Vol. 37 No. 9　　情　報　処　理　　Sep. 1996

解　説

2. ディジタル図書館実現のための要素技術と環境要素†

杉　本　重　雄††

1. はじめに

現在，ディジタル図書館（Digital Library）の開発プロジェクトが様々な国で，また様々な分野で進められている[21]．図書館には公共図書館，大学図書館，専門図書館，保存図書館等異なる役割のものがあり，それぞれの図書館には多様な分野の資料が蓄積されている．ディジタル図書館のプロジェクトを理解するには，図書館が持つ機能を理解する必要がある．おおまかに言って，図書館はコレクションを形成し，それを利用者に提供するとともに利用者の情報アクセスを支援する．図書館はコレクションの形成のために資料を収集・組織化し，蓄積するとともに，自身でも情報アクセスのための情報（2次情報）を作りだす．利用者は適切に整備された2次情報を頼りにして所望の情報を探す．書架で本を探すように，1次情報に直接アクセスすることもある．また，図書館員は利用者にとって重要な情報源である．

ディジタル図書館で扱う1次情報は図書や雑誌文献，写真や地図，オーディオビジュアル資料等多様である．また，2次情報は従来の図書館システム以上に整備する必要がある．コレクションの形成から利用に至る過程で様々な情報技術が必要であること，利用される情報技術は蓄積される情報の種類や利用者の特性に依存することも明らかである．このように，ディジタル図書館を実現するには様々な情報資料を扱い，かつ多様な利用者を満足させるための情報技術を総合する必要がある．

本稿では，はじめにプロジェクトの目的に基づいてディジタル図書館をいくつかのタイプに分け，その後，ディジタル図書館プロジェクトにとって重要な技術的話題について述べる．また，個別のプロジェクトについて述べるのではなく総合的な観点から述べる．そのため，誌面の都合で断片的にならざるを得ない部分もあるが，本特集の他の記事の説明と合わせて理解していただけるとありがたい．

2. ディジタル図書館のタイプ

ディジタル図書館のプロジェクトを大別すると，NSF/ARPA/NASAの共同助成による米国の6大学のプロジェクトに代表される従来の図書館の枠組みにとらわれず新しい情報の蓄積と提供環境の試作を目指したものと，大学図書館や国立図書館を中心として進められている既存の資料のディジタル化と蓄積・提供を目指したものに分けることができる．

2.1 米国6大学の研究プロジェクト[22]

1994年秋に発表されたNSF他による下記の6大学への研究助成の決定は，国家情報基盤（National Information Infrastructure, NII）上の新しい図書館像を作りだす研究プロジェクトとして非常に注目された．この研究助成プログラムでは，計算機科学，図書館情報学他の複数の分野からの研究者が参加することと，大量のデータを持つ機関（出版社，放送局，地方自治体，政府機関，図書館等）との共同プロジェクトを進めていることが特徴的である．また，これらは将来の真に大規模なディジタル図書館の構築を進めるための実験台（testbed）となり得るシステムの構築を目指している．したがって，新しい要素技術の開発を目指すだけではなく，利用可能な様々な技術を総合し，大量でかつ多様なデータを多様な利用者に提供することを目指している．また，知的財産権や電子商取引（Electronic Commerce）等

† Component Technologies and Environment for Digital Libraries by Shigeo SUGIMOTO (University of Library and Information Science).

†† 図書館情報大学

社会制度的な問題とも関連するので，社会科学的観点からの研究もプロジェクトに含まれている．

（1） カーネギーメロン大学（CMU）：Informedia Interactive On-line Video Digital Library. 放送局と協力し，対話的に検索と視聴ができるビデオ映像ライブラリを作り上げる．

（2） ミシガン大学（University of Michigan）：The University of Michigan Digital Library（UMDL）. 宇宙・地球科学分野の多様な資料を高校生から研究者まで幅広く多数の利用者に提供することを目的としている．エージェントモデルに基づいてシステムを実現する．

