Adaptation of Journal Article Tag Suite XML for Japanese humanities papers

Hidehiko Nakanishi, Tsuyoshi Yamamoto, Nao Hattori, Satoshi Taga

Nakanishi Printing, Kyoto, Japan

Abstract

Out of East Asian languages which do not use the Latin alphabet, Japanese is a very complicated writing system that uses “kanji,” which are ideograms, and “kana,” which are phonetic characters. Most of the Japanese papers published so far using Journal Article Tag Suite (JATS) are science, technology, and medicine fields adapting horizontal writing systems, which are structurally consistent with English papers. Most of them only replace Latin letters with Japanese characters. In this presentation, we suggested method of presenting vertically oriented Japanese humanities articles in JATS XML. For vertical description of Chinese numeric, we would like to propose the introduction of an element which specifies description direction. Alternatively, <styled-content> could be used as a hidden command when creating a document. We propose the following notation in the part of the number that can be converted: <styled-content style-type=“numeric”>六五</styled-content>. Chinese numeric 六五 is a Arabic numeric 65. With this, it is shown that 六五 of Chinese numerals can be converted to 65 in Arabic numerals. For vertical text description with JATS, we would like to suggest adding @ writing-mode as an attribute to <article>: <article writing-mode=“vertical”>. Furthermore, note and references should be differentiated for example, between a <mixed-citation> and a <note> in the future. As Kanji are ideograms, there are variations that cannot be expressed with UTF-8. If these difficult Kanji are included in the JATS text, it will be necessary to decide on their description method. For the propagation of use of JATS XML for non-Latin characters articles, the structure of the document for example, vertical description, and special presentation should be considered more widely.

Keywords

Japanese language; Journal Article Tag Suite; Humanities; Vertical writing; XML

Received: May 21, 2018
Accepted: May 30, 2018

Correspondence to Hidehiko Nakanishi
hidena@nacos.com

ORCID

Hidehiko Nakanishi
<http://orcid.org/0000-0003-3348-6708>
Tsuyoshi Yamamoto
<http://orcid.org/0000-0002-7425-8333>
Nao Hattori
<http://orcid.org/0000-0003-2937-0075>
Satoshi Taga
<http://orcid.org/0000-0003-2205-9612>

This article was presented at the Journal Article Tag Suite Conference (JATS-Con) 2018 and posted on JATS-Con 2018 proceedings website available from: <https://www.ncbi.nlm.nih.gov/books/NBK493141/>.

Introduction

Around the world, many studies in the sciences and humanities are made outside of English speaking countries, and many articles are written in languages other than English [1]. Languages and letters, moreover, are not necessarily used as one unit, and one character can often

render multiple languages. For example, the Latin letters used in English also transcribe Turkish and Vietnamese, as well as German and French. In East Asia, Kanji, which is a unique ideogram system based on vertically written Chinese characters, is widely used for language notation. The population using Kanji is 1.5 billion people in Japan, Taiwan, Korea, and China, while close to 1.8 billion people use Latin letters. Also, in Japanese, in addition to Kanji, we use the original Japanese characters comprising the Kana phonetic writing systems of Hiragana and Katakana.

Since Journal Article Tag Suite (JATS), developed from United States National Library of Medicine Document Type Definition (NLM DTD), can describe multiple languages, even Japanese and other languages using Kanji can be expressed in JATS XML [1,2]. However, JATS, which originally developed from the representation of Latin letters, is accomplished with difficulty with regard to the notation of letters other than Latin letters. Despite this difficulty, we have been using JATS XML for publication of Japanese online journals since 2012. This background and detailed method was announced at Journal Article Tag Suite Conference (JATS-Con) 2015 [3].

This paper reports on our efforts to make a Japanese journal of humanities called “the Journal of Indian and Buddhist Studies,” which incorporated vertical direction notation via JATS, in 2017. In addition, We would like to propose the future use of JATS for the adaptation of East Asian languages.

