Multilingual Support in GSDL

K.T.Anuradha

NCSI, Indian Institute of Science,

Bangalore 560 012, Karnataka, India

Email: anu@ncsi.iisc.ernet.in

OR

anu_ncsi@yahoo.co.in

http://vidya-mapak.ncsi.iisc.ernet.in

Multilingual Support in GSDL

- Topics covered:
  - What are char set, encoding & fonts
  - Creating interface in your language for GSDL
  - How to set default interface language
  - How to handle multilingual content
  - Unicode text
  - Non-Unicode text
  - Limitations

Charsets, encoding and fonts

Charset:- is a bunch of characters, in the way a human would understand them.

Ex: A, B, C so on are charset of Latin English

Encoding:- is a way of storing characters on a computer as bits.

Ex: ASCII, EBCIDIC, Unicode etc.

Font:- is an Image or glyph for a particular character.

Charsets & Encoding

- Widely used charsets
  - ISO 8859 series, windows series, gbk,
  - ISO 10646 (utf-8, utf-16, utf-32), ISCII, user-defined, etc..
- Unicode
  - Emerging encoding standard
  - Assigns unique hexadecimal numbers for more than 65,000 characters

    Ex: U+0041 is hexadecimal number of “A”

UNICODE..

- Encoding chosen by operating systems for supporting various non-English languages on computers.
- Though it supports all the languages of the world, the operating systems such as Windows, Linux may not have implemented all the languages yet. On Windows, the support for Indian languages was not available until Windows 2000 was released.

Unicode..

- Covers almost all international standards
  - First 125 characters are same as ASCII
- Synchronized with the corresponding versions of ISO-10646 (utf-8, utf-16, utf-32)
- Groups the characters together by scripts in code blocks

Enabling Indian Languages

- In Windows XP and Windows 2000, you should first enable the Indian languages before you can use Indian language UNICODE. You can enable the Indian language support using **Control Panel --> Regional Options** applet. If you are using an operating system that does not support the Indian language UNICODE, then you will not see UNICODE text properly. Instead you will see 'square box' or 'question mark' characters.

Language Fonts..

- Kannada, Telugu, Gujarati, Punjabi
  - UNICODE for Kannada, Telugu, Gujarati and Gurumukhi scripts is supported only in Windows XP and later operating systems. Following are different language fonts:
  - Open Type font "**Tunga**" to display Kannada UNICODE text.
  - Open Type font "**Gautami**" to display Telugu UNICODE text.

Language Fonts..

- Kannada, Telugu, Gujarati, Punjabi
  - Open Type font "**Shruti**" to display Gujarati UNICODE text.
  - Open Type font "**Raavi**" to display Gurumukhi UNICODE text.

Language Fonts..

- Oriya
  - Oriya script is not yet supported by Windows operating systems.
  - Windows XP shows Oriya UNICODE, provided a Open Type font that has Oriya characters (such as Arial Unicode MS) is installed in the system.

Language Fonts..

- Malayalam, Bengali
  - UNICODE for Malayalam, Bengali scripts is supported only in Windows XP SP2 and later operating systems.
  - Open Type font "**Kartika**" to display Malayalam UNICODE text.
  - Open Type font "**Vrinda**" to display Bengali UNICODE text.

Language Fonts..

- Devanagari, Tamil
  - UNICODE for Devanagari, Tamil scripts is supported only in Windows 2000 and and later operating systems.
  - Windows 2000/Windows XP uses an Open Type font "**Mangal**" to display Devanagari UNICODE text.
  - Windows 2000/Windows XP uses an Open Type font "**Latha**" to display Tamil UNICODE text.

Unicode support in GSDL

- Unicode is used throughout the process
- All major charsets are internally converted into utf-8 through mappings
- Different charsets mapped from and to Unicode are as follows
  - ISO 8859 series, windows series, gbk, big5, ShiftJis, uhc, ISCII, kio8_r, kio8_u, euc_jp, dos866

GSDL mapping files are stored in gsdl/mappings folder

Interface Customization

- Requirements
  - Operating system and browser should support your language
  - Install IME (Input Method Editor) for your language (Unicode supported)

Try out!!!!