（3） イリノイ大学アーバナ・シャンペイン校（UIUC）：Interspace. 大量の科学技術分野の学術文献を非常に多数の利用者に提供する．IEEE他の出版社と協力し，雑誌論文をSGML（Standard Generalized Markup Language）に基づく全文データベースとして蓄積する．

（4） カリフォルニア大学バークレイ校（UCB）：Electronic Environmental Library. カリフォルニア州が持つ大量の環境情報に関する大規模データベースを構築する．研究プロジェクトには公共図書館も含まれ，航空写真等を含む多様なデータを多様な利用者に提供する環境を構築する．

（5） スタンフォード大学：Stanford Integrated Digital Library Project. ネットワーク上に提供される様々な情報を仮想的な1つの図書館として利用できるようにするための技術を開発する．

（6） カリフォルニア大学サンタバーバラ校（UCSB）：Alexandria Digital Library. 地図や航空写真等の空間的な情報の相互利用性を高め，大規模データベースの構築を可能にするため，メタデータに関する研究を進めている．

2.2 国立図書館のプロジェクト

国立図書館は各国の様々な図書館の要であり，保存図書館として働いているので非常に多数の蔵書を持ち，多くの貴重資料を所蔵している．国立図書館のプロジェクトでは著作権に関する問題がなく，かつ資料の保存性とアクセス性の両方を高めることができるため，貴重資料のディジタル化から始めているところが多い．たとえば，我が国の国立国会図書館のプロジェクト，米国の議会図書館によるNational Digital Library Project，フランス国立図書館による資料ディジタル化プロジェクト[5),6)]等がある．下にフランス国立図書館の例を簡単に述べる．

フランス国立図書館では，新しい図書館に移転するにあたり，資料のディジタル化を進め，館内のみならずネットワークを通じた資料の提供を進めようとしている．ディジタル化の対象となる資料は図書，写真や絵画，オーディオ情報等様々であり，1997年までに10万件の図書，30万枚の写真，1000時間分の録音資料のディジタル化を計画している．図書は300 dpiのページイメージ形式でディジタル化し提供する．ディジタル化計画の当初，人文科学分野の大きなコレクションを持っていることもあって全文テキスト化することが望ましいと考えていたが，OCR（光学的文字認識システム）の読み取り精度が十分でなく（特に古書），入力コストが高くなるためページイメージ方式を選択している．写真や絵画等のイメージ資料は性質に応じてスキャナやカメラ，マイクロフィルムから取り込んでいる．

2.3 ドキュメントデリバリ

従来，文書の複写依頼と送付を郵送によって行うドキュメントデリバリ（Document Delivery）が広く利用されてきた．検索と複写依頼をインターネットで行い，Faxや郵送で文書を送付するという新しい形のサービスもある．さらに進めて，ネットワークを利用した文書の検索・閲覧環境を提供するシステムが現われてきた．たとえば，AT&TがRightPagesに医学関係雑誌記事を載せたカリフォルニア大学サンフランシスコ校（UCSF）のRedSage，ミシガン大学のTULIPやJSTORがある．また，本特集にある学術情報センターの電子図書館システムNACSIS-ELSは情報処理学会，電子情報通信学会他の学協会が出版している雑誌記事を提供するシステムである．奈良先端科学技術大学院大学図書館は，自然科学系の大学院大学であるという特色を活かし，雑誌記事を学内でディジタル化し，利用者に提供する[1)]．これらはいずれも，印刷物として出版されている雑誌記事をスキャン入力したページイメージとして蓄積したもので，利用者は端末上で書誌情報（および全文テキスト）を利用して記事を検索し，記事を閲覧できる．また，

NACSIS-ELSを除くといずれも大学内でのサービスを前提としている．

2.4 電子テキスト

電子ジャーナルや全文データベース等の電子的に作成されたテキスト（電子テキスト，Electronic Text）は，図書館が提供する重要な情報資源である．電子テキストはCD-ROMで配付されるものもあるが，ここではオンライン利用するもののみを考える．電子テキストを提供する機関である電子テキストセンター（Electronic Text Center）の多くは大学図書館の中に設けられ，人文科学分野の電子テキストを提供しているものが多い[14]．

