XMLに基づく多言語化学論文全文データベースによる電子出版

○李　穎*
石塚　英弘**

XML-based Electronic Publishing of Chemical Article's Multilingual Full-text Database

○Ying li
Hidehiro Ishizuka

Abstract: In an electronic publishing field, attention is being focused on XML. We constructed chemical article's multilingual (Japanese, Chinese and English) full-text database based on XML, and distributed it in the Web. The On-line Journal XML-DTD of NII (National Institute of Information) is adopted as the DTD of the database. Unicode is used as an encoding character code to represent those languages. To display the database in Microsoft Internet Explorer (IE) 5.0, the data of the standard style sheets for XML (i.e. CSS and XSL) are added to the XML document. Applying VB Script and DOM, the elements selected by a user, such as title, author, abstract, bibliography, are extracted from the full-text database, and displayed in IE 5.0. Finally, we discussed the problems related with Unicode, expression of chemical formula, as well as schema.

1. 研究の目的と背景

現在、インターネットのWebで情報を配布・閲覧するための言語としてはHTMLが普及している。HTMLは簡単な方法で情報を作成できるため、便利なマーク付け言語である。しかし、文書構造が表現できないなどの理由で、応用範囲が限られる。

一方、SGMLは階層的な構造化文書として、情報を表現できる言語であるが、Web向きの仕様になっていないため、Web対応言語としては事実上使用できない。

XMLは、HTMLやSGMLがもつ問題を解決、緩和するために登場した機能拡張性の高い仕様の言語である。XMLの主な特徴は以下の通りである。

- 構造化文書を表現することができる。
- スタイルシート言語を利用し、XML文書をWeb上に表示することができる。
- XMLは、文書の構造・内容と、書式データが分離していて、ソフトウェアで文書内容を扱いやすい。
- XML文書は、Unicodeを始めとして、様々な文字コードを使って書くことができる。

* 図書館情報大学大学院情報メディア研究科博士後期課程
**図書館情報大学図書館情報学部

・ 全文検索や属性検索だけでなく、構造検索や意味検索を含めた高度な複合検索が可能である。

出版業界は、SGML、HTML、XML 等を用いて、電子出版を行っている。電子出版について、欧米、日本、中国の進行状況を調査した。調査の対象は、Elsevier[1]、アメリカ化学会[2]、ChemWeb.com[3](The World Wide Club for the Chemical Community)、国立情報学研究所 NII[4][5]（旧名：学術情報センター）、科学技術振興事業団[6][7]、日本電子出版協会[8]、日本化学会[9]、中国科学雑誌社[10]である。調査結果によると、日本、中国、アメリカには、XML に基づく全文データベースの実例はまだ少ない。特に、多言語に対応する XML データベースは見つからなかった。

そこで本研究では、複雑な文章構造を持つ化学領域を選んで、XML に基づく多言語論文全文データベースの構築とその Web による配布と利用を目指し、研究した。

2. 本システムの構成と構成手順

本システムにおける電子出版の工程は、XML に基づく全文データベースを構築する工程と、それを Web によって提供する工程の２つに大別される。

XML データベースの言語を表記する文字コード体系として Unicode を用いるため、全世界の言語を対象にできるが、今のところ、中国語、日本語、英語の文献を対象とした。構築された XML 多言語化学論文全文データベースを Web 上で表示するため、XML の標準スタイルシート言語 CSS と XSL を用いた。

また、Web 利用者の要求はさまざまで、それに応じて XML データベースからデータの抽出を行い、得られた結果は Web ブラウザで表示した。これらの処理はスクリプトと DOM を用いて行った。

次に、本システムについて、詳しく説明する。

本システムの概念図を図１に示す。

図１に示す①から⑥までの作業を簡単に説明する。詳細は 2.1-2.4 に述べる。

①NII のオンラインジャーナル用 XML-DTD を本研究の DTD として、多言語データベースに加える。

②印刷物に付いた図をイメージスキャンすることと CS Chem Office を利用して描画することにより、XML 文書の化学図を表示し、多言語データベースに加える。

