

情報処理学会情報学基礎研究会資料(1993年5月18日)

## 学術文献を対象とした電子図書館システムの構成法

安達 淳 橋爪 宏達 高須 淳宏

学術情報センター研究開発部

学術論文を対象とした電子図書館システムのための一実現方法を概括している。論文のページを画像情報として蓄積し、これを利用者が直接検索表示できるようなシステムであるが、これは従来の文献の二次情報検索機能をも統合した、総合的な学術情報サービスシステムの実現を考えることもできる。筆者等はこのシステムを client-server モデルにそった分散処理システムとして、高速の広域ネットワーク上で稼働することを前提として実装を進めている。冊子体のページのブラウズを重視した画像データベースを基本設計方針として採用し、また既存の二次情報データベースや雑誌目録所在情報データベースとの調和のとれた検索機能の実現をめざしている。プロトコルには Z39.50 を拡張したものを使用し、利用者インターフェースは Open Look を用いて実現している。

## Design of an Electronic Library System for Scientific Documents

Jun Adachi Hiromichi Hashizume Atsuhiro Takasu

NACSIS, Research and Development Department  
(National Center for Science Information Systems)

This paper describes a design of an electronic library system for dissemination of scientific and technological documents. Pages of journals are stored in the form of digitized images and the images are sent to user's workstations via internet on requests. This system is a prototype of next-generation information service system that integrates and replaces conventional online information retrieval systems.

The authors are implementing the system in a distributed processing environment including high-speed networks, based on a client-server model. Featuring page-browsing capability, the prototype enables harmonized utilization of multiple conventional databases and images databases. An extended version of ANSI Z39.50 protocol is employed and Open Look is used as a basis of the graphical user interface.

## 1 まえがき

電子媒体による情報蓄積とその処理の一般化が進行していく中で、「電子図書館」に対する期待が高まっている[1]。この背景には、電子出版の発展や CD-ROM を代表とするオンライン情報検索システムの普及とともに、インターネットなど広域研究ネットワークの発展によるネットワーク経由のデータ配布の容易化もあり、両者をどう統合するかはまだ定説はない。米国におけるプロジェクトを見ても、「electronic library」、「digital library」、「virtual library」、「electronic document delivery」などさまざまな言葉が使われている。

本稿では、筆者らが現在試作している情報システムについて、その設計思想と稼働中のプロトタイプシステムを紹介する。このシステムでは、主に学術的な情報に限って考えており、機械可読化された学術論文が「電子図書館」での蓄積の対象となる。また、蓄積された文書情報は、高速の広域ネットワークを介して直接利用者のワークステーション等で検索・表示・印刷することが可能のようにシステムを設計している。したがって、従来からの document delivery service や図書館で行われている文献複写サービスを包含し、代替する機能を有するものといえる。

本稿で述べるオンライン電子図書館システムの基本的特性は以上のようなものであるが、まず第2節ではシステムの置かれる環境と基本的な構想について述べる。第3節では核となる画像データベース、第4節ではシステム実装方針について紹介し、最後に今後の進め方について述べたい。

## 2 システムのサービス機能

### 2.1 基本コンセプト

基本的な構想としては、センターの中核システムと利用者のワークステーションをネット

ワークで結合した client-server 型の構成をとっている。サービスセンター側のシステムは、学術図書や雑誌等のすべてのページを画像データベースに蓄積し、オンラインで利用者に情報提供するような画像データベースを中核に位置付けており、一次情報の配布には広域の広域ネットワーク (WAN) を介し、利用者に対して直接的に行えるような形態を考えている。図1にシステム全体像を示している。図中 B の部分に学術雑誌等のページをラスター情報で蓄積し、さまざまな広域ネットワークや LAN を通して利用者に直接原文献を送り届けようとするものである。

### 2.2 一次情報の取り扱い

雑誌のページなどの一次情報を取り扱うには、一般に、

- i) scanning によってデジタル化されたラスター情報。欧米では 300dpi がよく使われるが、日本の国情では 400dpi の方が適していると考えられる。
- ii) PostScript などのページ記述言語。
- iii) ASCII テキスト情報に基づくもの。TeX や SGML 等。

