Implementing XML for Japanese-Language Scholarly Articles

2012. 10. 16
Soichi Tokizane
Aichi University

Whom Do I Represent?

• Scholarly XML Publishing Association
  – Newly-formed organization to promote XML for scholarly publishing in Japan
SGML/XML for Scholarly Publishing in Japan

- There were attempts to implement SGML to scholarly publishing in early 90’s.
- XML was tried, too, in early 2000’s

J-STAGE

- J-STAGE, an ejournal platform of JST
  - Launched in 1999
  - Developed J-STAGE SGML DTD
    - Based on JICST-DTD
  - FrameMaker+SGML production workflow
  - Investigated XML in 2000, but found it too early to implement it
JATS 0.4 for New J-STAGE

• New version of J-STAGE launched in May, 2012
  – Implemented JATS 0.4

But all three languages are different from each other

• Japanese language came from Mongolia
• Japanese language owes
  – characters to China
  – old words to Korea
In any case …

• We, three, all needed multi-language extension of NLM DTD

Characteristics of Japanese Writing (1)

• Scripts
  – 野田 佳彦 (Kanji – Chinese)
  – のだ よしひこ (Hirakana)
  – ノダ ヨシヒコ (Katakana)
  – Yoshihiko Noda (Romaji – Roman)
Characteristics of Japanese Writing (2)

• Name Order
  – 野田 佳彦 (Kanji – Chinese)
  – Yoshihiko Noda (Romaji – Roman)
  • (Family names in red)

• In JATS
  name-style="eastern"

Characteristics of Japanese Writing (3)

• Vertical, Right-to-Left
  – Traditional (also in historic Chinese)
  – Used in newspapers, magazines, books and humanity/social science publications/textbooks

• Western style
  – New
  – Exclusively used in business
  – Used in science/technology publications
Vertical, Right-to-Left

Parentheses rotated in Vertical Writing

あいうえお

かきくけこ
Ruby

- Used to express pronunciations
- Used to teach children how to read
- Needed, e.g., for personal names as well

- 時実 象一
  - can be pronounced as, “Tokijitsu Zoichi”

- 時実 象一
  - so that it is read correctly as, “Tokizane Soichi”
Emphasis

- *Italics* or **bold faces** do not exist in Japanese writings
- Emphasis are used instead

Warichu

- Short note inserted within a sentence in two lines
Non-Gregorian Years

• Japanese traditional year designations
  – Now called, “Emperor year”, and is formal
• Rare in scholarly publications
  – Still appear in citations of old references

昭和四〇年 = 1965

Japanese STM Publishing (1)

• Number of journal titles (2008)
  – 3,047 total in STM
  – 2,673 Japanese titles (774 ejournals, 29.0%)
  – 374 English titles (298 ejournals, 79.7%)
Japanese STM Publishing (2)

- English-language journals are the same as Western ones in styles
- Japanese-language journals have:
  - Article titles, author names, affiliations – always in Japanese and English (Romanized)
  - Abstracts and keywords – in two languages mostly
  - Captions – either in Japanese or English
  - Cited references – either in Japanese OR English
  - Emphasis and Warichu are not used

Example

日本発行の科学技術分野の電子ジャーナル数
2005年から2008年への変遷

Electronic journal titles in science, technology and medicine published in Japan
Changes from 2005 to 2008

TOKIZANE Soichi

Faculty of Letters, Aichi University

Faculty of Letters, Aichi University

Machihata-cho, Toyohashi-shi, Aichi 441-8522
Example

• Citations either in Japanese or English
  – Sometimes both for a single citation


Characteristics of Humanity/Social Science Publishing

• Vertical writing
  – Always for humanities
  – Often for social sciences
• Use of Empahsis and Warichu
• English-language article metadata are not common
Ejournal Publishing in Japan

- Mostly in PDF
- Difficult to add value to ejournals
- XML is needed
  - not only for English-language journals
- Multi-lingual XML DTD is needed
  - to meet the Japanese-language journal needs to describe meta data in two languages
Limitations of NLM DTD (3.0)

- Does not allow @xml:lang for authors and affiliations
- Bypassing example using @name-style

```xml
<contrib-group>
  <contrib contrib-type="author">
    <name name-style="western"><surname>Nihon</surname><given-names>Taro</given-names></name>
    <name name-style="eastern"><surname>日本</surname><given-names>太郎</given-names></name>
  </contrib>
</contrib-group>
```

