JATS XML for Japanese language

The 3rd Asian Science Editors’ Conference and Workshop 2016

Hidehiko Nakanishi Ph.D.
Nakanishi Printing Co., Ltd. Kyoto Japan
I am from Kyoto, Japan
Kyoto was a former capital

Where my company, Nakanishi Printing, is located.
Our Tradition

Former building of Nakanishi Printing in Taisho era (1912-1926)

Current building of Nakanishi printing
Founded in 1865 by our ancestor.
150 year old family business.
One of the oldest printers.
Our history

A brazier made by Woodcut print plate in 19c

Type picker 1960’s

Today
Contents

1. Introduction
2. JATS adoption in Japanese XML expression
3. Creating Japanese XML articles in JATS
4. Creating PDF using AH Formatter
5. Challenges of Applying JATS to Japanese language texts
6. Future
7. Conclusion
Introduction
Many countries use Non-Latin script
Not all research articles are written in English.

- Many articles don’t even use the Latin alphabets
What languages are used in articles written in Japan?

STM
- English
- Japanese

Articles published in J-Stage, E-journal platform operated by the Japan Science and Technology Agency (JST).

All areas
- English
- Japanese

University journal articles indexed in NDL-OPAC.
The importance of articles in native tongue

- English is the world’s common language.
- But there are many merits of describing and reading in one’s native tongue.
The change of world common language

- **English**
  - As natural language, common language especially in STM domain.
  - The evolution of translation software will change this position.

- **XML**
  - Becoming the world common language.
  - Prescribing common structure → Common understanding
The importance of the XML in the native tongue

- Common structured articles in the native language → XML in the native language.
- So far English was thought to be essential for XML description, now it is not!
- XML articles in Japanese is ideal.
- We need suitable schema.
JATST adoption in Japanese XML expression
We wanted schema applicable to Japanese

- Even for Japanese-language articles, e-articles are essential.

- We were looking for schema for Japanese-language articles.

- Such schema had to accept English as well.
In 2011 JATS 0.4 enabled to express Japanese-language articles in XML.

J-STAGE supported JATS 0.4 immediately.

We started creating JATS XML for Japanese-language articles.
This is a Japanese e-journal
Sustained Elevation of Serum Tumor Markers after Resection is an Important Prognostic Factor for Pancreatic Cancer

Purpose: Curative resection is the only chance for long-term survival for pancreatic carcinoma, however, the surgical results remain poor. Even after resection with a microscopically negative margin (R0), prognostic factors and predictors of recurrence are needed. Several reports revealed that sustained elevation of tumor markers (TM) after
3. 切除線維増殖マーカーによる生存期間の比較

294例の生存曲線を、Fig. 1に示す。生存期間中位値は684日で、6年生存率は24.4%であった。3群で生存曲線を比較すると、各群の1、3、5年生存率は、非上昇群で64.8、48.0、43.8%、正常化群で78.9、42.0、31.7%、非正常化群で61.6、18.4、9.9%であり、生存期間中位値は、非上昇群で1,092日、正常化群で736日、非正常化群で505日であった。非正常化群で、有意に生存期間が短く、予後に不良であった（Fig. 2, P<0.0001, log-rank test）。

Fig. 1
Kaplan-Meier survival curve of entire cohort.

Fig. 2
Kaplan-Meier survival curves by tumor marker status. There is significant difference between the survival curves by log-rank test (P<0.0001).

4. 多変量解析による予後因子の抽出

非切除・術前増殖マーカー・術式・進行度・組織型・細胞増殖・術後増殖マーカー・術後無残症の各因子で、Cox比例ハザードモデルを用い解析した。多変量解析では、術前増殖マーカー（CA19-9上昇・非上昇）、術式（肝全摘・切除切除）、進行度（stage 4> stage 3> stage 2> stage 1）、細胞増殖（R1>R0）、術後増殖マーカー（R1正常化・非正常化）が有意な予後因子として抽出された。それら多変量解析で有意であった因子を用いて多変量解析を行うと、進行度、術後増殖マーカーが有意な予後有意因子として選択された（Table 3）。
Expressing Multiple Languages

- Alternate expressions for a single object are necessary
- Simple repetition of a tag can be confusing
  - Two name expressions of the same person?
  - Or two different persons?
- JATS introduced “alternatives” tags for such cases
“Alternatives” Tags

• Two name expressions of a single person

```
<name-alternatives>
  <name name-style="eastern" xml:lang="ja-Jpan">
    <surname>中西</surname>
    <given-name>秀彦</given-name>
  </name>
  <name name-style="western" xml:lang="en">
    <surname>Nakanishi</surname>
    <given-name>Hidehiko</given-name>
  </name>
</name-alternatives>
```
“Alternatives” tags
How multiple language can be expressed in JATS