などの記憶形式が考えられる。それによるとデータ量の概数を表1に示した。

**全文情報の取り扱い** 今後は第三の全文型の一次情報が着実に増加すると考えられる[2]。したがって、将来の電子図書館システムではこの形式を基本とすること期待される。しかし、現在のところ SGML の普及の見通しに始まり、図表の取り扱いなど難問が多いのも現実である。また、大規模データベースとしての需要や利用方法についても不明の点が多い。本プロトタイプでは今のところ二次情報の延長線上に全文データを位置付けている。

**画像情報の取り扱い** 現在のプロトタイプでは、ページの情報をラスターデータの形式で扱うのが基本としている(文献[3]のシステムでも同様のアプローチである)。前述のラスターと PostScript の違いは、画像・図形情報

## センター側電子図書館サーバシステム

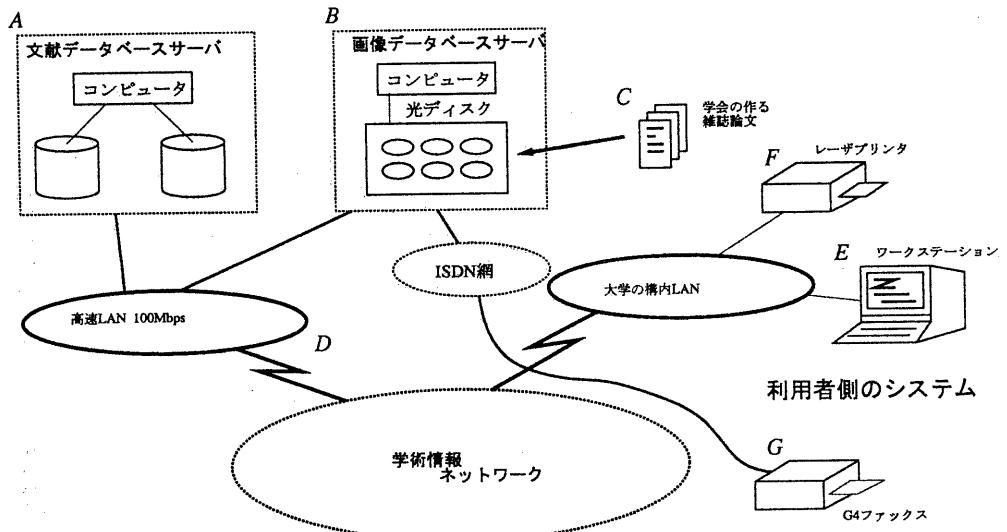


図1 システムとネットワークの概念図

表1 画像データの大きさの例

蓄積形態	データ量
ラスターデータ(1ページ) (A4, 400dpi, 非圧縮)	約 2 MB
ラスターデータ(1ページ) (A4, 400dpi, G4 MMR)	約 100 KB
本稿(全ページのPostScript)	約 950 KB
本稿(ASCII部分のみ)	約 27 KB
本稿 (すべての図のPostScript)	約 740 KB

の蓄積形式(メディア)の相違と考えられ、プロトタイプでは両方に対応可能にする方針である(蓄積形式としてはTIFFを採用し、現在はラスターデータのみを実装)。

現実的な対応としては、ある程度広範囲の情報源を対象にデータベース構築しようとした時、画像を採用する方法以外に実用的な手

法がないことによる。すなわち、すべての学術雑誌や会議録をすぐにTeXやSGMLにすることは難しい。さらに、サービス実用化段階にいたっても、過去の雑誌の遡及情報の入力を行う場合には、ラスター入力以外の方法は運用上難しい。以上により、電子図書館システムの運用開始後も相当長期にわたってラスター情報を蓄積維持せざるを得なくなると判断される。すなわち、いずれにせよラスター情報を取り扱えるシステムであることが要請されるのである。

