3．電子図書館に要求される基本機能

電子図書館には、２章で述べたような特徴がある。それを実現するためには、優れたコンピュータシステムと通信システムの支援が要る。高速の処理速度・通信速度、充分な記憶容量、精度の良い入力（スキャナ）と表示（ディスプレイ）の機能等をもつハードウェアと、それらの性能と機能を有効に活かして、柔軟なユーザインターフェイスを与えるソフトウェアとの協調が大切である。

ここでは、電子図書館の機能を実現するために要求される事項についてまとめる。そして、それらの事項が本学の電子図書館においてどのように実現しているかを、主に、システムのソフトウェアの面から述べる。本学では、まず、電子図書館の基本的な機能を確実に実現することを目標にシステムが構成されている。稼働と利用の実績を基にして、機能の充実と性能の向上を、常に計っている。

3.1 様々な形態の一次情報を入力する作業（電子化作業）の支援

電子図書館の基本は組織的に蓄積される電子化データである。様々な形態の学術情報を統一された検索対象として効率良く構成する方法が必要である。

3.1.1 冊子体からスキャナで読み取り、OCRによってテキストに変換する

スキャナには、高速で精度が良いだけでなく、両面のページを自動的に読む機能、カラーページを読む機能、綴じた冊子体から自動的にページ送りをして読む機能が要求される。

本学では、スキャナで読み取ったデータから、表示用のイメージデータ（GIF, TIFF形式）と、縮小ページ用のイメージデータ（GIF形式）を作成している。それらは、一頁ずつ一つのファイルとして扱っているが、一つの論文を構成するページを順にまとめて、一つのディレクトリを作成している。

イメージデータからOCRを通して、検索用の文字テキストデータを作る。文字認識は、英文および和文のものが要る。一段組および二段組の平文を、高い精度で認識できることは、基本の条件である。字体やサイズが変わっている、図面が入っている、あるいは、テキストの位置が規則的でないなどに、出来るだけ対応するものが望ましい。

テキストへの変換の誤りは、表示や印刷には影響しないが、検索に当たってその文献が候補に上がらない結果となる。多量のデータに対して、変換の誤りを人手で直すことは、殆んど不可能なことである。英文の場合はスペルチェックをして、自動的に修正することが可能である。

本学では、英文・和文を問わず、認識出来なかった部分の前後の文脈を見て、自然言語処理の技術により認識率を向上させる研究を始めている。（参考資料5-6参照）

入力作業は人手による労力と時間が最も多く掛かる部分である。とくに、冊子体から

の入力は、人手の作業を軽減して、機械による処理の部分を少しでも多くする工夫が必要である。

3.1.2 ビデオ情報のディジタル化入力

ビデオテープで提供されるビデオ情報は、ディジタル情報に変換し、一般には、画面の順に、MPEG形式で情報圧縮を行なって蓄えられる。

本学のシステムでは、テープの開始からの時間を指定して、画面を選び出して表示する機能がある。また、早送りや、停止表示が可能である。画面の特徴を表すキーワードなどで、ランダムな検索をして、表示ができると良いが、これは今後の課題である。

3.1.3 CD-ROM のテキストや画像の入力と情報形式の変換

学術情報の記憶媒体として、CD-ROMの比重が大きくなっている。今後その傾向は益々伸びると予測される。学術雑誌、会議録など、動画像やプログラムを含めた柔軟な編集ができる。そのため、内容（コンテンツ）の形式に統一性がなく、検索・読み出しには、個々のCD-ROMに備わっている検索ソフトウェアに頼ることになる。また、電子図書館システムのデータとして取り込む場合には、大量の記憶容量を必要とする。

このような理由から、本学では、現在のところ、キーワードによりCD-ROMを選び出して、読み取り駆動装置にマウントするに留まっている。今後、個々のCD-ROMの利用価値を高めるためには、内容まで含めて、統一的に検索できることが必要である。