<table>
<thead>
<tr>
<th>element name</th>
<th>multi-language tag</th>
<th>Note</th>
</tr>
</thead>
<tbody>
<tr>
<td>article title</td>
<td>&lt;trans-title&gt;</td>
<td></td>
</tr>
<tr>
<td>article subtitle</td>
<td>&lt;trans-subtitle&gt;</td>
<td></td>
</tr>
<tr>
<td>names</td>
<td>&lt;name-alternatives&gt;</td>
<td></td>
</tr>
<tr>
<td>affiliations</td>
<td>&lt;aff-alternatives&gt;</td>
<td></td>
</tr>
<tr>
<td>collaborators</td>
<td>&lt;collab-alternatives&gt;</td>
<td></td>
</tr>
<tr>
<td>abstract</td>
<td>&lt;abstract&gt;</td>
<td>&lt;abstract&gt; is repeatable with different &quot;xml:lang&quot;. &lt;trans-abstract&gt; is for articles later translated.</td>
</tr>
<tr>
<td>keyword group</td>
<td>&lt;kwd-group&gt;</td>
<td>&lt;kwd-group&gt; is repeatable with different &quot;xml:lang&quot;.</td>
</tr>
<tr>
<td>generic</td>
<td>&lt;alternatives&gt;</td>
<td>any component which need multi-language data</td>
</tr>
</tbody>
</table>
The proposal from Japan

- It is not enough to describe Japanese only by multiple language expansion of JATS

- XSPA propose some new elements for JATS standing committee. Some of them were adopted in JATS1.1
  - era elements
  - ruby elements
era elements

Example

...<date date-type="received" calendar="Japanese"
  iso-8601-date="1947-07-01">
  <day>1</day>
  <month>7</month>
  <year>22</year>
  <era>昭和</era>
</date>
...

Example 1

畿内や西日本一帯では麦を裏作とする二毛作が普及していた。

... 畑内や西日本一帯では<ruby><rb>麦</rb><rt>むぎ</rt></ruby>を<ruby><rb>裏</rb><rt>うら</rt></ruby>と<ruby><rb>作</rb><rt>さく</rt></ruby>という二毛作が普及していた。
...
Creating Japanese XML articles in JATS
Creating XML articles in JATS

- We don’t have tools readily available for creating Japanese XML files.

- Our method
  1. Convert Microsoft Word to Microsoft Office Open XML
  2. Convert Microsoft Office Open XML to JATS XML
  3. Validate XML
(1) Converting Microsoft Word to Microsoft Office Open XML

MS Open XML tags
(2) Converting Microsoft Office Open XML to JATS XML

- Through XSLT, removing unnecessary tags.
- Perl program processing.

- We faced the difficulty of Asian languages
  - A word connects to the next word without space.
  - Computer cannot distinguish word separation.
  - Even in given name and surname separation.
(3) Validating XML

- Use the Oxygen XML editor
- Final JATS XML is obtained to be uploaded to J-STAGE
PDF is still necessary

- For paper publishing.
- For readability.
43.0, 31.7%。観察期間において、184.9%であり、生存期間中央値は、寿命差群で1,692日、正常化群で738日、寿命差群で505日であった。寿命差群で、有意に生存期間が短く、予後不良であった（Fig. 2, P＜0.0001, log-rank test）。

4. 多変量解析による予後因子の抽出

年齢（性）、既往歴（病状）・施行状況・治療、脳腫瘍マーカー・脳内動脈瘤の各因子で、Cox比例ハザードモデルを用いて解析した。多変量解析では、脳腫瘍マーカー（C-19-9上昇または未上昇）、脳腫瘤（C-19-9上昇または未上昇）、脳内動脈瘤（C-19-9上昇または未上昇）、脳内動脈瘤（C-19-9上昇または未上昇）の存在が有意な予後因子として抽出された。それら多変量解析で有意であった因子を用いて多変量解析を行った。さらに、脳腫瘍・脳腫瘤・脳内動脈瘤が有意な予後変動因子として選ばれた（Table 3）。
Creating PDF using AH Formatter

Diagram:
- Word OpenXML → JATS XML → XSLT → XSL-FO → PDF
- J-Stage
- Original programme
- AH Formatter Interface
- XSLT Processor
- AH Formatter
XSLT

The XSLT converts a JATS file into XSL-FO which expresses page model format for PDF.
For Japanese rendering
AH Formatter extension
Challenges of Applying JATS to Japanese language texts

- But in Japan, exquisite type settings are requested.
- Automatic type setting by AH formatter may not be sufficient.
Avoiding Word Breakup

- Some words, such as personal names shall not be broken-up between lines

- We use “Zero Width Joiner” code (&#x200D;)
  - e.g. 中&#x200D;西

- 「中」 and 「西」 are torn up by “&#x200D;”
  - →「中西」cannot be searched
Avoiding Line-Top Punctuations

- Punctuation marks shall not come at the top of a line ⇒ Also in English
- 「っ」or「ッ」 (to mark a geminate consonant) does not come in a head of a line ⇒ Japanese rule
- AH Formatter can handle these rules
Positioning Figures/Tables

- Figures and tables should be positioned in the SAME page that the corresponding texts appear.
- This requires customized XSLT, sometimes for each figures and tables.
- This increases costs.
Positioning Figures/Tables