人文科学分野では，文献の中に現われるある単語の出現頻度や語の変化を調べることといった要求があるので，正確でかつ検索性に優れた全文テキストを必要とする．これには，単に文字列としてのテキストを提供するだけではなく，章や節といった文書の構成に関する情報，注釈や書き込み等も含めて電子化する必要がある．この点は閲読主体のドキュメトデリバリとの大きな相異点である．また，電子テキストの作成にあたって文書構造の定義のためにSGMLが広く利用されている．このようにして電子化されるテキストは高品質であるが，それゆえに生産コストも高くなる．TEI（Text Encoding Initiative）は電子テキストの共有を進めるため，人文科学分野の資料のためのDTD（Document Type Definition）を定義し配付している[23]．また，多国語や古代語を扱うことを考慮しなければならないので文字コード系も大きな問題であり，Unicodeを利用している例もある[17]．

2.5 そ の 他

●総合的情報環境としてのディジタル図書館

総合大学には様々な分野の研究者，学生がおり，専門分野に応じたいくつかの図書館が設けられている．各図書館は互いに協力し総合的な情報環境を形成している．同じように総合的なディジタル情報環境を形成するには様々な分野のディジタルコレクションを構築する必要がある．たとえば，本特集にあるミシガン大学では様々なディジタル図書館プロジェクトを進めている．いくつものプロジェクトがキャンパスネットワークと組み合わさり，これら全体で総合的なディジタル図書館環境，いわばキャンパス知識情報基盤を形成することになる．そこでの利用者の情報アクセス活動は，従来の図書館を基盤としたものとは大きく異なる可能性を持っている．

●公共図書館での活動

公共図書館のディジタル図書館に関連した活動は，現在のところ主として利用者のインターネットへのアクセスポイントとして働くことおよび情報アクセスの援助にある．たとえば，カリフォルニア州では公共図書館を利用して州の情報へのアクセスを進めており[18]，ミシガン州ではミシガン大学と協力して住民のためのMichigan Electronic Libraryと呼ぶプロジェクトを進めている[19]．

3. ディジタルコレクションの作成

3.1 全文テキスト（FullText）とページイメージ（Page Image）

印刷された図書・雑誌をディジタル化しディジタルコレクションを作成することを主としたプロジェクトが最も多い．こうしたプロジェクトには過去に遡って資料を入力（遡及入力）することを含むものもある．ディジタル化に際して全文テキストを採用すべきか，あるいはビットマップイメージによるページイメージを採用すべきかという点が設計上の大きな別れ目となる．閲読が主であるかそれとも検索が主であるか，入力に関するコストと道具が用意できるかという点が重要である．たとえば，OCRは英語や日本語等需要の多い言語には用意されているが必ずしもすべての言語には用意されていない．また，古書の場合認識率が著しく低下する場合もある．下に両者の性質を対比する．両者の中間的なものとしてPostscript形式があり，テクニカルレポートなどの配布用に多く使われている．しかしながら，遡及入力には向かないこと，フォントなど利用者側の環境に依存する部分が多いことといった問題がある．

FullText

- 検索などテキストそのものを計算機で利用する用途に向く．
- 文書の論理構造を反映することができる(SGML)．
- データサイズが小さい．

ページイメージ
- (ページ単位の) ブラウジングに向く.
- フォントを必要とせず, かつ印刷レイアウトを反映することができる.
- 作成コストが低い.

3.2 マークアップ言語

SGMLテキストの特徴は文書の構造定義に基づいた検索が可能な点であり, SGMLとその検索ツールは電子テキストのコレクションを形成する上では必要不可欠な道具である. 一般的に文書の構造は著者や出版社が決めるため多様である. 一方, ディジタル図書館の立場からは多様な種類の文書を対象とする検索機能を実現する必要がある. たとえば, 雑誌記事の全文データベースの場合, 複数の雑誌にまたがる検索を行うことは当然の要求であろう. また, SGML化のコストは低くないので文書そのものを共有することも重要である. そのため, 前述のTEIのようにDTDを共有化 (あるいは共通化) することが望まれる.

