Japanese support on eSciDoc solutions

I18N and other issues

Masao Takaku
National Institute for Materials Science (NIMS), Japan
TAKAKU.Masao@nims.go.jp
Key issues

- Japanese language has many differences with Western languages.
  - What is characteristics of Japanese language?
- I18n/m17n overview in PubMan
  - How to provide I18n/m17n capabilities for PubMan?
- Cultural difference (beyond the language processing)
  - What is needed for Japanese social adoption of PubMan other than just the language itself?
I18N

- Internationalization (i18n) enables software to provide solutions in users’ native language environment.
  - Multilingualization (m17n) enables software to work with multiple mixed language environment.

- I18N/M17N issues in PubMan:
  - GUI (labels and messages)
  - Browsing Japanese text (styles and fonts)
  - Searching Japanese text (Indexing and retrieval)
  - Input support for Japanese
Japanese characters

- One of the most important differences of linguistic feature between Japanese and Western languages are its unique character set.

- Japanese characters are comprised of three types of mixed scripts:
  - Kanji (Chinese character) \(\geq\) Over 10,000
  - Katakana, Hiragana (Japanese original characters) \(\approx\) about 140
  - Romaji and other symbols (Alphabetical scripts) \(\approx\) about 50

- Unicode character encodings
  - Java: character processing is based on Unicode.
  - UTF-8: most of single Japanese character take three bytes.
    - “あ” \(\equiv\) ¥xe8¥x87¥x92
Example of Japanese characters...

- Japanese have to learn a lot of characters...
- From my niece’s notebook.
New videos out on PubMan user workspaces
by nek - 2009年6月04日
The workspaces in PubMan serve as a kind of role specific desktop. Dependant on the role you have in the system you will have ...

New videos out on how to submit publications to PubMan
by nek - 2009年6月04日
PubMan gives several possibilities to submit publications to PubMan. In the following two easy submission methods are being ...

PubMan and Browsers
by nek - 2009年5月08日
PubMan video to support all browsers.
GUI i18n (Labels/messages)

- Modern programming frameworks support a mechanism of switching user locales.

- For eSciDoc framework (PubMan etc.), Java resource bundles work well.
  - From PubMan-R4, Japanese GUI has successfully been provided as well as English and German.
  - `pubman_presentation/src/main/resources/Label_{en,de,ja}.properties`
  - `<Key> = <Value>`
  - I18n how-to documentation is available at `[[User:Masao/i18n]]` on CoLab.
GUI i18n (Labels/messages)

PubMan supports the dissemination and re-use of publications and supplementary material. The solution PubMan is a component of the eResearch infrastructure of the Max Planck Society and is based on the service-oriented architecture of eSciDoc.

(Read more...)

pubman_presentation/src/main/webapp/ HomePage.jsp

...<h1><h:outputText value="#{lbl.HomePage}" /></h1>...

National Institute for Materials Science
GUI i18n (Labels/messages)

```jsp
<h1><h:outputText value="#{lbl.HomePage}" /></h1>
```

```properties
... 
HomePage = Home 
... 
```

```properties
HomePage = ¥u30DB¥u30FC¥u30E0 
(Unicode escaping)
```
ホーム

一時保存 (0)

PubMan - eSciDoc

PubManは研究者、技術者、研究機関などを対象としています。

PubManリソースは、研究者、技術者を支援するための、効率的な情報管理システムとして設計されています。

(詳細はCoLab上の説明をご覧ください)

文検査

... また、組織一覧からも探せます

新刊お知らせ

New videos out on how to submit publications to PubMan

PubMan gives several possibilities to submit publications to PubMan. In the following two easy submission methods are being ...

PubMan and Browsers

PubMan tries to support all commonly used browsers. A detailed list of supported browsers can be found in CoLab.

eSciDoc Days 2009
National Institute for Materials Science
Searching issue

- Indexing part
  - Tokenize Japanese text

- Retrieval part
  - Displaying text

- Sample sentence:
  - 私たちをeSciDoc Days 2009にご招待いただき、ありがとうございました。
  - Thank you very much for your invitation to eSciDoc Day 2009.
Japanese tokenizer

- Japanese have no space between words.
- 私たちをeSciDoc Days 2009にご招待いただき、ありがとうございます。
  □ 13 words
- There are two approaches to tokenize Japanese sentence:
  □ N-gram based approach
  □ Morphological analysis (NLP-based approach)
Lucene and its variance

- eSciDoc uses Lucene as backend search/retrieval engine
- In Lucene, there are two tokenizer for Japanese text:
  - CJKAnalyzer / JapaneseAnalyzer
  - (n-gram based / morphological based)
- With a tokenizer setting, eSciDoc can handle Japanese text searching correctly.
Thank you

E-mail: TAKAKU.Masao@nims.go.jp