XML について

中西印刷株式会社

2014/1/30

XML とは

現在、J-STAGE や Pub Med Central といったメジャーなオンラインジャーナルに掲載するには論文が XML で記述されていることが必須です。

実際の XML による記述はこんな感じになります。学術論文の冒頭のほんのすこしをごらんにいれましょう。

```
Journal Publishing DTD v0.4 20110131//EN">
- <article xml:lang="ja" dtd-version="0.4" article-type="research-article" xmlns:xlink="http://www.w3.org/1999/xlink"
xmlns:mml="http://www.w3.org/1998/Math/MathML" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
  - <front>
    - <journal-meta>
        <journal-id journal-id-type="j-stage">johokanri</journal-id>
        <issn pub-type="ppub">0021-7298</issn>
        <issn pub-type="epub">1347-1597</issn>
      </journal-meta>
    - <article-meta>
      - <title-group>
          <article-title>国産電子ジャーナルの著作権と ライセンス:</article-title>
          <subtitle>J-STAGEジャーナルの現状に見る課題と可能性</subtitle>
        - <trans-title-group xml:lang="en">
            <trans-title>Copyright and licensing of E-Journals:</trans-title>
            <trans-subtitle>Perspective from E-journals in J-STAGE</trans-subtitle>
          </trans-title-group>
        </title-group>
      - <contrib-group>
        - <contrib contrib-type="author">
          - <name-alternatives>
            - <name name-style="eastern">
                <surname>林</surname>
                <given-names>和弘</given-names>
              </name>
            - <name xml:lang="en" name-style="western">
                <surname>HAYASHI</surname>
                <given-names>Kazuhiro</given-names>
              </name>
            </name-alternatives>
            <xref rid="aff1" ref-type="aff">1</xref>
            <xref rid="aff2" ref-type="aff">2</xref>
          </contrib>
        - <contrib contrib-type="author">
          - <name-alternatives>
            - <name name-style="eastern">
                <surname>和田</surname>
                <given-names>光俊</given-names>
              </name>
            - <name xml:lang="en" name-style="western">
                <surname>WADA</surname>
                <given-names>Mitsutoshi</given-names>
              </name>
            </name-alternatives>
            <xref rid="aff3" ref-type="aff">3</xref>
          </contrib>
```

XML は eXtended Markup Language の略です。拡張された Markup 言語ということになります。これだけではなんのことだかわかりませんよね。ここではあえて翻訳しなかった Markup と言う言葉が鍵になります。まず Markup を解明しなければ、XML もその意味がつかめません。Markup は

辞書では「タグ付け《テキスト中に書体などの標識情報を埋め込むこと》」となっています（研究社英和）。また別の言葉がでてきました「タグ付け」。これはもう実際に見た方が早いですね。

Markup Language で一番よく目にするのは、インターネットホームページを作るのに使われる HTML（Hyper Text Markup Language）でしょう。下は実際に中西印刷の住所を書いた部分です。

```
<strong>京都本社</strong><br />
〒602-8048　京都府京都市上京区下立売小川東入ル<br />
TEL：　075-441-3155　infos(a)nacos.com<br />
<strong>東京営業部</strong><br />
〒113-0033　東京都文京区本郷2-26-11 浜田ビル５階<br />
TEL：　03-3816-0738　info_tokyo(a)nacos.com
```

この や
 が Markup のタグといわれる部分です。このタグの部分はホームページ画面には現れてきません。 は強調ですから、実際に見ると「京都本社」という文字が太字で表示されることになります。

このように、テキスト本文ばかりではなく、それを修飾する記号をいれるのが Markup 言語です。しかしそれならば、そんな Markup とかややこしいことをやらなくても WORD でも Indesign でもマウス操作で太字にするくらいできるではないかということを言われるかもしれません。そうです。その通りです。それを Markup でわざわざやることの利点こそが、XML の本質になってくるのです。

まず、WORD や Indesign で作った場合の不都合を考えてみましょう。

色々な不都合はありますが、最大の不都合は特定のソフトで作った文書は特定のソフトでしか読めないということです。WORD で作った文書は WORD でしか読めません。Indesign で作った文書は Indesign でしか完全には読めません。だから、いろいろな文書に応用しようと思えば、WORD にも Indesign にも、あるいはホームページにも共通に使える文書形式である必要があるのです。さらに現在では文書は紙への出力だけを前提にできません。画面の上でいかに読みやすくするかということも考慮する必要があるのです。