WWWのドキュメントの記述形式を決めているHTML (HyperText Markup Language) はSGMLに基づいて定義された文書構造の1つである. しかしながら, HTMLは見栄えを決めることが中心で, 文書構造を表すようには作られていないため, 全文を対象とする検索が可能ではあっても, 文書の構造を反映した検索は困難である.

3.3 イメージデータ, マルチメディアデータ

前述の国立図書館のプロジェクトでは写真や絵画, 地図等のイメージデータをスキャン入力してディジタルイメージコレクションを作っている. 音声データやビデオデータに関しても同様にディジタルコレクションの形成が進められている. CMUのInformediaは音声認識技術, 画像認識技術, ビデオ画像のシーンの切り出しや索引付けの自動化技術, ディジタルビデオ技術を組み合わせた新しいLibraryの代表であろう. UCSBのAlexandriaプロジェクトはいくつかの機関と共同して地図や航空写真をディジタル化し大規模なコレクションを形成しようとしている. 大規模なイメージライブラリを作るためにいくつもの機関で作成されたイメージデータの相互利用性を高めることが重要である. そのため, イメージデータおよび付随するデータに関するデータ (メタデータ) が重要な役割を果たしている.

4. 情報アクセスのための利用者支援

検索と閲読のためのユーザインタフェースはディジタル図書館の利用性を高めるための重要な要素である. Informediaは音声認識と自然言語理解機能を組み合わせた対話機能を持つ. UMDLでは, 利用者に応じた対話と検索を実現するため, 利用者のタイプに応じて対話する知識を持つエージェントを実現しようとしている.

情報の可視化 (Information Visualization) は情報アクセスを助けるため, また利用者のナビゲーションのために非常に重要な技術である. UIUCのInterspaceではINSPECのシソーラスをグラフィカルに表示し, 適切な検索語の選択を支援する機能を持っている. また, 諸橋[3], 谷[10], Rennison[4]やLieberman[2]等, 情報の可視化ツールはディジタル図書館にとって重要なツールである.

ディジタル図書館にとって図書館員は重要な知的情報資源である[12]. ディジタル図書館では遠隔地にいる利用者が図書館員の助けを得ながら作業すること, また利用者同士が共同して作業することもある. そのため協調作業支援ツール (CSCWツール) も重要な要素技術と考えることができる. 筆者などは, GUIを持つオンライン目録システム (OPAC) の利用方法を教えるという一種の参考サービスを協調作業支援ツールを利用して遠隔地にいる利用者に提供するという実験を行い, 肯定的な結果を得た[9]. また, 機械翻訳技術を利用した情報アクセス支援ツールへの試みもある[11].

5. インターネット上の情報の組織化と利用

現在インターネット上では全文検索サービスやメタインデックスを利用することで所望の情報の所在情報 (URL: Uniform Resource Locator) を手に入れることができ, インターネットをあたかも図書館のように利用することも可能ではある. しかしながら, サービスの提供者が自らコレクションを形成していない点で従来の図書館とは大きく異なっている. 従来の図書館では, 基本的に図書館が与える資料の識別子は長期間有効であり, 資料の2次情報 (書誌情報) も一定の規則の

下に作成されている．一方，インターネットの場合には，資料の信頼性など出版に関わる問題以外に，長期間有効な資料と短期間しか有効ではない資料が混在していること，実質的に資料の識別子として使われているURLが資料の格納場所のアドレスでしかないため資料の同定が安定的に行いにくいこと，2次情報を作成するための規則が確立されていないこと等の問題がある．こうした問題に関して，インターネット上の資料を識別するための記法URN（Uniform Resource Name）に関する検討が進められている[16]．また，OCLCを中心として，2次情報を作成し情報アクセスをよりシステマティックに行えるようにするため，インターネット上で提供される情報に関する情報（メタデータ）に関する検討が進められている[15]．

