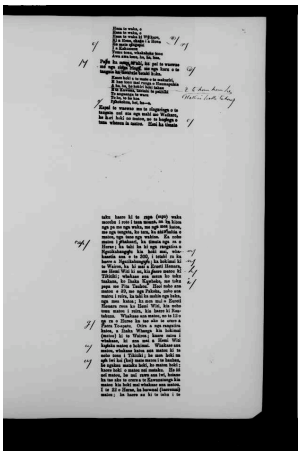
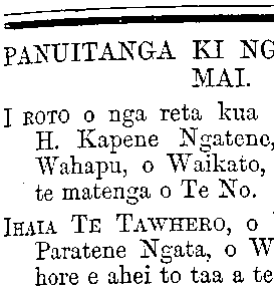


Image Capture

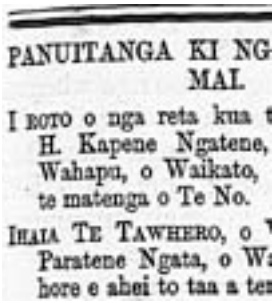
- Quality or fidelity of image depends on:
  - density of dots (dots per inch - dpi)
  - sensitivity of dot values (bits/pixel)
- Large variation in quality and size of original captured images
- Experiments showed that 300dpi best for OCR
- Images stored as "tagged image format" (.tif)
- Two images captured at the same time:
  - bitonal image for Internet delivery (220k)
  - grey-scale image for OCR (5-10Mbytes)
- 8 CDs for entire collection in bitonal form
- 90+ CDs for entire collection in grey-scale form

Bitonal and grey-scale images


PANUITANGA KI NG
MAI.

I ROTO o nga reta kua
H. Kapene Ngatene,
Wahapu, o Waikato,
te matenga o Te No.

IHAIA TE TAWHERO, o
Paratene Ngata, o W
hore e ahei to taa a te


PANUITANGA KI NG
MAI.

I ROTO o nga reta kua t
H. Kapene Ngatene,
Wahapu, o Waikato,
te matenga o Te No.

IHAIA TE TAWHERO, o
Paratene Ngata, o W
hore e ahei to taa a te

Cropping and Splitting

- Many images were double-page spreads
- For consistent granularity, 1-page-per-file, these images were split into two separate images prior to text extraction
- Extraneous image material (outside the text) was cropped reducing size
- Skewed images straightened

Text capture

- OCR software identifies shapes in the image to create a corresponding page of text
- Recognition accuracy depends on
  - image quality
  - visual quality of original
  - font
  - software sophistication
- OCR is time-consuming, depending on quality of images captured
- Manual checking is needed
- Electronic text has low storage requirements
  - say 4KByte for single spaced A4 page

OCR software

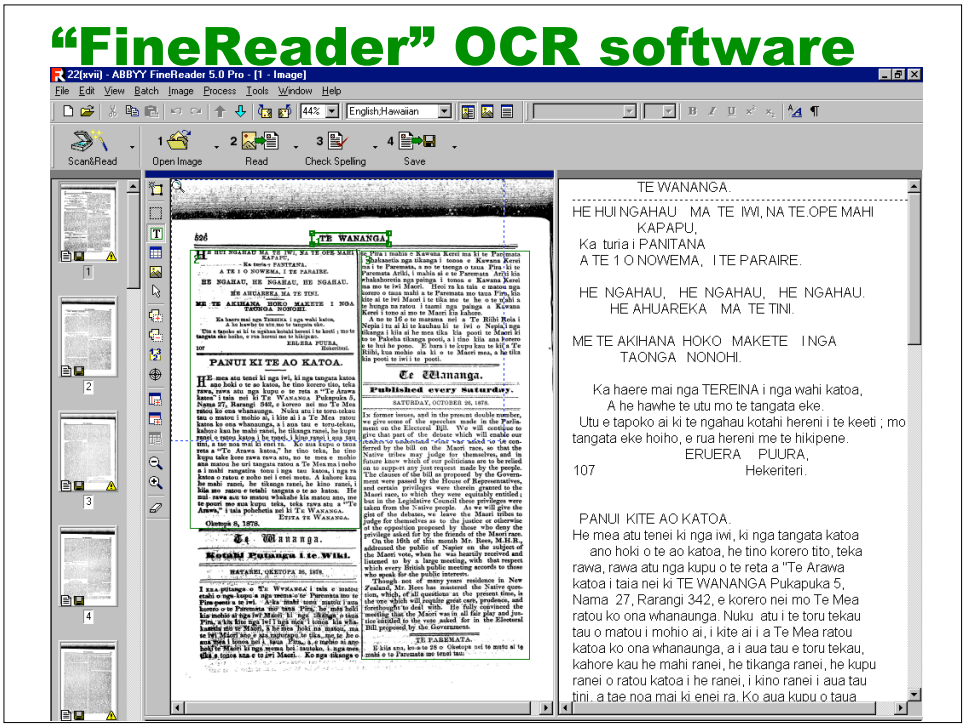
After testing many OCR programs we settled on ABBYY FineReader 6.0 Corporate because:

- Supports 175 languages, including Maori (though not fully)
- Does not "correct" non-English words to English
- Very accurate
- Cost effective (US$400)

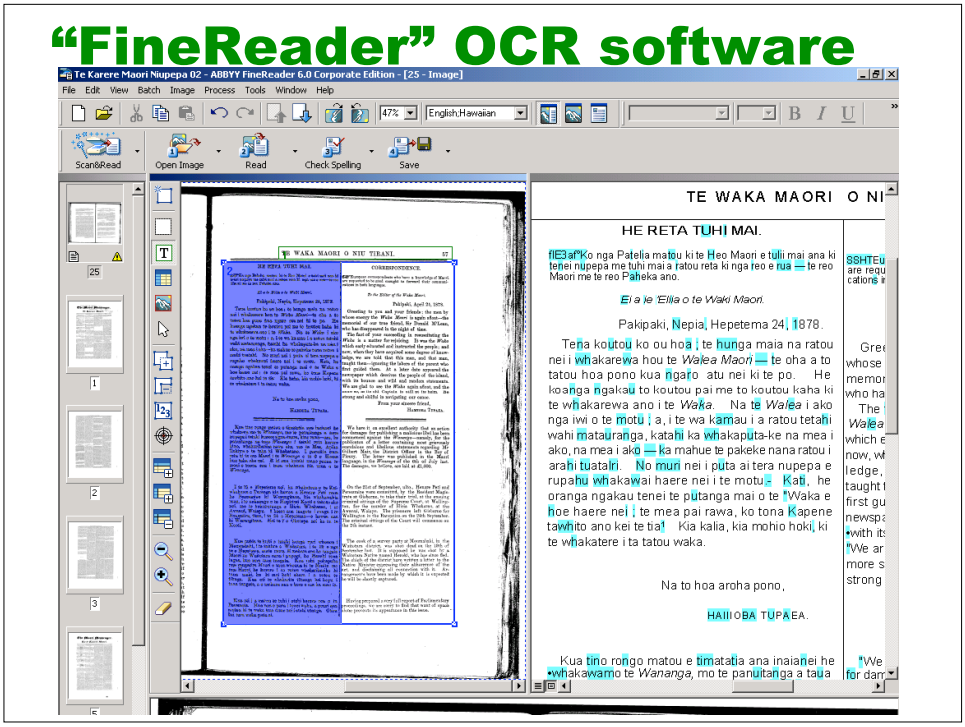
Also:

- Can train a "recognition pattern"
- Can input dictionaries and maintain them in the proofing process
- Can define the character set
- User friendly and easy to run
- New staff require minimum training