### 2.3 検索のシナリオ

プロトタイプでの検索機能は大きく二つに分けられる。従来の二次情報データベース流のアプローチと、雑誌文献への直接的アクセスの二つである。

図1からわかるように、オンライン電子図書館システムは、従来の二次情報文献データベース(A)を基本として、これを一次情報と直結させたものと見ることもできる。

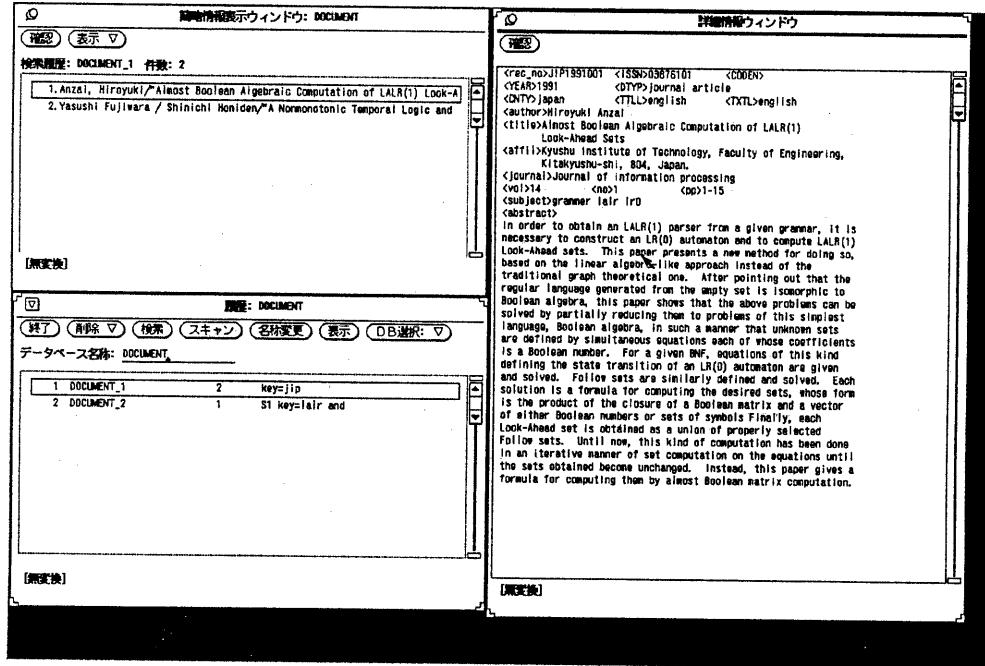


図 2 二次情報検索結果の例

オンライン二次情報データベースでは、伝統的にキーワード等を用いて主題によるアクセスを中心に利用者インターフェースや検索機能が考えられてきた。プロトタイプでは、このような形での論文の探索も基本機能として実装している。現在稼働しているプロトタイプでの検索例を示したのが図2である。キーワードによる文献検索を行い論文情報の簡略表示とその詳細情報を示す。

しかし、研究者の日常の活動を考えると、多くの場合、入手したい論文名、雑誌名等は明らかになっていることが多く、主題からよりも、書誌情報に基づく原文献への直接アクセスが主な関心事である。現在のところ、学術情報センターの持つ学術雑誌総合目録[4]が、唯一この情報を提供するデータベースであり、プロトタイプでもこれを内蔵している。このような観点から、文献書誌情報からの一次情報の直接的な入手プロセスを全面的

に電子化することが、本プロトタイプでの主要な実現項目になる。この例を示したのが図3である。ここでは書誌情報の検索から、必要な巻、号、ページ番号等を指定して、直接画像情報を検索している。

#### 2.4 データベースサーバの蓄積するもの

以上をまとめると、本システムのデータベースサーバでは、

- 二次情報データ (論文単位の文献情報)
- 雑誌に関する書誌データ (雑誌目録所在情報)
- 画像データ (雑誌のすべてのページの画像情報、ラスターなしし PostScript)
- 全文データ (論文単位のテキスト情報)

の情報を統合的に扱えるようなシステムが望まれる。特に、画像情報や全文情報をアクセスするために必要となる二次情報(文献情報および書誌所在情報)を、一次情報へのアクセスのためにいかに再構成するかが、電子図書館システムの成否を分ける。また、各種の