6. その他の関連技術

6.1 多言語文書

図書館には古代語を含む多様な言語で書かれた資料が蓄積されてきているので，ディジタル図書館にとって，多国語の問題は非常に重要である．日本の場合，これに加えて外字や異体字の問題もある．ところが，これまでのコンピュータとインターネットは英語（ASCII文字）を基本とし，それに各国（地域）が自国語を加えるという形で発展してきたため，多国語を扱う環境は整備されてこなかった．

現在，ISO-2022-JP-2やUnicode（ISO-10646-1）といった多国語対応の文字コード系があり，議論も盛んである．多国語テキストに関する問題は，テキスト編集，清書・印刷，検索，閲読といった機能に分けて考える必要がある．ディジタル図書館の観点から必要とされる基本的機能は検索と閲読である．前に示したように資料の利用のしかたは分野によって異なるので文字コードに対する要求も異なる．テキスト検索を基本とする電子テキストやOPACの場合，多言語対応の文字コードとそのフォントを利用できることが重要である．一方，閲読を中心とする場合にはページイメージが利用できれば，多言語であることは問題にならない．

WWWでは多言語文書に関する規定作りを進めており，送信データのヘッダの中にコード系の指示を含めることを提案している[13]．しかしながら，すべてのクライアントがすべての文字コードセットに対応するフォントを持つことは，現在のところ実際的であるとは思えない．そのため，WWWのように閲読中心の応用にはページイメージを送る方法，DeleGate[8]が採用している非ASCII文字列をインラインイメージにして送る方法，動的に必要なフォントグリフだけをテキストと一緒に送る方法[7]等が有用であると考えられる．

6.2 課 金

従来の図書館では資料の閲覧が一時には1人の閲覧者に限られており，また資料の閲覧と複写サービスとは別々である．ところが，ディジタル図書館の場合は同時に何人もの利用者が遠隔地からでも利用でき，かつ閲覧と複写の境目があいまいである．そのため，図書館が無償で利用者に資料を提供することに対して出版社からの許諾が得られることは一般的ではない．従来のドキュメントデリバリの場合は，サービスの提供者が出版社と契約を結び利用者は有償でサービスを受けてきた．前述のネットワーク上でのドキュメントデリバリは評価段階であったり，大学内に限ったサービスをしていることもあって，現在のところ無償でドキュメントが利用者に提供されているが，今後有償のサービスが一般的になると考えられる．その際，機関利用者と個人利用者といった課金の対象に関する問題，また特に個人を対象とする場合に，「小額を広い範囲から電子的に集めること」といった問題がある．たとえば，UMDLやInterspaceではCMUのNetBill[20]を基礎にして個人利用者への課金方法を検討している．

7. おわりに

ディジタル図書館は，世界情報基盤（Global Information Infrastructure, GII）上での重要な応用と位置づけられたこともあって現在非常にホットな話題である．図書館と一口に言っても，その役割は多様であるため，必要とされる技術的要素もディジタル図書館の目的に依存する部分が多い．いずれの場合にも大量でかつ多様な情報の中から必要な情報をネットワークを介して多様な利用者に提供することを目的とし，かつ単に大規模データベースの開発ということに陥ることなく「図書館」が持つ機能を反映したシステムの構築

を目指していると思える．ディジタル図書館の発展による総合的な情報環境ができあがるまでにはまだ時間がかかると思われるが，本特集のミシガン大学の記事にもあるように21世紀を目指したいろいろな試みが進んでいくことと考えられる．