Every articles need these XSLTs
Using Formatter for STM articles

- There are no major problems
- The basic style of writing STM papers do not differ greatly between western countries and Japan.
- Word separators should be inserted in XML in advance

- Possible if disregarding profitability
Current Japanese XML journal
<table>
<thead>
<tr>
<th>No.</th>
<th>Organization</th>
<th>Journal Title</th>
<th>Language</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>日本コンピュータ化学会</td>
<td>Journal of Computer Chemistry, Japan</td>
<td>混</td>
</tr>
<tr>
<td>2</td>
<td>日本育種学会</td>
<td>至種学研究</td>
<td>和</td>
</tr>
<tr>
<td>3</td>
<td>日本熱帯医学会</td>
<td>Tropical Medicine and Health</td>
<td>混</td>
</tr>
<tr>
<td>4</td>
<td>社団法人 日本産業衛生学会</td>
<td>産業衛生学雑誌</td>
<td>和</td>
</tr>
<tr>
<td>5</td>
<td>特定非営利活動法人 日本歯周病学会</td>
<td>日本歯周病学会会誌</td>
<td>和</td>
</tr>
<tr>
<td>6</td>
<td>日本緩和医療学会</td>
<td>Palliative Care Research</td>
<td>混</td>
</tr>
<tr>
<td>7</td>
<td>日本応用糖質科学会</td>
<td>Journal of Applied Glycoscience</td>
<td>混</td>
</tr>
<tr>
<td>8</td>
<td>一般社団法人 日本鉄鋼協会</td>
<td>鉄と鋼</td>
<td>和</td>
</tr>
<tr>
<td>9</td>
<td>独立行政法人 海洋研究開発機構</td>
<td>JAMSTEC Report of Research and Development (JAMSTEC-R)</td>
<td>混</td>
</tr>
<tr>
<td>10</td>
<td>日本官能評価学会</td>
<td>日本官能評価学会会誌</td>
<td>混</td>
</tr>
<tr>
<td>11</td>
<td>一般社団法人 日本消化器外科学会</td>
<td>日本消化器外科学会雜誌</td>
<td>混</td>
</tr>
<tr>
<td>12</td>
<td>公益社団法人 日本看護科学学会</td>
<td>日本看護科学会誌</td>
<td>和</td>
</tr>
<tr>
<td>13</td>
<td>日本行動医学会</td>
<td>行動医学研究</td>
<td>混</td>
</tr>
<tr>
<td>14</td>
<td>日本社会心理学会</td>
<td>社会心理学研究</td>
<td>混</td>
</tr>
<tr>
<td>15</td>
<td>日本公衆衛生看護学会</td>
<td>日本公衆衛生看護学会誌</td>
<td>和</td>
</tr>
<tr>
<td>16</td>
<td>科学技術振興機構</td>
<td>情報管理</td>
<td>和</td>
</tr>
</tbody>
</table>

混: English Japanese mixed up  
和: Pure Japanese  
April 2016 JST provided
Few Japanese XML Journals

- The total number of pure Japanese journals in J-STAGE is 568
- The number of pure Japanese XML Journal is only 7
  - After 4 years have passed
The difficulties of creating Japanese XML

- Labor and time are necessary
- Which require additional costs
- It is hard for publishers to find more merits than PDF journal
The most difficult problem
No separating spaces

- Typical in East Asian languages
- No separating spaces between words
Inserting word separators.

- we insert separators manually.
  - surname, “中”, given name, “吉昭”, are attached as "中吉昭" in an article
  - It is separated as "中＠吉昭"
- Possible alternatives are “中＠吉昭”, and “中吉＠昭”, but only human can eliminate them
- There is no algorithm to determine it correctly.
Tools necessary for us

- XML tool of the East Asia cultural sphere
- The inputting tool is necessary at first

- Not MS WORD
- We need input tools for Japanese XML

- A common tools for the East Asia cultural sphere Japan, China and Korea is needed!
Future

- What is to be done next
- Vertical writing
- Emphasis or “Knten”
- Warichu
Vertical writing

Traditionally, Japanese (and Chinese and Korean) writes from top to bottom
Vertical Writing

- Vertical Writing causes some interesting problems, orientation of Arabic numerals and Latin alphabets.

New element for direction is necessary. as `<writing-direction="vertical">`
Emphasis

- Emphasis or “Kenten”
- It is like bold faces and italics in English
- We use `<styled-content>` for these expressions
- We need a generic tag, like as `<emphasis>`
Warichu

- Vertical writing texts sometimes contain notes called “Warichu”.
- Warichu uses 2 lines within a parent line.
Suggestion

- Additional tags for
  - Vertical writing
  - Emphasis or “Kenten”
  - Warichu
Conclusion

- JATS opened a new horizon in processing Japanese-language articles
  - No major difficulties
  - UTF-8, encoding for XML, also enables to express most Japanese characters correctly
Conclusion

- Still there are remaining issues in processing non-Latin languages such as Japanese.

- Challenges
  - Word separators have to be inserted manually
  - Line break issues
  - Positioning figures and tables correctly

- Some requests to JATS
  - for applying Asian languages
Only English is not globalization

Thank you