それには、出版に当たってのCD-ROMへの記録形式と検索方法の標準化が求められる。

3.1.4 広域ネットワークからの情報の受け入れと形式の変換

ネットワークを通して提供される学術情報には、冊子、ビデオテープ、コンパクトディスクなどの情報媒体は一切不要である。しかも、全てディジタル化されているため、システムへの取り入れも、殆んど機械的に行なえるので、人手の作業は僅かで済む。

ただし、情報の表現形式が、文字テキスト（plain-text）、HTML、SGML、PDFなどさまざまなものがあるため、表示のための形式や、全文検索に適した形式への変換が必要となる。ときには、TeX 形式や、PostScript形式のこともある。日本語による文書の形式変換は面倒である。

たとえば、本学では、学内で作成された博士論文や修士論文は、PS形式で提供されている。日本語の論文については、全文検索の対象とするために、日本語テキストファイルに変換する必要がある。（詳細は5章参照）

3.2 入力された一次情報についての書誌情報の作成と、目録の管理

ディジタル化して入力された情報に対して、多様な検索要求に応えるためには、次のような情報を付け加えなければならない。これらは、入力情報から、機械的に抽出することができる部分もあるが、人手の作業によるものが残されている。

3.2.1 書誌情報の作成

図書、雑誌、および、論文には、蔵書目録として書誌情報が付けられる。これは検索の最小の糸口でもある。論文には、アブストラクトも含められる。

図書、雑誌の書誌情報は、学術情報センターからダウンロードしている。個々の論文については、一部電子データがすでに提供されているが、一般には、題目、著者名などをタイトル部分から抽出して認識しなければならない。しかし、これらの部分に対するOCRでの認識結果は、現在のところ、信頼性があまり高くないので、論文のタイトルや著者名からの検索については、次項による目次情報に頼ることになる。

3.2.2 目次情報の作成

冊子体の図書や雑誌の目次の部分には、検索に有効なキーになる語が多く含まれている。しかし、目次を構成する文字フォントや表現の形式が様々であるため、OCRによる認識の成功率が低い。OCRで得られた結果を訂正するか、最初から目次部分の文字入力をするかを、人手によって行なわなければならない。

目次から、対応する本文のページにリンクを付けておくことにより、目次単位での飛び飛びのページ送りが可能となる。

3.2.3 キーワードの添付

学術論文には、多くの場合、内容に関連したキーワードが付いている。しかし、図書や雑誌（論文集）には、全体にキーワードが付いていることは稀である。キーワードは、題目とは違って、少し広い関連する分野を表していることがあるので、漠然とした検索の有力な手掛かりとなる。

しかし、人手によって、新たにキーワードを付けることは、高度な判断力を必要とする。本文を全文検索して、出現頻度の高い語を抽出するなど、機械的な方法でキーワードの候補を選択することができれば良い。それをそのままキーワードとしてもよいが、それらを包含する分野あるいは概念を表す語を、概念体系の辞書から、選ぶことも考えられる。それは、これからの課題の一つである。

3.2.4 入力作業状況の追跡、作業終了における登録・通知

情報の入力（イメージ入力、OCR、書誌情報・目次・キーワードの添付）から、検索の対象となるファイルへの登録までの一連の作業では、蓄積されるデータの関連付けに誤りのないようにしなければならない。

たとえば、100ページの一つの雑誌を、一つのスキャナで入力していれば、ページの順序の乱れや、ページ抜けは起こらない。入力を速く済ませるために、二つのスキャナで50ページずつ読み込むときには、それらを正しく順序付けしなければならない。また、カラーページは、カラースキャナで読んだものに置き換えたり、マスクをかける必要があるページは、その結果のページで置き換えたりすることが必要となる。

本学のシステムでは、入力作業の進行を監視して、冊子、ビデオテープ、CD、ネットワークなどの媒体に依らず、イメージファイルとテキストファイルの作成がどのような

状態であるか、最終的にどのように登録されたかを、常に把握している。入力が完了して、システムの中で検索と表示の対象となったとき、新着情報として利用者に通知できる。

3.3 検索に適した大量の電子化情報の蓄積

蓄積される情報は増加する一方である。しかも、増加量は、年々大きくなる。表示用のイメージファイルは多量の記憶域を必要とし、検索用のテキストファイルは高速の検索に対応するという問題がある。電子図書館は、蓄積された情報の増加によって、年と共に利用価値も増大していくが、そのためには、これら２つの課題に対処しなければならない。

