ユニバーサル図書館に向けての図書入力システム「情報ファクトリ」の試作

神谷 俊之 *1、大門 秀章 *1、瀬川 哲*2、
中島 昇*3、石田 和生 *1、波内みさ *4
*1 ＮＥＣ 関西Ｃ＆Ｃ研究所、*2 ＮＥＣ 機能エレクトロニクス研究所
*3 ＮＥＣ 情報メディア研究所、*4 ＮＥＣ Ｃ＆Ｃ研究所
〒 540 大阪市中央区城見 1-4-24
TEl: 06-945-3213, Fax: 06-945-3096
E-Mail: {kamiya,daimon,ishidakz}@obp.cl.nec.co.jp, segawa@mech.cl.nec.co.jp,
noboru@pat.cl.nec.co.jp, nami@swl.cl.nec.co.jp

概要

ディジタル図書館の構築において、既に電子化されている情報を蓄積し、提供することは当然であるが、今までの知識、情報が大量に保存されている既存の図書も同様に利用者から検索可能であることが必要となる。

我々は誰でもがどこからでも必要な情報にアクセスできるディジタル図書館『ユニバーサル図書館』構築の要素技術として既存図書の電子化遡及入力のためのシステム「情報ファクトリ」を提案する。本システムは図書のスキャナ入力、ＯＣＲ、構造化、データベース蓄積を統合的なシステムとして提供することで図書の電子化を容易にすることを目標とするものである。

キーワード

ディジタル図書館、遡及入力、文書認識、文書構造

A Development of books digitizing system "Information Factory" for the "Universal Library"

Toshiyuki KAMIYA, Hideaki DAIMON, Kazuo ISHIDA,
Satoshi SEGAWA, Noburu NAKAJIMA and Misa NAMIUCHI
NEC Corporation, Kansai C&C Research Laboratories
1-4-24 Shoromi Chuou-ku Osaka, 540, JAPAN
Phone: +81-6-945-3213, Fax: +81-6-945-3096
E-Mail: {kamiya, daimon, ishidakz}@obp.cl.nec.co.jp, segawa@mech.cl.nec.co.jp,
noboru@pat.cl.nec.co.jp, nami@swl.cl.nec.co.jp

Abstract

It is natural that digital libraries storage and supply documents in digital format from the beginning, but It is also necessary that a huge amount of old "paper documents" information can be searched by digital library.

We propose and develop a integrated digitaizing system "Information Factory" for a part of digital libary concept "Universal Library". it integrates scanning process, OCR process, structurizing process and storing process and it aims at easily digitizing documents for non-professional computer users.

Keywords

Digital Library, Retroactive input,Document Recoginition, Document Structure

1 はじめに

図書館内の目録情報の電子的な管理から始まったディジタル図書館の研究は現在、国内外で要素技術開発やシステムインテグレーションの研究が盛んに行なわれてきている。[NSF],[柿本 95],[藤澤 96],[高橋 96] ディジタル図書館の研究領域は大きく以下のような研究領域にわけることができる。

- データの入力方法に関する研究 (OCR、図形理解、タグ付け等)
- 情報の検索、フィルタリング、抄録に関する研究 (検索理論、情報可視化等)
- 高速ネットワーク技術、ネットワークセキュリティなどに関する研究 (ネットワークセキュリティ、データ圧縮、課金手段等)

これらの研究領域にうち、検索手法に関しては、我々の研究 [市山 96],[神谷 95] などを含め、国内外で多くの研究が行なわれている。また、ネットワーク技術に関してもマルチメディアデータの流通を可能とするATM などの高速なネットワークシステムの研究が行われている。これに対してデータの入力手法に関しては電子図書館では既に電子化されたデータを対象とする場合が多く、既存の図書を検索のために遡及的にデータ化する研究は比較的少ない。[石川 94]

しかし、現在の図書館に蓄積されている情報のほとんどは紙、書籍の形であり、これらの情報を電子的に検索、閲覧可能とすることの意味は小さくない。我々は、大学、企業の研究者による研究、調査のための図書館、公共の図書館を日々利用するような一般の利用者のための図書館などを統合し、ネットワーク上のディジタル情報として流通させることを目的としたユニバーサル図書館について検討、試作を行なっている。本稿ではそのなかで、既存の資料のディジタル遡及入力のためのシステムである「情報ファクトリ」の検討と試作について報告する。

2 ユニバーサル図書館

従来の図書館、とくに一般向けの公共図書館 (ここでは一般図書館と呼ぶ) は市民が誰でも利用できるという性質を持っていた。これに対して企業内、大学内での図書館、あるいは特殊な文献を扱う図書館 (ここでは専門図書館と呼ぶ) では、特定の人々に対して (有料の) サービスを提供している。この 2 種類の図書館、一般図書館と専門図書館は異なったサービスを異なった利用者層に提供しているもので、ディジタル化された情報を扱うディジタル図書館においても一般の市民に広く情報を提供する情報センター的なディジタル図書館と特定のユーザ層に情報の提供だけでなく利用を支援するタイプのディジタル図書館の両方が必要とされると考えられる。

ディジタル図書館全体をネットワークを介してディジタル化された情報をどこからでも誰でもが閲覧、利用することができるようにするシステムとした時に、ディジタル専門図書館とディジタル一般図書館の特色は以下のように想定される。

・ディジタル専門図書館

- 研究利用などの、情報の高度な利用
- 利用者層は研究者など特定のグループ
- 蓄積するデータは論文、特殊な文献等
- ユーザインタフェースは効率的な検索用を主目的

