

SGMLによる『情報管理』誌の冊子体・電子版同時作成の開始と全文検索の試み

新名真紀子*1, 森田歌子*1, 高木徹*2, 木谷強*2

1 科学技術振興事業団科学技術情報事業本部「情報管理」編集事務局
〒102-0081 千代田区四番町5-3

Tel:03-5214-8415, Fax:03-5214-8417, E-mail:editor@johokanri.jst.go.jp
2 (株) NTTデータ情報科学研究所

〒210-0913 川崎市幸区堀川町66-2

Tel:044-548-4606, Fax:044-548-4693, E-mail: {takaki,tkitani} @lit.rd.nttdata.co.jp

「概要」

SGMLによる文書作成・文書処理・編集・印刷の一貫システムを実現する『情報管理』誌 SGML 編集システムについて、実際の動作を中心に紹介する。このシステムは、『情報管理』誌 1999年4月号の編集・印刷作業から運用を開始する。また、4月号発行時より、冊子体と電子ジャーナルの同時提供を予定している。全文DBについては、96年4月号より冊子体発行後にSGML文書として蓄積しており、現在、全文検索の試みを行っているので、併せて紹介する。

「キーワード」

SGML, 編集システム, 全文検索

Parallel Publishing of Printed and Electronic Form of "JOHO KANRI" Using SGML, and a Tentative Full-text Search

SHIMMYO Makiko*1, MORITA Utako*1, TAKAKI Toru*2, KITANI Tsuyoshi*2

1 Japan Science and Technology Corporation, Information Center for Science and Technology, Editorial
Office Journal of Information Processing and Management

5-3, Yonban-cho, Chiyoda-ku, Tokyo, 102-0081

Phone:+81-3-5214-8415, Fax:+81-3-5214-8417, E-mail:editor@johokanri.jst.go.jp

2 Laboratory for Information Technology, NTT Data Corporation

66-2, Horikawa-cho, Saiwai-ku, Kawasaki-shi, 210-0913

Phone:+81-44-548-4606, Fax:+81-44-548-4693, E-mail: {takaki,tkitani} @lit.rd.nttdata.co.jp

「Abstract」

We introduce the workflow of "JOHO KANRI" publishing system which realizes SGML publishing system, straight line system of document production, document processing, editing, and printing. This system starts operation with the publication of "JOHO KANRI" April, 1999 issue that is supposed to be published both in printed form and electronic journal. We introduce a tentative full-text search system, since "JOHO KANRI" has stored up its full-text DB from 1996 issue in SGML document format.

「Keywords」

SGML, publishing system, full-text search

1. はじめに

科学技術振興事業団 (JST) は、SGML (Standard Generalized Markup Language) による情報の標準化の役割が重要であるとの見地から、1996年にJICST-SGML文書処理システムを開発した。『情報管理』誌では、このシステムを編集・印刷へ応用する試みとして、商用の組版ソフトウェア Interleaf5<SGML>で組版の試作を行い、96年10月の第8回「デジタル図書館」ワークショップにおいて報告した[1]。

その後、SGMLによる文書作成・文書処理・編集・印刷の一貫システムを実現する『情報管理』誌 SGML 編集システムを開発したので、本稿では実際の動作を中心に紹介する。

このシステムは、1999年4月号の編集・印刷作業から運用を開始する。また、4月号発行時より、冊子体と電子ジャーナルの同時提供を予定している。当初からの目的であった全文DBについては、96年4月号より冊子体発行後にSGML文書として蓄積しており、現在、全文検索の試みを行っているので、併せて紹介する。

2. 『情報管理』誌 SGML 編集システム

このシステムは、JICST-SGML文書処理システムと、今回開発した『情報管理』誌 SGML 編集システムの二つから構成されている。システムの全体構成を図1に示す。

