

学術論文 CD-ROM システム

齋藤 雅

大日本印刷株式会社

Abstract

近年、大学の図書館における学術情報の新しいサービス形態の模索が行われつつある。今回、その一つの試みとして学術論文の電子メディアサービスによる学術情報の流通と活性化を目的とした学術論文 CD-ROM システムを開発した。従来より印刷物として流通している学術論文のフルテキストデータベース化を行い CD-ROM に格納化した。このシステムにより、論文の調査、情報収集を迅速に行い、本来の研究時間を有効に活用できるようになる。ここでは、論文のフルテキストデータベースの構築と今回開発したシステムの検索・表示の概要を述べる。

The CD-ROM system for paper

Tadashi Saito

DAI NIPPON Printing Co.,Ltd.

Abstract

New services of a science information have recently been critical matters of discussion in the library of universities. Through out of new trials, we developed the CD-ROM system for paper. We selected a CD-ROM to suit for the purpose of distribution and activation of a science information. This paper describes the construction of a full-text database and the outline of retrieving and browsing a paper in this system.

1 はじめに

大学図書館における学術情報のサービス機能として、各種の情報の収集・保存・管理を主体とした冊子体による情報提供から、蓄積された情報の有効利用と新しい学術情報の情報提供サービスが重要視されるようになってきた。図書館に集められる学術情報が、研究者の研究情報としてより柔軟に活用できる機能の提供を行い、新しい情報発信を担うために図書館の役割は、ますます重要になっている。

従来から、研究論文は、活字として印刷された冊子体によって配布されてきた。しかしながら研究者が、活字化された学術情報の中から自分の研究テーマに関連する情報を探し出すことは、非常に困難になってきている。一方、分野によっては、書誌情報のデータベース化が進み、オンラインでの情報検索が行われてきている。このような状況の中、情報化社会の急速な進展に対して、図書館が主体となって行える新しい情報提供・情報発信機能が要求されるとともに新しい図書館のあり方が模索されつつある。

我々は、図書館の新しい情報サービスとして、慶應義塾大学図書館（研究・教育情報センター）と共同で学術論文の全文（フルテキスト）データのサービスが可能な学術論文 CD-ROM システムを開発した。学術論文 CD-ROM は、活字として提供されている学術情報に対して、CD-ROM を利用した電子出版による学術情報の流通と活性化を目的として開発を行った。このシステムにより、研究者が自分の必要とする情報をより早く得られる機能を持った論文データの 1 次情報（全文）サービスの実用化を可能にした。

2 学術論文 CD-ROM の設計

学術論文の CD-ROM 化は、学術論文のフルテキストデータベース化を前提とした CD-ROM システムとして設計した。大きな方針としては、データの管理・分類に重点を置かず、実際に使用する研究者にとって必要な機能に焦点を合わせた。設計方針としては、以下の 3 項目の実現を目指して行った。

- フルテキストデータベースの構築と CD-ROM 化
- 論文データの電子化と標準化
- 研究者向けの論文調査手段の提供

学術論文 CD-ROM は、従来のデータベースより、さらに書籍そのものに近い情報データベースといえる。このデータベースでは、書籍の持つ品質・機能より、情報入手の方法を提供し、論文調査の効率を向上させ、さらに新しい論文調査手段を与えてくれるものである。よって、このシステムでは、論文検索と該当論文の内容調査が効率よく、早く行えるための手段の提供を目指しているが、書籍と同様に論文を読むためのシステムではない。

論文データは、テキストデータとイメージデータに大きく分けて電子化した。テキストデータは将来国内の学術情報が、電子メディアとして流通し、また海外にも提供されることを考慮し、国際標準としての規格が整備されつつある SGML (Standard Generalized Markup Language) の記述形式に従い電子化を行なった。図・表などは、イメージデータとして TIFF 形式により電子化を行った。データは、各論文単位に管理をし、CD-ROM 上においても検索インデックス以外は、論文ごとに管理している。

テキストデータは、書誌データと本文データを合わせて一つのファイルにした。このファイルから検索に必要な各種キーワードファイルの作成および検索実行後に表示されるテキストデータの作成とイメージデータのリンク情報の抽出や編集を行った。