"FineReader" OCR software



"FineReader" OCR software



OCR Unit

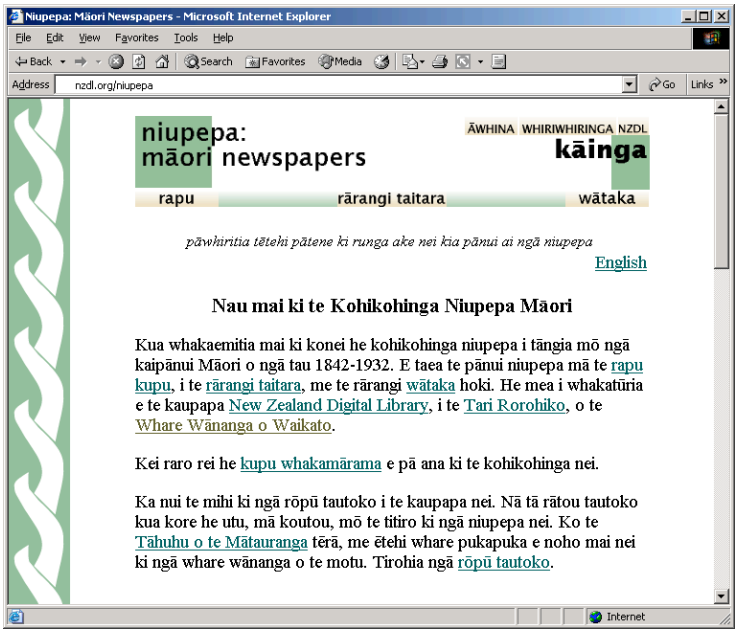
- Skills developed led to creation of Waikato OCR Unit
- Employ a supervisor and OCR technician
- Undertaken general OCR contracts specializing in Polynesian languages
- Large contract with Hawaiian Newspapers

User Interface

`nzdl.org/niupepa`

- Default language is Maori
  - collection is primarily in Maori
  - funded as a Maori language resource
  - makes a statement about the use of Maori
  - (can easily be switched)
- Three access methods
  - Browse by Title
  - Browse by Date
  - Search by Content