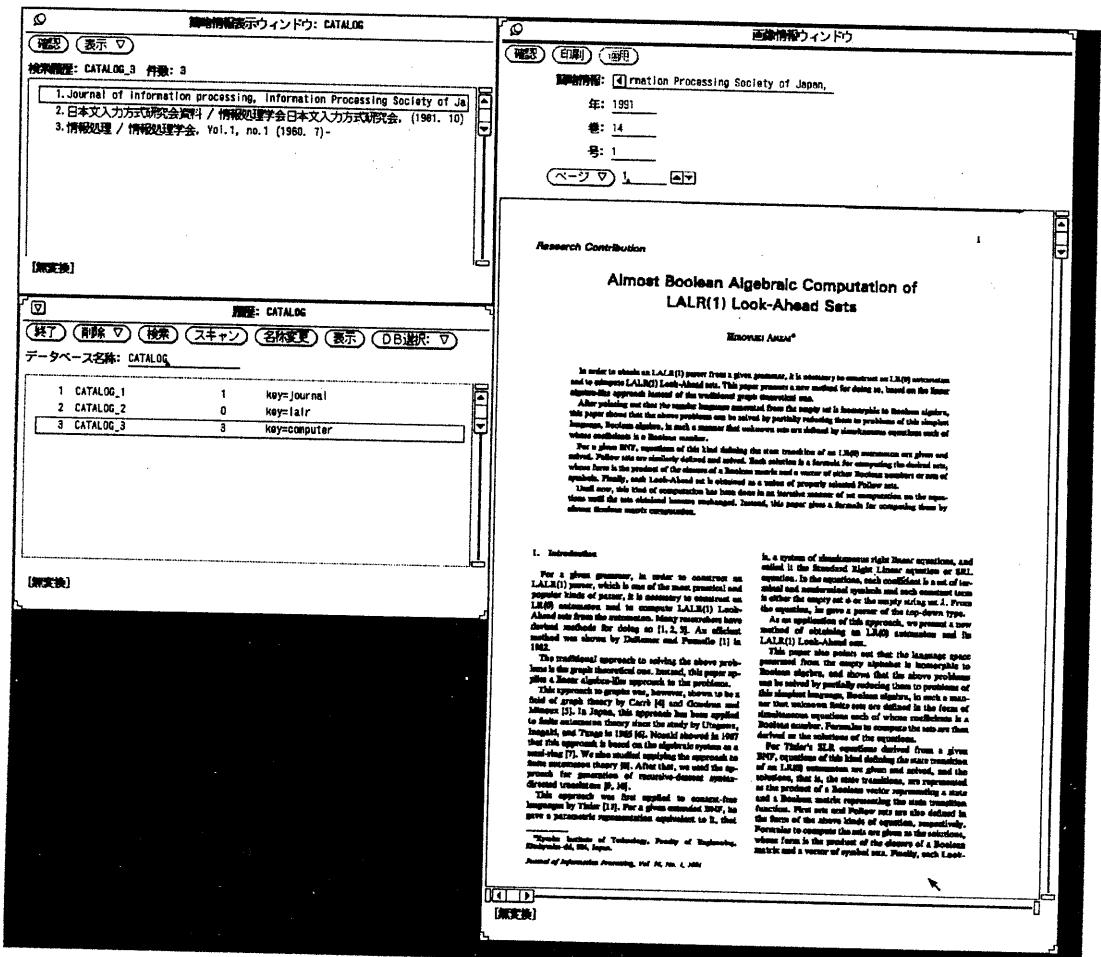


図3 画像の表示例

二次情報データベースの統合型の利用者インターフェース機能などもシステムの基本として重要である。

### 3 画像データベースの構成

#### 3.1 情報へのアクセス

学術雑誌のページとしての一次情報へのアクセスには次の四種が考えられる。

- 雑誌や会議録のタイトル・巻・号とそのページから所望の論文を得る。
- 二次情報を検索し、キーワード等から。

- 目次からの探索。
- 引用関係による探索。あるいはハイパーテキスト流のアプローチ。

後の二つは従来の情報検索システムには余り見られない機能であるが、今後的情報サービスシステムでは重要な機能であろう。理想的にはこれらの四つのアクセスに対応可能な画像データベースとして設計されていることが望まれる。

本プロトタイプでは、なるべく現行の紙媒体の冊子の感覚を活かすように利用者インターフェースを実装する方針を採用了した。すなわち

- 雑誌や会議録のタイトル・巻・号とそのページから所望の論文を得る。
- 二次情報を検索し、キーワード等から。

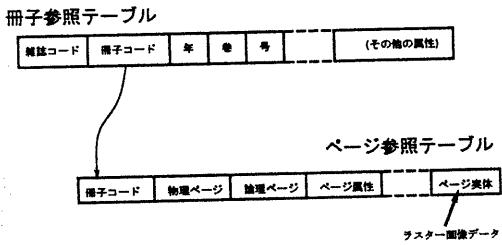


図4 画像データベースのスキーマ

ち、利用者は、目次のページから順々に冊子をブラウズしていくことも可能であるし、特定の論文の第一ページに位置付けた後、ひとつ前の論文の最後のページをめくることもできるようと考えている。