③日本語、中国語、英語最新の雑誌論文を、NII の XML-DTD に基づき、XML 文書に書き直し、多言語データベースに加える。

④構築された XML 多言語全文論文データベースを Web 上で出版するため、CSS と XSL を利用して、スタイルを付けて表示する。

⑤XML 全文論文データの一部分のみ（例えば、タイトル、著者、抄録など）を表示するため、VB Script と DOM を利用して、その部分を抽出する。

⑥抽出された部分を表示する。

図１に示した目的を達成するため、2.1 から 2.4 までの４つのことを行った。

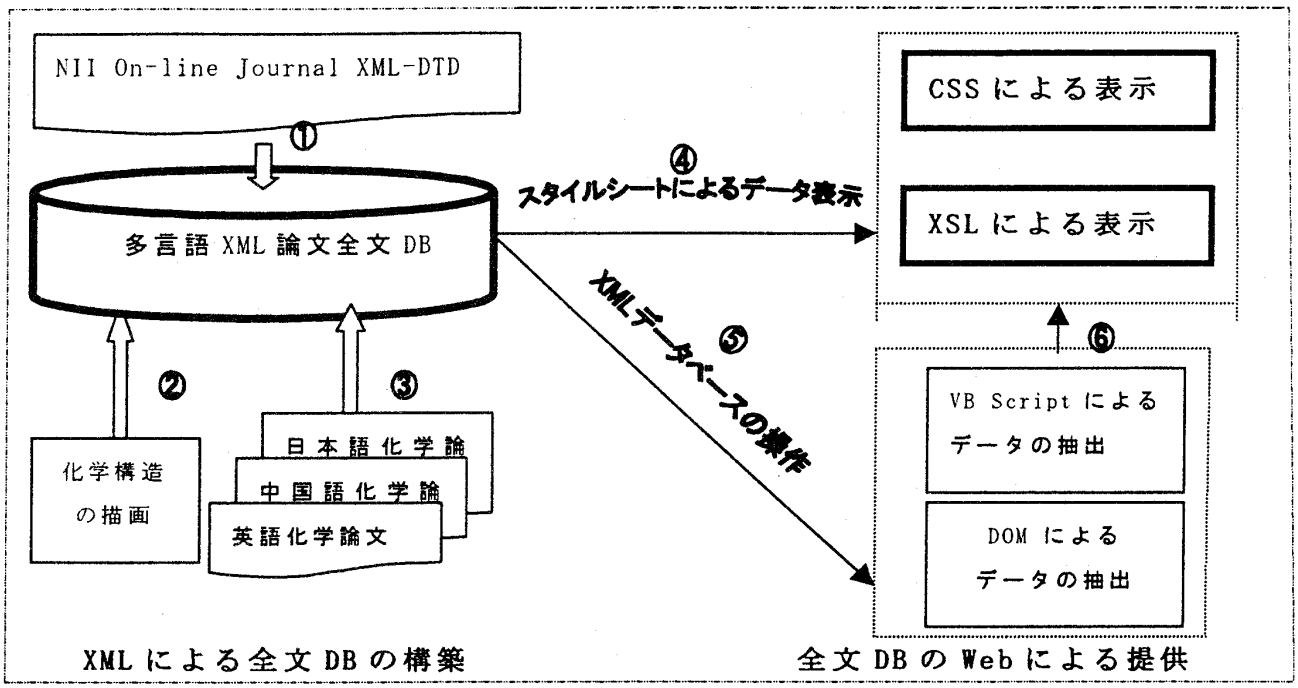

図 1　電子出版システムの概念図

2.1　NII の論文 DTD [11]の検討と移植

NII のオンラインジャーナルにおける XML DTD は、詳細版と簡易版の二通りがあり、本研究では、サブシステム間でテキストを交換するための共通フォーマットとしての簡易版 XML DTD を用いた。

本システムでは XML 文書を Windows PC 上の XML エディタで作成する。簡易版 XML DTD は UNIX 上で開発されたものであるため、この DTD を PC 上に移植した。495KB の巨大な量の DTD であるため、Near & Far Designer3.0 を用いて DTD の構造を検討しつつ、PC 上に移植した。