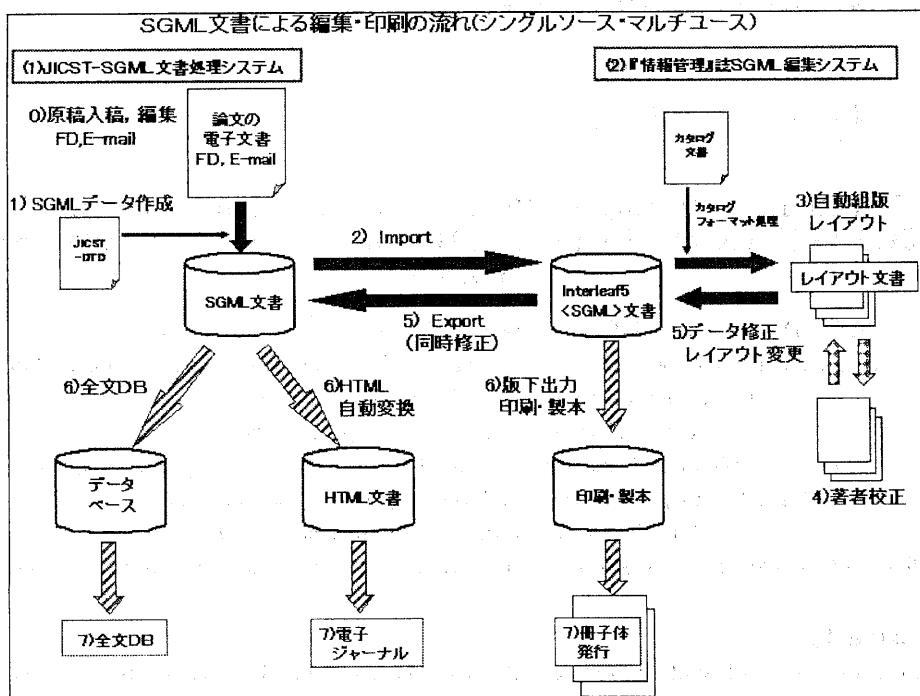


図1 SGML文書による編集・印刷の流れ

2. 1 SGML 移行の目的

SGML 移行の目的は以下のとおりである。

- (1) 学術論文流通の標準化
- (2) データの一元的管理：冊子体の編集・発行工程で作成したデータを、全文 DB、電子ジャーナルなどに展開する（冊子体、全文 DB、電子ジャーナルの同時作成）。
- (3) JICST-SGML 文書処理システムの実用化
- (4) 編集・印刷工程の作業量軽減・期間短縮：レイアウト、校正の繰り返しを減らし、校正ゲラの出力回数を減らすことで物理的な「もの」と「ひと」の動きを少なくする。これは印刷経費の削減にもつながる。
- (5) 経費の節減：冊子体作成経費の節減を図るとともに、データの同時作成を行うことにより、全文 DB、電子ジャーナルの作成に関わる経費を最低限に抑える。

2. 2 『情報管理』誌の編集・印刷に必要な組版ソフトウェアの機能

『情報管理』誌の編集・印刷に必要な組版ソフトウェアの機能を分析した結果、以下のような機能が必要であることが明らかになった。

- (1) データ作成の同時性：著者校正やレイアウトの変更などの編集・印刷データの修正を、元の SGML データに反映させることができること。
- (2) レイアウト機能：二段組みと一段組みの混在が可能で、図表のレイアウトが自由にできること。
- (3) 柱、脚注が扱えること。
- (4) コラムごとに異なったレイアウトデザインに対応可能のこと。
- (5) 日本語の学術文献に合ったフォントを使用できること。
- (6) イメージデータのみのページが混在できること。

以上の六つの機能について、SGML データ対応の日本語組版ソフトウェアを、調査・検討した結果、編集・印刷の基本的な条件に適応できるものとして、Interleaf5<SGML>を導入した。

2. 3 『情報管理』誌 SGML 編集システムの機能概要

『情報管理』誌 SGML 編集システムは、WS (ワークステーション) 上で動作する Interleaf5 の制御言語である Interleaf Lisp で構築したシステムである。また、開発には Interleaf5<SGML ToolKit>を利用している。ソフトウェア構成は、システム構築のために導入した市販ソフトウェアと、JICST - DTD に対応するために開発したソフトウェアからなる。導入ソフトウェア一覧を表 1 に、開発ソフトウェア一覧を表 2 に示す。

このシステムの機能としては、次の三つがある。

- (1) Import 機能：JICST - DTD に沿った SGML 文書 (DTD の規約に基づいて記述された SGML インスタンス) から、Interleaf5<SGML>文書 (Interleaf5 文書に SGML タグ情報が埋め込まれている) に変換する機能。
- (2) カタログによるフォーマット変換機能：Import した Interleaf5<SGML>文書を、記事の種類別に書式情報を変換する機能。このシステムでは、「論文用」、「その他の記事用」；「奥付け用（編集後記）」、「目次用」の 4 種類のレイアウト用カタログを作成した。
- (3) Export 機能：Interleaf5<SGML>文書を、JICST - DTD に沿った SGML 文書に変換する機能。