### 3.2 画像入力

大量の一次情報の蓄積に必要な記憶容量は巨大なものになる。それに加えて、学術資料のデータベース入力を組織化する際に問題になるのは、個々の情報への平均アクセス頻度が相当程度低いと予想されることである。つまり、ページの画像入力コストは相対的に高く、また蓄積のためのコストも高くなる。そこで実用システムのためには合理的な画像入力ストラテジーとアーカイブ手法の開発が鍵になる。

第一の方法は、センターなどが定期的に雑誌の毎号のページを全数入力するという方法である。比較的リクエスト数が期待できる雑誌を選択して入力していくことになろう。

第二の方法は、利用者からリクエストされた特定の論文だけ入力するという、on-demand 入力である。システム上で利用者がページの表示を指示した際に、画像がまだ蓄積されていない場合、入力指示が適当な入力拠点(例えば図書館など)に送られる。実際の冊子を保持しているところで入力が行われ、システムのデータベースに登録の後、利用者には数時間から数日遅れて入力完了の通知が届くことになる。この場合は、現在の文献複写サービスをより合理的にしたものとも見る

ことができる。

本システムでは、両方法を併用することを想定して設計している。特に、遡及分は deferred on-demand 入力になり、結果的にアクセス頻度の高い論文の入力が促されることを期待している。

### 3.3 画像データベースのスキーマ

以上の考察に沿って、ページイメージでの文献の取り扱いを実現するため、プロトタイプでは、書誌的な意味での雑誌という実体、個々の冊子という実体、ページデータの集合としての冊子の三つを配慮してスキーマを定義している(図4のBに相当)。図4の「ページ参照テーブル」が個々のページを格納する画像データに対応している。実際の格納に関しては1ページの画像情報が大きくなる場合を考慮し、リレーション DBMS と普通のファイルを併用してデータベースサーバとして機能するようにしている。

ページは冊子という実体に合わせて管理し、個々の論文情報とのマッピングは雑誌コードと冊子コードを介して行われる。二次情報データベースからはふつう雑誌コードが得られることになるが、これは図4の「冊子参照テーブル」を使って、巻号等の情報と合わせて冊子コードに対応付け、最終的に当該ページの画像データを得る。

情報検索システムを、単一の非正規リレーションに対する restriction/projection 操作を実現するものと概念化した場合、本プロトタイプでの画像データベースとは図4中のふたつの表を join したページの集合を格納しているものと見ることができる。さらに情報検索との対応では、あらかじめ冊子単位でレコード(ページ)の集まりがグループ化されており、これが検索結果集合として常に標準的に獲得されるという形態になっているとみなすこともできる。

したがって、利用者の行う論文のブラウズは、雑誌の検索の結果得られた冊子全体の

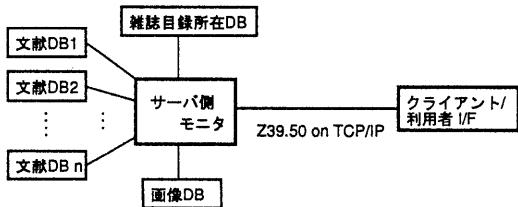


図5 データベースとサーバ・クライアント

ページ集合の中の適当なページにポインタを位置付け、次々とレコード（ページの実体）を転送・表示するという操作に相当することになる。このような設計方針は、第4.3節で述べるプロトコルZ39.50との整合性を探るために必要となったものである。

なお、論文毎に得られる全文情報に関しては、プロトタイプでは、データベーススキーマ上は個々の論文に付随する属性情報のひとつとして（ちょうど長いアブストラクトと同じように）とらえている。

## 4 システム構成

### 4.1 採用した構成方法のねらい

現在の図1のプロトタイプでシステム構成上特徴としている点は、

- client-server モデルにそった分散処理指向のシステム設計
- scalability (データや能力の規模に応じてシステムを大型化できる)
- 新たなサービスを容易に組み込むことが容易で柔軟な機能構成

である。

プロトタイプの設計方針としては、すべての要素を UNIX operating system のもとで動かし、エンドユーザのアプリケーションとの間は TCP/IP 上で応用層レベルのプロトコル Z39.50[5] に基づいて通信し合う Client-Server モデルに基づくように構成する。これによって、ハードウェアとソフトウェア間お

よびサーバ・クライアント間の相互干渉を少なくし、scalability と flexibility を出そうとしている。

### 4.2 データベースサーバ

サーバ側には、図5のように、性格の異なるデータベースが登載される。モニタ部はクライアントとの通信管理、利用者の認証、統計情報の採取、セッション管理などを行う。中でもセッション管理には情報検索特有の検索処理と結果集合の保持等の機能が含まれる。

