ODBと全文検索エンジンの連携による人文系DB構築システムと電子図書館

丸山勝巳†

国文学，史学などの人文系研究においても，本格的なデータベース構築とその検索閲覧サービスが望まれている．この分野では情報間の複雑なつながり関係や多様な非数値データの記録が要求されるので，表形式を基礎とする関係データベース（RDB）では扱い難い場合も多い．そこで，オブジェクト指向データベース（ODB）を活用してDB構築から電子図書館サービスまで統合したシステムを開発し効果を上げた．本稿では，(a) 情報間に複雑なつながりを含む人文系データベースのODBによる容易な実現，(b) 多数の一般ユーザに簡便な検索閲覧サービスを提供するために，ODB内容をSGMLファイル化して全文検索エンジンによりサービスする電子図書館システムの有効性，(c) 統合システム内インタフェースとしてのSGMLの効用，を具体的に述べる．

Humanity Database Compilation and Digital Library Service by Cooperating ODB and Fulltext Search Engine

KATSUMI MARUYAMA†

Even in humanity research fields, large scale Databases and Digital Library service are strongly required. In these fields, intricate inter-data relations and various non-numerical data are treated, and relational database (RDB) systems are sometimes not flexible enough to handle them. To answer these humanity requirements, an object-oriented database (ODB) system and a digital library system using a full-text search engine have been developed. It has been proved that (a) ODB is very suitable to handle mutually related humanity data, (b) a digital library system using an SGML-based full-text search engine is easy to use and convenient for public service, and that (c) SGML is a suitable interface to interwork an ODB system and a digital library system.

1. はじめに

国文学，史学などの人文系研究においても，本格的なデータベース構築とその検索閲覧サービスが望まれている．この分野のデータベースは，目録DB，数値文字DB，フルテキストDB，画像DBほか多種多様である．スプレッドシートや関係データベースが効果的に適用できるものも多いが，複雑なものになると固定的なスキーマでは扱い難い．また，特に資料と資料の間の関係を適切に表現でき，関連する資料を次々と効率良く検索できることが重要視される[1)]．

また，大きなDBの内容構築は非常に大変な作業であるので，人文系の人にも使いやすいDB内容構築環境が重要である．さらにDBの検索閲覧に関しても，ネットワークを介してどこからもアクセスでき，だれでも簡単に検索でき，関連資料を効率良く検索しまわれるような検索閲覧サービス環境が望まれる．

著者は，巨大で複雑な内容を持ち，資料間の関係の表現が特に多いDBについて，オブジェクト指向データベースを適用したDB内容構築システムを開発して，その効果を確認した．また，このDBの検索閲覧サービスとしては，全文検索エンジンとインターネットWWW環境を利用したシステムを開発し，電子図書館サービスの一部に組み入れ，ユーザの好評を得ることができた．

本稿では，複雑な人文系DBを開発しようとする人の参考になるように，オブジェクト指向DBの利用例，DB構築環境およびDB検索閲覧サービス環境の実装経験を述べる．

2. 背景と新データベースの目標

国文学研究資料館（文部省大学共同利用機関）では，CTS（computer typesetting system）出版の内容作

† 学術情報センター
National Center for Science Information Systems

成のために，長年にわたって大型計算機と関係データベースシステムを用いて我国の古典籍に関する著者，著作，書志のデータを蓄積してきている．その規模は約400 MB（著者7万件，著作43万件，書誌17万件）に達する．ここに“著作”とは源氏物語といった作品を意味し，“書誌”とは各所で保存されている個々の写本，刊本を意味する．データは非常に項目豊富で込み入っており，たとえば書誌は書名，別書名，形態，書写者，出版者，注記，所蔵者など数十項目を持ち，その大部分は可変長文字列であり，繰返し項目も多く，また関連情報に結びつけるための情報も多い．

このシステムはCTS出版が目的であったため，一般ユーザへのオンライン検索サービスの機能は持っておらず，DB構築環境もラインコマンドベースで今となっては旧態である．今回，メインフレームからワークステーションへの移行を機会に，システム機能としては以下の改善を行うことにした．

- DBの内容構築者のための環境（以降，“DB構築環境”と呼ぶ）の抜本的改善．
- DBの内容利用者のための使いやすい環境（以降，“DB検索閲覧サービス環境”と呼ぶ）の提供，特に電子図書館サービスとしての不特定多数へのDB検索閲覧サービスの提供や，DB内容のCD-ROM出版の容易化．