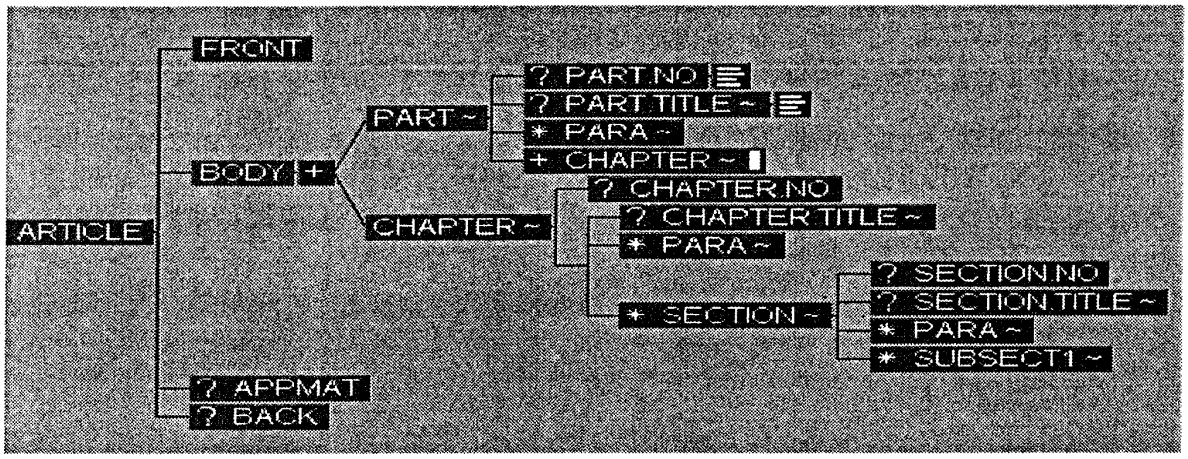

図 2　NII　XML-DTD の階層構造の一部

図 2 は、“Near & Far Designer 3.0 ” を用い、NII の XML-DTD の階層構造を表示した結果である。図 2 に示すように、論文の構造は大きく分けると、front、body、appmat、back の 4 つで構成されている。全ての構造を表示すると複雑なため、こ

こではbodyの部分を4段階まで表示した。

2.2 NIIのDTDによる日本語、中国、英語の化学論文全文データベースの構築

最新の日本語、中国語、英語の化学論文を選び、NII論文XML-DTDに従い、XML1.0によるサンプル全文データベースを構築した。

以下の2.2.1-2.2.4節では、XMLによる化学論文全文データベースの構築について、説明する。

2.2.1 多言語の文字コード

文字コードはUnicodeを用いた。そのため、Unicode対応のWebブラウザ、例えば、MS Internet Explorer5.0を用いれば見ることができる。

ただし、特殊な記号などでUnicodeに含まれていない文字が在る場合に関しては、幾つの方法を検討した。本システムではデータベースの規模が小さいため、外字はイメージとして、扱った。

2.2.2 XML多言語文書の編集

テキストの編集には、XMLエディタ、Unicode対応のXML Spy2.5を使用した。中国語の場合は、フロントエンドプロセッサIMEを利用して入力し、XML Spyの機能によりUnicodeの内部表現へ変換された。

本研究では、拡張グリッドビューを利用し、XML文書の全体構造を階層的に表示し、編集した。ソースビューを利用し、直接ソースを編集する方法で作業した。ブラウザビューは、CSSとXSLを完全にサポートするため、これを利用して適宜XML文書を表示してXML文書の形式を確認しつつ編集作業を行った。

2.2.3 XML文書にある表の扱い

XML文書にある表については、2つの処理方法を検討した。

- 名前空間を利用し、HTMLにおける表の関連タグを使い、表を表示する。
- XSLの構造の変換機能（XSLT）を利用し、HTMLの表に変換して、表示する。

本研究では、前者は、Microsoft Front Page 2000を利用して、表を編集した。後者は、XML Style Wizardを使って、XMLを一般的な表形式で表示するためのXSLファイルが簡単に作成できた。

2.2.4 論文における化学構造図の編集

図の編集については2つの方法を検討した。

- スキャナーを利用し、化学論文の印刷物から図をデータベースに入力する。
- CS Chem Office（化学構造式描画、分子モデリング等の機能を持つ。Cambridge Soft社）を使用し、新たに描画する。

2.2.1から2.2.4に述べた手法で、XMLに基づく多言語化学論文全文データベースを完成した。図3に1999年の「中国科学」誌上の化学論文のXMLソースデータを示す。