Japanese Online Journal Using JATS

When Japanese papers are published online, J-STAGE, the Japan Science and Technology Agency’s platform, is most often used. JATS was adopted as DTD from J-STAGE version 3 and has been in use since May 2012. Since then, it has become possible to publish Japanese journals online in XML. The first example was the Japanese journal “the Journal of Gastroenterological Surgery”, published in July 2012.

Some actual screen images displayed on J-STAGE are shown in Fig. 1. This screen can be displayed anywhere, in any computer environment equipped with Kanji fonts (This was the case at that time; the interface has since been changed. https://www.jstage.jst.go.jp/browse/jjgs/45/7/_contents/-char/en/).

Even papers using Kanji are displayed in the form of full text online journals like this. English is used only for the caption of the figure. Abstracts and captions are written in English to promote international understanding and is often done in Japanese papers.

Previously in Japan, only PDF was used to display Japanese language papers. As a result, it was not possible to use various functions of online journals in Japan, so the convenience of online journals was not well understood and therefore not widely utilized. It seems that this prevented the development of the Japanese online journal.

In our presentation at JATS-Con in 2015, we presented methods and pointed out issues that arose with this first Japanese online journal publication. We also made a proposal for JATS tag for application in the East Asian language environ-

Fig. 1. First online journal written in Japanese, “the Journal of Gastroenterological Surgery” (J-STAGE 2012).

ment. Several of these proposals have been adopted in JATS 1.1 and JATS 1.2 d.

For example, although the <emphasis> tag, which represents general purpose emphasis, was not adopted as a new element, it was solved by introducing @style-detail in the existing element called <styled-content>. With this and existing attribute @style-type, detailed emphasis description became possible, and it became possible to emphasize in Japanese such as the Kanten (Fig. 2). NISO JATS Standing Committee recommended changes between NISO/JATS 1.1 and JATS 1.2d1 (Committee draft) in response to comments on NISO Z39.96-2015 (JATS V1.1) through April 10, 2017.

In addition, we appreciate the JATS Committee's concession that "Many languages (Japanese, Korean, Thai, Chinese, Arabic, Hungarian, and Armenian, to name but a few) use stress marks and similar typographic conventions (such as dots or sesamis) in the same way that English (as one example) uses <bold> or <italic> emphasis." in the concluding comment of this denial of the <emphasis> tag adoption.

Although it was not our proposal, East Asian documents including those written in Japanese became easier to describe, such as <ruby> element and <era> element (Fig. 3). We appreciate the JATS Committee's understanding of the circumstances of non-Latin languages.

あいうえお
かきくけこ

Fig. 2. Example of Kanten emphasis.

