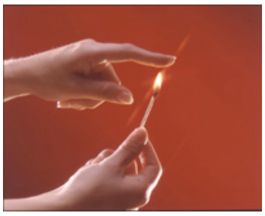
Examples: Multimedia and scanned images



Course material prepared by

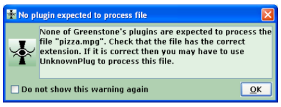
Greenstone Digital Library Project
University of Waikato, New Zealand

Agenda

- ❖ Multimedia and UnknownPlug
- ❖ Scanned images

Multimedia and UnknownPlug

- ❖ Greenstone has plugins for these multimedia file types:
  - – Images (.bmp, .gif, .jpg, .png, .tif, …): ImagePlug
  - – Audio (.mp3, .ogg): MP3Plug, OggVorbisPlug
  - – Video (.rm): RealMediaPlug
- ❖ What about files that Greenstone doesn't have a plugin for?



No plugin expected to process file

None of Greenstone's plugins are expected to process the file "pizza.mpg". Check that the file has the correct extension. If it is correct then you may have to use UnknownPlug to process this file.

☐ Do not show this warning again

  - – Use UnknownPlug, a plugin that will process *any* type of file

---

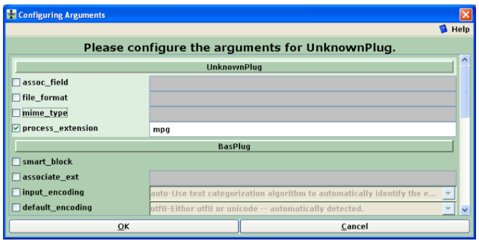
Multimedia and UnknownPlug

- ❖ UnknownPlug knows nothing about any type of file, so it can only extract *very* basic metadata: file name and file size
  - – Metadata must be manually assigned
- ❖ UnknownPlug is also unable to extract any text, so the files cannot be searched
  - – Files are accessed by browsing or searching on assigned metadata

---

Multimedia and UnknownPlug

- ❖ In the GLI, add UnknownPlug to the collection
  - – Use the "process_extension" option to tell UnknownPlug which type of files to process



Configuring Arguments

Please configure the arguments for UnknownPlug.

UnknownPlug

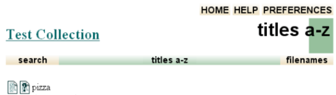
☐ assoc_field
☐ file_format
☐ mime_type
☑ process_extension: mpg

BasPlug

☐ smart_block
☐ associate_ext
☐ input_encoding: auto-Use text categorization algorithm to automatically identify the e...
☐ default_encoding: utf8-Either utf8 or unicode -- automatically detected.

Multimedia and UnknownPlug

- ❖ When the collection is built, the file will now be included in the collection:

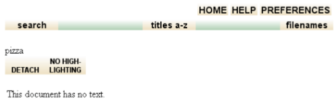


Click to view the Greenstone version of the file

Click to download/view the file

Multimedia and UnknownPlug

- ❖ Since UnknownPlug cannot extract any text, the Greenstone version of the document is very boring!



pizza

This document has no text.

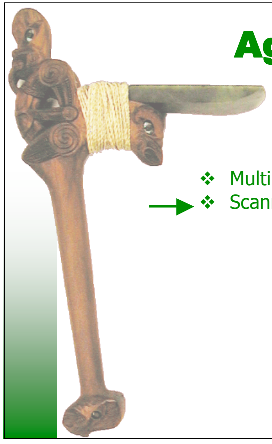
- ❖ Let's remove the link to this version

  Edit the VList format statement, and replace

  [link][icon][/link]

  with

  {If}{[ex.Plugin] ne "UnknownPlug",[link][icon][/link]}



Agenda

- ❖ Multimedia and UnknownPlug
- → ❖ Scanned images

Scanned images

- ❖ Need to be converted to an electronic form – scanning produces a set of images
- ❖ To add each page as an individual image, process using ImagePlug
- ❖ To group them into a single document, process using PagedImgPlug
  - – Requires an 'item' file which lists all the pages and gives additional metadata

---

Scanned images

- ❖ Can add metadata to the images to enable searching
- ❖ If full text searching is desired, use OCR (Optical Character Recognition) to generate an electronic version of the text
- ❖ Alternatively, if the documents are small and few, manually type the text into a file
- ❖ Text files can be included with the images in the item file