・ディジタル一般図書館

- 情報消費利用の場
- 利用者層は一般の市民
- 蓄積するデータは一般的な書籍、ビデオなど
- ユーザインタフェースは操作が容易で情報の閲覧を主目的

この2種類のディジタル図書館のうち、今後まず実用化されていくのは専門図書館であると考えられる。

これはディジタル図書館の基盤となるのはディジタル情報を流通させるネットワークであり、現在、急速な普及を見せているインターネットにおいても、情報取得のために要する時間や利用するためにかかる費用の問題のため、一般の市民に普及しているとは言い難いためである。

このため、ある程度の費用負担をしてもディジタル図書館を利用することにメリットのある層がディジタルドキュメント(インフォメーション)デリバリーとしてのディジタル図書館を利用するようになると思われる。

これに対して、現在の公共図書館の役割を持つディジタル一般図書館はインターネットなどのネットワークが広く普及した後に、行政などのサービス機能として安価(無料)での市民へのベーシックな情報サービスとして提供することが考えられる。この場合の形態としては単に現在の図書館の置き換えだけではなく、現在の博物館や美術館などうち、一般に市民に利用されている部分を含むマルチメディアサービスとして提供される。

これらの全体を含めたディジタル図書館利用のイメージを図1に示す。我々はこれら全体を含む誰でもが使うディジタル図書館像をユニバーサル図書館と呼び、各サブシステムの検討・試作を行なっている。

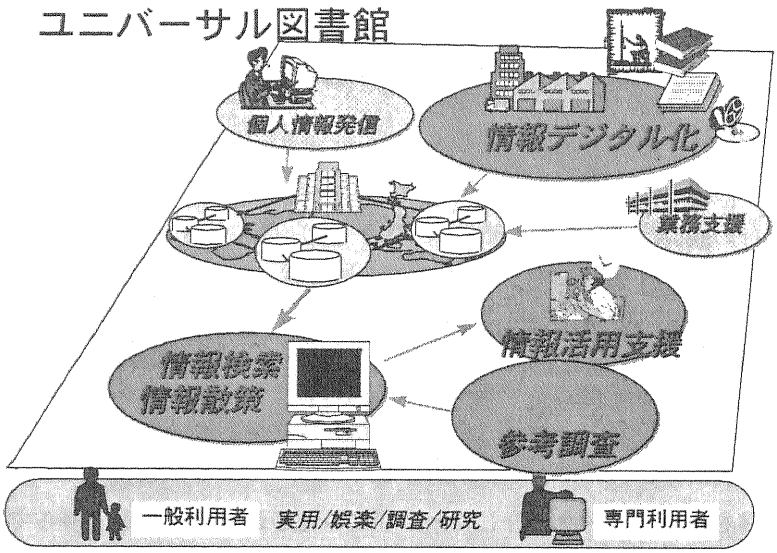


図1　ユニバーサル図書館のイメージ

3 図書データの遡及入力システム

情報の入力に関しては、現在多くの情報が既に電子化され流通し、また、電子化されていない印刷物やビデオなどにおいても、データの作成、編集段階においてはディジタルな情報として扱われることが多くなってきてる。しかし、従来から図書館等に蓄積されている蔵書や企業内での既存の書類も現在大量に存在する。

我々はディジタル図書館での要素技術としての情報ディジタル化においてまず、ディジタル専門図書館の既存文書遡及ディジタル化を対象としたシステム「情報ファクトリ」の開発を行なっている。[大門 96]

「情報ファクトリ」では、書籍、雑誌、文書など紙に印刷されたもの(「紙文書」)を全て「ディジタル文書」化する入力工場を開発することが目標である。紙文書をディジタル化し、再利用が可能な情報とするには、スキャナによるディジタルイメージ入力、レイアウト解析/文字認識、構造化、蓄積の各手順を経る必要がある。これらの処理を行い、文字情報は電子的なテキストとして、画像についてはディジタル画像として保存する。また、論文、書類、書籍はその中に、章や節といった構造をもっており、この構造情報を再構成する処理を行う。また、紙文書内の文字、画像等のレイアウト情報についても蓄積する。

情報ファクトリでは現在、専門図書館で扱われているような文書 (雑誌、文献、新聞等) を大量に入力することを目的とし、システムの利用者 (入力作業者) は必ずしもコンピュータによる作業に慣れているとは限らない。このためシステムは以下のような特徴・機能を持つ。

・ 統合システム

- スキャナ入力から蓄積までの各プロセスを統合して一括処理できるようにする。

- 幅広い対象を (単行本、文庫本、論文、特許等) をカバーする。

- 様々な検索方法 (書誌事項、レイアウト等) に対応する。

・ 入力作業のナビゲーション

- 作業状況を常に画面上に表示し現在の作業をナビゲートするユーザインタフェース。

- 煩雑な各モジュールのパラメータ設定を一括して行なえるようにし、自動的あるいは容易に適切なパラメータ設定が行なえるようにする。

・ 分散/協調環境での入力作業

- 大量の文書を入力するためには、多地点にある入力作業環境で入力対象を分担して入力することが必要になると考えられる。

- 入力の各プロセス自体も各モジュール、入力作業が効率的に行なえるように各プロセスの所要時間に比例して、複数のマシンで協調的動作するシステムとする。

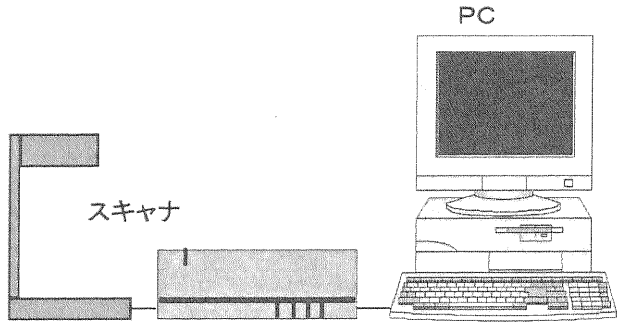
3.1 システム構成

我々は現在、前節で述べた機能を満たすことを目的とした図書ディジタル化システム 情報ファクトリの試作を行なっている。但し、現段階ではシステムはPC単体に実装され、分散/協調の枠組に関しては考慮していない。