電子文書の世界では、読まれる画面によって適切な表現が違うという問題が生じます。電子文書では、PC で読むのとスマホで読むのとでは適切な表現が異なってくるのです。PC で読むときは、行の幅が少々長くても 1 画面内で読めますが、スマホでは、行は短くないといちいち横にスクロールしないと読めません。1 行はスマホの小さい画面ひとつに収まるよう短くないといけないのです。または最近ではタブレット端末で読む人も多いでしょうし、まだまだ紙で読みたいという人もいるでしょう。こんなとき、WORD や Indesign でそれぞれにふさわしい文書をいちいち作っていたのではたまったものではありません。

まずは、Markup で文書を作り、あとはどのようにも自動変換して利用できるようにしておく方がデジタル時代にはふさわしいのです。デジタル時代にあっては、文書はインターネットや電子書籍というかたちでも広く利用できなければならないのです。Markup 言語で書いてあれば、変換プログラムを通せばどのように形式にも変換することができます。いわゆる OneSource Multi Use です。

XML はこうしたどのような出力形態でも読みやすいように、文書の中に Markup タグを埋め込

む際の文書形式です。ここで重要なのは文書に書かれた <タグ> の中身でどのような表現が実際にされるかは、XML では規定していないということです。 と書いて HTML なら太字になるという意味ですが、XML では別にこれは太字でなくても色を変えるでも字を大きくするでもいいのです。この <タグ> と実際の表現をどうするかは、スキーマで定義します。このスキーマとしてよく使われるのが DTD（Document type Definition）です。XML とはその定義の定義、どういう <タグ> がくればどういう表現にするかを決める方法を決めているにすぎないわけです。

いわば、XML とはどのような文書形態にも利用できるようにするためのタグ付けの規則を決めるための規則です。実際の XML 文書を見ると冗長なぐらい <タグ> がはいっています。

構造化文書

ここで XML が表現を規定しないのなら、一体何を規定しているのかという疑問を抱かれることと思います。XML とは一体何を規定しているのか。XML は実は文書構造を規定しています（厳密には文書構造以外も定義することは可能ですが、ここでいう学術情報用 XML では文書構造といってさしつかえありません）。

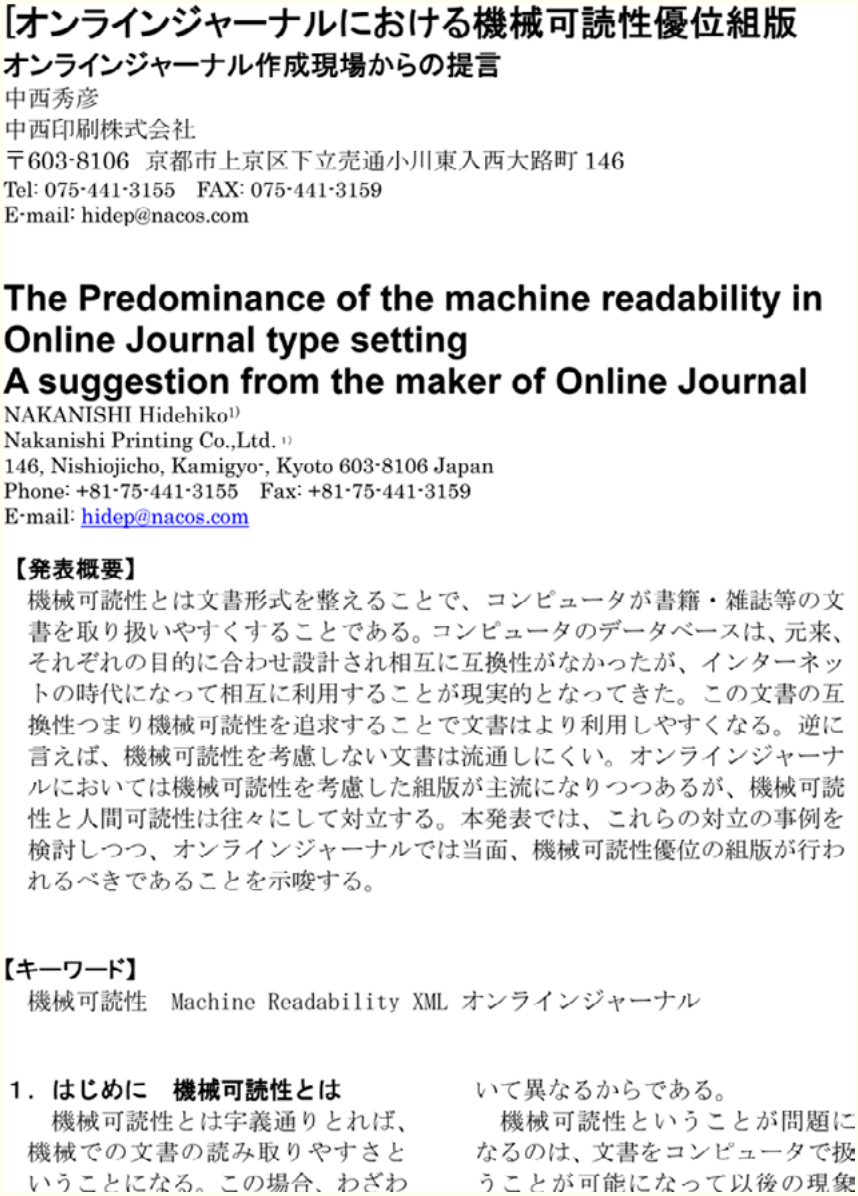
文書構造とは聞き慣れない言葉かもしれません。普通、意識はされていませんが文書には構造があります。この構造がはっきりわからないと文書はとても読みにくいものになってしまいます。たとえば次の文書を考えましょう。