Limitations of NLM DTD (3.0)

- Many elements not repeatable
  - `<kwd-group>`, `<publisher-name>`, etc.
  - unable to describe in multiple languages
- Some repeatable, but lack envelopes
  - `<name>`, `<aff>`
  - cannot indicate that descriptions in two languages belong to a single identity
Scholarly Publishing Japan (SPJ) Working Group

- Established by stakeholders in early 2010
  - Publishers
  - Typesetting companies
  - JST (observer)
  - NII (observer)

- Purpose
  - To enhance NLM DTD (JATS) so that it would fully support multi-language contents
  - Worked with NLM DTD (JATS) Working Group

SPJ Proposals (1)

- NLM DTD should be as much as structural
- Support both multiple languages as well as multiple scripts using IETF RFC 5646, for example `xml:lang="ja-Kana"`
- Stop using the default language as English for all elements
- Allow describing `@xml:lang` for `<name>`
- Introduce `<subbody>` to allow multiple language body texts
SPJ Proposals (2)

- Devise ways to describe cited references in multiple languages, for example, `<compound-element-citation>`
- Allow describing `<journal-meta>` data such as `<journal-title>`, `<journal-subtitle>`, and `<abbrev-journal-title>` in multiple languages
- Allow specifying `<xref>` for individual `<name>` elements so that the institution name in the same script may be linked directly

NLM DTD Version 3.1 Draft

- 2010/9
  - Multi-lingual support
- 2011/3
  - Became NISO JATS 0.4
Multi-lingual Support of JATS 0.4


Multi-lingual Support of JATS 0.4

- xml:lang is now usable for almost all the elements
- xml:lang is inherited down the XML document tree
- Both language and script may be recorded
- Most elements may be repeatable to describe multiple language expressions
- Wrapping tags are introduced to concatenate repeating elements for a single data, such as `<name-alternative>` or `<aff-alternatives>`
@xml:lang with Scripts

- 野田 佳彦  xml:lang="ja-Jpan" (Kanji)
- のだ よしひこ  xml:lang="ja-Hira" (hirakana)
- ノダ ヨシヒコ  xml:lang="ja-Kana" (katakana)
- Yoshihiko Noda  xml:lang="en" (Romaji)

Most Elements Repeatable

<kwd-group xml:lang="en">
  <kwd>heated air</kwd>
</kwd-group>
<kwd-group xml:lang="ja">
  <kwd>加温空気</kwd>
</kwd-group>

- Except article titles
  - Use <trans-title>
Wrapper

• `<name-alternatives>`

```xml
<name-alternatives>
  <name name-style="eastern" xml:lang="ja-Jpan">
    <surname>中西</surname>
    <given-names>秀彦</given-names>
  </name>
  <name name-style="western" xml:lang="en">
    <surname>Nakanishi</surname>
    <given-names>Hidehiko</given-names>
  </name>
  <name name-style="eastern" xml:lang="ja-Kana">
    <surname>ナカニシ</surname>
    <given-names>ヒデヒコ</given-names>
  </name>
</name-alternatives>
```

Wrapper

• `<aff-alternatives>`

```xml
<aff-alternatives>
  <aff id="aff1">
    <institution xml:lang="ja-Jpan">国立言語学博物館</institution>
  </aff>
  <aff id="aff-en">
    <institution xml:lang="en">National Museum of Linguistics</institution>
  </aff>
</aff-alternatives>
```
Additional Proposals

• Coding Ruby
  `<rubigrp>多武峰<rubi>とうのみね</rubi></rubigrp>`
  or
  `<ruby>多武峰<rt>とうのみね</rt></ruby>`
  (as in HTML5)

• `<ref-alternatives>`

```xml
  <ref-alternatives id="B30">
    <ref xml:lang="ja">
      <mixed-citation publication-type="journal" publication-format="print">
        <person-group person-group-type="author">
          <string-name name-style="eastern">
            <surname>柿崎</surname>
            <given-names>一郎</given-names>
          </string-name>
        </person-group>. <year>2000</year>. 『<source>タイ 経済と鉄道 1885 年～1935 年</source>』<publisher-name>日本経済評論社</publisher-name>
      </mixed-citation>
    </ref>
    <ref xml:lang="en">
      <mixed-citation publication-type="journal" publication-format="print">
        <person-group person-group-type="author">
          <string-name name-style="western">
            <surname>Kakizaki</surname>, <given-names>Ichiro</given-names>
          </string-name>
        </person-group>. <year>2000</year>. <source>Thai Economy and Railway 1885-1935</source>. <publisher-name>Nihon Reizai Hyoronsha</publisher-name>
      </mixed-citation>
    </ref>
  </ref-alternatives>
```
New J-STAGE

