Reducing costs and expanding XML submissions with PDF to JATS conversion

Katoh Keishi・Kobayashi Tokushige・Kitazawa Mitsuru

Digital Communications Co. Ltd.
Antenna House, Inc.
Japan Science and Technology Agency (JST)

The paper presents a brief overview of the challenges facing institutions with the XML-ization of academic journals and the steps being taken in Japan to meet those challenges with the new J-STAGE implementation and a solution for automatically analyzing and converting PDF into XML for JATS metadata and bibliographic information. J-STAGE has fully adopted the metadata and bibliographic JATS format. The automated solution is currently achieving more than a 90% accuracy rate and future plans are to expand it to be able to produce full-text XML from PDF.

Key Words
XML, PDF, E-Journal, Bibliographic information, JATS, Data extraction

Overview

JST has developed the bibliographic metadata creation tool for academic societies using J-STAGE and JST has already presented a paper about an overview of this tool at a conference in Japan, so we introduce this tool from technical aspect in this paper.

The XML-ization and full-text submission of academic journals has been slow to spread due to the high cost of authoring content in XML. Even now in Japan this is the situation within many institutions in spite of the fact that XML-ization of papers was tackled early on.

Many people understand the merits of expressing papers in XML format, but preventing this from happening are:

- The fact that it is difficult to actually author in the XML format. For that reason academic journals quite often are not authored in XML.
- The many various tools used for writing the papers, most of which don't produce XML.
- Some printing companies who only produce print and don't have capabilities to work with XML.
- No standard layout of documents by the various institutions.
- Writing, reviewing and printing/production considered as one process and then XML creation as another separate process.
- The higher perceived costs of authoring in XML and the higher skills required for production people when using XML.

J-STAGE (Japan Science and Technology Information Aggregator, Electronic) solution

J-STAGE, one of the major e-journal publishing platforms of Japan, has started a new J-STAGE service in May of this year[1)2)]. Japan Science and Technology Agency (JST) provides this service. The objective of J-STAGE is to strive for the acceleration and internationalization of Japanese science and technology information transmission and circulation by building on the Internet a uniform flow - from the submission to release of science and technology information.

A major enhancement of J-STAGE is that it has fully adopted JATS and now requires every journal to include the metadata and bibliographic information in JATS format. It is also possible to include the full-text body in JATS format, but the gradual transition without the burden of full-text has been selected.

The BibTeX-like format was used in the previous versions of J-STAGE. There were restrictions on the content of metadata expressions which came from the format, character set depending on the encoding (Shift JIS), etc. These restrictions were drastically removed with the adoption of JATS.

In order to reduce the burden on the institutions who provide the information or printing production companies who make a service to create the information, J-STAGE now offers a web

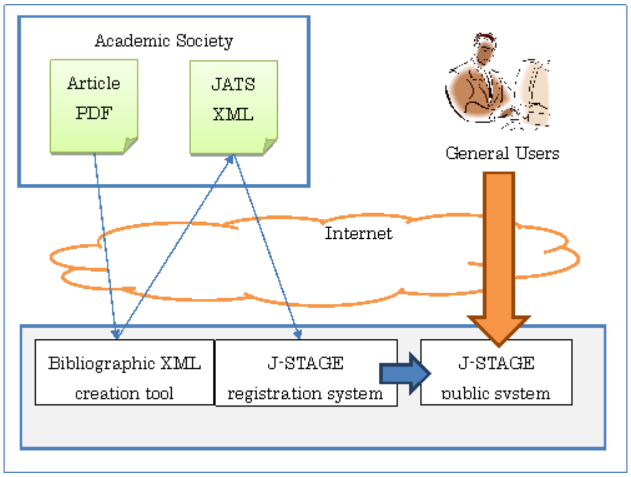


FIG 1: J-STAGE system configuration diagram

service to assist creation of JATS metadata and bibliographic information by JST.

Positioning of bibliographic XML creation tool in J-STAGE

Currently J-STAGE still supports registration using the existing BIB format in addition to the JATS format, but it is planned to stop support for registration using the BIB format. This has made it urgent to support JATS by the academic societies and the printing companies for registration of papers in the future. The bibliographic XML creation tool was developed so that JATS bibliographic information metadata, required for registration could be easily created from PDF data for print publications as a preemptive measure to address the needs of the smaller academic societies and printing companies that would have trouble on their own dealing with the costs and technical issues. The tool has been available for academic societies since April, 2012.