参考文献

1) Imai, M. et al.: Design of a Digital University Library: Mandala Library, Proc. of ISDL'95, pp. 119-124 (1995).
2) Lieberman, H., Power of Ten Thousand: Navigating in Large Information Space, Proc. of UIST '94, pp. 15-16 (1994).
3) Morohashi, M. et al.: Information Outlining —Filling the Gap between Visualization and Navigation in Digital Libraries, Proc. of ISDL'95, pp. 151-158 (1995).
4) Rennison, E.: Galaxy of News: an Approach to Visualizing and Understanding Expansive News Landscape, Proc. of UIST'94, pp. 3-12 (1994).
5) Renoult, D.: Digitizing Program of the French National Library, Proc. of ISDL'95, pp. 87-90 (1995).
6) Renoult, D. (杉本重雄訳)：フランス国立図書館における資料ディジタル化計画，情報管理，Vol. 38, No. 11, pp. 981-985 (1996).
7) Sakaguchi, T. et al.: A Browsing Tool for Multi-lingual Documents for Users without Multi lingual Fonts, Proc. of DL'96, pp. 63-71 (1996).
8) Sato, Y.: What is DeleGate?, http://www.etl.go.jp:8080/etl/People/ysato@etl.go.jp/DeleGate/
9) Sugimoto, S. et al.: Enhancing Usability of Network-based Library Information System—Experimental Study on User Interface for OPAC and of a Collaboration Tool for Library Services, Proc. of DL'95, pp. 115-122 (1995).
10) Tani, M. et al.: User Interfaces for Information Strolling on a Digital Library, Proc. of ISDL'95, pp. 167-174 (1995).
11) Yamamoto, H. et al.: W3-PANSÉE: WWW Machine Translation System that Supports the Comfortable Internet Surfing, Proc. of ISDL'95, pp. 159-166 (1995).
12) 山本毅雄：電子図書館員の仕事とその道具，ディジタル図書館，No. 1, pp. 29-38 (1994).
13) Yegeau, F. et al.: Internationalization of the Hypertext Markup Language, Internet Draft. 37 p. (1995), http://www.alias.com:8085/ietf/html/draft-ietf-html-i 18 n.txt
14) Directory of Electronic Text Centers, http://www.ceth.rutgers.edu/info/ectrdir.html
15) D-lib Working Groups, Metadata, http://www.dlib.org/groups/metadata.html
16) D-lib Working Groups, Naming Object in the Digital Library, http://www.dlib.org/groups/naming.html
17) Electronic Text Center—University of Virginia, http://www.lib.virginia.edu/etext/ETC.html
18) InFoPeople Project, http://www.lib.berkeley.edu:8000/
19) The Michigan Electronic Library, http://mel.lib.mi.us/
20) NetBill Project Home Page, http://www.ini.cmu.edu/NETBILL/home.html
21) Special Issue on Digital Libraries, Communications of ACM, Vol. 38, No. 4 (1995).
22) Special Issue on Digital Library Initiative, IEEE Computer, Vol. 29, No. 5 (1996).
23) The Text Encoding Initiative, http://info.ox.ac.uk/bnc/tei.html

(平成8年4月9日受付)

付録 ディジタル図書館関連の情報資源

1) D-libのホームページ，http://www.dlib.org/，D-lib magazine (Corporation for National Research Initiatives が電子的に出版している月刊誌) ほかのDL関連情報．
2) 米国議会図書館，http://www.loc.gov/，National Digital Library や米国政府等の情報が提供されている．
3) IFLA: International Federation of Library Associationのホームページ，http://www.nlc-bnc.ca/ifla/home.html: DLに関するメタインデックスがあり，研究・開発プロジェクト等のホームページにアクセスしやすい．
4) UKOLN: UKOLN (UK Office for Library and information Network) のホームページ，http://ukoln.bath.ac.uk/：英国のJISC (Joint Information Systems Committee) の下に進められているディジタル図書館関連プロジェクトを知ることができる．
5) DLnet，http://www.DL.ulis.ac.jp/：図書館情報大学で開催しているディジタル図書館ワークショップの情報，およびそこで発表された講演論文を収めている．

杉本 重雄（正会員）

1953年生．1982年京都大学大学院工学研究科博士後期課程情報工学専攻修了．工学博士．現在図書館情報大学図書館情報学部助教授．ディジタル図書館，ユーザインタフェース，計算機言語等に興味を持つ．IEEE-CS, ACM，電子情報通信学会，ディジタル図書館学会，日本ソフトウェア科学会，人工知能学会各会員．