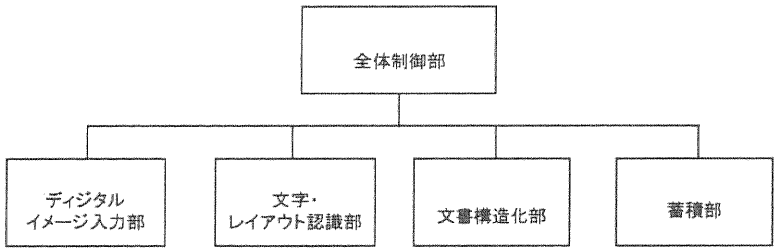
試作したシステムの構成を図2に示す。ハードウェアはパソコン1台に入力機器を接続した構成である。(データベースはネットワークを介して別のPCを利用することもできる。) 全体制御部が、各モジュールを管理する形態をとり、各モジュールのパラメータの起動、データの受渡しを行なう。

情報ファクトリでは以下に示すような手順で図書のディジタル化入力を行なう。

- [ディジタルイメージ入力]「紙文書」をスキャナで画像としてとりこむ。一般的な文書では文字認識を行なうために300～400dpiが必要である。
- [文書レイアウト解析]「紙文書」のページレイアウトを解析し、文字認識可能な文章領域と図表領域の切り分けなどレイアウト情報の抽出を行なう。



(a) ハードウェア構成



(b) ソフトウェア構成

図2 システム構成

- [文字認識] 文章領域に対して文字認識を行ない、目次や本文、書誌事項記載部分のテキスト化を行なう。
- [文書構造化] 入力した文書の文章構造を解析し、構造化テキスト (SGML) に変換する。
- [蓄積] 構造化テキストおよび書誌事項をデータベースに蓄積する。

以下では、各部の機能と特徴について述べる。

3.2 ディジタルイメージ入力部

ディジタルイメージ入力部では、紙文書のイメージをスキャナを用いてシステムに取り込む処理を行なう。ディジタルイメージ入力部では、後のレイアウト・文字認識部において文字認識可能な程度の解像度 (通常の文書では 300～400dpi) をできるだけ高速に取得可能であること。また、通常の綴じられた形式の文書が容易に入力可能であることが望ましい。

このため、情報ファクトリのスキャナ入力としては、通常のフラットベッドタイプのスキャナおよびデスクスタンドタイプのスキャナを接続して用いている。このうちフラットベッドスキャナは Windows での標準的なスキャナインタフェースである TWAIN を用いており、TWAIN 対応の各種スキャナが利用可能である。フラットベッドスキャナではカラーの表紙画像、背表紙画像などを高精細に入力することができる。また、本文ページの入力においては、書籍の見開きページの入力が可能であるデスクスタンドタイプのスキャナ [柏谷 95] を利用する。(図 3)

デスクスタンドスキャナは通常の見開きの原稿をそのまま入力可能であり、「紙文書」のうちに特に書籍の形態のものを分解せずに容易に入力作業が可能であるという特徴を持つ。また、より大規模な入力には自動給紙機構を持ち、ページの裏表入力の可能なスキャナの利用が考えられる。

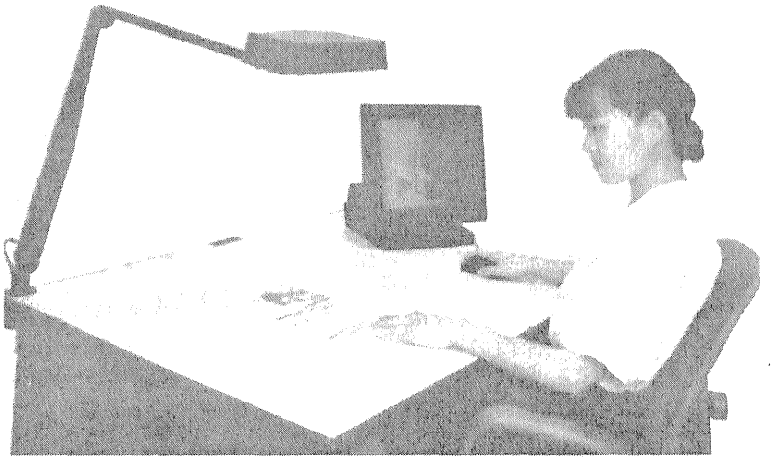


図3 デスクスタンドスキャナ

スキャナのパラメータは全体制御部の指示によって設定され、入力されたディジタルイメージは文字・レイアウト認識部へ送られる。

3.3 文字・レイアウト認識部

文字・レイアウト認識部ではディジタルイメージ入力部で入力されたディジタルイメージを対象にレイアウトの解析および文字認識を行なうモジュールである。

本システムの文字・レイアウト認識部では従来の手法に比べ、以下の2点について改良を行なったシステムを利用している。

・ 動的二値化方式

デスクスタンドスキャナ入力画像に見られる背景部の照明むらに影響されない2値画像を生成するため、新規に動的二値化を導入した。(図4)

・ レイアウト認識では従来、文書画像の水平・垂直方向への過疎を投影し、投影パターン上での空白を検出することで領域分割を行い、文書のレイアウト構造を抽出していた。[辻91] しかし、この方法では文書の傾き、周囲に現れるノイズ、囲み枠が存在する場合にレイアウト解析が行えない。そこで、文字の辺、旁の間隔、行間距離等のレイアウト上の規則に基づいて、統合して行く処理に基づくレイアウト解析を行なう(図5)