Figure 1 shows the positioning of the bibliographic XML creation tool within the J-STAGE service.

Overview of bibliographic XML creation tool

Candidate for conversion

The target input data for this tool is a PDF of the journal article. This PDF is analyzed and the bibliographic information metadata is extracted as XML in JATS format. The bibliographic information contains a title, authors, affiliations, an abstract, and references of an article. In addition, it contains the accompanying information on the journal, including the journal title, ISSN, the date of issue, etc. which is associated with the tool user's account.

The conversion is targeted only for the bibliographic information; the conversion of a full text to JATS is under consideration.

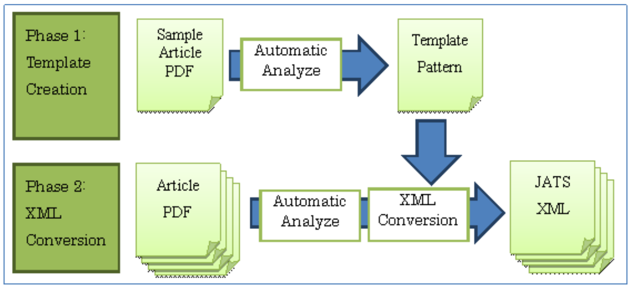


FIG 2: Workflow using a bibliographic XML creation tool, and its functional components

Why PDF?

There are two reasons why a PDF file is used in the bibliographic XML creation tool for extracting bibliographic information:

First, the wide variety of formats used to both prepare papers for submission and then for publishing. Ideally we would use the author's original input for the most accurate conversion results, but authors tend to use a variety of word processors in addition to ASCII editors, XML authoring tools, and even TeX. The printing companies also use a variety of tools ranging from desktop publishing, TeX, semi-automated publishing systems like 3B2, and even word processing. Extracting the bibliographic information from each of these tools depends on the associated underlying format. It is unrealistic to try and support all the various formats of all the different tools and systems.

The second reason is that content expressed in XML structure can also vary significantly depending on the context in which it is used. For example, if it is the formatting from XML or the manuscript data based on TeX, the structure of the content will be included in the format and it is thought that mapping to the bibliographic information structure in JATS format should be comparatively easy. However, if the data is only arranged onto pages from the designing intent, the layout design itself expresses the meaning and it would be extremely difficult under the present circumstances to extract the required elements.

For these two reasons and based on the target group of users and the intended purpose of this tool, the method of the bibliographic information extraction from visual layout of the PDF was selected. This also enables the tool to deal with articles where the original workflow was not designed for structured data going forward. This approach bypasses the many unique processes used to gather the data and lay it out for the PDF before the public presentation.

Tool workflow

Figure 2 shows the two phases of the tool's workflow, (1) template creation and (2) registration of PDF and conversion to XML.

Template creation

In Phase 1 a template is created from a sample PDF of the article. The template identifies layout patterns such as characteristics of text blocks and metadata blocks. The charactaristics include font family, font size, list style, and character string of headings in the PDF that are then used to extract bibliographic information metadata from the actual PDFs of the articles.

Article registration / structure conversion

In Phase 2 the actual conversion of the PDF articles to JATS XML takes place. An individual article PDF is registered into this tool. The bibliographic information is then extracted based on the patterns identified by the article template from each registered PDF and JATS XML is generated. The tool also supports registering and converting multiple PDFs of the same layout pattern as a package.

Editing / outputting XML

The tool also has an XML editing feature. With the XML editing feature, the automatically transformed XML contents can be checked and edited. Incorrectly recognized content can be corrected if needed and if there are elements that the tool has difficulties extracting automatically, it's possible to input the elements manually in the edit window. After performing such editing work, the completed XML data is downloaded. Moreover, before downloading the XML data, the created contents of the JATS XML are automatically checked according to the XML data format guideline of J-STAGE.

Each of the tool’s features is accessible through a web browser and is discussed below in more detail and how it relates to the overall workflow creation of an article template and conversion to JATS from PDF.