<title>オンラインジャーナルにおける機械可読性優位組版－オンラインジャーナル作成現場からの提言</title><author>中西秀彦</author><affiliation>中西印刷株式会社</affiliation><abstract>機械可読性とは文書形式を整えることで、コンピュータが書籍・雑誌等の文書を取り扱いやすくすることである。コンピュータのデータベースは、元来、それぞれの目的に合わせ設計され相互に互換性がなかったが、インターネットの時代になって相互に利用することが現実的となってきた。この文書の互換性つまり機械可読性を追求することで文書はより利用しやすくなる。逆に言えば、機械可読性を考慮しない文書は流通しにくい。オンラインジャーナルにおいては機械可読性を考慮した組版が主流になりつつあるが、機械可読性と人間可読性は往々にして対立する。本発表では、これらの対立の事例を検討しつつ、オンラインジャーナルでは当面、機械可読性優位の組版が行われるべきであることを示唆する。</abstract><keyword>”機械可読性”, “Machine Readability”, ”XML”, “オンラインジャーナル”</keyword><subtitle1>はじめに　機械可読性とは</subtetle1><p>機械可読性ということが問題になるのは、文書をコンピュータで扱うことが可能になって以後の現象であることは論を待たない。機械可読性はまず書誌データベースの問題としてたちあらわれてきた。 MARC(MAchine Readable Cataloging)である。まさに MAChine Readable(機械可読)なカタログであり、 図書館カードで行われてきた書誌データベースをコンピュータで作成し、検索や整理をコンピュータを通じて行うようにしたものである。

当初書誌データの作成は、文書そのものの作成とは独立して考えられてきた。文書は文書として作られ、書誌は図書館等において独立して文書整理のために作成されることが普通だった 1)。やがて、この Machine Readability を文書の作成当初から考慮し、書誌データとして役にたつような形式での提供が考えられるようになった。特に、学術文献においては、膨大な資料の中から適切な情報を探し出す必要性と資料そのものがより引用されることが重要視されるため、みずから積極的な書誌情報の提供を行うようになっている 2)。

これではとても読めません。小学生でも今時こんな文書は書かないでしょう。たぶん、小学校の作文でも「題名」は上から 3 字あけて、「名前」は行をかえて、下の方に。「本文」は 3 行目で一字下げてと形式を整えることを習うでしょう。印刷屋でも「題名」は文字を大きくして中揃えで、「名

前」は題名より小さく、本文より大きく、書体をゴシックにして「本文」は、左詰明朝体でとかして読みやすくします。こんな感じ



[オンラインジャーナルにおける機械可読性優位組版
オンラインジャーナル作成現場からの提言
中西秀彦
中西印刷株式会社
〒603-8106 京都市上京区下立売通小川東入西大路町 146
Tel: 075-441-3155 FAX: 075-441-3159
E-mail: hidep@nacos.com

