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?

NIMS Digital Library
.e SciDoc

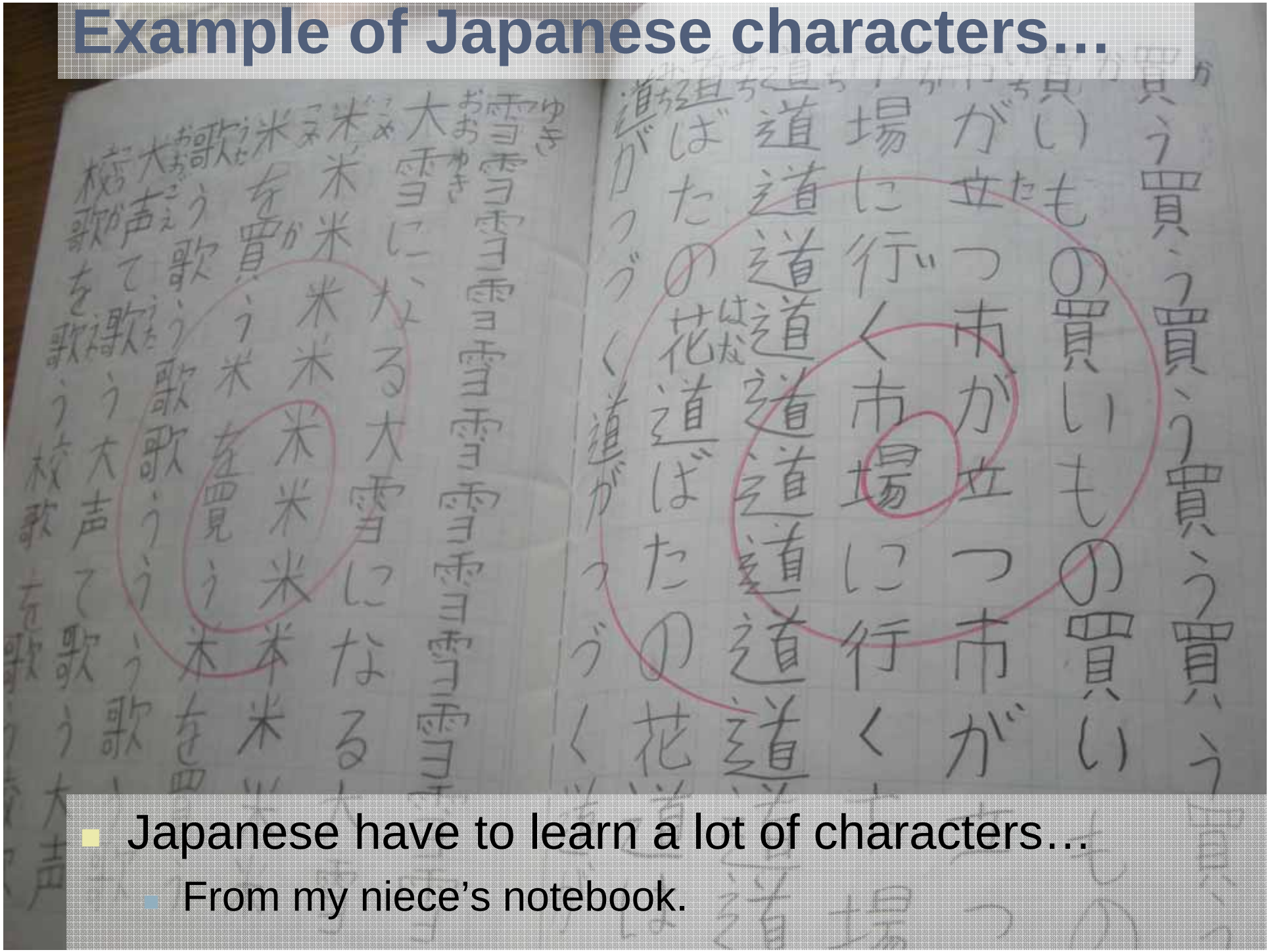
I18N

- 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 compriesed of three types of mixed scripts:
  - Kanji (Chinese character) → Over 10,000
  - Katakana, Hiragana (Japanese original characters)
    → about 140
  - Romaji and other symbols (Alphabetical scripts)
    → about 50
- Unicode character encodings
  - Java: character processing is based on Unicode.
  - UTF-8: most of single Japanese character take three bytes.
    - "**あ**" → ¥xe8¥x87¥x92

Example of Japanese characters…



- Japanese have to learn a lot of characters…
  - From my niece’s notebook.