Conversion from PDF to JATS

Creation of an article template

Margin settings

A user of the tool first prepares one sample article PDF which becomes the master when creating the article template pattern. When the number of bibliographic items that are extracted varies depending on actual content of articles, it is desirable to use the typical article PDF in which all the items are included. Once a user uploads the sample article PDF to the tool, margins will be set up first. Then, the body region and the header / footer regions are separated based on the margin settings.

Block recognition of PDF layout

The content of the PDF file is then analyzed by PDF Analyzer which is a standard library developed by Antenna House. In this phase headings, paragraphs, columns and other text blocks are recognized per page (Figure 3).

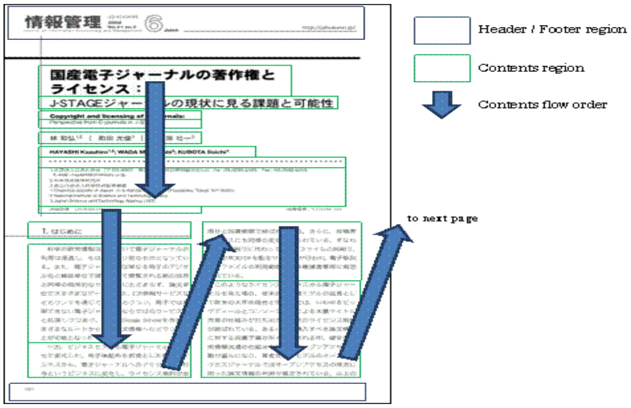


FIG 3: Block recognition with PDF Analyzer

First, lines are constructed based on the arrangement and position of the characters on the page of PDF. At this time, superscripts are constituted, lines in the columns and column gap are judged by the white space between blocks. Based on the line endings and the white space at the end of lines, the indenting information, the text block that constitutes the paragraph will be created.

Simultaneously with the paragraph recognition, PDF data will be cleaned, the hyphens removed and the insertion of white spaces is performed. At this stage, the information included in the PDF is arranged and it is divided into the text blocks. Finally, the contents are rearranged in order of the context of the text block, the text flow of the whole contents containing columns will be decided.

Selection of a recognition block

A block with characteristic information like font size, etc. is selected from the recognized text blocks. Then JATS items, which correspond to JATS XML elements, are assigned to the selected text block. This work can be done by placing the page image of PDF and setting item form side-by-side. (Figure 4). In the block selection the conditions to select the range that includes the candidate bibliographic item from the automatically recognized text block are specified. The font size, font family or text pattern of a regular expression, etc. are also specified as identifying information for conditions to select the block. For example, font family or font size of a title, fixed titles, such as 'Abstract' can be used as the identifying information of the block selection.

By dragging and marking the range of a selection on the page image of the PDF on the left side, the setting information of a font family, font size, and text settings can be acquired based on PDF data. For this reason, it is not necessary to refer to the formatting instruction information or to investigate the PDF format information separately using a PDF viewer, the operation can be done very easily within the tool.

If the text block that matches conditions is found, the matched paragraph will be one processing block, or the range until the next text block that matches the different block selecting condition is found will be extracted as one processing block.

Assignment of the JATS items

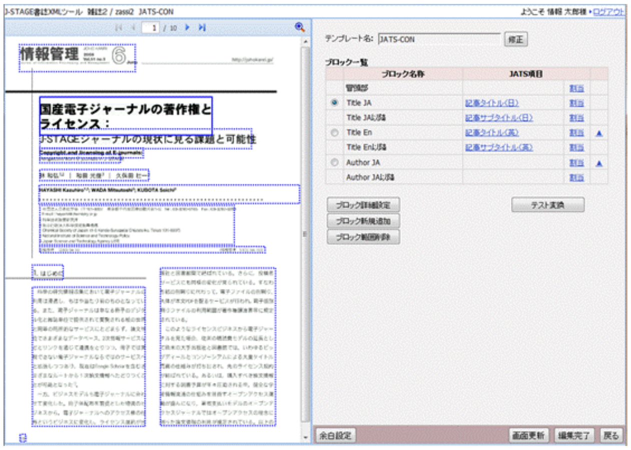
Next, assign the JATS items to be included to a processing block. Items include an article title, a subtitle, an author's name, affiliation, a cited reference, etc. If a text block is divided so that it may become one JATS item against one processing block, the recognition rate can be raised comparatively easily. When suitable conditions to divide a text block are not found, it is also possible to assign multiple JATS items to one processing block.