表1 導入ソフト一覧

利用ソフト	分類	内容	備考
Interleaf5 基本	製品	Interleaf5 の基本操作に関する製品	
Interleaf5 パック放り出オプション	製品	大量文書管理オプション	
Interleaf5<SGML オプション>	製品	SGML の Import/Export オプション	Leafdoc のみ
Interleaf5<SGML Toolkit>	製品	DTD 組み込み、 Import/Export 処理拡張用オプション	
Perl V5.001	Free ware	Import の前処理、Export の後処理時に キート変換処理で利用	

表2 開発ソフト一覧

利用ソフト	分類	内容
情報管理誌 SGML 文書編集ツール	開発	JICST-DTD 用 SGML 文書からの Import 处理 JICST-DTD 用 SGML 文書への Export 处理 JICST-DTD の Interleaf5 フィルネットへの割付 カタログによるフォーマット処理 Import の前処理 Export の後処理

2. 4 編集作業の流れ

編集作業は、以下のような流れで進められる。

- (1) PC で作成した 1 号分の SGML 文書を、Interleaf5 の編集システムで利用するために、編集作業用の WS に集める (PC からのファイル転送、FD など補助記憶装置の利用等)。
- (2) SGML 文書を 1 号単位に管理するため、発行月のフォルダを作成し、その中に記事単位のフォルダを作成する (図 2)。
- (3) 記事単位に Import 处理を行い、SGML 文書を Interleaf5 文書に変換する。
- (4) 記事単位に記事の種類別のカタログによるフォーマット変換をし、レイアウト・編集を行う (図 3)。
- (5) 著者校正用に、レイアウト・編集したものに出力し、著者校正を依頼する。
- (6) 記事単位に、Interleaf5<SGML>の基本編集機能を使い、著者校正の赤字や図表等の変更、レイアウトの変更などの修正を行う。

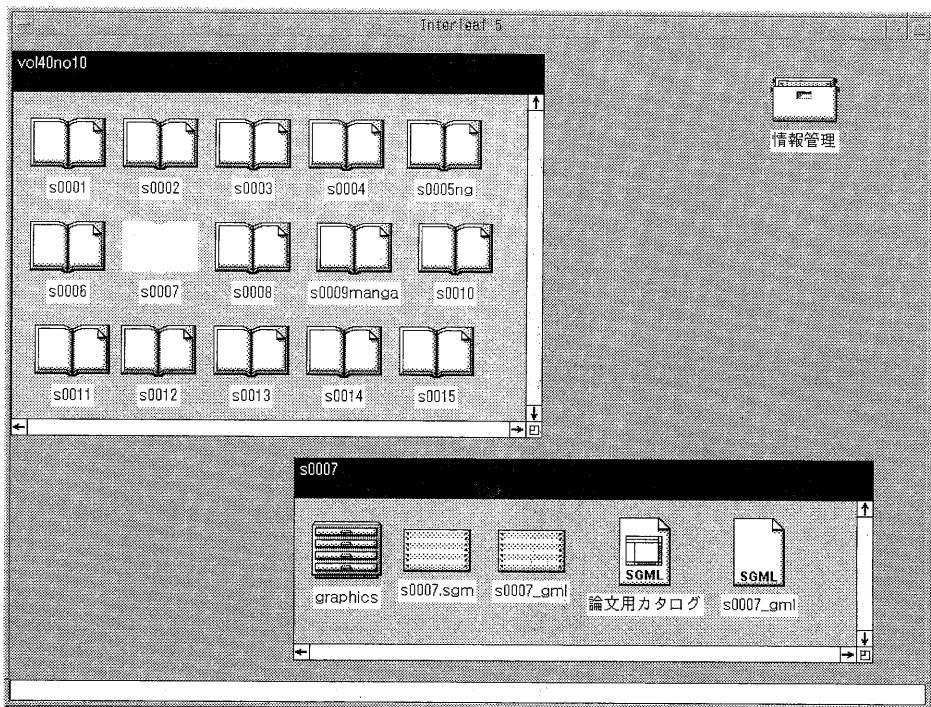


図 2 1号分の各記事単位の SGML 文書

(7) 最終版下データ管理のための版下用フォルダを新たに作成し、編集作業がすべて終わった記事を、このフォルダに集める。ここで台割（雑誌 1号分の記事配置）を行う。