レイアウト認識部では、各ページの文字、行などの領域構造を抽出し、文字領域と図表領域の分離を行なう。このレイアウト情報と、文字認識結果の組が全体制御部に送られる。

全体制御部では、一文書を単位としてレイアウト情報と文字情報をまとめて文書構造化部を呼び出す。

3.4 文書構造化部

文書構造化部はテキストの持つ、文字種情報と、文字のページレイアウトの情報から文書の論理構造情報の抽出処理を行なうモジュールである。[石田96] 情報ファクトリでは論理構造の記述として、SGML[ISO_SGML]を採用し論理構造の記述を行なう。

(a) 証明むらのある入力画像

(b) 単純な局所二値化
(画像の右と下の波形は
画素投影パターン)

(c) 本手法による二値化結果
(画像の右と下の波形は
画素投影パターン)

図4 動的二値化処理

既存の文書を電子化して保存する場合には、文書を単にイメージとして保存するのではなく、文書の持つ文字情報を計算機可読なテキストに変換することにより、テキストの全文検索などの各種の検索が可能となる。さらに、文書、特に例えば議事録、論文などのように定型的なレイアウト、構造を持った文書についてその論理構造を抽出することで、構造を利用した検索が行なるなど電子化による特徴をより生かすことができる。

・ レイアウト情報の利用

「紙文書」は用紙中への文字や図表それ自身のもつ情報に加えて、タイトル、章、節などの文章の(階層的な)構造に関する情報を持つ。さらに、紙面上のレイアウトも文章の見易さや理解し易さに大きな影響を与える情報であり、重要な役割を担っている。また、文書の検索の側から見た場合にも、テキスト自身の全文検索や、文章の章、節の構造に関する検索に加えて、紙面のレイアウトの記憶によって検索が行なうことが考えられる。

通常、既存の文書のSGML化においては、文書の持つ論理構造のみを記述し、SGML自身にはレイアウトに関する情報を含めない。しかし、上で述べたようにレイアウトに関する検索や、検索後の文書を画面上あるいはプリントアウトした時に読み易く表示するためにはレイアウト情報についても保存する必要がある。

本モジュールでは図6のように、通常の論理構造を示すタグに加えてページ中でのレイアウト構造を示すタグ<Layout>をタグとして定義し、文章、図表のブロック(段落)単位でタグ付けを行なうことで検索時にレイアウトによる検索、表示が行なえるようにしている。

・ 論理構造の抽出

本モジュールでは各ページの文字情報およびレイアウトの情報を入力として、論理構造の抽出を行なっている。レイアウト情報として利用可能な情報は、行(文字)の大きさ、行の間隔、インデントなどが利用可能である。文字に関する情報としてはページ番号、章、節番号部分の数字を利用している。また、今後、章・節の構造の認識に目次についての認識結果を利用する予定である。

認識手法としては、まず全文を行単位で読み取り、前後の行との相対的なレイアウト(インデント量等)、行先頭の文字種などにより章・節タイトル、箇条書項目などのタイプらしさについて各行に評価値を与える。その後、文の先頭から評価値に基づいてSGMLタグを埋める処理を行なっている。

3.5 蓄積部 - 図書管理クラスライブラリ -

蓄積部は文書構造化部によって作成されたSGML文書を蓄積・保存するためのモジュール(データベース)である。SGMLやODA[ISO_ODA]などの構造化文書を活用する上で、それらを格納・管理するデータベース(DB)には次の機能が要求される。

- 題名、著者、発行日などの書誌情報によって文書を検索できること
- 利用者による任意/特定のキーワードによって、文書を検索できること
- 文書要素の構成に基づく検索が可能なこと
- 文書全体のみばかりでなく、文書の構成要素を部品として管理し、それらを再利用するために検索・提供できること

このうち1,2は、非構造化文書を含めた文書DB一般に要求される機能であるが、3, 4は、構造化文書の持つ構造情報を利用することによって実現可能な機能である。

これらの機能を実現するために、我々はオブジェクト指向データベースPERCIOを利用して、構造化文書を蓄積するためのクラスライブラリを作成した。[波内96]オブジェクト指向データベースではクラス階層と継承を用い、文書の構成要素をオブジェクトして管理することにより構造を持つ文書を素直に表現が可能である。

本構造化文書DBシステムは、構造化文書の中でもSGML文書の管理を目的とし、「DBスキーマ生成部」,「文書登録部」,「書誌情報管理部」,「文書構造管理部」の4つの部分から構成される(図7)。これらは、以下の機能を持つ。

- [DBスキーマ生成部] SGML文書を規定するDTD (Document Type Definition)を解釈し、それに則ったクラスを自動生成することにより「文書構造管理部」のDBスキーマを構築
- [文書登録部] SGML文書をDBに登録するためのインタフェース; 通常のC++インタフェースを持ち、Microsoft VC++プログラムなどから直接呼び出すことが出来る。
- [書誌情報管理部]「文書」として実際に取り扱われるものの構成情報、書誌情報を管理
- [文書構造管理部]「文書」を構成するファイルの1つに対し、その内容構造を管理

以下、本システムの中心となる文書構造管理部について述べる。

文書構造管理部では、SGML文書の内容を実際にDB中に格納する。SGMLでは、文書の構成要素(文書要素)とそれらの間の関係、すなわち文書構造が、文書型定義(DTD)によって規定されている。このDTDに則った実際のSGML文書(文書インスタンス)中では、文書要素の境界をタグで区切ることにより、文書要素とテキストを対応付けている。

文書構造管理部では、文書の構造情報をDB構造に反映させるために、このタグで区切られた文書要素ごとにオブジェクトを生成し、そのオブジェクトに文書の属性、対応するテキストなど、文書要素の情報を格納する。

本システムでは、基本的にDTDのELEMENT定義に忠実にDBクラスを生成する。例えば、DTD中にELEMENTとして“Document”,“Author”が定義されている場合には、その構造定義をそのままクラス構造として持つDocumentクラス、Authorクラスを生成する。したがってこの方式では、DTDごとに異なるDBスキーマを生成する必要がある。