2.3 XMLデータベースのスタイルシート言語CSSとXSLによるWeb上電子出版

構築されたXML多言語化学論文全文データベースをWeb上で表示するため、XMLの標準スタイルシート言語CSSとXSLを用いた。ブラウザとしてはMS Internet Explorer5.0を使用した。

図4に1999年の「中国科学」誌上の化学論文を本研究の方法によりIE 5.0で

表示した例を示す。

2.4 論文の一部分を抽出するための Visual Basic Script と DOM の利用

Web 利用者の要求はさまざまなため、それに応じて XML データベースから利用者が必要とする要素のみを抽出し、得られた結果を Web ブラウザで表示するようにした。これらの処理は Visual Basic Script と DOM を用いて行った。

3. 考察

本研究で得られた結果に基づき、以下の点について考察する。

①システム開発プラットフォームの Unicode 対応の問題

使用したプラットフォームは日本語 Windows98 であり、OS レベルでは Unicode に対応していない。Unicode への対応は個別のソフトウェアに依っている。たとえば、WWW ブラウザ：MS IE 5.0 は Unicode で書かれた Web 文書を表示できる。また、XML エディタ：XML Spy はシフト JIS コードで入力した日本語テキストを Unicode に変換する機能と、GB2312 コードで入力した中国語テキストを Unicode に変換する機能を持っている。これらのソフトウェアの機能を使って、本研究では Unicode の XML 文書である論文データベースを作成し、WWW ブラウザ上で表示した。論文全体を表示するだけではなく、タイトルなど特定の部分のみを表示することもできた。

しかし、全文検索はまだ実現していない。その原因は日本語 Windows98 が Unicode に対応していないためである。全文検索では検索語を入力して、データベースの中の語とマッチングする必要があるが、検索語はシフト JIS コードあるいは GB2312 コードで入力されるため、Unicode で書かれた語とマッチングできない。この問題を解決する方法は、OS レベルに Unicode への変換機能を持たせるか、あるいは WWW ブラウザ上にコード変換機能を持たせるかである。マイクロソフト社は Windows2000 で Unicode に対応したため、今後実験するつもりである。

②数式と化学式の表現

NII 簡易版 XML-DTD では数式と化学式については規定していないため、本研究では数式と化学式はイメージで扱った。詳細版 XML-DTD では、数式のためには MathML を組み込んでいる。また NII は、化学式については ISO chemical formula element set を取り込んでいるが、将来的には CML の XML DTD を取り込む予定であると言っている。

CML には化学式だけではなく数式表現も含まれているので、我々は化学論文の場合は CML を使って化学式と数式を表現するのが良いと思う。

③ データベース構造表現としての DTD の問題

本研究は、NII オンラインジャーナル XML-DTD に従って、多言語化学論文 XML データベースを構築した。しかし、文書構造を表現する枠組みとして、DTD は以下の問題点があるという指摘がある。

・独自の構文
・拡張性の欠如
・構文解析がしにくく、ツールが作りにくい

・データ型がない

そこで、新しい枠組み：XML Schema が提案された。1998 年 8 月、W3C は XML Schema WG(Working Group)を結成した。1999 年 5 月 6 日に、第一版の作業ドラフトが W3C に公開された。9 月 24 日、11 月 5 日、12 月 17 日に、次々新バージョンが公開された。XML Schema[12]の役割は以下の通りである。

・ DTD と同様にマークアップ語彙を定義する。
・ 文書の意味に関する情報を付加する。
・ 文書化支援

XML Schema は今注目されているが、制定が遅れ、巨大な仕様であり、理解が極めて難しく、実装はまだ問題があることなど困難が多い。そのため、本研究では、XML Schema に関する検討は行っていない。

参照文献

[1]Elsevier Science <http://www.elsevier.nl/>

[2]http://www.acs.org/

[3]http://chemweb.com/

[4]http://www.nacsis.ac.jp/olj/index.html

[5]大山敬三, 神門典子, 佐藤真一. NACSIS オンラインジャーナルプロジェクト. 情報の科学と技術, 49(6), p295-300. (1999).

[6]http://www.jstage.jst.go.jp/JA

[7]吉田幸ニ. J-STAGE：科学技術情報発信・流通総合システム電子ジャーナル作成とインターネットによる流通. ディジタル図書館, No.16, p.50-59 (1999).