Niupepa home page



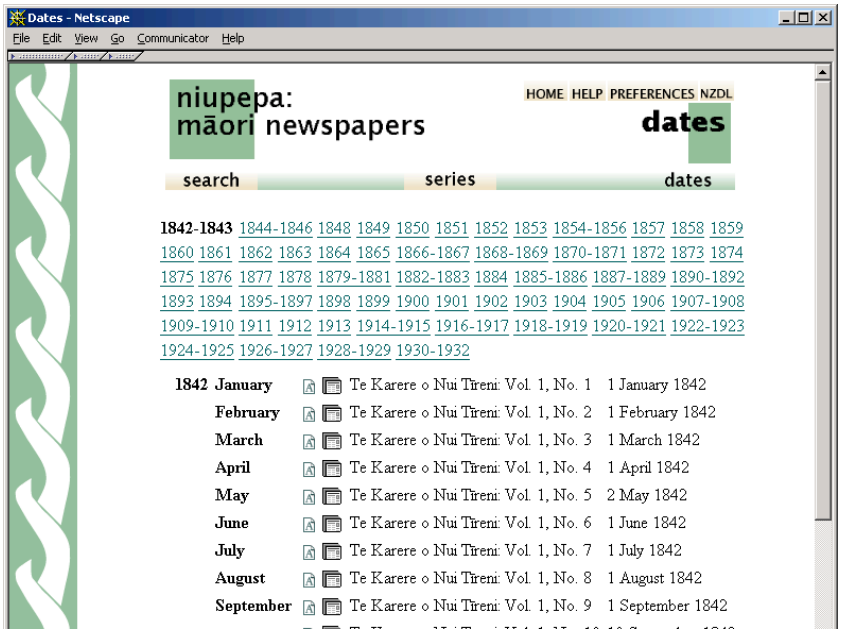
English home page



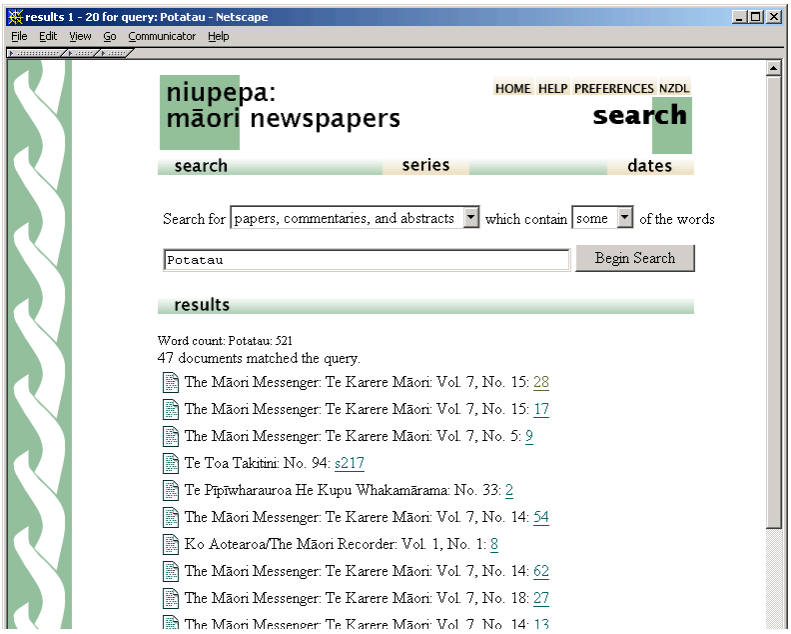
Browse by Series



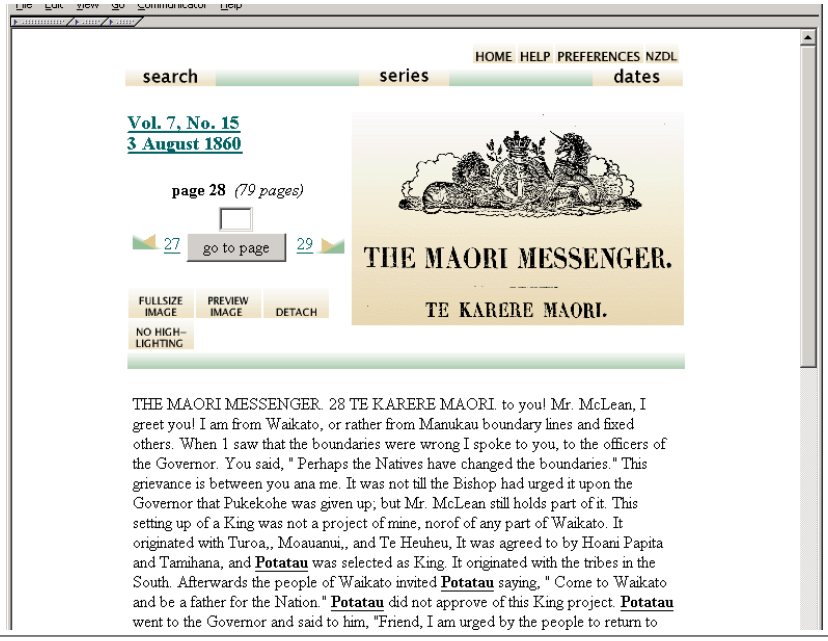
Browse by date



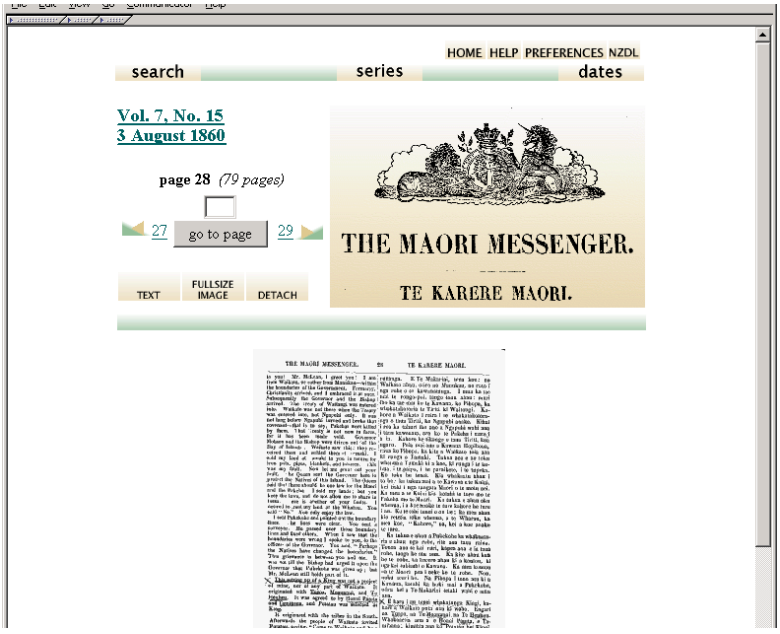
Full-text search



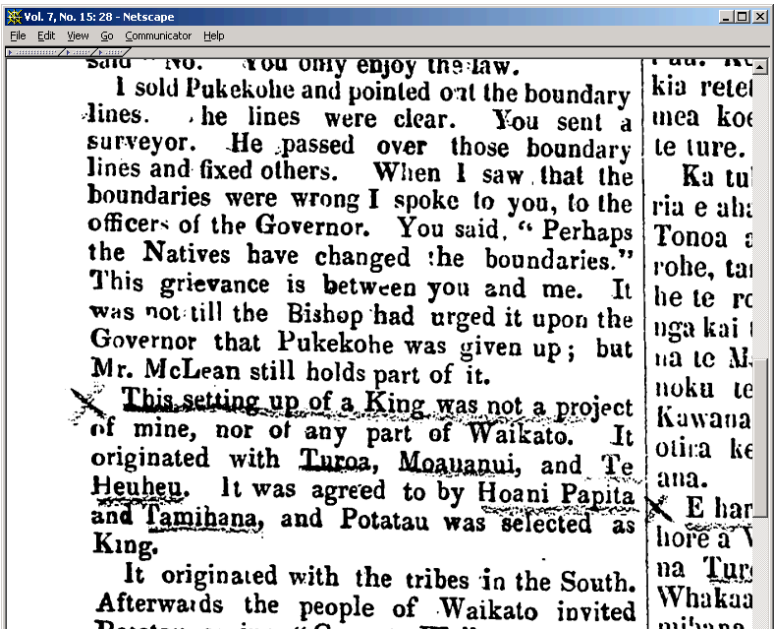
Text page



Small image page



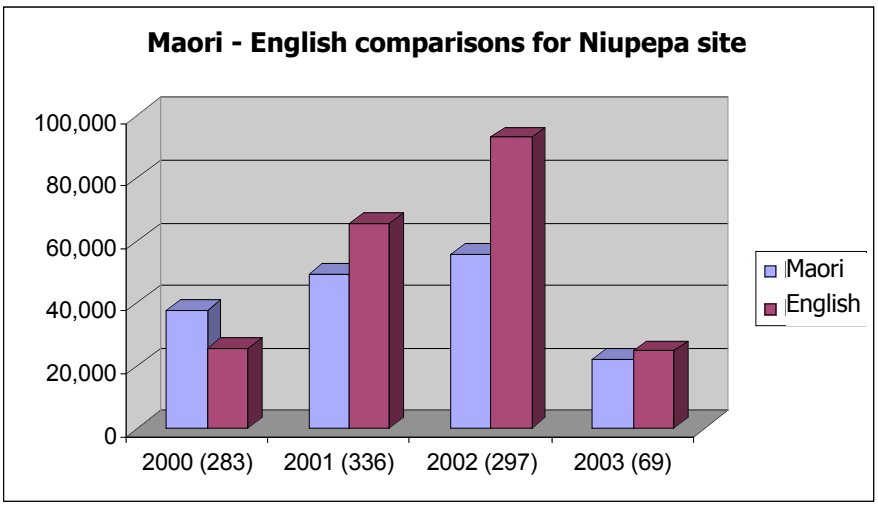
Large image page



User interface issues

- Full text search **not** available for <2% of total series (poor quality originals)
- Unicode displays non-ASCII characters correctly
- Had to generate new words for the user interface

Maori - English usage



- Top 50 search terms — 48 are in Maori
- Next 50 — 41 are in Maori

Conclusion

- Unique collection of historical indigenous language newspapers
- Greatly enhances access, preservation too
- Full-text search adds significant utility and value (though costly)
- Can overcome OCR difficulties with minority languages
- Experience and techniques apply to a wide range of legacy text collections
- Contributes significantly to the promotion and preservation of language and culture

Questions?

Comments?

Discussion?

Feedback form!