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>
...
```

Fig. 3. Elements <era>. Eastern Asian calendars can be displayed with historical eras.

The Present Status and Breakthrough of Japanese Vertical Writing Paper

Since then, the number of Japanese papers created in JATS XML has been gradually increasing, and it can also be applied in the other East Asian countries such as Korea and China. However, it has not yet become mainstream because of the difficulty of Asian language XML expression. In Japan, for a long time, it has been taken for granted that online journals are to publish via PDF or images. This is remarkably behind the trend of the rest of the world.

XML expression, especially in the field of Japanese humanities, is considered impossible, and even attempts to publish have been abandoned due to this limitation. Among the J-STAGE, of 2103 journals, less than 100 journals can be said to be humanities journals. This number is far fewer than the 716 journals of the total J-STAGE expressed in Japanese. Most of these, with a few exceptions, are PDF publications.

In humanities journals, the structure of the articles are not fixed, in contrast to the document structure in STM journals; moreover, they are, fundamentally different from the STM in the cited documents and notes, as well as the vertical writing. These facts make it difficult to establish as an online journal. It is especially problematic that many of the Japanese papers use vertical writing. The orientation of East Asian writing was originally vertical, and horizontal writing was only adopted in recent

Fig. 4. "The Journal of Indian and Buddhist Studies" in J-STAGE.

years to create consistency with Western documents. STM papers are currently almost always horizontally written, while those in the humanities almost always use vertical writing. However, notwithstanding such obstacles, image PDF method is no longer accepted and, at minimum, bibliographic information such as cited documents must be provided in XML. Let us now turn to the history and technical problems of the publication of the humanities journal, “the Journal of Indian and Buddhist Studies” as an online journal (Fig. 4). URL is as follows: <https://www.jstage.jst.go.jp/browse/ibk1952/-char/ja/>.

Characteristics of Japanese Humanities Papers and of “the Journal of Indian and Buddhist Studies”

“The Journal of Indian and Buddhist Studies”, published as online journal via J-STAGE, is an academic journal that studies the fields of Indian philosophy and Buddhist studies, and an institutional journal of “the Japanese Association of Indian and Buddhist Studies” founded in 1951. Characteristically, along with English written articles, vertically oriented Japanese articles about Indian philosophy and Buddhism studies are listed (Fig. 5). Paper journals with vertical writing are not unusual in the humanities, but among J-STAGE journals this number is also extremely small, only 15, and even if we add “the Journal of Indian and Buddhist Studies” it counts just 16 journals. Many of these vertical writing online journals do not

even include bibliographic descriptions, and only two journals offer citations in XML.

In addition, quotations used in the Buddhist scriptures are very numerous in “Indian Science Studies Buddhist Studies”, and various Kanji characters are used. Finally, non-Latin, non-Kanji characters such as Bon-ji (Siddham script), which are ancient Indian characters, are also used (Fig. 6).

As is common in humanities journals, notes are frequently used, but they are used for both citations and supplements of the contents of the paper, and the list of cited documents is not an independent item. The composition of these articles tends to be seen most often in Japanese humanities literature journals, in contrast to the JATS online journal developed based on similarly structured STM journals in Europe and the US style.

“The Journal of Indian and Buddhist Studies” is devoted exclusively to the humanities and it is therefore difficult to apply JATS XML developed for STM use in English. In the first place, it may seem that this was not the original purpose of JATS. A schema called TEI is said to have been developed for this purpose in the humanities system [4]. However, J-STAGE is the only platform that has a wide range of online journals in Japan, and there J-STAGE has adopted JATS. Although J-STAGE itself originally started as an online journal platform for STM, it absorbed the NII-ELS of the National Institute of Informatics, which had handled a broader field, with the exception of STM, in 2017. The exact reason for the absorption is outside the scope of this paper; however, because of this absorption, documents originally published in NII-ELS have been flowing into J-STAGE.

