


## Scanning and OCR: Niupepa collection



**Course material prepared by**

**Greenstone Digital Library Project  
University of Waikato, New Zealand**

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

## 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



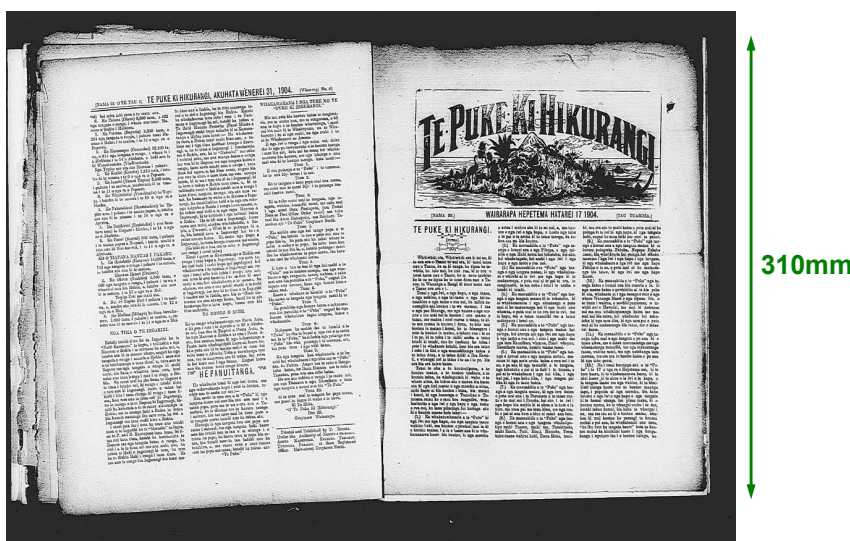




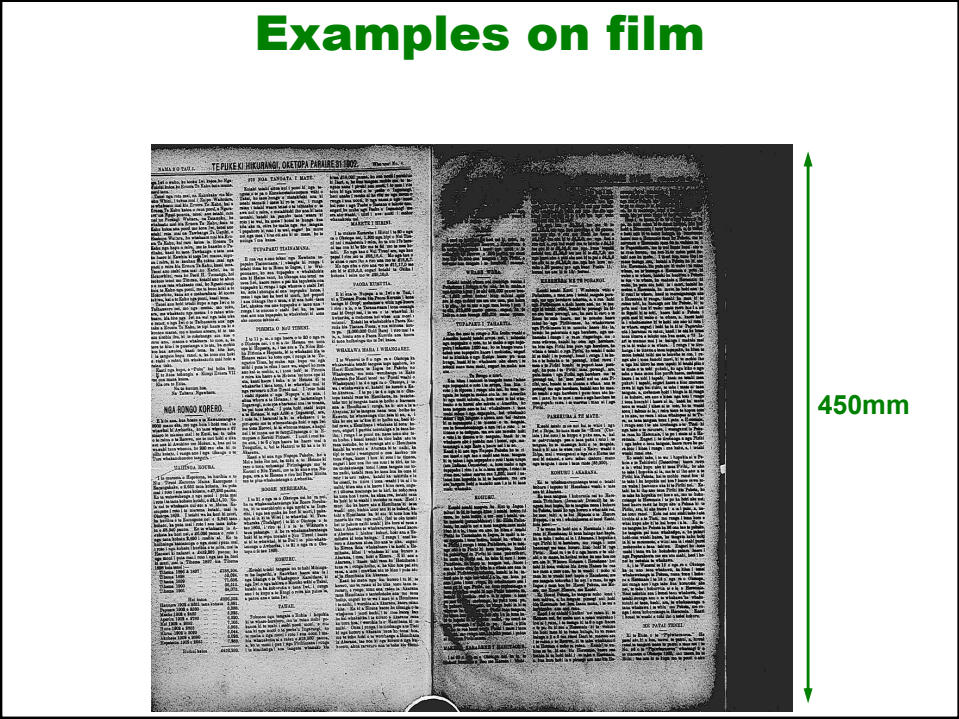
## Capturing print material

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

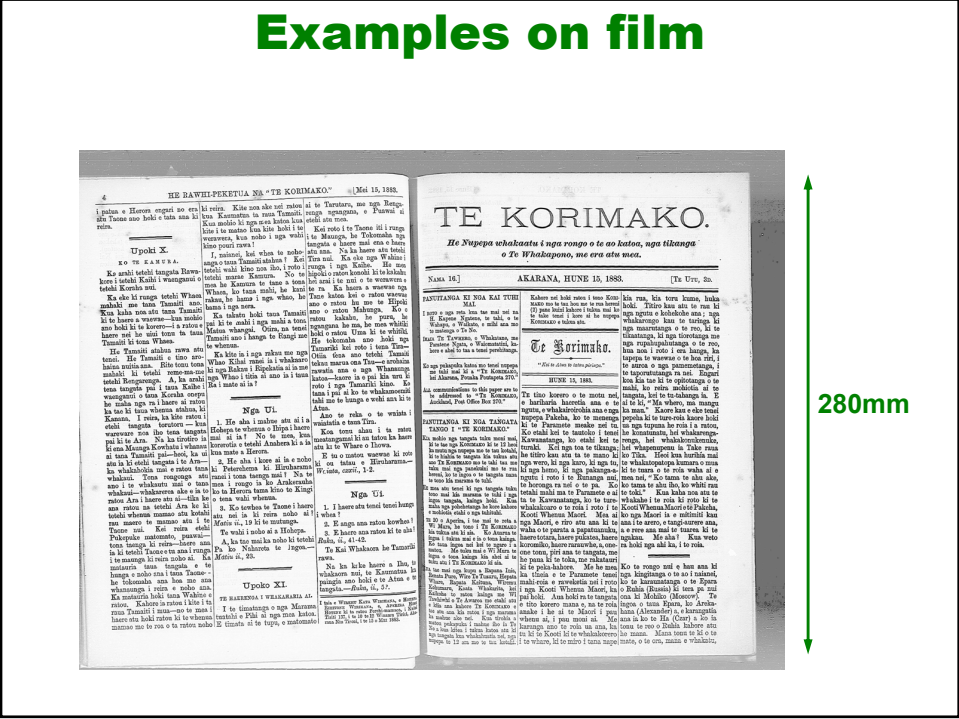
## Examples on film



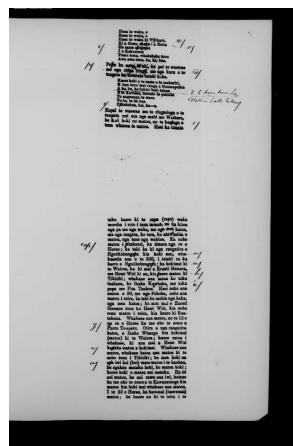
Examples on film



Examples on film



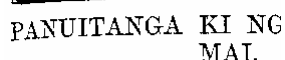