[8]http://x.jepa.or.jp/jepax/spec/jepaxspec09.xml

[9]http://wwwsoc.nacsis.ac.jp/csj/journals/csj-journals/sakurai.html

[10]http://www.scichina.com/mainsc.htm/ ; http://www.chinainfo.gov.cn

[11]http://www.rd.nacsis.ac.jp/olj/access_dtd-j.html

[12]村田 真. XML Schema.「XML とデータベースの接点をめぐる最新技術動向」. 電子情報通信学会, データ工学研究専門委員会. (1999 年 12 月 17 日).

参考文献

・根岸正光・石塚英弘. SGML の活用. オーム社, 168p. 1994 年.

・XML/SGML サロン. XML/SGML 標準 XML 完全解説. 技術評論社. 東京, 354p, 1998 年.

・http://www.w3.org/XML/

・Two Tigers. <http://www.ufoc.com.tw/>

・薬師寺国安・聖. IE5.0 における DOM プログラミング. 1999XML フェスタ. (1999 年 11 月 11-12 日).

・http://wwwsoc.nacsis.ac.jp/csj/journals/csj-journals/sakurai.html

・XML Spy 日本語マニュアル：http://www.tas.co.jp/xml/download/eva/

```
<?xml version="1.0" encoding="UTF-8"?>
<?DOCTYPE nacsis:article SYSTEM "root5.DTD" ?>
<?xml:stylesheet type="text/css" href="例5nacsis5.css"?>
<nacsis:article xmlns:nacsis=http://www.rd.nacsis.ac.jp/
xmlns:HTML="http://www.w3.org/Profiles/XHTML-transitional">
<nacsis:front>
<nacsis:titlegrp>
<nacsis:titlegrp.title alphabet="LATIN" lang="cn">云南中、小盆地低演化天然气地球化学
特征 <HTML:BR/>
</nacsis:titlegrp.title>
</nacsis:titlegrp>
<nacsis:authgrp>
<nacsis:authgrp.author.lang alphabet="LATIN" lang="cn">
<nacsis:authname>
<nacsis:fname>徐</nacsis:fname>
<nacsis:surname>永昌</nacsis:surname>
</nacsis:authname>
</nacsis:authgrp.author.lang>
<nacsis:authgrp.author.lang alphabet="LATIN" lang="cn">
<nacsis:authname>
<nacsis:fname>沈</nacsis:fname>
<nacsis:surname>平</nacsis:surname>
</nacsis:authname>
</nacsis:authgrp.author.lang>
<nacsis:authgrp.author.lang alphabet="LATIN" lang="cn">
<nacsis:authname>
<nacsis:fname>沈</nacsis:fname>
<nacsis:surname>建京</nacsis:surname>
</nacsis:authname>
</nacsis:authgrp.author.lang>
<HTML:BR/>
<nacsis:authgrp.author.lang alphabet="LATIN" lang="cn">
<nacsis:authname>
<nacsis:fname>刘</nacsis:fname>
<nacsis:surname>文汇</nacsis:surname>
</nacsis:authname>
</nacsis:authgrp.author.lang>
<nacsis:authgrp.author.lang alphabet="LATIN" lang="cn">
<nacsis:authname>
<nacsis:fname>关</nacsis:fname>
<nacsis:surname>平</nacsis:surname>
</nacsis:authname>
</nacsis:authgrp.author.lang>
<HTML:BR/>
</nacsis:authgrp>
<nacsis:abstract>
Abstract<HTML:BR/>
    对云南若干中、小盆地(陆良、杨林、保山及景谷)第三系生物气及未-低熟油伴生气进行了研究.生物气组分以甲烷为主，在烃气中>99%.同位素组成轻，δ13C1为-60.0‰～-75.4‰，其中保山盆地相对较重(-60‰～-65‰)，陆良及杨林盆地较轻，δ13C1均小于-70‰，比较而言，可能意味保山生物气藏成藏时间较早.景谷盆地原油为未-低熟原油，其伴生气具有相对高的湿度，烃气中甲烷相对含量为58%～95%,就组分而言与正常石油伴生气相似，但其甲烷碳同位素值为-53.8‰～-57.8‰，明显较一般油田生气相对富集12C，具有与生物-热催化过渡带气相似的特征.乙烷碳同位素在-34.6‰～-29.‰之间，其源岩位素在-34.6‰～-29.‰之间，其源
```