(8) Interleaf5 の基本編集機能を使い、表紙、奥付け、目次を作成する。これで印刷用版下のデータ作成は終了する（図 4）。

(9) 編集工程での文字修正等を、元の SGML 文書に反映させるため、記事単位に Export を行う。この処理を行うことにより、冊子体と同一の SGML データができ、重複した修正作業を行うことなく、全文 DB・電子ジャーナルとして利用できる。

以上の編集作業はすべて WS 上で行うが、その作業画面の一部の例を紹介する。図 2 は、1号分の記事単位の SGML 文書を格納したフォルダの内容を示す画面、図 3 は、一記事のレイアウト編集作業を行っている画面、図 4 は、1号分の編集が終了した版下用フォルダの内容を示す画面である。

3. 『情報管理』誌全文検索プロトタイプシステム

SGML 編集システムで作成した冊子体と同一の SGML データに対して全文検索プロトタイプシステムを構築した。この『情報管理』誌全文検索プロトタイプシステムは、全文検索エンジンとして OpenText を使用し、Wed サーバを経由して全文検索機能を実現している。

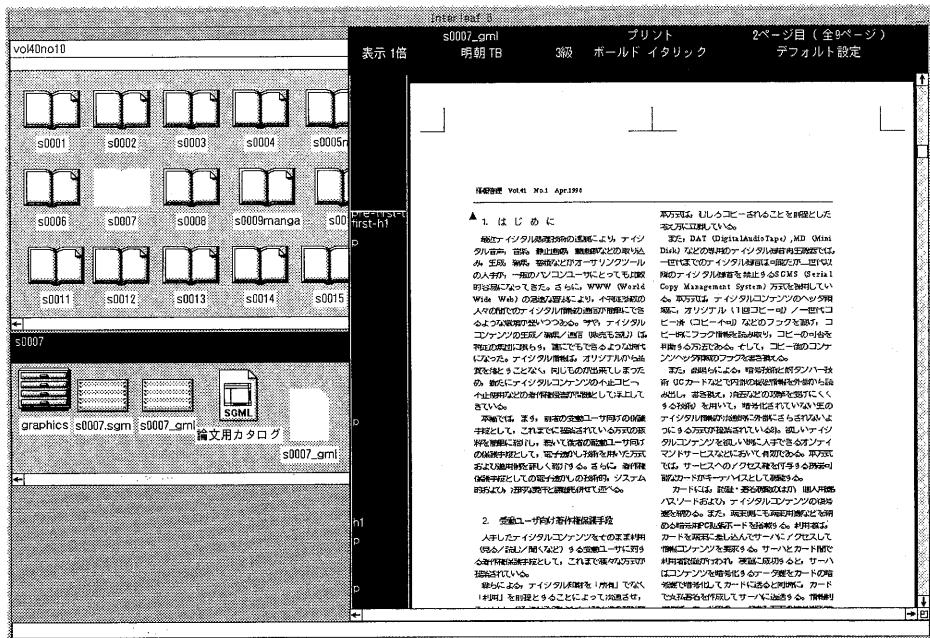


図 3 レイアウト画面

3.1 全文検索システム構築の前提条件

SGML データへの移行の目的として、データの一元管理がある。本システムも、文書作成からの一連のデータ処理で全文検索を実現できることを念頭に開発した。本開発の前提条件は次のとおりである。

- (1) JICST-DTD で定義された文書構造に基づく検索が可能であること。
- (2) 編集過程で生成された SGML データを加工することなく検索に使用すること。

3.2 全文検索システムの開発

SGML データのファイル構成は、図 5 に示すように発行年月、文献ごとに階層化された構造になっている。また、各文献ファイルは、SGML データ(拡張子 sgm), 一太郎ファイル(同 jbw), 図表などの tiff 形式のイメージファイル(同 tif) から構成されている。

本システムでは、『情報管理』誌 2 年分(1996~1997 年度発行分)の 304 文献を検索対象とした。システム構築に用いた文献データは、SGML データと tiff 形式のイメージデータを用いた。304 文献のデータサイズは、SGML データが 5.0MB(304 ファイル), イメージファイルが 281.9MB(1,006 ファイル)である。