---

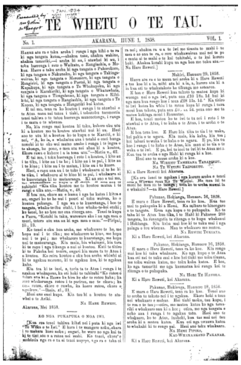
Sample document

4 newspaper page images and their scanned text

9 files:

```
10_1_1.item
images/10_1_1_1.gif
images/10_1_1_2.gif
images/10_1_1_3.gif
images/10_1_1_4.gif
text/10_1_1_1.txt
text/10_1_1_2.txt
text/10_1_1_3.txt
text/10_1_1_4.txt
```

10_1_1_1



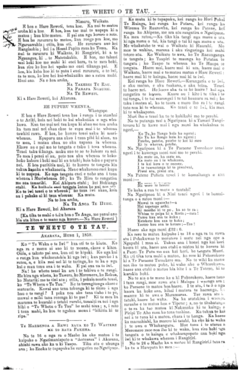
TE WHETU O TE TAU.
No. 1.
AKARANA, HUNE 1, 1858.
VOL. 1.

HAERE atu ra e taku aroha i runga i nga hihi o te ra ki nga tangata katoa,—ahakoa tane, ahakoa wahine, ahakoa tamariki,—i aroha ki au, i atawhai ki au, i toku haerenga i roto o Waikato, o Rangiaohia, o Mokau. Haere e taku aroha ki nga tangata o Pukorokoro, ki nga tangata o Nakunaku, ki nga tangata o Takingawairua, ki nga tangata o Meremere, ki nga tangata o Rangiriri, ki nga tangata o Paetai, ki nga tangata o Kupakupa, ki nga tangata o Te Whakapaku, ki nga tangata o Karakariki, ki nga tangata o Whatawhata, ki nga tangata o Tiongahemo ki nga tangata o Kihikihi, ki nga tangata o Rangiaohia ki nga tangata o Te Kopua, ki nga tangata o Hangatiki hui katoa.

E tai ma, tena ra ko koutou i runga i te atawhai o te Atua nana nei tatou i tiaki, i tohutohu taeanoatia tenei takiwa o to tatou haerenga manenetanga, i runga i te mata o te whenua.

Na, kia rongo mai koutou ki taku, kahore aku utu ki a koutou mo ta koutou atawhai nui ki au. Heoi ano te utu ki a koutou ko te kupu o te Karaiti, e ki nei, " A, ko ia, e whakainu ana i tetahi o enei hunga nonohi ki to oko wai matao anake i runga i te ingoa o te akonga, he pono, e mea atu nei ahau ki a koutou, Kore rawa e kahore i a ia tona utu."—Matiu x 42.

10_1_1_2



TE WHETU O TE TAU.
Nimaru, Waikato
E hoa e Hare Reweti, tena koe Ka nui to matou
aroha ki a koe E hoa, kua tae mai to nupepa ki a
matou, kua kite matou. E pai ana nga korero o roto
E hoa he raruraru to matou, kua rongo pea koe, mo
Ngaruawahia otiia, kua oti He raruraru ano kei
Rangiaohia kei i a Hoani Papata raua ko Toma Ka
nui te raruraru ki Waikato ki Rangiriri, ki a te
Ngaungau, ki te Matetakahia E hoa, me tahuri
mai hoki koe me mahi ki enei hara, ta te mea hoki
kua mo ko koe hei upoko mo enei tikanga pai. E
hoa, kia kaha koe ki te patu i enei raruraru, o to iwi,
ta te mea, ko koe hei kai whakatika mo matou mahi
Heoi ano .Na o hoa aroha,
NA TAKEPEI Te Rau
NA PANAPA NOAUMU
NA TE REWETI
Ki a Hare Reweti, kei Akarana
HE PUPURU WHENUA
Whangape
E hoa e Hare Reweti tena koe i runga i te atawhai
o te Ariki, koia nei hoki te kai whakatika a o nga wha
kawa Kua tae mai nei tau kupu ki ahau me tou aroha
ka taea mai noi ahau e koe te rapu mai i te whenua
tawhiti rawa E hoa, ho korero tenei naku u ki muri
whenua, Epupuru ana ahau i taua pihi ki ahau, ua
te mea, i matau ahau. no aku tupuna ia whenua
Kore au e pai ma te tangata e tuku i tena whenua
Tenei taku tikanga maku ano te ae te kahore ra nei
Te mea i penei ai au, poto noa aka whenua te hoko