3 学術論文の全文データベース化

論文をデータベース化し、情報検索に利用するために、まず論文データをフルテキストの形式で電子化することにした。フルテキストのデータベース化は、SGMLの規格に基づいてデータベース化した。各論文は、その論文を構成する論理的な要素ごとにSGMLのタグを設定し、これらのタグにより内容の識別を与えフルテキストの電子データとした。図・表は、イメージデータとしてスキャナから取り込み TIFF形式のファイルとした。論文テキスト内の図・表番号とイメージデータとして取り込んだ図・表データとは、SGMLのタグによりテキストデータから図・表データが参照できるようにリンク情報を持たせている。

図1に、論文を構成する要素で表した論文データの構造例を示す。

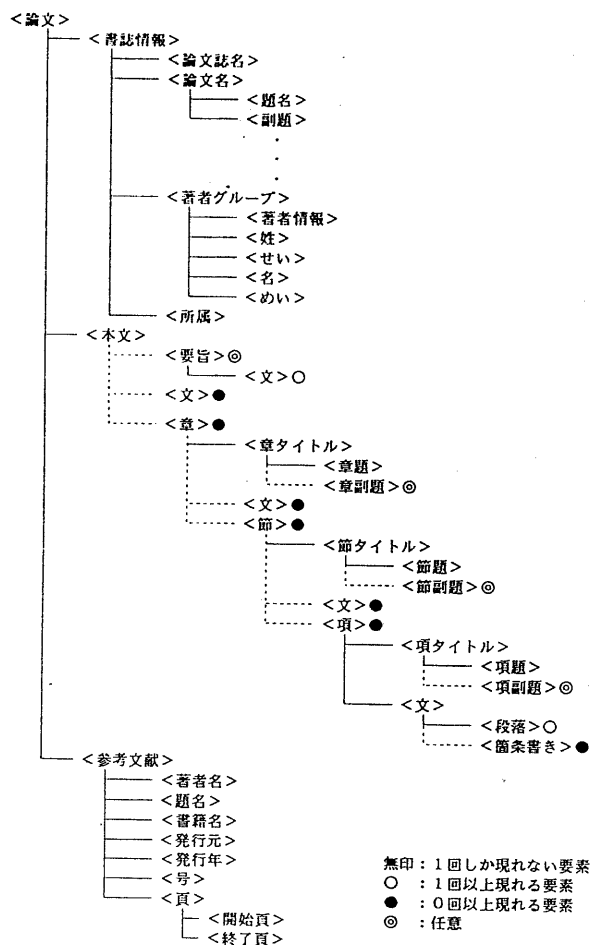


図1 論文構成要素のデータ構造例

図2には、論文データをSGMLで記述したテキストデータの例を示す。先頭の部分は、この論文の書誌データの記述であり、その後本文が続いている。テキストデータは、論文を構成する要素であるタイトル、章、節、箇条書き、脚注などの論理的な構造が、`< >` や`</ >`のタグによって囲まれ、構造化されている。`< >`が始めのタグを表し、`</ >`が終わりのタグを表している。