3.3.1 検索インデックスの付加、情報間リンクの付加

全文検索は、与えられたキーワードに対して、本文全体を探索して、該当する図書、論文を取り出すものである。最もきめ細かな検索方法を提供するが、それに要する時間は膨大なものになる。

本学のシステムでは、要求のたびに、与えられたキーワードについて、全てのテキストファイルの中身を検索・照合して全文検索を実現している。そのため、現在でも、応答時間はかなり長くなっている。できれば応答時間が蓄積されている情報の量に余り依存しない、高速な検索方法をもたせて、将来に備えなければならない。

たとえば、検索の分野（情報、バイオサイエンス、物質・材料など）を区分して、キーワードによって検索対象（通信、遺伝子、高分子など）を限定するなどの方法が考えられる。

汎用的には、キーワード表を作成して、キーワードを含む情報へのリンクを付けておく方法が有効である。リンク付けは、たとえば、次の２段の手順で行なえる。

(1) 新たに文献を登録するとき、キーワード表にある全てのキーワードについて、リンク付けの可能性を調べる。

(2) 表にないキーワードに対して、最初に検索要求が発生したとき、全文検索を行って、その結果をキーワード表に記録する。

文献を新たに登録するとき、および、キーワード表にないキーワードに対して最初に検索するときは、現在よりも長い応答時間が掛かる。同じ検索は繰り返して行なわないので、２回目以降は応答が速くなる。キーワード表とリンク付けに余分の記憶域を必要とするが、スペースと時間のどちらを選ぶかの問題である。

3.3.2 記憶媒体の階層化と情報の柔軟な移動（マイグレーション）

情報を大量に蓄積することと、それらに統一的に高速にアクセスすることとは、コストの面での釣合の問題である。

本学のシステムでは、アクセスの高速性と、大量の記憶容量とを実現するために、ファ

イル装置を３層（磁気ディスク、光磁気ディスク、磁気テープ）で構成している。蓄積された情報量の増大に応じて、アクセス実績の古いものを順に、低速の記憶装置に移す。それに対して、新たなアクセス要求が発生すると、それをディスクに取り出す。システムは、装置の間で、情報の移動を矛盾なく効果的に行っている。

一般には、検索要求と、その結果による表示要求とは、いつも連動して要求が出されるとは限らない。検索対象のテキストファイルに比べて、対応する表示用のイメージファイルは、低速の装置にある率が多くなる。

3.4 多様な情報検索方法の実現

大量の学術情報の中から、本当に欲しいと思っているものを探し当てるのは労を要する。論文の名前や、出処が判っている場合は、それを単純に取り出せば良い。漠然とした検索要求に応えるには、絞り込み検索法など、多様な機能が要望される。それが、高性能なコンピュータシステムによる電子図書館の大きな特徴のひとつである。

3.4.1 キーワード検索、書誌に基づく検索

検索の最低の機能として、書誌情報（タイトル、著者名、出処など）と、キーワードによる検索がある。これは、従来の図書カードに相当する項目を、電子ファイル化したものである。ただし、タイトルだけでは、内容の適応性を充分に表しているとは言えないので、検索のヒット率は余り良くない。

3.4.2 論文のアブストラクト、目次、全文検索による抽出

次に、与えられたキーワードに対して、アブストラクトの検索を含めるものがある。論文や技術報告書（研究会資料など）で、アブストラクトが言葉を選んで、丁寧に書かれているものに対しては、ヒットする可能性が大きくなる。商用のデータベース検索の多くがこの機能を備えている。

さらに、目次と、本文の全文検索機能が含まれると、与えられたキーワードを含むものは、全て抽出できることになる。

本学のシステムでは、全文検索機能を含めて、すべての検索を選択して指定できる。目次を作成する労力と、全文検索する時間とを、対価として費やしているが、より高度な電子図書館に発展させるためには、基礎となる不可欠な機能である。