全文検索エンジン OpenText により全文検索用のインデックスを作成した。インデックス生成は SGML データのみを対象とし、次の二つのステップで実施した。

(a) 文字単位のインデックス生成

テキストの文字情報を高速に検索するためにインデックスを作成する。

(b) DTD の構造のインデックス生成

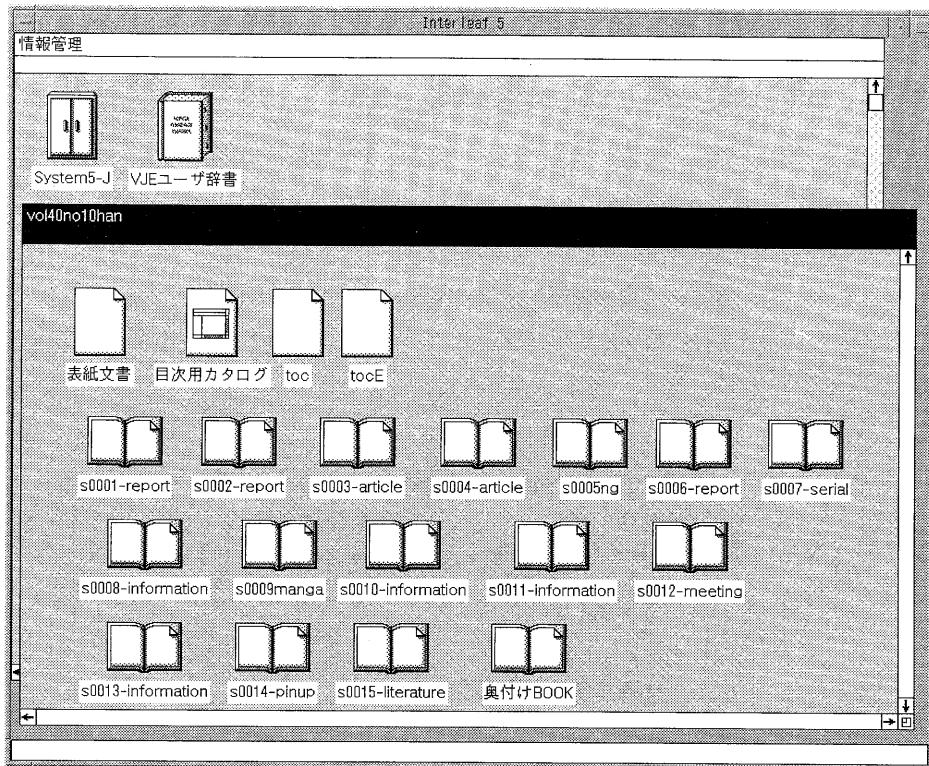


図4 1号分の編集が終了した版下

DTD 内に定義されている文書構造を自動的にインデックスに反映させ、(a)でインデックスしたテキストが、文書構造のどの部分に出現しているかという情報をインデックスとして保持する。

(a) と (b) のインデックスを生成することにより、文書構造を考慮した検索が可能となる。なお、本システムのインデックス作成時間は、(a)20秒、(b)16秒の合計36秒であった(SUN Enterprise3500、メモリ1GBを使用)。また、作成されたインデックスのサイズは10.7MBである。

検索時は cgi プログラムから OpenText をコールし、生成された全文インデックスを検索することで全文検索を実現した。なお、テキスト表示時にはインデックス前の SGML データを参照する。図6 に本システムの全文 DB データの構成を示す。