図・表のイメージデータは、1論文1ファイルのTIFF形式のファイルになっている。1つのTIFFファイルには、複数のイメージデータを格納してある。

```

<論文>
<書誌情報>
<論文名><题名>文書作成環境</题名><副題>-</副題></論文名>
<著者グループ>
<著者情報><姓>斎藤</姓><せい>さいとう</せい><名>雅</名>
<めい>ただし</めい></著者情報>
<所属>大日本印刷株式会社 CTS事業部</所属>
</著者グループ>
</書誌情報>
<本文>
<要旨>
<文><段落> 著者が文書を作成する作業を出版、印刷の立場から支援する文書作成環境をワークステーション上に実現した。一般のWP、DTPが文書作成するためのツールとして実現された機能では、正しい日本語の文書編集を行う知識をユーザに求められる。よって、著者は文書を作成する作業とともに編集、レイアウトの煩雑な作業をすることになる。ここでは、文書を作成する作業を支援し、高品質な日本語文書を作成するシステムを開発したので、その特徴を述べる。</段落></文>
</要旨>
<章>
<章タイトル><章題>1. はじめに</章題></章タイトル>
<文><段落> 文書を作成するという作業は、人間における一つの知的生産作業である。人間は知識の伝達手段として情報を正しく正確に伝える為に文書を作成する。そして印刷による情報の大量複製と伝達が可能になり、文書として作成された内容をより正確にまたより読み易くするための研究が編集という作業を通して行われてきた。
文書の編集とは、単に文字列の組版の機能ではなく、日本語の文書としてより美しく読み易いページ編集が必要である。このような背景のなかで文書を作成する知的作業を編集の知識と印刷の機能をあわせて支援する環境として「やまぶき」の開発を行った。
</段落></文></章>
<章>
<章タイトル><章題>2. 文書作成の要件</章題></章タイトル>
<文><段落> 一般に文書を作成すること特に出版される書籍を例に考えれば、文書の作成とは、文書の内容を作成する著者、作成された情報を紙面上でいかに読み易く、分かりやすく表現するか、また文書の番号の付け方や言葉の表現が論理的にまちがっていないかをチェックし、編集レイアウトを決定する編集者、そして実際に文書を組版し紙面上に印刷を行う印刷所の3者の共同作業ともいえる。よって、文書を作成する著者自ら使うことのできる文書作成システムは、編集と印刷の双方を兼ね備えることが必要となる。われわれは日本語の編集知識を持たなくても十分に美しい文書、特に学術論文が作成できる環境として次のような必要項目を考えた。
<箇条書き>
・誰でも簡単に作業ができるユーザインタフェイスを持つこと</箇条書き>
<箇条書き>
・文書の編集レイアウトの知識をシステムがサポートすること</箇条書き>
<箇条書き>
・ユーザの作業を自動化し支援すること</箇条書き>

```

図2 論文データのSGML記述例

4 論文検索の方法

論文の検索は、タイトル、要旨、および著者が付与したキーワードからのキーワード検索と論文の書誌データからの検索が行える。各検索項目ごとに、ヘルプ機能を設けている。ヘルプ機能は、キーボード入力されたある語からその語を含むキーワードを検索インデックスからサーチして一覧表示する機能である。このヘルプにより検索キーワードの指定を的確に行うことが可能になる。ヘルプ機能は、検索対象となるキーワード、著者名、などの検索条件入力フィールドごとに一覧表示を行い、該当キーワードの選択をサポートする。たとえばキーワード検索において、「企業」という言葉に対してヘルプ機能を行うと、「企業戦略」・「日本企業」・「後発企業」・「企業防衛」などの「企業」を含むキーワード文字列の一覧を表示する。よってユーザは、このキーワード一覧の中から一つまたは複数のキーワードを選択して検索を行う。検索は、完全一致検索を行うが、AND検索・OR検索の論理検索も行える。

検索実行後、対象論文のタイトル一覧が表示される。ここで、さらに検索論文の候補を絞り込むために2次検索の機能を設けた。2次検索では、さきに検索された論文を対象として任意の文字列をキーワードとして、このキーワードが論文の本文に存在するかどうかをサーチする。その結果は、論文タイトル一覧の画面上にマークとして表示する。ユーザは、このマークを参考にして調査する論文の絞り込みと調査の順序を事前に決めることができる。

5 システムの機能

検索された論文を表示する手段として、次のような機能を開発した。

1. 書誌データの表示
2. 要旨データの表示
3. 目次の表示と目次からの本文表示
4. 本文中からの図・表、脚注表示
5. 図・表目次の表示と図・表目次からの図・表データ表示
6. 文中文字列サーチ
7. 表示メディアに合わせたテキスト編集機能