It may be possible for languages such as English, widely used in many countries, but in many small countries including Japan, it is difficult to create separate schema systems for every non-STM academic field for each language. It is impossible to provide many online journal platforms for every language and every field, mainly for budgetary reason. JATS has already been widely adopted and used, so it is thus likely that from now on, making online journals with schemas other than JATS will be difficult regardless of the field. JATS has already left NLM and is also getting out of English. It is more realistic to expand JATS than to look for another schema, with, of course, respect to its

Fig. 5. Traditional expression of “The Journal of Indian and Buddhist Studies.” PDF files are available on J-STAGE.

Fig. 6. Bon-ji or Siddham script which was used in original Buddhist Scripture. Classical Indian language was expressed in these written characters.

limitations and appropriate application methods.

Practice of Bibliography XML

There are two methods of loading to J-STAGE by XML. One is a technique to create the whole paper in XML. Although this is indispensable in HTML publication, there are not many examples, even in STM fields in Japan, because XML tools for Japanese are not fully developed and its cost is high. The other is the bibliography XML, a method we have adopted, in which the main text is not written in XML, but only the bibliography. The citation reference link, which is the one of the most advantageous merits of the online journal, can be used. For these reasons, at this time we have adopted bibliography XML.

In “the Journal of Indian and Buddhist Studies”, abstracts are described in English. There is also a horizontally written thesis. In that case, tagging with XML is done first, and automatic formatting is done using dedicated software, but in the case of vertical writing this technique cannot be used, as it ba-

（参考文献）
山辺習学・赤沼智善『教行信証講義』教行の巻、法蔵館、一九五一
星野元豊『講解教行信証』教行の巻、法蔵館、一九七七
岡亮二『教行信証口述五〇講』教・行の巻、教育新潮社、一九九三
紅樺英顕『「教行信証」における「往生礼讃」引用文について』『印度学仏教学研究』第四七巻第二号、一九九九年
武田晋『本典引用の「往生礼讃」と「集諸経礼懺儀」』『龍谷大学論集』第四七四・四七五号、二〇一〇
（キーワード）『教行信証』、『往生礼讃』、『行巻』、大行としての念仏

Fig. 7. Original vertical writing citation reference.

山辺習学・赤沼智善『教行信証講義』教行の巻、法蔵館、一九五一

星野元豊『講解教行信証』教行の巻、法蔵館、一九七七

岡亮二『教行信証口述五〇講』教・行の巻、教育新潮社、一九九三

紅樺英顕『「教行信証」における「往生礼讃」引用文について』『印度学仏教学研究』第四七巻第二号、一九九九年

武田晋『本典引用の「往生礼讃」と「集諸経礼懺儀」』『龍谷大学論集』第四七四・四七五号、二〇一〇

Fig. 8. An example of converting vertical cited documents in Fig. 7 to horizontal writing.

sically writes out an article formatted by Adobe InDesign in XML and converts it using XSLT to JATS XML.

Bibliography of vertical written articles

Although JATS adopted in J-STAGE is capable of multilingual correspondence, the premise is that left to right horizontal writing is essential. The direction specification of description is not currently supported. Also, J-STAGE itself does not support vertical HTML display. Of course, since the tag structure also supports only left to right horizontal writing, in the case of vertically written articles, we face the problem of horizontalization of vertical writing. The horizontalization of the citation reference description is particularly important for XML as a journal cannot be retrieved unless it is written in horizontal contents with Arabic numerals (Figs. 7, 8). It will be retrieved if the original paper archive is written in horizontally written Chinese numerals, which was very rare.

In this case, it is unnatural that we simply rewrite vertical written text to horizontal written text because there is a fundamental difference in notation in vertical writing and horizontal writing. Arabic numerals, especially become a problem, as in Japan, Arabic numerals are used for horizontal writing, and Chinese numerals are used for vertical writing. In newspapers et cetera, Arabic numerals are sometimes used in vertical writing, but only when there are few digits. This is a combination of vertical and horizontal writing, and Japanese InDesign supports this style (Fig. 9). However, at least in “the Journal of Indian and Buddhist Studies”, this style of notation is not used. Normally, in Japanese typography, when changing vertical writing to horizontal writing, conversion of Chinese numerals to Arabic numerals by substitution as a group is performed.

However, we cannot replace all Chinese numbers with Arabic numerals here; this is because there are many expressions

①	②	③	④	平成三十年は特別な年
⑤	平成30年は特別な年			
	平成30年は特別な年			
	平成30年は特別な年			
	平成30年は特別な年			
		①	correct	
		②	wrong	
		③	acceptable	
		④	acceptable	
		⑤	correct	

Fig. 9. Acceptance of numerical notation of vertical writing and horizontal writing

that must remain Chinese digits even in horizontal writing. For example, numbers are also commonly used in people's names, where Chinese numerals cannot be replaced with Arabic numerals characters. The popular Japanese name “一郎 (Ichiro),” which means “first-born boy,” provides a good example; the name “一 (ichi)” means “1” or “first,” but is never written as “1郎.”