A special setup can be made for every JATS item according to the contents of the element. Figure 5 shows an example of setting an author's name and keyword. For author's name, the delimiter in case multiple authors are in a line, the pattern of a reference number character to an affiliation organization and the delimiter of a reference number can be specified. As for the pattern of a delimiter or a reference number character string, the specification by the character string and a regular expression is prepared in addition to patterns used with high frequency and preset conditions. In addition, you can set up whether to separate automatically between an author's family name and first name with a space. The pattern of a reference number character also exists as well in the JATS item of the affiliation organization which exists independently. When a reference number character of both an author's name and an affiliation organization matches, the reference ID is generated within the XML. Similarly, when multiple keywords are located in a line, it is divided into each keyword by specifying a delimiter.

JATS element Japanese/English language support

For the JATS element in which the multilingual notation is available[3)4)], the item to allocate Japanese and English is prepared. Thereby, it's possible to process a Japanese article, an English article, and an article that has Japanese text in the body part but has a title and an abstract in English statement. In this case, processing the JATS item in Japanese and English can be satisfactorily performed, even if these are assigned to the processing block placed at a distance.

As for JATS items, such as an article title, the contents are simply extracted for each of the <article-title> element and the <trans-title> element. As for the items where multiple contents are enumerated, such as an author's name or an affiliation organization, and when Japanese and English are written together in the item, the pairing is performed based on the appearance order or the reference number information.



J-STAGE書誌XMLツール 雑誌2 / zassi2 JATS-CON
ようこそ 情報 太郎様 • ログアウト
1 / 10
情報管理
国産電子ジャーナルの著作権とライセンス：
J-STAGEジャーナルの現状に見る課題と可能性
1. はじめに
余白設定
画面更新
編集完了
戻る

テンプレート名: JATS-CON 修正

ブロック一覧

	ブロック名称	JATS項目		
	冒頭部		割当	
●	Title JA	記事タイトル(日)	割当	
	Title JA以降	記事サブタイトル(日)	割当	
○	Title En	記事タイトル(英)	割当	▲
	Title En以降	記事サブタイトル(英)	割当	
○	Author JA		割当	▲
	Author JA以降		割当	

ブロック詳細設定
ブロック新規追加
ブロック範囲削除
テスト変換

FIG 4: Example of recognition of a block and a block selection setting list

Although JATS items for each Japanese and English keywords are prepared, one to one correspondence of each keyword for Japanese and English is not required according to the XML guideline of J-STAGE, then the pairing using the <compound-kwd> element are not performed differently from an author's name, etc.

Confirming the template settings

In the middle of a template setting, a structure conversion result can be checked on a trial basis (Figure 6).

The confirmation screen has a web preview display and structure display. In the web preview the screen imitates how recognized JATS data will be displayed on the J-STAGE public system website. In the structure display the quasi-structure with the element names outputted, the division of a keyword or a cited reference, and the pairing belonging to an author can be confirmed.