この「文書要素対応クラスを生成する方式」を利用することにより、以下の利点が得られる。

- 文書要素名 (タグ名) がクラス名に、属性・構造がメンバ変数として DB クラスが実現されるため、利用者が直観的に AP 開発を行うことが可能
- 各文書要素に特化した操作手続きを新たに定義できるので、柔軟なデータ操作が可能
- 文書部品 (タグ名) が検索条件に含まれる場合に、その文書部品を管理するクラスの Extent を利用した高速検索が可能

しかしその一方で「DTD が変更されると大幅なスキーマ更新が必要な場合あり」といった問題点があるが、既存の文書を SGML 化する場合には DTD をあらかじめ決まっているもとしてほぼ扱えるので大きな問題とならないと考えている。

3.6 全体制御部

全体制御部では、既に述べた各モジュールの起動、およびデータの受渡しを仲介し、各モジュールの設定パラメータの管理を行なう。また、システム全体のユーザインタフェースを提供する。

・ 情報入力ユーザインタフェースの設計方針

情報ファクトリは、図書館等における蔵書の大量入力を支援するシステムであるため、以下の点を考慮してシステム構成とユーザインタフェースの設計を行った。

- [スキャナ入力からデータベース蓄積までを統合したシステム] 以下のプロセスを統合し、従来、各プロセス間で前処理/後処理等でデータの整合をとらなければならなかった点を改善する。

1 入力パラメータ設定

2 スキャナ入力

3 レイアウト解析/文字認識

4 文書構造化

5 オブジェクト指向データベース蓄積

入力から蓄積まで一連の作業を行える統合システムにより、各処理間のデータ変換などの手間を減少させ、全体をひとつのユーザインタフェースで操作することによって、操作性を改善する。

- [入力手順のナビゲート] 利用者が現在の作業内容、状態を把握できるよう作業状況を常に画面上に表示し、作業手順をナビゲートする。

- [パラメータ設定の簡易化] 利用者の経験等により、入力品質にばらつきが出にくい、誰が操作しても一定品質が得られるよう、各処理部での設定パラメータを適切なものに自動的、あるいは簡易的に設定する機能が必要である。

・ 簡易入力用ユーザインタフェース

情報ファクトリのユーザインタフェースとしては、利用者がスキャナ等の各種の設定を行なえる「一般入力用インタフェース」と特にコンピュータに不慣れな人への入力手順のナビゲートを主眼とした「簡易入力インタフェース」を試作した。以下ではウィザード形式で入力が行なえる簡易入力インタフェースについて述べる。

簡易入力用インタフェースはコンピュータに不慣れな人を対象にしたユーザインタフェースである。入力から蓄積までを、「図書入力ウィザード」に従って行えばよいようデザインした。図書入力ウィザードとは、Windows のインストール時に使用されるウィザード等を参考とした入力ガイドである。入力作業を数

段のステップに分け、1ステップずつウィザードに従って操作を行えば入力作業が完了する仕組みである。画面図を図8に示す。

画面左上にステップ番号を表示し、画面左側に簡単な操作説明と操作ボタンを配置している。ボタンは、「読込」「次へ」「戻る」などで、極力数を減らしている。「読込」ボタンによりスキャナから画像を入力し、これを画面右側に縮小表示する。画面下部には入力作業の手順を示し、全体のうち、どの部分の操作を行っているかを一覧できるようにしている。また、余計な操作ができないよう画面一杯にウィザードを表示している。

ウィザードのステップは本を表紙から順にページをめくって画像の入力を行なった後、書誌情報の設定、認識/構造化/蓄積までを繰り返す形式になっている。

・パラメータ設定

入力する本を、その図書の入力に用いるパラメータの共通性から分類し、分類毎に各モジュールの設定パラメータ情報を持たせることによりパラメータ設定を簡易化した。各モジュールは例えば本の読み取りサイズ、紙質、使用言語などによりパラメータ設定を行なう必要がある。同じシリーズになっている書籍等は同じようなパラメータで入力が可能である。また、よく似たパラメータを持つ書籍に関してはパラメータセットの一部のみを変更して利用できるようにすることにより設定の手間を減少させている。

4 おわりに

本稿で述べたディジタル遡及入力システム「情報ファクトリ」では、ディジタル図書館で既存の図書館における膨大な知識、情報を検索、閲覧可能とするための仕組みの検討を行なった。また、既存の図書の入力について、データのディジタル化、テキストデータ化、データベース化を一括して誰でもが入力作業を行なえるPCベースのシステムの試作を行なった。

ディジタル図書館実現のためには、検索や蓄積、ネットワーク、データ入力などの技術的な要素の他に、従来から指摘される、著作権の問題があり、現在の図書館の所蔵図書あるいは出版社の出版する書籍を本格的にディジタル化するにはしばらくの時間がかかりそうである。しかしながら、今回、検討を行なった情報入力のディジタル化は、学会誌、論文誌など著作権の問題が比較的容易にクリアできそうな

分野や、企業組織、交響団体などでの文書管理システムを手始めとして着実に普及していくものと考えられる。

これらの分野に対応するためには、今後、今回開発を行なったシステムについては、スキャナ部分、文字認識部分などの高速化、高精細化などの各部の機能向上の他、分散入力環境への対応、実際のユーザの評価に基づくユーザインタフェースの向上などを行なっていく。また、その他にも高精細な画像の入力や立体物、ビデオ、音声等のマルチメディアコンテンツの入力についての開発を積極的に行なっていく必要がある。

また、出版社や新聞社などのコンテンツを持つメディア業種と連携したシステムの検討、著作権を中心とする知的財産保護に関する研究を進め、実用的なシステム運用のための技術を確立、実証していく必要がある。

参考文献

[NSF] http://www.nsf.gov/nsf/press/pr9452.html

[柿本95] 柿本 俊博、吉田 哲三,「電子図書館実験システムの開発」, FUJITSU,Vol.46,No.3,pp.276-284,1995.