## Examples on film



## 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

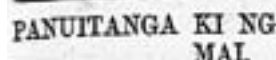
## Bitonal and grey-scale images



PANUITANGA KI NG  
MAL.

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

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



PANUITANGA KI NG  
MAL.

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

IHAI TE TAWHERO, o  
Paratene Ngata, o W  
hore e ahei to taa a tes

## 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

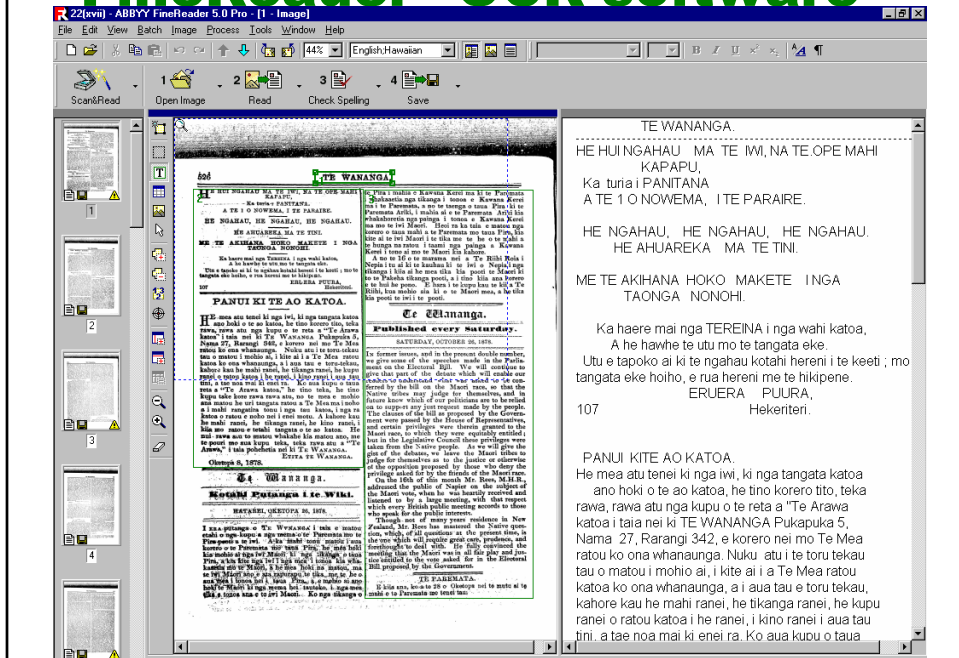
After testing many OCR programs we settled on ABBYY FineReader 6.0 Corporate because:

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

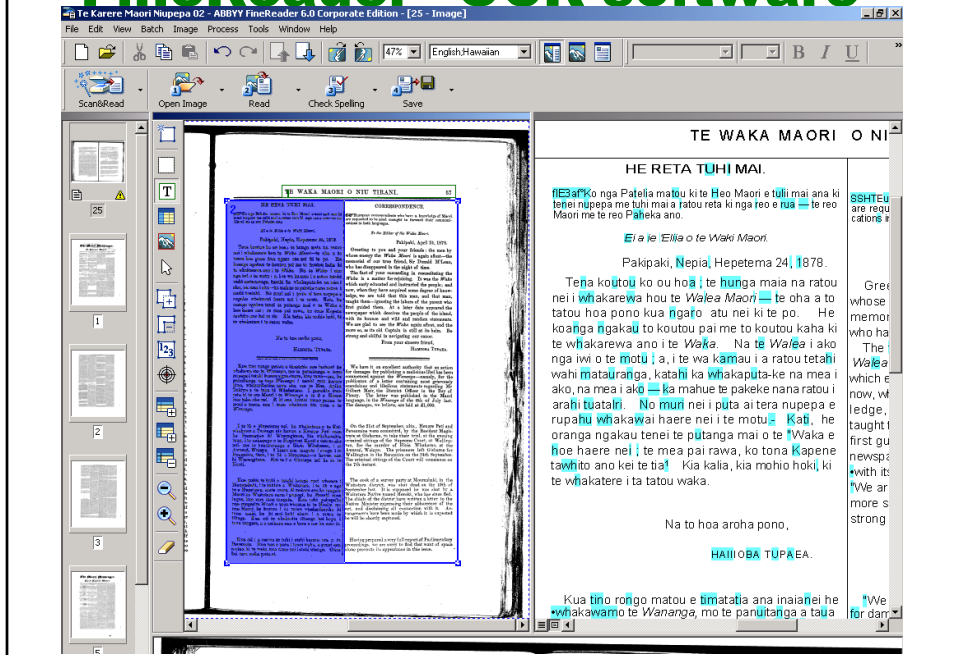
Also:

- 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

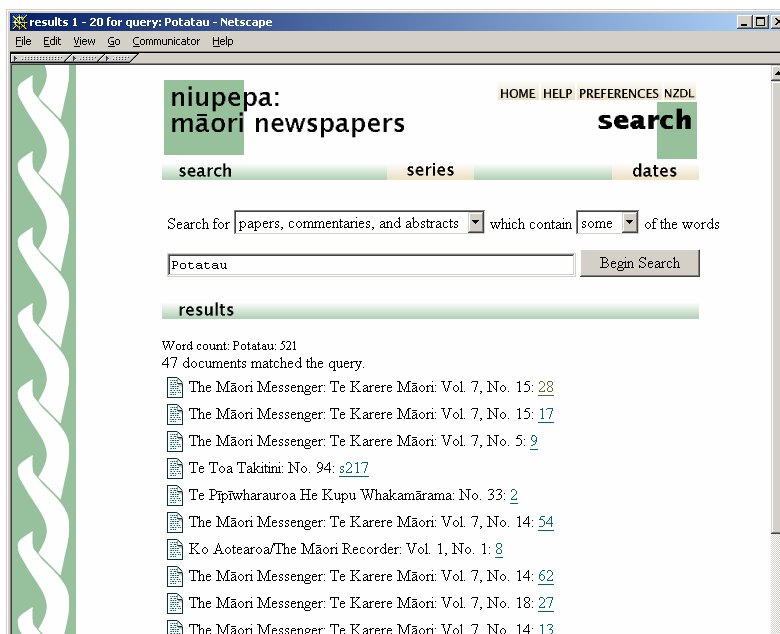


## User Interface

[www.nzdl.org/niupepa](http://www.nzdl.org/niupepa)