現在のプロトタイプでは、データベースをまたいだ検索処理は限定的かつ ad hoc に実装されている。文献から書誌、文献から画像、書誌から画像の間でデータベースを越えたレコードの連結が可能であるが、文献データベース間の組み合わせ検索は今のところ実装されていない。これは Z39.50 プロトコルの使用による制約（第4.3節参照）とサーバ・クライアント間での独立性向上を図ったことに起因している。

サーバシステムのデータベース管理システムとしては、現在 SUN ワークステーション上の ORACLE を使用している。現時点では、画像データベースおよび二次情報データベースサーバをリレーショナル DBMS の上に構築しているが、本格的なシステムでは、二次情報検索時に検索集合の大きさ等で相当の機能拡張が必要になると判断している。

### 4.3 client-server 通信

クライアント・サーバ間の通信は情報検索向きのプロトコルである ANSI Z39.50 を機能拡張したデータベースアクセスプロトコルを採用している。画像データベースを取り扱う都合上、いくつかの点で機能拡張を行った。ひとつは、検索結果集合を転送する際に、個々のレコードのフィールドに条件を付与し選択的に送るようにする機能であり、これは画像情報転送時に特に使われる。また管理情報やエラー情報の転送機能などの強化を予定している。

#### 4.4 利用者インターフェース

利用者インターフェースとしては標準的な GUI (Graphical User Interface) を積極的に採り入れていく方針である。目下のところ Open Look により実装しているが、今後 Motif や X の上での実装の必要性も出てくる可能性がある。

稼働しているプロトタイプソフトウェアの動作例は図2と図3に示した通りである。なお、所望の文献が見つかった場合はラスター情報を手近のプリンタに出力することになるが、図1の G に描くようにサーバ側は G4 ファクシミリへの出力機能も持っている。

#### 4.5 システム開発計画

電子図書館システムのプロトタイプはおおむね次のようなフェーズに分けて開発作業を進めている。

**フェーズ1** 1992年10月完了。クライアント側利用者インターフェース制御ソフトウェアとサーバのエミュレーションを主体。SUN RPC による Z39.50 の実装。

**フェーズ2** 1993年6月完了予定。画像データベースサーバを中心実現。

**フェーズ3** 1994年6月完了予定。サーバ側の各種データベースサーバ、ネットワーク、利用者インターフェースの性能を含めた評価を行う予定である。また、通信回りを拡充し、特に Z39.50 をソケットインターフェースで書き直す予定である。

### 5 むすび

本稿で述べたオンライン電子図書館システムは、従来の二次情報検索システムの機能を包含し、しかも欠けていた原情報の直接入手機能と統合した新しい情報サービスを実現するものと考えている。現在フェーズ2をとりまとめようとしている段階であるが、引き続くフェーズ3では、実用システム実現の基礎として特に重要なデータベースアクセスプロ

トコルの完成度の向上を第一の課題としている。また具体的なデータ作成・入力方式を念頭に置いたデータベース構成の詳細化を急ぐ予定である。

また画像の圧縮等の個別の技術的課題を解決しながら、あと一年余り開発作業を行い、本稿で述べたような分散処理体系に準拠した情報サービスシステムの実現可能性を実証したいと考えている。

**謝辞** 本研究は一部文部省科学研究費補助金の助成を受けております。試験データとして論文誌の使用許可を快諾していただいた情報処理学会に感謝いたします。ソフトウェア作成でご協力いただいている東海ソフト(株)の諸氏に感謝いたします。

### 参考文献

- [1] 安達淳, 橋爪宏達: 欧米における「電子図書館」プロジェクト, 情報処理, Vol.33, No.10, pp.1154-1161 (1992).
- [2] 根岸正光: フル・テキスト・データベースの応用動向, 情報処理, Vol.33, No.4, pp.413-420 (1992).
- [3] Story, Guy A., et al. "The Rightpages Image-Based Electronic Library for Alerting and Browsing," Computer, September, 1992.
- [4] 根岸正光: 学術情報センターのデータベース, 情報処理, Vol.33, No.10, pp.1144-1153 (1992).
- [5] "American National Standard Z39.50-1988 Information Retrieval Service Definition and Protocol Specifications for Library Applications," National Information Standards Organization (1989).