10_1_1.item

```
<Title>Te Whetu o Te Tau
<Date>18580601
1:images/10_1_1_1.gif:text/10_1_1_1.txt:
2:images/10_1_1_2.gif:text/10_1_1_2.txt:
3:images/10_1_1_3.gif:text/10_1_1_3.txt:
4:images/10_1_1_4.gif:text/10_1_1_4.txt:
```

Metadata

Page number

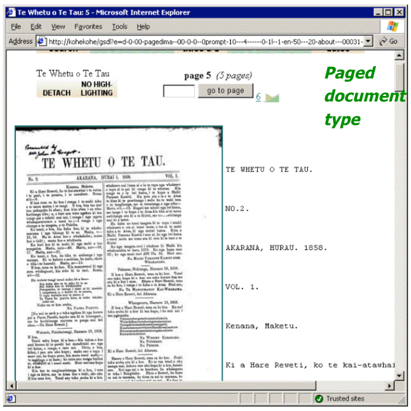
Image file

Text file

PagedImgPlug

- ❖ Processes item files and their corresponding image and text files
- ❖ Options:

screenview (screenviewsize, screenviewtype)
produce a preview image
thumbnail (thumbnailsize, thumbnailtype)
produce a thumbnail image
documenttype paged or hierarchical


Te Whetu o Te Tau: 5 - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address http://kohekohe/gsdl?e=d-0-00-pagedima--00-0-0--0prompt-10---4------0-1l--1-en-50---20-about---00031-
Go
Te Whetu o Te Tau
DETACH
NO HIGH-LIGHTING
page 5 (5 pages)
go to page
6
TE WHETU O TE TAU.
Trusted sites

TE WHETU O TE TAU.

NO.2.

AKARANA, HURAU. 1858.

VOL. 1.

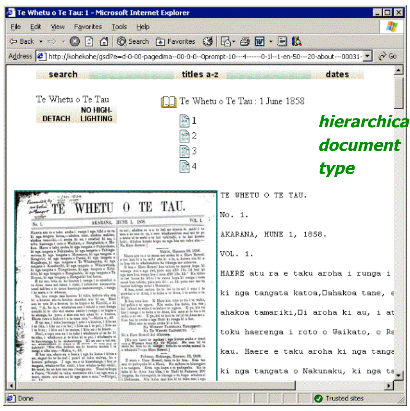
Kenana, Maketu.

Ki a Hare Reweti, ko te kai-atawhai

Preview image

Paged document type

Text


Te Whetu o Te Tau: 1 - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Search
Favorites
Address http://kohekohe/gsdl?e=d-0-00-pagedima--00-0-0--0prompt-10---4------0-1l--1-en-50---20-about---00031-
Go
search
titles a-z
dates
Te Whetu o Te Tau
DETACH
NO HIGH-LIGHTING
Te Whetu o Te Tau : 1 June 1858
1
2
3
4
TE WHETU O TE TAU.
Done
Trusted sites

TE WHETU O TE TAU.

No. 1.

AKARANA, HUNE 1, 1858.

VOL. 1.

HAERE atu ra e taku aroha i runga i

ki nga tangata katoa,□ahakoa tane, a

ahakoa tamariki,□i aroha ki au, i at

toku haerenga i roto o Waikato, o Ra

kau. Haere e taku aroha ki nga tanga

ki nga tangata o Nakunaku, ki nga ta

Preview image

hierarchical document type

Text

Extended item format

```
<PagedDocument>
  <Metadata name="Title">The Title of the entire
      document</Metadata>
  <Metadata name="Subject”>A Document level
      Subject</Metadata>
  <Page pagenum="1" imgfile="image1.jpg"
      txtfile="page1.jpg">
    <Metadata name="Title">The Title of this
      page</Metadata>
    … more metadata
  </Page>
... more pages
</PagedDocument>
```