逆に、全文検索では、関係の薄い、必要以上の検索結果を出力する傾向にある。一旦得られた検索の結果から、内容を絞り込んで選択を段階的に行なう機能が求められている。

3.4.3 利用者の資格に応じた多様なアクセス権の設定と制御

検索操作の結果、閲覧の要求に対しては、電子化の許諾の条件に合わせて、表示、あるいは、印刷の許可の判断が出来なければならない。

所在の検索については、一般には、学内・学外を問わず、全ての利用者に対して、その対象（論文、雑誌、図書、辞書・辞典、ビデオ、CD-ROM、学内情報）の制限はされない。そのため、検索システムは、利用者の資格を区別せずに検索要求を受け入れれば良いが、全文検索に対応する負荷が、利用者の人数と情報の増加の両面に伴って増大していくことを意識しておかねばならない。

閲覧のための本文へのアクセス操作については、縮小ページ画面の表示、拡大画面の表示、印刷などの可否を判断する必要がある。利用者の資格（本学で閲覧を許された者と、そうでない者）と、対象とする情報の公開の範囲（広く学外にも公開を可とするものと、そうでないものなど）との組合せに応じて判定される。

電子図書館を少しでも有効に利用してもらうために、できるだけ多くの利用者に、できるだけ多くのものを公開したいが、著作権に絡む許諾の条件を尊重しなければならない。

3.4.4 ユーザプロファイルの登録による新着情報の通知

検索は、利用者が何かを探したい要求を持ったときに発せられるものである。しかし、研究・教育を進めて行く中で、関心のある話題について、新しい情報が発生したときに即時に知らせて貰えると、研究・教育活動が促進される。

たとえば、登録した雑誌の新着号の目次情報を知らせること、登録したキーワードを含む論文の到着を知らせることなどがある。これは、新しく電子化された情報が登録されると、各利用者の登録内容を見て、自動的に通知の電子メールを発送する機能である。一旦キーワードを登録すれば、人は誰も介入することなく、コンピュータが手順を進めてくれる。

3.4.5 各種の利用記録、アクセス統計の収集

電子図書館の利用状況（検索、閲覧、印刷要求など）の記録をとることが大切である。これは、電子化の許諾を得ている出版者の要望に応えるためだけではなく、電子図書館の機能の充実のためでもある。情報へのアクセスの傾向や、アクセス応答時間などの記録から、記憶装置の増強とファイルの構成法、全文検索方式、通信と印刷の時間の短縮などシステム全体の性能向上の方策を考える資料となる。

蓄積される情報の量の増大と、形態の多様化に対して、コンピュータシステム（CPU、ファイル記憶装置、入出力装置、通信機能など）の性能の向上と、ソフトウェアシステム（入力・変換、蓄積、検索などの方法）の改良を計らなければ、電子図書館の役割を充分に発揮できなくなる。

3.5 図書館の運用と業務の支援

電子図書館であっても、これまでの図書館と同じ業務がある。まず、電子化の基になる冊子やCDは、購入をしなければならない。また、電子化できない多くの図書や、雑誌は、購入から管理までの手続きが必要である。

ここでは、これまでの図書館業務を支援するシステムの機能を、本学の場合を例として、取り挙げる。

3.5.1 図書購入の希望や、文献複写の依頼の受け付け

本学全体から、広く図書や雑誌の購入希望を集めること、学外への情報検索や文献複写の依頼を受け付けることなど、資料の収集についての要望を電子メールで受け取っている。

3.5.2 発注から、納入作業までの連係

送られてきたメールの情報を利用して、購入や複写の発注、納品の確認、登録、複写の到着通知の発送など、一連の作業を、業務支援サブシステムのもとで進めている。このとき作成される書誌情報は、所在検索のデータとして利用される。

3.5.3 貸出／返却の追跡と管理

本学では、図書の貸出しおよび返却の手続きは、利用者自らが行なっている。身分証明書の裏面の磁気コードと、図書に張られたバーコードを読み取るだけで、キー入力を伴わない操作である。貸出し状況と統計を取ること、期間を過ぎたときの返却催促のメールを出すことなど、さまざまな処理を自動的に行なえる。