また，DB内容としても，膨大な量の古典籍や著者の情報をいろいろな観点から容易に検索でき，自在にリンクをたどって関連情報にトラバース（traverse）ができ，そこからさらに原本画像や翻刻テキストも閲覧できる統合データベースとして抜本的に改善することとした．

たとえば，表題に“蝦夷”を含み種別が“地誌”の著作を検索すると該当する著作リストが表示され，その中の“北蝦夷図説”を見ると作品の詳細な情報，著者“間宮林蔵”の詳細情報にアクセスするための情報（これを以降リンクと呼ぶ）や個々の刊本の書誌情報へのリンクが表示されている．著者リンクを選ぶと，著者の詳細な情報や全作品へのリンクが表示され，またいずれかの書誌リンクを選択するとその刊本の書誌情報，所蔵者情報，原本画像や翻刻テキストへのリンクが表示される．原本画像リンクを選択すると原本画像を見ることができるというようなシステムである．

DB内容構築時にもDB検索閲覧サービス時にも，このように検索論理式で目的物を検索し以降はリンクをたどってトラバースするといった使い方が多いので，関連情報のリンクを簡潔に表現できるDBでなければならない．

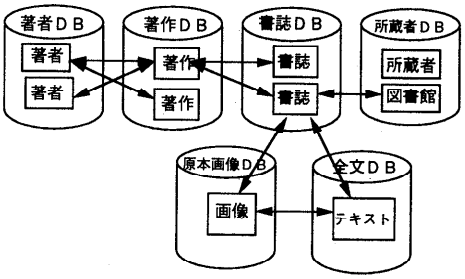

図1 新データベースシステムの目標
Fig. 1 A new database system plan.

3. 本システムに対する要求条件

本システムに対する要求条件は，統合データベース化，DB構築環境の改善，電子図書館サービスによる快適なDB検索閲覧サービスの実現に大別される．

3.1 統合データベース化への要求

図1に示すように著者DB，著作DB，書誌DBなどを含み，以下の機能を持つ統合データベースシステムを実現すること．

- 著者DBは，本名，別称，職業，出身地，生没年などの著者情報を有する．将来は歴史人物をも含み系図なども追加して総合人名DBに発展できることとする．
- 著作DBは，作品名，別書名，成立年，分類，注記などの著作情報を含む．
- 書誌DBは，個々の写本や刊本の表題，内題，刊年，写者や書肆名，形態，所蔵者情報などの書誌情報（個々の写本や刊本の属性）を含む．
- 著者と著作，著作と書誌の間などには両方向リンクが設定してあり，たとえば著者情報，著作情報，書誌情報，所蔵者情報を，任意情報から始めてどちらの方向にも自在にたどれる．1人の著者が複数著作を持ちうるし，共著もあるので，これらのリンクは一般に多対多の関係である．
- 叢書，合綴など，1つの表題の中に複数の子作品が含まれるものがある．たとえば勅撰和歌集の二十一代集は，古今和歌集，後撰和歌集などの21個の歌集の総称である．また子作品も叢書であって，さらに子作品を含むこともある．このような親子関係を著作構造や書誌構造と呼び，相互にリンクでたどれることとする．
- 原本画像DBがあり，個々の写本や刊本の書誌情報からは，その原本画像にリンクでたどれる．
- フルテキストDBがあり，個々の写本や刊本の書誌情報から翻刻テキストにリンクでたどれる．また，原本画像と翻刻テキストとは，ページ対応で

双方向リンクが張られている．

- 書名，備考をはじめ大部分が可変長文字列フィールドで，かつ中間一致検索が頻用されるので，この検索が簡便で高速であること．

3.2 DB構築環境への要求

DB内容の構築者が，パソコン端末からLAN経由でUNIXサーバ上のDBMSにアクセスして，新規データ入力やデータ修正が行える環境とする．この際の作業形態としては，以下の両手法が必要である．

- DB構築端末から直接DBMSに入力や修正を行うオンライン手法．
- DB構築端末で入力データや修正データのファイルを作り，後で一括してDBMSに送るオフライン手法．

コンテンツ構築作業の効率化のために，使いやすいGUI（Graphical user interface）を持った端末プログラム（以下DB構築端末と呼ぶ）が必要である．