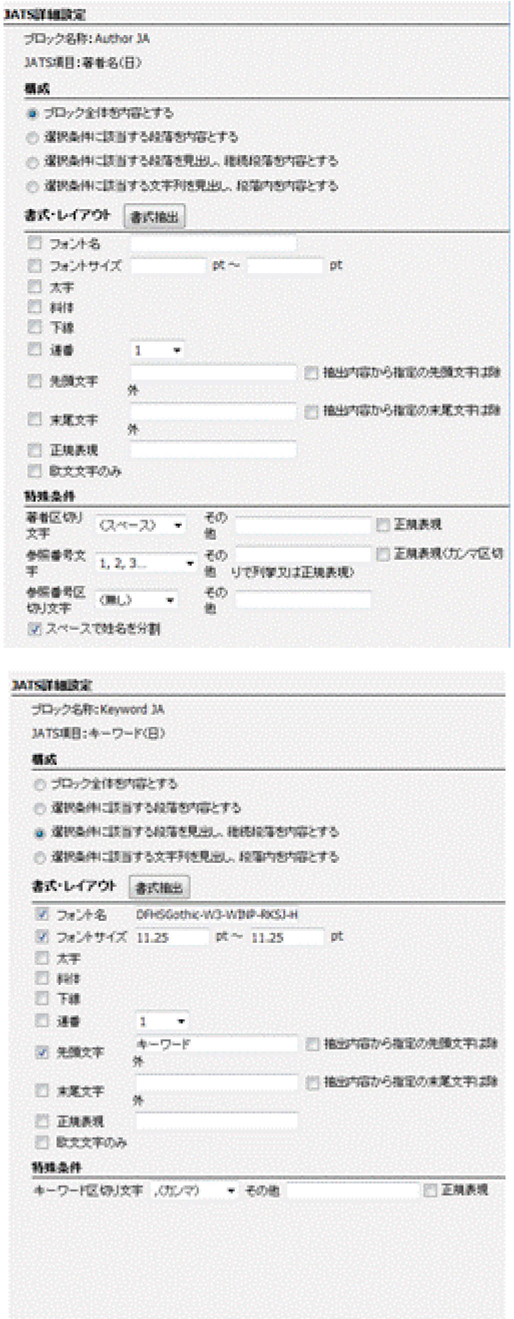
Converting article PDF to XML and editing XML

When the creation of a template is finished, the preparation for extracting the bibliographic information of JATS from the individual article PDF is completed.

Converting article PDF to XML

The conversion work from the article PDF to JATS bibliographic information XML is very simple. The template to be used for conversion is specified and the article PDF used as the candidate for conversion is uploaded to a server (Figure 7). Both single PDFs and/or a ZIP file containing multiple PDFs can be specified.

The conversion is performed by a background job based on the order received. As soon as conversion finishes, the user can edit the XML file with the extracted bibliographic information.



JATS詳細設定

ブロック名称：Author JA

JATS項目：著者名(日)

構成

● ブロック全体を内容とする
○ 選択条件に該当する段落を内容とする
○ 選択条件に該当する段落を見出し、継続段落を内容とする
○ 選択条件に該当する文字列を見出し、段落内を内容とする

書式・レイアウト [書式抽出]

□ フォント名
□ フォントサイズ pt ~ pt
□ 太字
□ 斜体
□ 下線
□ 連番 1
□ 先頭文字 □ 抽出内容から指定の先頭文字は除外
□ 末尾文字 □ 抽出内容から指定の末尾文字は除外
□ 正規表現
□ 欧文文字のみ

特殊条件

著者区切り文字 (スペース) その他 □ 正規表現
参照番号文字 1, 2, 3... その他 □ 正規表現(カンマ区切りで列挙又は正規表現)
参照番号区切り文字 (無し) その他
☑ スペースで姓名を分割

JATS詳細設定

ブロック名称：Keyword JA

JATS項目：キーワード(日)

構成

○ ブロック全体を内容とする
○ 選択条件に該当する段落を内容とする
● 選択条件に該当する段落を見出し、継続段落を内容とする
○ 選択条件に該当する文字列を見出し、段落内を内容とする

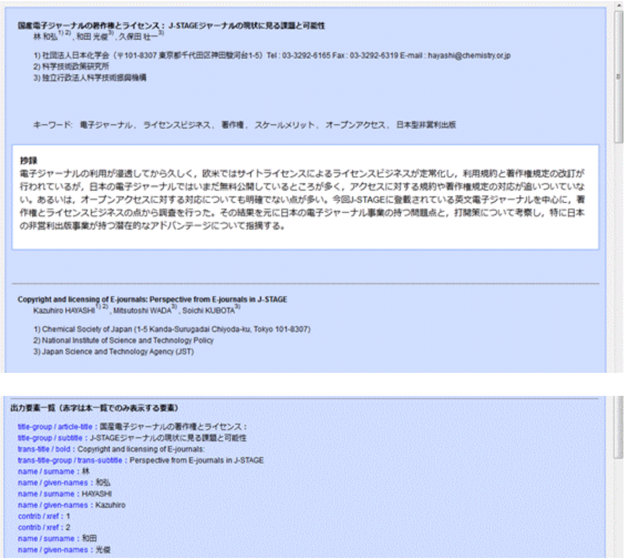
書式・レイアウト [書式抽出]