- **J-STAGE**
  - E-journal platform operated by the Japan Science and Technology Agency (JST)
  - 2,387,426 articles/1,658 titles
- **New J-STAGE launched in May, 2012**
  - Supports JATS 0.4
- **Several journals are loaded using JATS XML already**

JATS-Con 2012
J-STAGE XML Guideline (1)

• Characters
  – Characters are in UTR-8.
  – Entity references such as, &amp;, &lt;, &gt;, &apos;, &quot;, and characters of ISO8879(SGML), MathML characters, and JATS specific characters such as &gcaron;, &Hmacr;, €, and &franc;. may be used.

• Font attributes
  – &lt;**bold**&gt;, &lt;**italic**&gt;, &lt;**monospace**&gt;, &lt;**roman**&gt;, &lt;**sans-serif**&gt;, &lt;**sc**&gt;, &lt;**overline**&gt;, &lt;**strike**&gt;, &lt;**sub**&gt;, &lt;**sup**&gt; and &lt;**underline**&gt; may be used.

J-STAGE XML Guideline (2)

• XML declaration and DOCTYPE
  – version="1.0" encoding="UTF-8"
  – <!DOCTYPE article PUBLIC "-//NLM//DTD JATS (Z39.96) Journal Publishing DTD v0.4 20110131//EN"
    "http://www.jstage.jst.go.jp/dtds/JATS-journalpublishing0.dtd"
J-STAGE XML Guideline (3)

- Journal meta
  - `<journal-id>` and `<issn>` must exist.
- Article meta
  - “publisher-id” should be JOI (JST Object Idnetifier, not publisher’s own id)
  - `<contrib-group>` must exist
  - `<name-alternatives>` and `<aff-alternatives>` required
  - `<xref>` id has to be in `<aff-alternatives>`

J-STAGE XML Guideline (4)

- Mathematical formulas
  - Text, graphic, MathML and Tex/LaTeX are supported.
- Figures and Tables
  - Should appear where mentioned
  - OASIS/CALS not supported for a while
- References
  - `<mixed-citation>` is supported, not `<element citation>`.
Production Workflow (1)

• Typesetter A
  – MS Word 2007 to JATS XML (own program)
  – XML to XSL-FO
  – XSL-FO to PDF
  – XML Editor to make corrections

Production Workflow (2)

• Typesetter B
  – TeX to JATS XML (own program)
  – MS Word to XML using eXtyles
    • Japanese-language XMLs need much human editing
  – XML to 3B2
  – 3B2 to JATS XML and PDF
  – Corrections on 3B2
Production Workflow (3)

• Typesetter C
  – Pastes texts to FrameMaker
  – FrameMaker-generated texts processed by eXtyles, then to JATS XML using XSLT
  – FrameMaker will be replaced by Typefi/InDesign soon

Production Workflow (4)

• Typesetter D
  – MS Word to XML using eXtyles
    • Without cleanup for Japanese texts
  – XML to InDesign
  – InDesign XML to JATS XML using stylesheet
  – Corrections on InDesign
InDesign/Typefi


Problems and Issues with J-STAGE Guideline

- `<name-alternatives>` and `<aff-alternatives>` are must
- `<element-citation>` not supported
- "Publisher-id" should be a publisher's own id, rather than JOI.
- JST is working on them now.
Scholarly XML Publishing Association

- Extension of SPJ
- Established on June 28, 2012
- Soichi Tokizane named as President
- Symposium on September 19, 2012
- Future contact point of JATS Working group in Japan

What are next?

- J-STAGE has to be enhanced to fully take advantage of XML
- J-TAGE XML Guideline should be fixed
- Processing humanity/social science articles is very challenging
- Encourage scholarly book publishers to use JATS XML
Acknowledgement

• Soichi Kubota, JST
• Typesetters who gave me the information
• All SPJ members

Questions?

• Soichi Tokizane
  – tokizane@aichi-u.ac.jp