端末は，Windows，Macintosh，UNIXのいずれも使えることが望ましい．

3.3 DB検索閲覧サービス環境への要求

電子図書館サービス[2),3)]による多数ユーザへの検索閲覧サービスは，本DBの最重要目的の1つである．

人文系ユーザが対象であるので，検索はDB内のデータ構造を意識せずに検索できること，複雑な検索論理式が不要であること，長文の中の部分一致検索が簡単で高速であること，関連項目へのトラバースが簡単で高速であることが必要である．

さらに，小中規模のサーバマシンで同時に多数ユーザにサービス提供できることも必要である．

人文系研究者は自宅書斎で仕事をすることが多く，CD-ROMや印刷本としての出版も望まれるので，CD出版やCTS出版用にタグ付けしたSGML[4)]ファイルも出力できる必要がある．

4. システム設計方針

前述の要求条件に対して，以下の設計方針をとった．

4.1 DB管理システム：オブジェクト指向DBの利用

DBMSの種別としては，関係データベースRDBとオブジェクト指向データベースODBとが考えられる．今回目標のDBは，原理的な表現力からいえばRDBでも対応できなくはないが，以下の理由でODBを採用した．

- 本DBの著作，書誌などの論理資料単位は，繰返し項目，可変構造項目などを含むので，RDBでは複数テーブルを使った込み入った結合（join）操作が必要になるが，ODBでは可変長配列や複合オブジェクトを用いて簡潔に表現できる．
- 本DBは，関連資料間の多段にわたるリンクが頻繁に表現される（例：著者と著作の関係，著作と書誌の関係，叢書における親子関係，人物の系図など）．RDBでは複雑な構造を持つ識別子キーが必要になるが，ODBではオブジェクト間リンクとして直観的かつ簡単に表現できる．
- 本DBでは，関連資料間のリンクをたどるトラバースが頻繁に使われる．これはRDBでは結合操作で実現できるが，人文系ユーザにとっては検索式の記述が難しく，また計算機にとっても計算量の多い処理である．ODBでは，ユーザは目的リンクを選定するだけでよく，計算機にとってはアドレスをたどるだけの簡単な処理なので性能も上がる．
- 各オブジェクトクラスに特有な変換や操作をメソッドとして用意することで，高度の機能を使いやすい形で提供できる．

関係データベースにオブジェクト指向概念を追加したオブジェクト指向関係DB（ORDB）も表現力は高いので候補となる．しかし，ORDBでは応用プログラムはSQL-APIを使ってデータベースにアクセスしなければならないので，プログラム記述が長くなる．これに対しODBでは，応用プログラムはデータベース内の永続オブジェクトも普通のオブジェクトと同様にアクセスできるので，プログラム記述が容易である．この理由で，本システムではODBであるObjectivity/DB[5)]を用いた．

4.2 DB構築環境の構成

システム構成は，**図2**の左側に示すようにUNIXサーバ上のDBMSとパソコンのDB構築端末からなるClient-Server型の分散システムとした．

DB内容の構築者は，DB構築端末のGUIを用いて以下作業ができるようにした．

- 新規データ入力とチェック：GUIの各項目に入力されたデータに対し，構文および値の厳格な検査を行い，DBMSに送る．
- 各種の検索と表示：DBMSにコマンドを送り，検索条件に合致したオブジェクトの表示や，リンクをたどった表示を行う．
- DB修正：DBMSから読み出したデータを表示画面上で修正して返送することで，簡単にDBMS内の既存データの修正を行う．

要求条件で述べたように，この際の作業形態としてはオンラインとオフラインの両手法をサポートする必

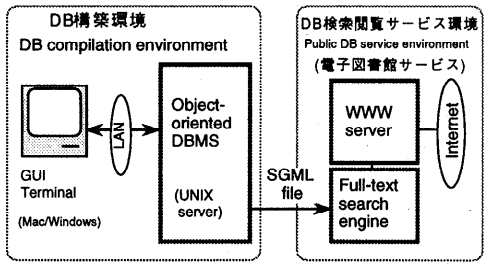

図2 システムの全体像
Fig. 2 Total system outline.