3.3 全文検索システム機能

本システムは、以下の検索機能を実現している。

(1) 書誌情報や、文書内の構造を指定した全文検索

検索項目として、「タイトル」、「本文」などを文書の構造単位で指定でき、その部分に出現する単語を指定することで、条件に適合する文献を検索する。

(例) タイトル中に検索語「情報検索」を含む文献を検索する。

(2) SGML タグに付与されている属性に対する検索

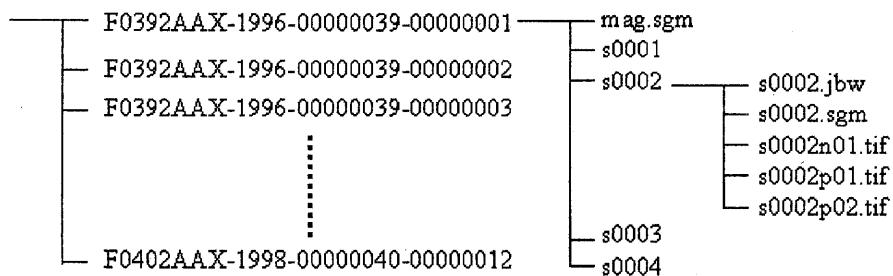


図 5 「情報管理」誌 SGML データのファイル構成

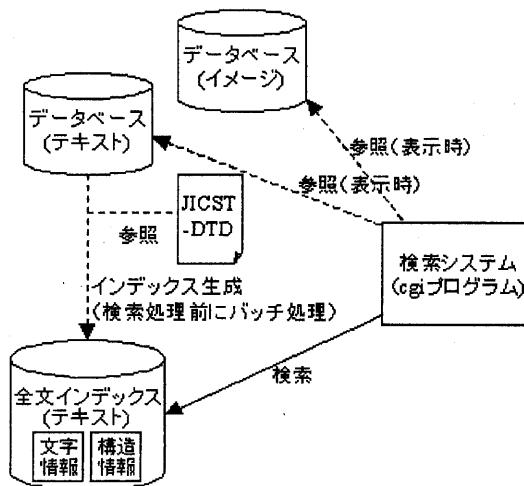


図 6 全文 DB 検索システムのデータ構成

『情報管理』誌では、論文の種別が paper タグ内の属性 type に記述されている。

(例) 論文種別が「講演」である文献のうち、検索語「電子新聞」を含む文献を検索する。

(3) 検索論理条件を指定した検索

和 (OR), 積 (AND), 差 (NOT) の各論理演算を実現した。

(4) 図表等のイメージファイルとのリンク

SGML データ内に記述されているイメージデータのファイル名を用いて、本文表示時にダイナミックにイメージファイルとリンクさせ表示を可能としている。

図 7 に本プロトタイプシステムの検索指示画面イメージ、図 8 に検索結果の表示画面イメージを示す。

3.4 検討課題

プロトタイプシステムの開発の中で明らかになった検討課題を述べる。

(1) 発行年月と巻号データの欠落

検索実行	検索履歴	概念検索	クリア																																						
<table border="1"> <thead> <tr> <th>論理条件</th> <th>検索項目</th> <th>検索語</th> <th></th> </tr> </thead> <tbody> <tr> <td>OR</td> <td>本文</td> <td>全文</td> <td>概念</td> </tr> <tr> <td>AND</td> <td>タイトル</td> <td>情報検索</td> <td>概念</td> </tr> <tr> <td></td> <td>全体</td> <td></td> <td>概念</td> </tr> </tbody> </table> <table border="1"> <tr> <td>論文種別:</td> <td> 全て(all) 講演(lecture) インタビュー(interview) 海外通信(overseas) 講座(serial) JICST通信(JICSTnews) </td> </tr> </table> <table border="1"> <tr> <td><input checked="" type="radio"/> 発行年月:</td> <td>1996</td> <td>年</td> <td>10</td> <td>月</td> <td>～</td> <td>1997</td> <td>年</td> <td>3</td> <td>月</td> </tr> <tr> <td><input type="radio"/> 発行巻号:</td> <td>39</td> <td>卷</td> <td>1</td> <td>号</td> <td>～</td> <td>40</td> <td>卷</td> <td>12</td> <td>号</td> </tr> </table>				論理条件	検索項目	検索語		OR	本文	全文	概念	AND	タイトル	情報検索	概念		全体		概念	論文種別:	全て(all) 講演(lecture) インタビュー(interview) 海外通信(overseas) 講座(serial) JICST通信(JICSTnews)	<input checked="" type="radio"/> 発行年月:	1996	年	10	月	～	1997	年	3	月	<input type="radio"/> 発行巻号:	39	卷	1	号	～	40	卷	12	号
論理条件	検索項目	検索語																																							
OR	本文	全文	概念																																						
AND	タイトル	情報検索	概念																																						
	全体		概念																																						
論文種別:	全て(all) 講演(lecture) インタビュー(interview) 海外通信(overseas) 講座(serial) JICST通信(JICSTnews)																																								
<input checked="" type="radio"/> 発行年月:	1996	年	10	月	～	1997	年	3	月																																
<input type="radio"/> 発行巻号:	39	卷	1	号	～	40	卷	12	号																																