検索実行から各データの表示に至る検索・表示系のシステム概要を図3に示す。

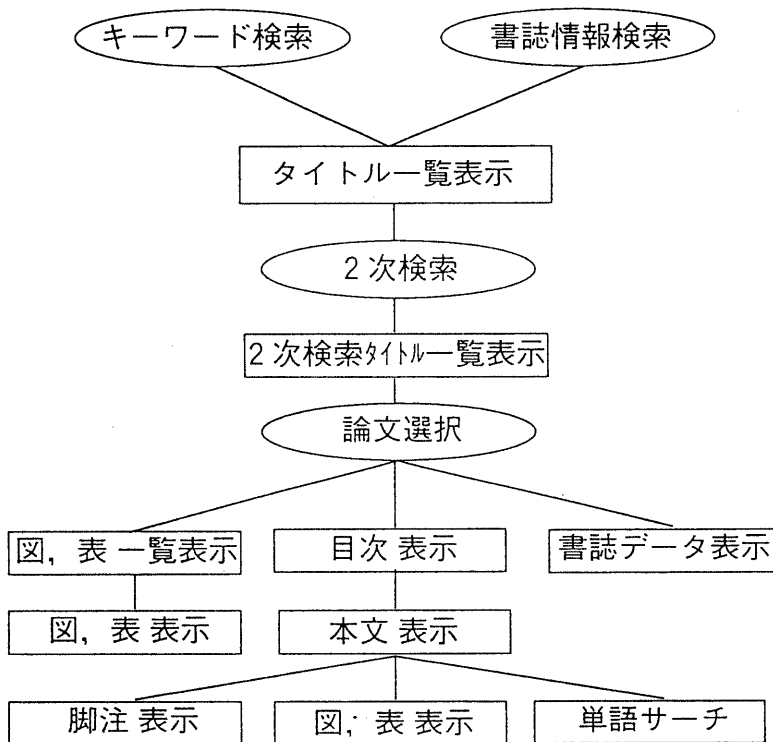


図3 検索・表示系のシステム概要

論文検索の実行により、論文タイトル一覧が表示される。この論文の中から任意の論文を選択すると、論文の書誌データや要旨の表示が行える。また論文の本文表示では、まず論文の目次が表示される。この目次から自分が見たい章や節の見出しを選択すると、その該当本文が表示される。ここで、今表示されているテキストデータに図・表・脚注がある時は、図・表・脚注のアイコンが画面下に表示され、このアイコンを選択すると図・表は、イメージとして、脚注は、テキストとして本文の上に表示される。

論文の概要を大ざっぱに見る手段として、図・表一覧の表示機能がある。これは、図・表の番号とキャプションを図・表目次として表示し該当論文に含まれる図・表を一度に見ることができる。そして必要なら図・表目次の任意の図・表データの表示や図・表データの連続した表示を行い、短時間で全体の概要を把握する手段として使用される。

表示メディアに合わせた編集機能とは、論文の表示は、できるだけ活字で印刷された時と同じように編集された方が、読み易い。しかし画面で表示したり、プリントしたりする時、活字と同じ様な編集はできない。そこで、論理的な意味として付けたタグを利用して簡単な編集を行えるようにした。各タグごとに表示するときの位置（レイアウト）を指定できるようにした。例えば、タイトルのタグで囲まれた文字列はセンタリングする指定によって、タイトルの表示は、画面の左右中央に表示することが可能になる。この機能のもう一つの利点は、表示する画面のサイズや大きさが、使用する計算機のディスプレイにより異なったり、また印刷するプリンタの用紙サイズなど印字サイズが異なる

るときに一行当たりの文字数を変更して最適なレイアウトで表示したり、印字したりすることを可能にしている。但し、実際のシステム上では、パソコンの標準的な画面を想定してデフォルトの値を決めて編集している。

6 まとめ

学術論文 CD-ROM システムは、書籍の代わりではなく、書籍を補完する道具であり、機能であると思われる。すなわち今回開発した CD-ROM システムは、情報を読むためのシステムではなく情報をいかに早く、容易に調査する手段を提供するかに重点がおかれているためでもある。この CD-ROM システムは、学術情報の流通と活性化を図るためにメディアとして CD-ROM を利用した。そして、研究者自ら論文検索・調査が行えることを目指した。一方図書館の中での情報サービスの一つとして利用する公共的な使われ方もある。個人が使用するシステムと公共で使用するシステムは情報としての提供メディアは同じであっても利用する環境によってシステムの形態が色々に変化し適応していかなければならないと考えている。また、論文情報の検索・調査のために CD-ROM システムが普及し、有効に利用されるためには、やはり多くの研究論文の収録が重要である。

参考文献

- [1] Martin Bryan, SGML An Author's Guide to the Standard Generalized markup Language, (Addison-Wesley, 1988).
- [2] Eric van Herwijnen, 実践 SGML, (日本規格協会, 1992) .