[藤澤 96] 藤澤 浩道、絹川 博之,「「仮想個人図書館」と個人情報環境」, 第 6 回ディジタル図書館ワークショップ予稿集,pp.11-21,1996.

[高橋 96] 高橋 淳一 他,「Global Digital Museum(1) Concept」, 第 53 回情処全大,Vol.3,pp.423-424,1996.

[市山 96] 市山 俊治 他,「多様な情報源を対象とする WWW ベース電子図書館システム」, 第 7 回ディジタル図書館ワークショップ予稿集,pp.32-50,1996

[神谷 95] 神谷 俊之 他,「3 次元ウォークスルーと CG 司書を用いた電子図書館インタフェースの開発」, 情報処理学会研究会報告 IM 19-5,pp.27-34,1995.

[石川 94] 石川 達也, 電子図書館システムとデータ構築 -データ入力工場設置の必要性-, 情報文化学会, マルチメディア分科会 第 1 回マルチメディア研究発表会資料,1994.

[柏谷 95] 柏谷 篤 他, 平面ミラー回転走査型イメージスキャナ (第 2 報), 日本機械学会 通常総会講演会論文集 (IV),PP. 77,1995

[辻 91] 辻 善丈,「スプリット検出法による文書画像構造解析」, 信学論,Vol.J74-D-II,No.4,pp.491-499,1991

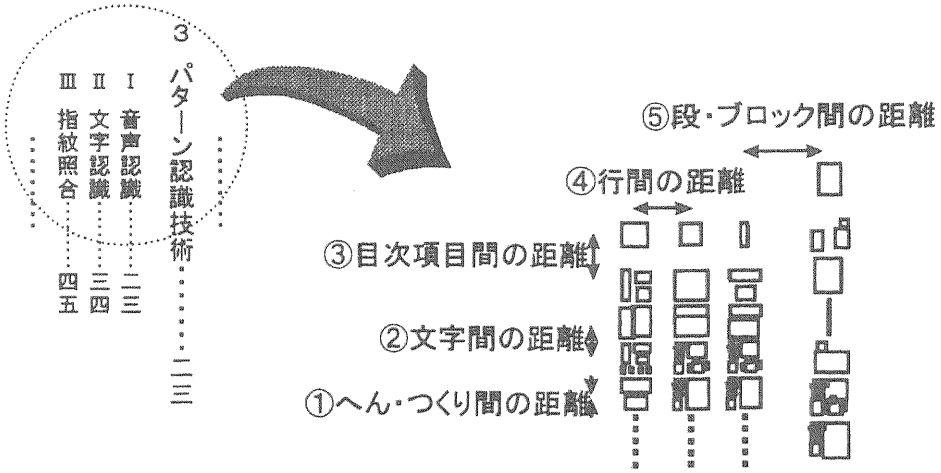
[石田 96] 石田 和生、市山 俊治,「既存文書のレイアウト情報付き構造化手法」, 第 53 回情処全大,No. 3,pp.121-122,1996.

[ISO_SGML] ISO 8879,"1986. Information Processing - Text and Office System -Standard Generalized Markup Language (SGML)," 1986.

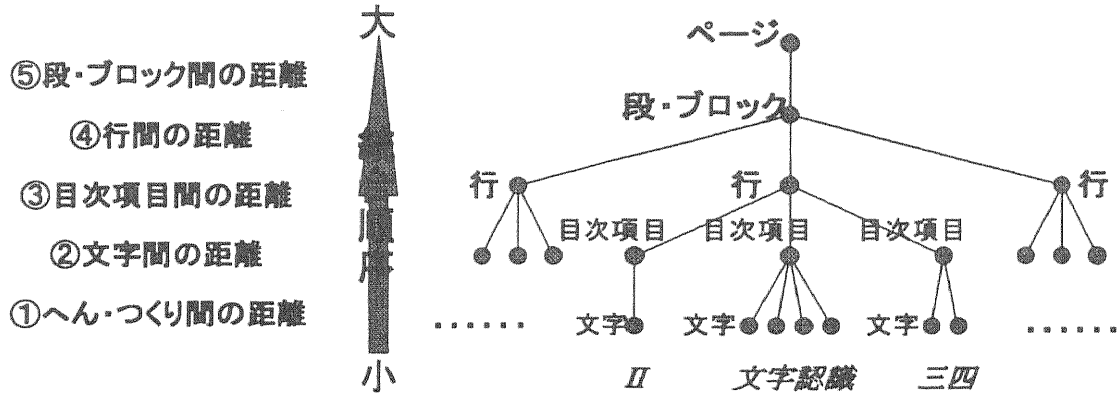
[ISO_ODA] ISO/IEC 8613,"Information Technology - Open Document Architecture (ODA) and Interchange Format," 1994.

[波内 96] 波内 みさ,「OODB による SGML 文書データベースの設計」, 情処 DBS 研究会第 109 回予稿集,1996.

[大門 96] 大門 秀章 他,「図書構造化入力システム「情報ファクトリ」の提案」, 第 53 回情処全大,Vol.3,pp.181-182,1996.