要がある．そこで，DBMSとDB構築端末はSGML[4]でタグ付けしたSGMLテキストで通信しあうようにした．たとえば，DBMSは検索要求を受けるとその結果をSGMLテキストに編集してGUI端末に返送する．DB構築端末はこれを受信して画面表示する．ユーザはその画面に修正を行ってただちに返送してもよく，また，いったんファイルに落として後でオフライン修正してDBMSに送ってもよい．DBMSから見れば，小さなSGMLテキストが小刻みに送られてくるか，オフライン作成された大きなSGMLテキストが一括して送られてくるかの差でしかなく，両者は同一処理で済む．

端末機種はWindows，Mac，UNIXが混用されているので，同一プログラムで全機種に対応させるためにJava[6]で開発することとした．

4.3 DB検索閲覧サービス環境

本DBの重要な利用の1つに，電子図書館サービスとしての多数ユーザへの検索閲覧サービスがある．電子図書館サービス環境としては，インターネットのWWW環境が適切な機能を使いやすい形で提供しているのでこれを利用することとした．

電子図書館システムのWWWサーバの下で検索閲覧サービスを提供する手法としては，DBMSの検索機能を用いる方式と，**全文検索エンジン**を用いる方式とが考えられるが，以下の理由で図2の右側に示すようにSGMLファイルを仲介とした全文検索エンジン方式とした．

- DB構築者はデータ構造を意識した複雑な論理式検索も必要とするが，電子図書館サービスではデータ構造を意識せずにだれもが直観的に使える全文検索とトラバースの方が好まれる．
- 電子図書館サービスでの検索では，完全一致よりも部分一致型の文字列検索が多用される．今回調査したオブジェクト指向DBMSの文字列検索機能は，完全一致・前方一致はインデックス機能により高速化できるが，中間一致は高速化されない．複数フィールドにまたがった検索は複雑な検索式が必要である．これに対し，全文検索システムならば，融通性も検索速度も満足できる．
- 構造化テキスト（たとえばSGML）対応の全文検索エンジンを使えば，簡単な検索式で文書構造に依存した検索も高速に行える．
- DB内のオブジェクト間リンクを電子図書館ユーザのブラウザ上にハイパーリンクとして実現することで，関連情報の閲覧も容易に行える．
- CD-ROM/CTS出版用のSGMLファイル作成プログラムに簡単な機能追加をするだけで，全文検索エンジン用のSGMLファイルを作成できる．
- 全文検索システムの方がより多数の同時アクセスに対応できる能力がある．
- データベース更新の結果は数時間で電子図書館サービスに反映できれば十分である．

5. DB管理システムの実装

DBMSは，DB構築端末の要求に応じてデータ入力，情報検索やトラバース，データ修正，各種ファイルやリストの出力などを行う．また，メインフレーム上の既存RDBからのデータ移植機能も含む．以下に実装技術を述べる．

5.1 スキーマとオブジェクト表現

本DBの主要オブジェクトと，その間の主要なリンク関係を図3に示す．著者，著作，書誌といった主要な論理実体は，それぞれAuth-obj，Work-obj，Biblio-objと条件に応じて付加されたオブジェクトからなる複合オブジェクトとして表現されている．共通属性と拡張属性の関係は継承機能により簡潔に表現でき，RDBで面倒な繰返しフィールドもODBでは配

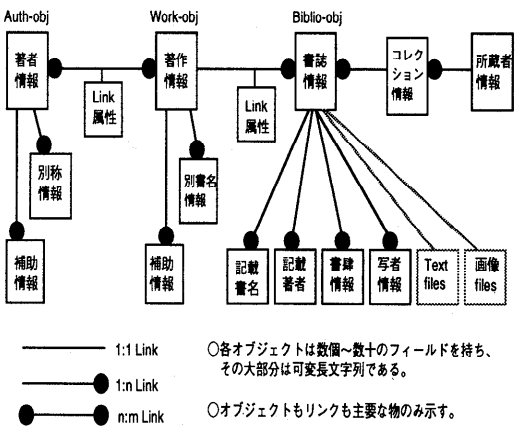

図3 主要オブジェクトとその間の関係
Fig. 3 Main objects and their relationship.