Interface Customization

- 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_{जी.एस्.डी.एल }

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

  - Ex.. _textimagehome_ {Home page}

Interface Customization

- If your language is in Non-ASCII or Non-Unicode format copy the encoding

  Ex… _textimagehome_{å¸®åŠ©é¡µ}

- Pass language argument in front of each macro Follow ISO 639 two letter language standard

  Ex.. _textimagehome_ [l=zh] {}

Zh=Chinese

Configuration in main.cfg

- Include newly created macro file in macro files list of main.cfg
- Configure your language by passing short name, long name, & default encoding arguments

Language shortname=hi

longname=Hindi

default_encoding=utf-8

- ***Lets try!!!!***

Default interface language

- For entire server:

  Add the following argument at the end of main.cfg

  ```
  cgiarg shortname=l argdefault=xx
  ```

(this is already added in GSDL 2.70 onwards)

- For particular collection:

  Use “Languages" format option in collect.cfg

  demo

Multilingual Content: Unicode content

- Operating system and browser should support your language
- Input filenames and its extensions should be in ASCII (English)
- Can build the collection using GLI or in command line mode

Settings to view collection

- Select respective language interface from preference page
- Select utf-8 encoding from preference page
- Enable auto select or select utf-8 encoding in browser

i.e, View -> encoding -> utf-8

Search

- Need to install respective IME (Input Method Editor) for your language or online keyboard
- Browser should support for passing query string
- Can search for a particular word or phrase
- Advance searching is bit complicated
  - i.e, Boolean and proximity operators should be typed in Latin language only
- Demo ….

Non-Unicode mapped content

- Collection building is normal
- Content should follow any one of the mapped native encoding format
- Respective encoding should be enabled in etc/main.cfg file
- GSDL will internally map the encoding into utf-8 and build the index
- Converts back to native encoding while displaying

Settings to view collection

- Select your language interface from preference page
- Select native encoding from preference page
- Enable Auto-select or select native encoding in the browser

Ex: Your collection is in windows-1256,

select Arabic (windows-1256) in preference page, Arabic(windows) encoding in browser

Searching

- Need to install IME (Unicode supported) of your language or use Online keyboard
- Browser should support for passing query string
- Search features are same as Unicode content

Demo….

Non-Unicode unmapped content

- Collection which uses encoding like
  user-defined
  X-user-defined etc..
- GSDL will index as it is (mapping won’t take place)

Settings

- Need to have particular fonts in fonts folder of operating system
- Enable auto-select option in browser
- Searching: query string should be entered in basic encoding format

Limitations

- Alphabetical sorting ?
- Stemming?
- Boolean and proximity search operators should be entered in ASCII form only
- Appears to support html, word files

For better results

- You need to Understand

> Operating systems multilingual support
> Browsers multilingual support
> Charsets, encodings, fonts
> Html properties
> Unicode

BARAHA

- Baraha allows you to convert the Indian language text into UNICODE format in various manners as follows.
  - You can copy the text in UNICODE format using *Edit / Copy Special* command.
  - You can type UNICODE text directly into any applications that support UNICODE using Baraha Direct.
  - You can save the Baraha documents as UNICODE text files using the *File / Export* command.
  - You can save the Baraha documents as HTML files which use UTF-8 encoding using *File / Export* command.

Resources

- GSDL developers guide
- Greenstone archives collection
- Greenstone wiki page
- http://www.unicode.org

Summarizing

- Creating multilingual interface
- Handling multilingual collection
- Unicode content
- Non-Unicode mapped content
- Non-Unicode unmapped content
- Limitations

Thank You!

To view sample collection visit:

http://vidya-mapak.ncsi.iisc.ernet.in/bahubhashi

Send your feedback to

anu@ncsi.iisc.ernet.in

anu_ncsi@yahoo.co.in