図3 「中国科学」誌の論文の XML ソースデータ (一部)

云南中、小盆地低演化天然气地球化学特征

徐 永昌 沈 平 沈 建京

刘 文汇 关 平

Abstract

对云南若干中、小盆地(陆良、杨林、保山及景谷)第三系生物气及未-低熟油伴生气进行了研究.生物气组分以甲烷为主，在烃气中>99%.同位素组成轻，δ13C1为-60.0‰～-75.4‰，其中保山盆地相对较重(-60‰～-65‰)，陆良及杨林盆地较轻，δ13C1均小于-70‰，比较而言，可能意味保山生物气藏成藏时间较早.景谷盆地原油为未-低熟原油，其伴生气具有相对高的湿度，烃气中甲烷相对含量为58%～95%,就组分而言与正常石油伴生气相似，但其甲烷碳同位素值为-53.8‰～-57.8‰，明显较一般油田伴生气相对富集12C. 具有与生物-热催化过渡带气相似的特征.乙烷碳同位素在-34.6‰～-29.‰之间，其源岩应为油源岩，但对低演化阶段石油伴生气而言，其组成偏重，同时，乙、丙、丁烷之间有同位素组成倒转现象，也许暗示着存在Ⅲ型有机质成气的贡献，δ13CCO2基本小于-10‰，应是有机成因产物.

低演化　天然气　碳同位素

1　样品采集与实验分析

在云南省的陆良、保山、杨林及景谷盆地采集天然气样品10个.样品分布见表1.

表1　云南新生代含油、气盆地样品采集层位

地区	井号	层位	深渡／m	备注
陆良盆地	陆1-1	N^2_2	453～462	钢瓶气样，25×9.8KPa
陆3		N^2_2	568～572	钢瓶气样，40×9.8KPa
杨林盆地	杨1	N_2	423～425	钢瓶样
保1		N_2	594～600	钢瓶样，5×9.8KPa
保山盆地	保2	N_2	456～468	钢瓶样，4×9.8KPa
保3		N_2	551～562	钢瓶样，5×9.8KPa
牛7		N^S_1	424～431	排水取气
景谷盆地	牛5-1	N^S_1	317～334	排水取气
牛2-1		N^{S3}_1	341～346	排水取气
牛2-6		N^{S3}_1	364～348	排水取气

样品在气体地球化学国家重点实验室完成了烃类组分和同位素组成分析.烃气组分分析采用HP-5880A型色谱仪完成，碳同位素分析是在GC-C-MS(MAT-252)上做在线分析完成，分析误差±0.3‰，分析结果见表2.

表2云南中、小盆地样品烃类气体分析结果

		天然气中烃气归一化组分／%							$\delta^{13}C_{CO2}$／‰							CO_2同位素特征
		C_1	C_2	C_3	iC_4	nC_4	iC_5	nC_5	C_1	C_2	C_3	iC_4	nC_4	iC_5	nC_5	
YN96-2	保1井	99.93	0.054	0.018					-60.0	-46.5	-32.3					
96-3	保2井	99.93	0.055	0.01	0.002	0.002			-65.4	-48.7	-30.6					-20.2
96-4	保3井	99.90	0.07	0.02	0.012	0.003			-60.7	-42.0	-30.5	-22.9	-21.9			-9.4
96-5	杨3井	54.60							-71.3		-11.1					
06-6	7井	58.3	5.7	16.3	4.2	8.9	3.9	2.7	-53.8	-31.6	-31.0	-30.5	-30.1	-28.0	-28.6	-12.2
96-7	牛5-1井	91.1	1.6	2.95	0.89	1.8	0.89	0.7	-55.6	-34.6	-34.8	-35.2	-34.2	-30.2	-32.4	
96-8	牛2-6井	95.5	1.35	1.68	0.35	0.63	0.28	0.19	-55.2	-33.0	-34.6	-35.7	32.1	-28.1		-18.3

図 4　XML と Unicode を用いて記述した「中国科学」誌の論文の IE 5.0 による表示