☑ フォント名 DFHSGothic-W3-WINP-RKSJ-H
☑ フォントサイズ 11.25 pt ~ 11.25 pt
□ 太字
□ 斜体
□ 下線
□ 連番 1
☑ 先頭文字 キーワード □ 抽出内容から指定の先頭文字は除外
□ 末尾文字 □ 抽出内容から指定の末尾文字は除外
□ 正規表現
□ 欧文文字のみ

特殊条件

キーワード区切り文字 ,(カンマ) その他 □ 正規表現

FIG 5: Example of JATS item advanced settings



国産電子ジャーナルの著作権とライセンス：J-STAGEジャーナルの現状に見る課題と可能性
林 和弘[1)2)], 和田 光俊[3)], 久保田 壮一[3)]

1) 社団法人日本化学会（〒101-8307 東京都千代田区神田駿河台1-5）Tel : 03-3292-6165 Fax : 03-3292-6319 E-mail : hayashi@chemistry.or.jp
2) 科学技術政策研究所
3) 独立行政法人科学技術振興機構

キーワード：電子ジャーナル，ライセンスビジネス，著作権，スケールメリット，オープンアクセス，日本型非営利出版

抄録
電子ジャーナルの利用が浸透してから久しく，欧米ではサイトライセンスによるライセンスビジネスが定常化し，利用規約と著作権規定の改訂が行われているが，日本の電子ジャーナルではいまだ無料公開しているところが多く，アクセスに対する規約や著作権規定の対応が追いついていない。あるいは，オープンアクセスに対する対応についても明確でない点が多い。今回J-STAGEに登載されている英文電子ジャーナルを中心に，著作権とライセンスビジネスの点から調査を行った。その結果を元に日本の電子ジャーナル事業の持つ問題点と，打開策について考察し，特に日本の非営利出版事業が持つ潜在的なアドバンテージについて指摘する。

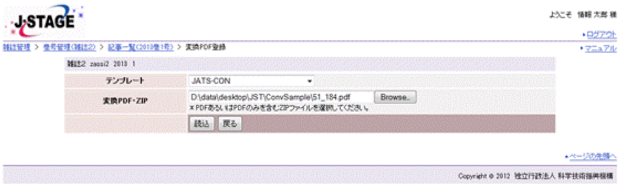
Copyright and licensing of E-journals: Perspective from E-journals in J-STAGE
Kazuhiro HAYASHI[1)2)], Mitsutoshi WADA[3)], Soichi KUBOTA[3)]

1) Chemical Society of Japan (1-5 Kanda-Surugadai Chiyoda-ku, Tokyo 101-8307)
2) National Institute of Science and Technology Policy
3) Japan Science and Technology Agency (JST)

出力要素一覧（赤字は本一覧でのみ表示する要素）

title-group / article-title：国産電子ジャーナルの著作権とライセンス：
title-group / subtitle：J-STAGEジャーナルの現状に見る課題と可能性
trans-title / bold：Copyright and licensing of E-journals:
trans-title-group / trans-subtitle：Perspective from E-journals in J-STAGE
name / surname：林
name / given-names：和弘
name / surname：HAYASHI
name / given-names：Kazuhiro
contrib / xref：1
contrib / xref：2
name / surname：和田
name / given-names：光俊

FIG 6: Displaying the test conversion result under the template setup (upper: Web preview, lower: structure display)



ようこそ 情報 太郎 様

雑誌管理 > 巻号管理(雑誌2) > 記事一覧(2013巻1号) > 変換PDF登録

雑誌2 zassi2 2013 1

テンプレート	JATS-CON
変換PDF・ZIP	D:\data\desktop\JST\ConvSample\51_184.pdf [Browse..] ※PDFあるいはPDFのみを含むZIPファイルを選択してください。
	[読込] [戻る]

Copyright © 2012 独立行政法人 科学技術振興機構

FIG 7: Operation for converting article PDF to XML

Using blank template for XML editing

As mentioned above, the basic use of this tool is to create the template for the pages with the layout of typical articles and automatically recognize the pages. In addition, we have received a request to use it for creating an article in XML that does not depend on a particular layout. For example, an author's original article in PDF for early publication, or accompanying information, such as news and book reviews.