GUI i18n (Labels/messages)

GUI i18n (Labels/messages)

- Modern programming frameworks supports 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_presentation/src/main/webapp/**HomePage.jsp**

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

GUI i18n (Labels/messages)

pubman_presentation/src/main/webapp/**HomePage.jsp**

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

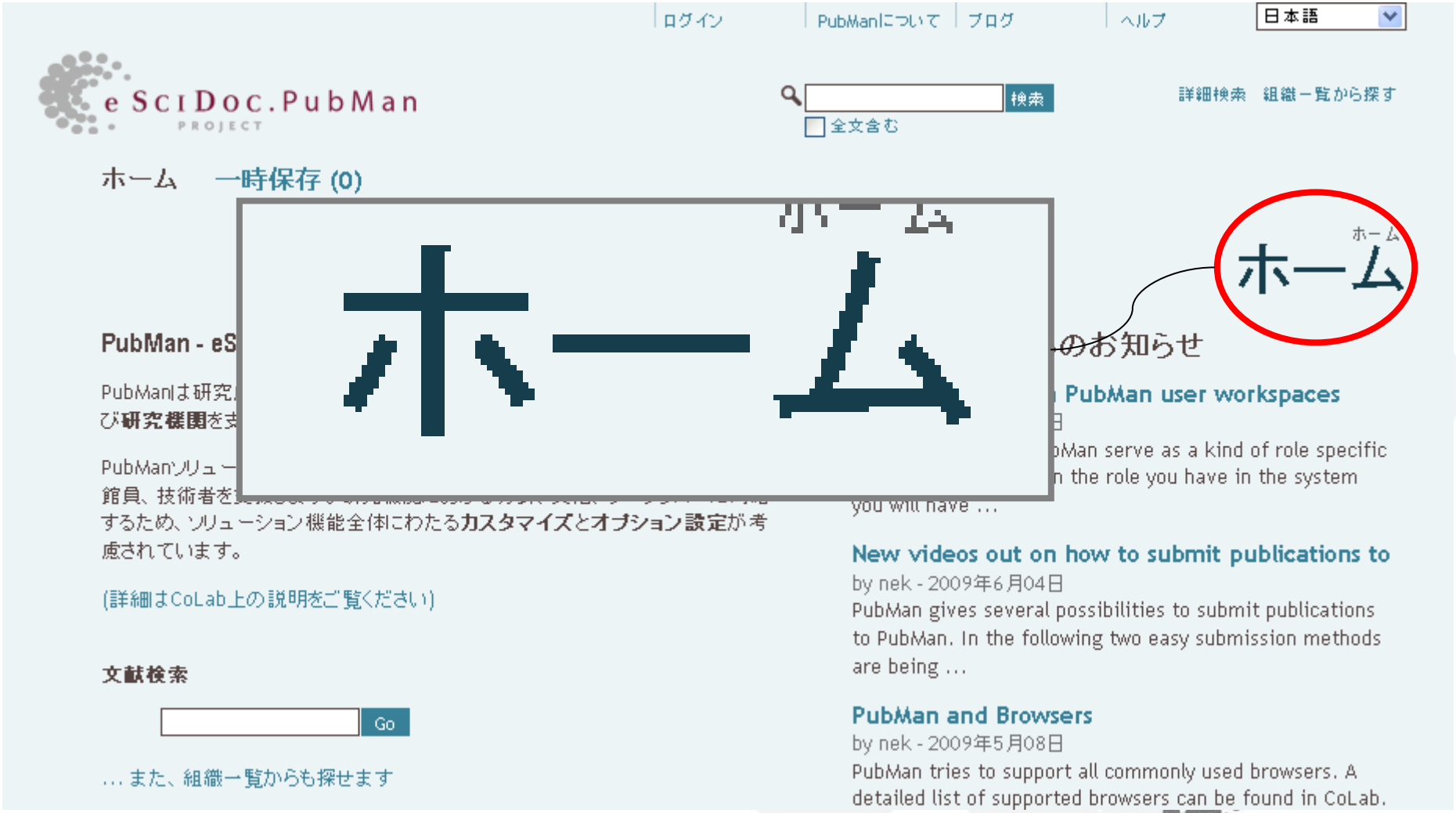
pubman_presentation/src/main/resources/**Label_en.properties**

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

pubman_presentation/src/main/resources/**Label_ja.properties**

```
HomePage = ¥u30DB¥u30FC¥u30E0
```

(Unicode escaping)


ログイン
PubManについて
ブログ
ヘルプ
日本語
eSciDoc.PubMan
PROJECT
検索
全文含む
詳細検索
組織一覧から探す
ホーム
一時保存 (0)
ホーム
ホーム
PubMan - eS
のお知らせ
PubManは研究
び研究機関を支
PubMan user workspaces
PubMan serve as a kind of role specific
the role you have in the system
you will have ...
PubManソリュー
館員、技術者を
するため、ソリューション機能全体にわたるカスタマイズとオプション設定が考
慮されています。
(詳細はCoLab上の説明をご覧ください)
New videos out on how to submit publications to
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 ...
文献検索
Go
PubMan and Browsers
by nek - 2009年5月08日
PubMan tries to support all commonly used browsers. A
detailed list of supported browsers can be found in CoLab.
... また、組織一覧からも探せます
NIMS Digital Library
.eSciDoc

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した。***
  → 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