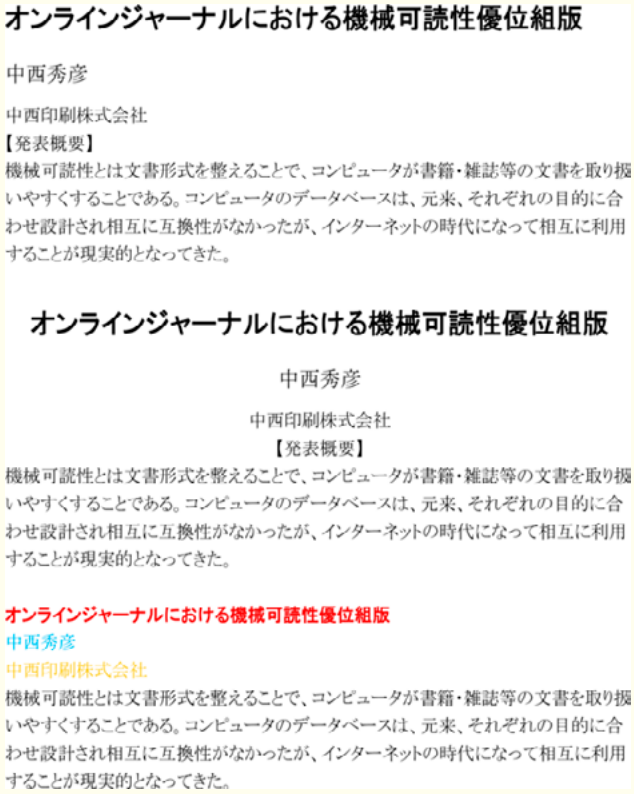
The Predominance of the machine readability in Online Journal type setting
A suggestion from the maker of Online Journal
NAKANISHI Hidehiko1)
Nakanishi Printing Co.,Ltd. 1)
146, Nishiojicho, Kamigyo-, Kyoto 603-8106 Japan
Phone: +81-75-441-3155 Fax: +81-75-441-3159
E-mail: hidep@nacos.com

【発表概要】
機械可読性とは文書形式を整えることで、コンピュータが書籍・雑誌等の文書を取り扱いやすくすることである。コンピュータのデータベースは、元来、それぞれの目的に合わせ設計され相互に互換性がなかったが、インターネットの時代になって相互に利用することが現実的となってきた。この文書の互換性つまり機械可読性を追求することで文書はより利用しやすくなる。逆に言えば、機械可読性を考慮しない文書は流通しにくい。オンラインジャーナルにおいては機械可読性を考慮した組版が主流になりつつあるが、機械可読性と人間可読性は往々にして対立する。本発表では、これらの対立の事例を検討しつつ、オンラインジャーナルでは当面、機械可読性優位の組版が行われるべきであることを示唆する。

【キーワード】
機械可読性 Machine Readability XML オンラインジャーナル

1．はじめに　機械可読性とは
機械可読性とは字義通りとれば、機械での文書の読み取りやすさということになる。この場合、わざわいて異なるからである。
機械可読性ということが問題になるのは、文書をコンピュータで扱うことが可能になって以後の現象

この「題名」「名前」「本文」と書いたのが実は構造です。ここで重要なのは、表現形式が変わっても、この構造は一定と言うことです。いくつか表現形式をかえてみますが、どこまで言っても「題名」は「題名」であり、「本文」は「本文」です。「題名」は本文とは違って目立つ必要がありますが、その表現形式は色々な形がありうるわけです。それでも「題名」であることにかわりはありません。表現形式としては、字を大きくする。太字を使う。中揃えにする。そしてそれらの併用。それはどんな表現形式でも間違いではありません。あるいは色を変えたっていいです。また視覚障がいのある方でしたら、自動読み上げソフトで声を大きくはりあげて、一時休止するといのうが「題名」の表現でありえます。それでも「題名」は「題名」です。


オンラインジャーナルにおける機械可読性優位組版

中西秀彦

中西印刷株式会社

【発表概要】

機械可読性とは文書形式を整えることで、コンピュータが書籍・雑誌等の文書を取り扱いやすくすることである。コンピュータのデータベースは、元来、それぞれの目的に合わせ設計され相互に互換性がなかったが、インターネットの時代になって相互に利用することが現実的となってきた。

オンラインジャーナルにおける機械可読性優位組版

中西秀彦

中西印刷株式会社

【発表概要】

機械可読性とは文書形式を整えることで、コンピュータが書籍・雑誌等の文書を取り扱いやすくすることである。コンピュータのデータベースは、元来、それぞれの目的に合わせ設計され相互に互換性がなかったが、インターネットの時代になって相互に利用することが現実的となってきた。