(a) 領域統合処理



(b) 領域統合順序

図5 領域統合処理によるレイアウト解析

```
<Section>
 <SecTitle>
  <Layout LEFT=0.02 TOP=0.1 WIDTH=0.2
          HEIGHT=0.07 PAGE=2>
    はじめに
  </Layout>
 </SecTitle>
 <Figure>
  <Layout LEFT=0.8 TOP=0.15 WIDTH=0.18
          HEIGHT=0.4 PAGE=2>
  </Layout>
 </Figure>
 <Paragraph>
  <Layout LEFT=0.02 TOP=0.2 WIDTH=0.7
          HEIGHT=0.3 PAGE=2>
    グラハム・ベルによって電話が発明されたのは
    １８７６年だった。日本では明治９年である。
  </Layout>
 </Paragraph>
```

図 6 レイアウト情報の埋め込まれた SGML テキストの例

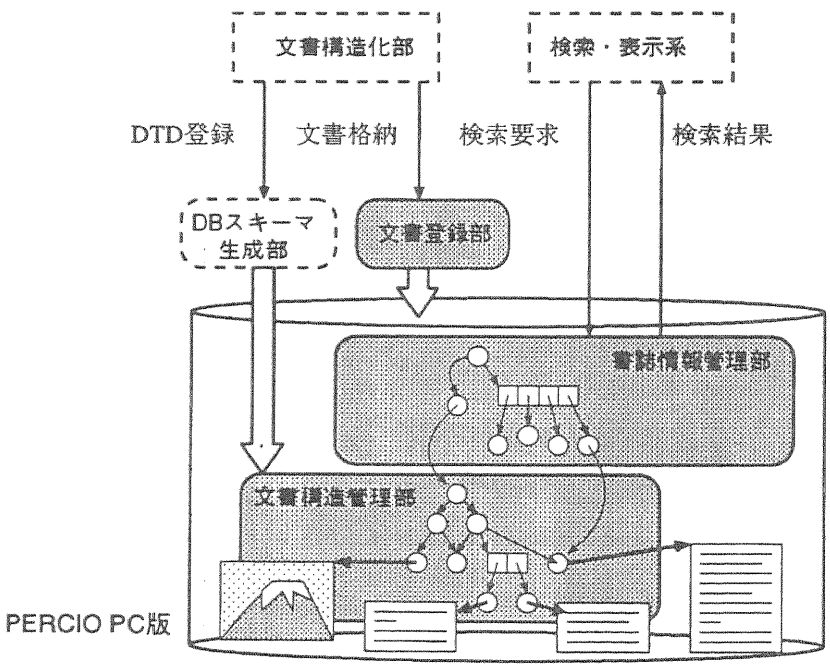


図 7 PERCIO を利用した構造化文書データベース・システムの構成

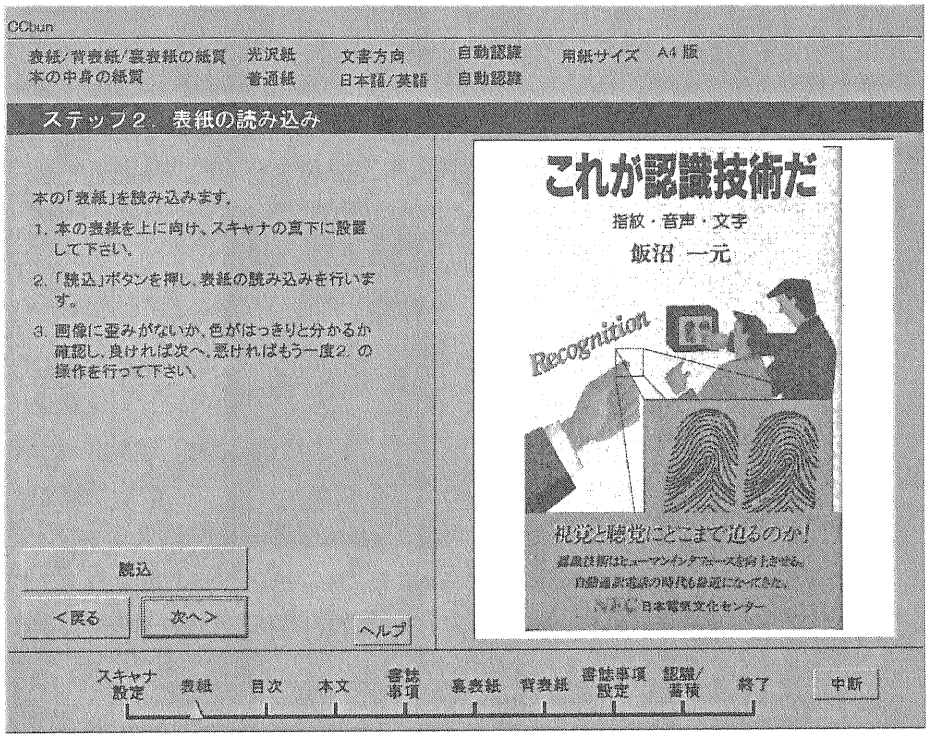


図 8 (a) 簡易入力用インタフェース (表紙読み込み)