- ❖ Full-text search available for > 98%.  
Other <2% poor quality images.
- ❖ Unicode displays non-ASCII characters correctly
- ❖ Document display – three choices
  - Text
  - Preview image
  - Full size image

## Full-text search



Text page

HOME HELP PREFERENCES NZDL

searchseriesdates

[Vol. 7, No. 15](#)  
[3 August 1860](#)

page 28 (79 pages)

27 go to page 29

FULLSIZE IMAGE PREVIEW IMAGE DETACH

NO HIGH-LIGHTING



THE MAORI MESSENGER. 28 TE KARERE MAORI to you! Mr. McLean, I greet you! I am from Waikato, or rather from Manukau boundary lines and fixed others. When I saw that the boundaries were wrong I spoke to you, to the officers of the Governor. You said, " Perhaps the Natives have changed the boundaries." This grievance is between you ana me. It was not till the Bishop had urged it upon the Governor that Pukekohe was given up, but Mr. McLean still holds part of it. This setting up of a King was not a project of mine, nor of any part of Waikato. It originated with Turoa, Moaunui, and Te Heuheu. It was agreed to by Hoani Papita and Tamihana, and **Potatau** was selected as King. It originated with the tribes in the South. Afterwards the people of Waikato invited **Potatau** saying, " Come to Waikato and be a father for the Nation." **Potatau** did not approve of this King project. **Potatau** went to the Governor and said to him, "Friend, I am urged by the people to return to

Preview image page

HOME HELP PREFERENCES NZDL


searchseriesdates

[Vol. 7, No. 15](#)  
[3 August 1860](#)

page 28 (79 pages)

27 go to page 29

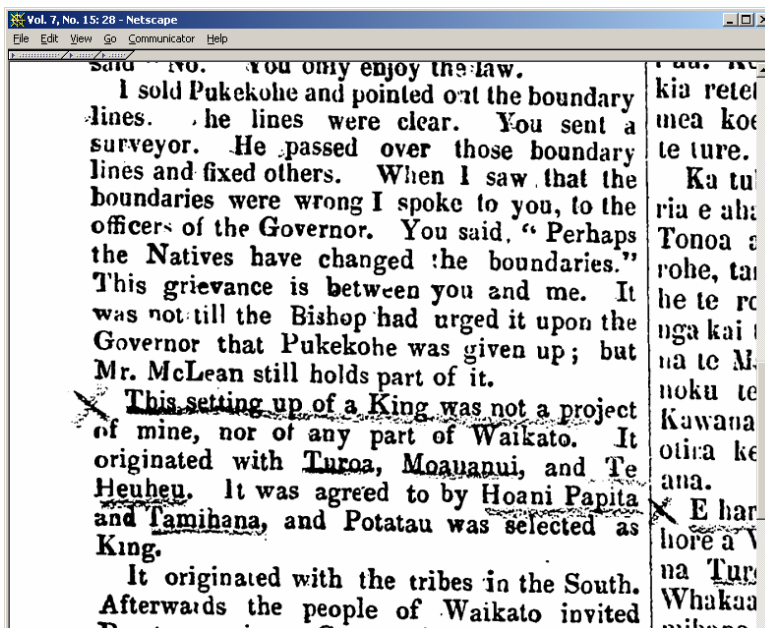
TEXT FULLSIZE IMAGE DETACH



THE MAORI MESSENGER. 28 TE KARERE MAORI

to you! Mr. McLean, I greet you! I am from Waikato, or rather from Manukau boundary lines and fixed others. When I saw that the boundaries were wrong I spoke to you, to the officers of the Governor. You said, " Perhaps the Natives have changed the boundaries." This grievance is between you ana me. It was not till the Bishop had urged it upon the Governor that Pukekohe was given up, but Mr. McLean still holds part of it. This setting up of a King was not a project of mine, nor of any part of Waikato. It originated with Turoa, Moaunui, and Te Heuheu. It was agreed to by Hoani Papita and Tamihana, and **Potatau** was selected as King. It originated with the tribes in the South. Afterwards the people of Waikato invited **Potatau** saying, " Come to Waikato and be a father for the Nation." **Potatau** did not approve of this King project. **Potatau** went to the Governor and said to him, "Friend, I am urged by the people to return to

## Large image page



## 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