For this purpose, you can use the blank template. The template has no configuration of automatic recognition and extraction. When you register a PDF, each item on the XML editing screen will be empty. It will take a method of extracting only the necessary bibliographic item from the page image panel and creating your XML. Since the checking based on the guidelines of J-STAGE, JATS schema specifications, it appears to be helpful in academic societies which are unfamiliar with XML.

Recognition accuracy

Automatic recognition rate

In this system, the automatic recognition rate as a criteria to calculate the conversion accuracy is defined as follows:

$$\mathit{Automatic\ recognition\ rate} = 100 \times \frac{\begin{array}{c}\mathit{Number\ of\ items}\\ \mathit{extracted\ automatically}\end{array}}{\begin{array}{c}\mathit{Number\ of\ items\ in\ paper\ image}\\ \mathit{to\ be\ recognized\ automatically}\end{array}}\ (\%)$$

If there is multiples of a JATS item on the same page image, the total number of items is reported as the number of items. An example of this would be if there are three authors automatically extracted, the number of items targeted to the automatic recognition is 3. If only two of the three authors are extracted, the correct number of items that have been automatically extracted is 2. In addition, acknowledgments, etc. that are not targeted for the automatic recognition are not recorded even if these are included in JATS specification. We choose this criterion based on the assumption that the actual amount of work required to confirm after the automatic conversion, is proportional to the number of items. That means, it's easy to fix if failing to recognize the article title, but if failing to divide the cited references, it requires the modification of the number of the cited references.

Verification results of automatic recognition rate

At the time of opening the service, we selected 10 journals that are registered in J-STAGE and requested 10 articles for each journal from the academic societies. These comprised the test sets. We then calculated the rate of automatic recognition and verified the results (Table 1). Some of the test sets provided by the journals included essays and errata papers. These were not included in the tests and thus some of the journals have less than 10 articles tested.

TABLE 1: Automatic recognition rate by sample article

Journal	Language*1	Automatic recognition rate			Number of articles
		Average value	Minimum value	Maximum value	
EL	J/E	91%	58%	100%	10
JO	J/E	97%	89%	100%	10
JE	J/E	98%	95%	99%	10
CL	E	93%	86%	100%	10
TR	E	90%	50%	100%	10
JI	J/E	91%	83%	96%	8
NI	J	91%	83%	100%	10
BU	J/E	93%	75%	98%	8
AD	E	100%*2	97%	100%	7
PJ	E	98%	90%	100%	9

Language J=Japanese only, E=English only J/E= Japanese-English parallel entry

99.53% due to rounding.

The worst results came from three journals: EL, TR, and BU. This was because the tool did not properly recognize the cited reference and the title of keywords went wrong resulting in none of the information being extracted. The deduction of points was done accordingly. This was caused because the PDF layouts of the articles were modified to accommodate a change to the page size of the PDF. If the article with the lowest score of each of the 3 journals is removed; the recognition rate exceeds 80% for all the remaining articles. Another failures in recognition is the fpage / lpage element. It is because there was a case where character strings other than the page number in a header/footer are extracted together. This is because there is no feature to extract just a part of a character string pattern.

Prospects for transforming XML from PDF

The current Bibliographic metadata creation tool makes the bibliographic information of JATS from PDF as previously described. The natural future of this tool is to enhance the accuracy of conversion and the full text conversion from PDF into JATS. The following section describes issues of conversion from PDF to JATS in order to put into perspective and make clear the future development theme.

This tool consists of two stages. The first stage is a process of recognizing the text block of PDF, the second stage is a process of assigning the recognized text block to JATS element.

Issues about the recognition process of text block

At first, we consider the first stage. The reason this tool selected PDF format as input data is summarized in the section: Why PDF. There are some problems regarding conversion from PDF.

Types of PDF that cannot be converted

There are two kinds of PDF classified by how they are created; generated and scanned. A generated PDF is created by a computer using data while a scanned PDF is created by scanning a page image from paper. The scanned page can then either be treated as an image in the PDF or converted to text using OCR. The tool has no capability to deal with scanned pages if they are kept as images. Even with generated and OCRed PDF, if text is outlined or character code mapping table is not embedded, those PDFs do not have the information necessary for extracting character codes.