Chinese numerals are also frequent in Buddhist scriptures. In the case of Buddhist terminology such as “念仏三昧,” “三” is a Chinese numeral of “3,” but this cannot simply be alpha-numerically represented as “念仏3昧.” This would change the fundamental terminology of Buddhism entirely or render it nonsensical. In English, this would be something like expressing “Trinity” as “3 nity” in writing about Christianity.

For this process, we used a script provided by Mr. Kiyonori Nagasaki of this society. Among them, the logic to not convert personal names and titles is written, but eventually we had to visually confirm them one by one to be sure, a painstaking process.

Problem with notes in bibliography XML

The largest issue with publishing journals online was the problem of “notes” widely used in humanities journals. Humanities journals are packed with various information such as supplements and acknowledgments of the text, not limited to references cited in the “Notes.” To make cited document links effective they need to be tagged, but it is impossible to tag them in this note description. Under present circumstances it is difficult to technically overcome this, and it is necessary to change the description method of the paper itself. In these examples, there are many other journals separating supplemental notes from cited references, and this method has been adopted from volume 65 in “the Journal of Indian and Buddhist Studies” as it was effective.

However, in the case of religious and historical papers, there are citations for sutras and historical materials as primary literature. This is a global description in religious studies, philosophy, and historical papers worldwide, for example in the case of Christianity the text position is done by section such as when citing the Bible, for example, “MAR.9.47, ACT.8.37.”

In “the Journal of Indian and Buddhist Studies,” it is widely practiced to cite sources with the page number of the “Taishō Revised Tripiṭaka,” which is the total collection of Buddhist texts translated into Chinese, when quoting a scripture [5]. However, there is no clear provision in the JATS citation stipulation method for this Buddhist scripture, and since there are too many documents to cite, Buddhist texts are currently not described as cited references. Instead, these references remain in the notes. Therefore, only some secondary documents published as articles are tagged with bibliography XML.

Proposal for Humanities Journal Full Text XML

Elements for vertical notation

Currently, there are no elements that specify the direction of description. This is thought to be because the vertical writing is an expression problem unrelated to its structure. Certainly, there is horizontal writing notation in Japanese, and even if vertical writing is converted to horizontal writing, the meaning itself, although not perfect, will not change very much. However, a final problem lurks in the margins of imperfect conversion, in terms of Japanese literature, Japanese history, et cetera, fields which, in particular, have source materials that can only be accurately represented in a vertical writing orientation by their very nature.

Of course, it is not right at present to expect this role of JATS, which is intended for STM. Also, if we need accurate rendering in order to stay faithful to source materials, we can use PDF. However, as we have seen this time, it is necessary to accurately express cited references and the like even if vertically written text is expressed horizontally. Otherwise, the advantage of the online journal cannot be used.

In this regard, we would like to propose the introduction of an element which specifies description direction. Alternatively, `<styled-content>` could be used as a hidden command when creating a document. We propose the following notation in the part of the number that can be converted:

`<styled-content style-type=“numeric”>六五</styled-content>`

Chinese numeric 六五 is a Arabic numeric 65. With this, it is shown that 六五 of Chinese numerals can be converted to 65 in Arabic numerals.

We are conceiving of a `<mixed-citation>` utilization method, in which we introduce an element that carefully specifies the volume or issue, and at the time of conversion, if the Chinese numerals are used in the `<volume>` and `<issue>` tags, they will be converted to Arabic numerals.

However, in order to eventually satisfy both the convenience of online journals and vertical rendering, vertical writing should be supported on the platform side, as in J-STAGE. Currently CSS 3 supports vertical writing, proving this is not impossible, and vertical writing sites have already begun to appear. Of course, it will only be in the distant future that the online journal platform like J-STAGE will fully implement its function. Technically there is probably no problem. But before the support for vertical writing, there are many things we must do; for example, J-STAGE does not support even JATS 1.1 yet.

Before J-STAGE realizes it, it will become important to be able to specify the original description direction of the document. Currently, we cannot write vertically on the platform, but if we do not designate it anywhere in the XML text, there