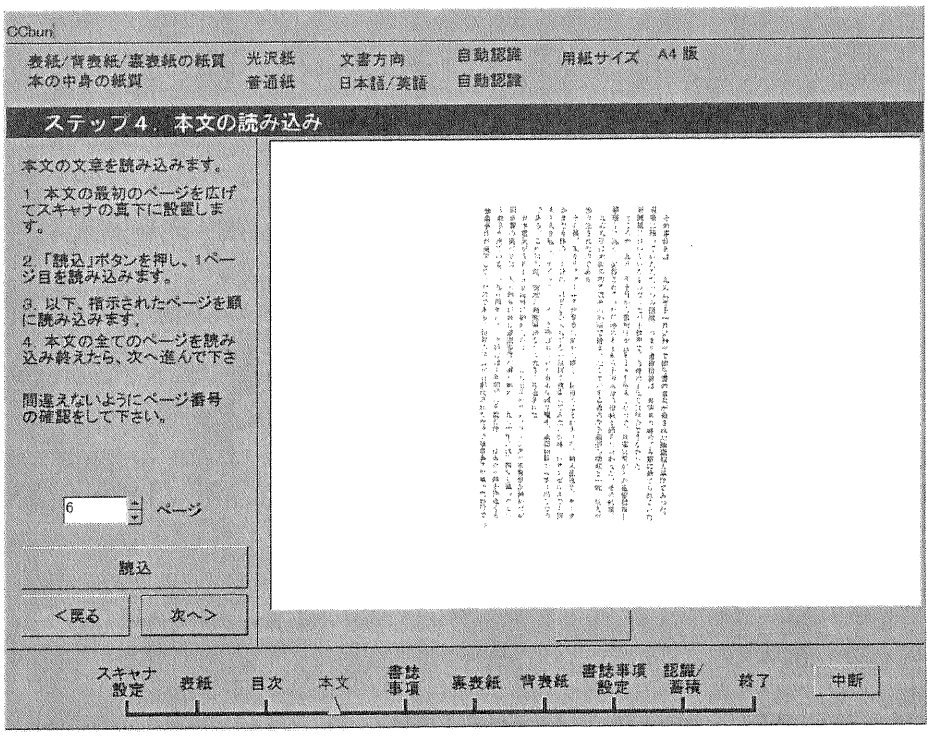


図 8 (b) 簡易入力用インタフェース (本文ページ読み込み)

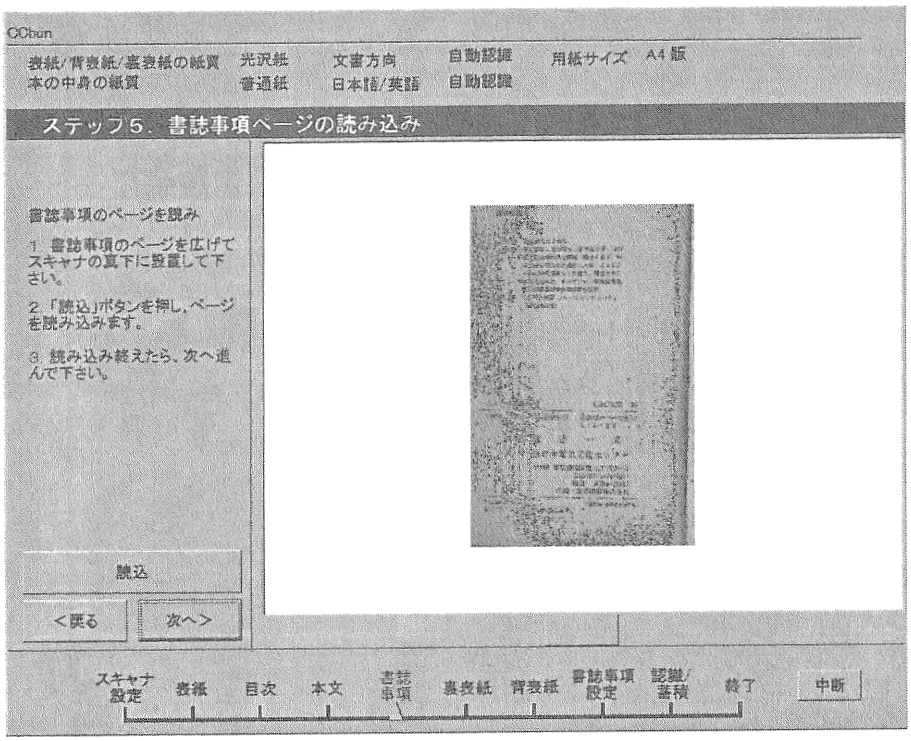


図8 (c) 簡易入力用インタフェース (書誌事項ページ読み込み)

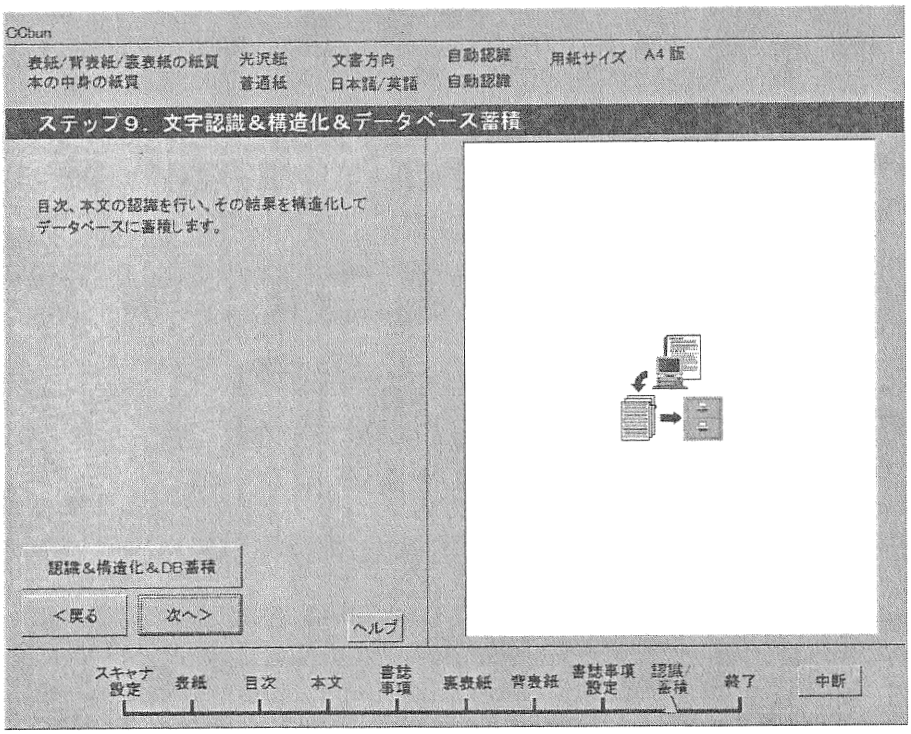


図8 (d) 簡易入力用インタフェース (認識/構造化/蓄積)