As this tool reads coded data in a PDF file and gets all information necessary from inside of the PDF, this tool is only applicable for a PDF that has the internal information necessary to extract the character code and recognize text blocks.

At the moment this tool is prepared for academic society users who typically understand this limitation. In the future, in order to service a wider range of JATS conversion from PDF it will be necessary to reduce the limits on the PDF that the tool can be applied to.

Improvement of PDF Analyzer

PDF is a file format to place characters in any position, orientation and size on a two-dimensional plane which simulates a paper media. Any PDF reader reads a PDF file and visualizes characters with location, orientation and size on the two-dimensional display as specified in the file.

The length of each character string and how to specify the location of them is dependent on the tool used to create PDF and varies a great deal depending on the tool. Extreme ones may specify the location of each character one by one on the page. As shown in the PDF reader, it is visible as a normal sequential string of text but internally in the PDF it is not arranged as the same sequential string of text as it appears in the reader. A PDF reader does not need to recognize any text block to display a page. A human reader recognize the text block through the display. The PDF Analyzer works like a human reader in that it recognizes text blocks in a PDF even if the characters are not arranged in a normal sequential string internally as they would appear on the visual page within a PDF file.

In fact, the software library that provides recognition of the text block within a PDF which is necessary for this tool is rare. The tool is realized by integrating a PDF Analyzer, which is a software library from Antenna House, to perform the block recognition of text. For this reason, the enhancement of accuracy of the PDF Analyzer is indispensable in the future in order to increase the accuracy of the service.

For example, the current PDF analysis algorithms have a particular problem in the recognition of the variable space between Latin words. Since the space between characters is reconstructed based on the space between words, it is affected by character spacing adjustments on DTP, etc., and the problem with accuracy of conversion has occurred from variable spaces in some articles. Antenna House is going to improve the function.

Improvements of semantics mapping

In the second stage of processing the tool maps a block of text corresponding to JATS elements. For example, there is a need to divide a text block into JATS elements for such items in a reference with author, title and name of the publication based on the meaning of the string. In other words, the second step is a process to give meanings to the text block.

Another issue identified with the test cases is where the publisher modified the layout of a article to accommodate the page size of the PDF. This resulted in very poor recognition. For a PDF that is created interactively with tools such as DTP, the layout may change depending on the situation. Therefore, it will be necessary to develop how to react when the layout has changed. To understand the meaning of the block, there is a need for improvement of the method corresponding to the elements of JATS.

For full text into JATS

There is a possibility to offer not only bibliographic information into XML but also full article JATS.

To implement the conversion of full article to JATS, the first challenge will be to recognize all objects in a PDF. PDF Analyzer extracts images, tables and vector graphics contained in PDF by recognizing the image borders when analyzing PDF. But it may not suffice. When targeting science and technology articles it will be impractical to ignore mathematical expressions. This requires the ability to recognize precise arrangements of characters and to understand automatically the context of the mathematical expression itself, this is considered very difficult.

Conclusion

In conclusion, it is a challenge to improve the recognition rate and to go forward with full text recognition of PDF, but the results will help the XML-ization of academic journals.

Acknowledgements

We would like to thank the academic societies and printing companies who cooperated in many ways to provide data for the system evaluation.

References

1) J-STAGE. Tokyo: Japan Science and Technology Agency; https://www.jstage.jst.go.jp/.

2) R Sato. New J-STAGE system accelerates digitization and distribution of academic journals from Japan. J Information Processing and Management. 2012 ; 55(2): 106-. DOI: 10.1241/johokanri.55.106. [in Japanese]

3) D.A. Lapeyre, B.T Usdin. Introduction to Multi-language Documents in NISO JATS. Journal Article Tag Suite Conference (JATS-Con) Proceedings [Internet]. 2011. Available from: http://www.ncbi.nlm.nih.gov/books/NBK62175/

4) S Tokizane. From NLM DTD to JATS: XML for scholarly articles in Japanese. J Information Processing and Management . 2011; 54(9): 555-. DOI: 10.1241/johokanri.54.555. [in Japanese]