will be a possibility of causing problems in the future. If we do not leave room for specifying whether the original text was premised on vertical writing or not, the XML document that we create will be inaccurately rendered and therefore insufficient for some future research purposes.

In CSS 3, there is a property that specifies the character direction. For JATS, We would like to suggest adding @writing-mode as an attribute to <article>.

<article writing-mode="vertical">

This article document may be expressed horizontally on a platform that cannot be written vertically. However, in the future, when vertical writing online journals become possible, they will be expressed in vertical writing all at once. If current articles are written in this style XML, the information that the article was originally intended to be written in a vertical orientation will be preserved.

Note and references

Problems of notes and cited documents also need to be addressed and solved. To use the cited documents described in the previous traditional note separately is to change their traditional description method. It is not preferable to change the tradition just because it has entered the online world. We need some solution. It is easy to imagine that Artificial Intelligence will learn to automatically tell the difference between a <mixed-citation> and a <note> and extract one from the other in the near future, even if cited documents are written in the notes.

Additionally, it is necessary to think about linking primary documents such as sutras from the notes or from the text to the scripture database. The text database of SAT "Taishō Tripiṭaka (大正新脩大藏經)" has already been created and reading the paper while linking with existing databases like this will undoubtedly be extremely useful for future research progress.

Difficult and rare characters

There are no examples of describing difficult Kanji or classical Bonji or Siddham script (as in the Hebrew text of the Old Testament of the Bible) inside of the cited references, but when the full paper XML becomes possible, that description problem will surely arise. In the present situation, character is expressed in UTF-8 and it is not thought to be difficult, but examples of difficult Kanji which cannot be described in UTF-8 will also certainly occur in the future. As Kanji are ideograms, there are historical variations that cannot be expressed with UTF-8. In that case, it will be necessary to decide on their description method.

Conclusion

It can be said that JATS currently holds the position of the world's common framework of academic information on behalf of English. Despite the differences between languages, the contents themselves are easier to understand due to the development of translation software and other tools. From now on, the most important consideration is not the language of scholarship itself, but the structure of the document containing it. If structured documents are common to the world, it will be easier to understand and collaboratively utilize the contents of scholarly documents.

JATS has already been used as a de facto standard all over the world in many research fields. Would it not be possible to expand the vast academic world by allowing papers of all languages and all fields to be expressed within the common JATS infrastructure?

The barriers of English and Latin letters have been removed. We wonder if the barriers of STM field can be lowered just a bit at this time. Our only choice, at the moment, is JATS.

Conflict of Interest

Hidehiko Nakanishi has been a President of Nakanishi Printing Company Limited, Kyoto, Japan since 2016 and Tsuyoshi Yamamoto, Nao Hattori, Satoshi Taga have been staffs of the company. This article is for research purpose not for advertisement of co-authors' companies.

References

1. Nakanishi H, Yamamoto T, Hattori N, Taga S. Japanese journal of humanities published in online journal in XML. Paper presented at: 14th Information Professional Symposium; 2017. https://doi.org/10.11514/infopro.2017.0_107
2. Tokizane S. Implementing XML for Japanese-language scholarly articles. Paper presented at: Journal Article Tag Suite Conference (JATS-Con) 2012; 2012 Oct 16-17; Bethesda, MD, USA. <https://www.ncbi.nlm.nih.gov/books/NBK100380/>
3. Nakanishi H, Naganawa T, Tokizane S, Yamamoto T. Creating JATS XML from Japanese Language Articles and Automatic Typesetting using XSLT. Paper presented at: Journal Article Tag Suite Conference (JATS-Con) 2015; 2015 Apr 21-22; Bethesda, MD, USA. <https://www.ncbi.nlm.nih.gov/books/NBK279832/>
4. Nagasaki K, Tomabeche T, Muller C, Shimoda M. Digital humanities in cultural areas using texts that lack word spacing. Paper presented at: Digital Humanities 2016; 2016 Jul 11-16; Krakow, Poland. <http://dh2016.adho.org/>

abstracts/416

5. Nagasaki K, Tomabeche T, Shimoda M. Towards a digital re-

search environment for Buddhist studies. Lit Linguist Comput 2013;28:296-300. <https://doi.org/10.1093/lc/fqs076>