図7 全文検索プロトタイプシステムの画面（検索条件入力）

発行年月と巻号が JICST-DTD 定義ではなく、SGML データに記述されていない。編集段階では、文献の発行年月は決定されるので、DTD 定義を追加する方法や、自動的に SGML 格納ディレクトリ名からデータを生成し、SGML データに追加することで実現可能である。

(2) 異なる複数の DTD による SGML データの、同一システムでの検索

JICST-DTD も、今回用いた「基本ドキュメント DTD」のほか、「二次情報 DTD」「雑誌 DTD」「予稿集 DTD」がある。一般的に、異なる DTD による SGML データを横断的に検索する場合、文書構造の規定が異なるため、特定の DTD には存在しない情報があったり、論理的には同一の情報であっても異なる SGML タグ名が付与されている場合がある。その場合、異なる DTD 間のタグ名対応テーブルを作成するなどの対処が必要となる。

(3) 全角(2バイト)文字・半角(1バイト)文字の混在

現状の「情報管理」誌 SGML データでは、英単語の記述が 1 バイト文字と 2 バイト文字の場合がそれぞれあり、その使い分け方法が統一されていない。英数字の 2 バイト文字は、1 バイト文字として全文検索用のインデックスを作成することにより、検索漏れは防ぐことができる。その際、検索語に 2 バイトの英数字文字列が入力された場合、1 バイト文字に変換する必要がある。

4. 今後の課題

『情報管理』誌 SGML 編集システムの運用を開始することで、編集者は作業量の軽減と期間の短縮になり、充実した誌面作りに多くの時間と労力を注ぐことができるようになる。原稿を執筆する著者にとっても、システム変更に対する抵抗がないよう、できるだけ簡単な執筆規定を作ることを考えている。

また SGML 編集・印刷の移行に伴い、印刷所の作業も大きく変化する。従来の、紙の原稿をもとにした写植、表組み、トレース、校正等の作業は、電子データを使ったコンピュータ上でオペレーション作業

検索条件		情報検索	次ページ
順位	スコア	タイトル	
21	500	インターネット活用法 The effective way of using academic databases the Internet	
22	500	NACISIS	
23	500	海外文献紹介	
24	500	JICS(化合物辞書データ)のMOLfileへの変換 Conversion of the data of JICS(Chemical Dictionary Database) into MOLfile	
25	500	参考資料 公共事業からの情報サイト	
26	500	海外文献紹介	
27	500	米国国立医学図書館(NLM)(National Library of Medicine)	
28	500	海外文献紹介	
29	500	海外文献紹介 Science Citation Index(SCISEARCH)の活用(2Romen1) Effective use of Science Citation Index (SCISEARCH)	
30	500	(2Romen1)	

図8 全文検索プロトタイプシステムの画面（検索結果一覧、本文表示）

になる。印刷した校正ゲラのやりとりもなくなり、ネットワークを介してデータを共有しながら工程を進めることになる。その変化に対応できるよう、印刷業界への啓蒙・普及を図ることが大きな課題である。

我が国における SGML 化の取り組みが、国レベルで検討されている状況から見ても、システムの一層のレベルアップと操作の簡便性を向上させることも重要である。今後、実際に『情報管理』誌 SGML 編集システムにより本誌の編集・発行を行いつつ、これらの課題に対処していきたい。

参考文献

- [1] 森田歌子, 新名真紀子, 鈴木政彦, 石黒裕康. 論文執筆と編集のためのツールとしての SGML -『情報管理』冊子体と全文 DB の同時発行に向けて. デジタル図書館. No.8, 33-43 (1996)
- [2] 石黒裕康, 千葉博, 森田歌子. SGML 文書作成プロトタイプシステム. 第 33 回情報科学技術研究集会発表論文集. p.153-162 (1997)
- [3] 森田歌子, 石黒裕康, 千葉吉一.『情報管理』誌 SGML 編集システム－冊子体, 全文 DB・電子ジャーナルの同時作成－. 情報管理. 41(6)445-459(1998)
- [4] 木谷強, 相原理, 高木徹. 全文データベースの事例紹介. 情報管理. 41(6)460-470(1998)