オンラインジャーナルにおける機械可読性優位組版

中西秀彦

中西印刷株式会社

機械可読性とは文書形式を整えることで、コンピュータが書籍・雑誌等の文書を取り扱いやすくすることである。コンピュータのデータベースは、元来、それぞれの目的に合わせ設計され相互に互換性がなかったが、インターネットの時代になって相互に利用することが現実的となってきた。

表現形式は実は構造とは別物です。文書はもちろん内容が一番大事ですが、構造も重要です。そして構造さえしっかり決まっていれば、表現形式はそのときそのときでふさわしいものが選ばれればいいのです。さきに述べたように、紙の本でふさわしい表現形式、PC で読むのにふさわしい表現形式、スマホで読むのにふさわしい表現形式。そして、読み上げるのにふさわしい表現形式。それぞれ自動的に変換して表現すればいいだけのことです。ですから、あとで実際の表現に変換しやすいように、構造が明確に表現できればいいわけです。実際、コンピュータが発展して紙よりも画面で読むことが普通になった今、重要な概念といっていいでしょう。

XML はこの構造を規定していきます。構造を表現するのに適した Markup 言語と言えるでしょう。これまで雑誌編集の現場では表現形式を整えることが何よりも重んじられました。字の大きさ、書体、空白の配置、タイトルの表示など、人間が読みやすいように読みやすいように細心の注意を払って作られてきました。しかし、これからは、表現形式は読む側のデバイスによっていくらでも変わりうるので、構造の方がより重要です。構造が整わない文書はコンピュータが迷ってしまって適切に表現できないからです。また、すでに文書は人間が読むだけのものではありません。機械つまりコンピュータが読んでコンピュータが処理する場合も増えてきています。その代表例が検索エンジンでしょう。検索エンジンにちゃんと拾ってもらわなければその文書はインターネット世界の中でないのも同然です。コンピュータが処理しやすいのはやはりこうした構造がしっかりした文書

であって、文書がいくら人間にとって綺麗であっても関係ないのです。
　XML はコンピュータの時代の言語といえます。

学術分野における XML

　学術分野では構造が表現できる言語として既に XML がデファクトスタンダードになっています。おそらくこれはもうこれから何十年という単位で変わることはないでしょう。学術分野では構造化が非常に重要です。それは使われる構造要素が非常に多いからです。小説だと、構造は「題名」と「本文」だけです。「著者名」は普通表紙に一回書いてあるだけでしょう。学術文書だと、「題名」「副題」「著者名」それも筆頭かそうでないか、「所属」、「キーワード」「抄録」「見出し」「掲載日」「受理日」「注」「引用文献」「図」「表」といくらでも規定しておかなければなりません。最近では「研究助成者」とかも重要になってきますし、「動画」とか「ソフトウェア」などというのもありえます。それらは、一緒くたにしてコンピュータに渡してもなんのことだかわかりません。人間なら当然気がつくことでも、コンピュータは懇切丁寧に教えてやらないとわかりません。

　またこれらをきっちり構造化しておく（<タグ>づけしておく）ことで、今までの紙の本では考えられなかった機能も付加できます。著者名で検索をかけて、同じ著者の論文だけを集めるとか、「掲載日」の近いものを比較する。または領域の近いものを横断検索する。最近では、論文の盗用がないか調べるといったことにもつかわれ始めています。また将来的にはどの「研究助成」が効率的に論文を生産したかなども検討できるようになるでしょう。そういう意味でももはや XML による構造化文書作成は基本です。

　冒頭にも書いたように、現在、J-STAGE や Pub Med Central といったメジャーなオンラインジャーナルに掲載するには論文が XML で記述されていることが必須になっています。オンラインジャーナルに載せようとすれば、XML で書かねばならないということです。もちろん、XML の記載はそんなに簡単なものではありませんから、著者が直接書くという性格のものではありません。これは印刷会社の仕事なのです。紙の印刷がなくなったら印刷会社の仕事はなくなる。そんなことはありません。XML の組版こそがこれからの印刷会社の仕事といえるのです。

　なお、XML は定義の定義といいましたが、XML ではスキーマでそれぞれの＜タグ＞がどういう意味であるかを規定してやる必要があります。<TITLE> というタグは「題名」をあらわしているといった定義です。これは用途に応じてさまざまなものが使用されますが、学術分野では JATS（Journal Tag Suite）が有名です。JATS についてはこちら。