列や集合として簡単に表現できる．

エンティティ間のリンクは，RDB では結合操作による間接表現となるが，ODB ではオブジェクトポインタで直接に表現できる．一般にオブジェクト指向 DBMS は，オブジェクトポインタを強化したアソシエーション（association）機能[5),12)]を持っており，これにより関連オブジェクト間を自在にたどれるリンクが組め，処理速度も速く，片方の修正を自動的に相手側に反映させてデータの一貫性も保証される．アソシエーションは 1：1，1：m，m：n のすべての関係を表現できるので，図 3 に示すように，オブジェクトの相互関係が直観的に表現されている．たとえば，複数著作や共著の関係は m:n のアソシエーションとして表現される．

これらのリンク（DBMS 上ではアソシエーションで実現）は，DB 構築端末や電子図書館サービスのブラウザではリンクボタンとしてグラフィカル表示されており，クリックするだけで瞬時にオブジェクト間をトラバースすることができる（7 章参照）．

図 4 に簡略化したスキーマ定義とプログラム例を示す．ooRef() および ooHandle() は，Objectivity/DB のオブジェクト参照子を意味する．このように，DBMS 内の永続オブジェクトも普通のオブジェクトと同様にアクセスできる．Auth 定義の

```
ooRef(Work) workLink[] <-> authLink[];
```

と Work 定義の

```
ooRef(Auth) authLink[] <-> workLink[];
```

は，著者と作品がフィールド Auth::workLink と Work::authLink によって相互関連していることを示す．配列記号 [] は複数収容を示すので，これは m:n 関係を意味する．

```
class Auth: public ooObject {
  int          aid;
  ooVString  name;
  ooVString  yomi;
  ....
  ooRef(Work) workLink[ ]<->authLink[ ];
  ....
  .....
public:
  void Auth(char* _name ...)
  {    .....      }
  void setName(char* _name, char* _yomi)
  { name = _name;  yomi = _yomi;
  }
  .....
  .....
}
```

(A) Author object

```
class Work: public ooObject {
  int          wid;
  ooVString  title;
  ooVString  yomi;
  ....
  ooRef(Auth) authLink[ ]<->workLink[ ];
  ....
  .....
public:
  void Work( )
  {    ....    }
  void setTitle(char* _title, har* yomi)
  { title = _title;  yomi = _yomi;
  }
  .....
  .....
}
```

(B) Work(作品) object

```
ooHandle(Auth)  a1;  // Author object handling
a1 = new Auth( );
a1->setName("紫式部", "むらさきしきぶ");
 ....
ooHandle(Work)  w1; // Work object handling
w1 = new Work( );
w1->setTitle("源氏物語", "げんじものがたり");
 ....
w1->add_authLink(a1); // Associate "w1" and "a1"
```

(C) Object handling example

図 4 永続オブジェクトの定義と操作
Fig. 4 Persistent object definitions and operations.

5.2 DB 間の関連情報リンク設定処理

本システムでは，著者 DB，著作 DB，書誌 DB の個別 DB にデータ入力が済んだ時点では各 DB 内に閉じたリンクのみが設定されおり，DB 間にまたがったリンクは未設定である．たとえば，著作 DB 内の“源氏物語 obj”には《name=“紫式部”, yomi=“むらさきしきぶ”》などの“相手検索情報”が記録されているのみである（相手検索情報としてはその他の属性も使われる）．

そこで，著作 obj に対して，著者名などの相手検索情報で著者 DB を検索して目的オブジェクトを見つけてリンクを設定するのが本処理である．この際に，相手検索情報を満足するオブジェクトの検索結果としては，以下の 4 ケースがある．

- ユニークに決定．
- 複数候補あり（たとえば同名の著者）．
- 類似候補あり（たとえば表記と読みのうち，一方だけが一致の場合）．
- 候補なし．

DB 構築者は，この結果を見て正しいものを選択したり，誤データの修正や不足データの入力をしたりする必要があるので，この作業の容易化が重要である．

このため，DBMS の DB 間リンク設定処理は，“リンク先候補検索フェーズ”と，“リンク設定実行フェーズ”の 2 フェーズで行っている．リンク先候補検索フェーズを実行すると，DBMS は検索結果を SGML テキストに編集して DB 構築者に報告する．この SGML テキストには，《リンク元の主要フィールドの内容とこのオブジェクトの ID 番号》と，リンク先候補の個数分の《リンク先候補の主要フィールドの内容とそのオブジェクトの ID 番号》とが対応して記録されており，DB 構築者はこの SGML テキストを見て，次の処理を行う．

- 候補がユニークに決定の場合は，それが正しい相手ならば了解マークを記す．正しい相手でないとしたらデータのエラーや未入力なので，データの修正もしくは新規入力を行う．
- 候補が複数の場合は，その中から正しい相手に了解マークを記し，それ以外は SGML テキストから削除する．
- 類似候補の場合は，データのエラーや未入力なの

で，データの修正もしくは新規入力を行う．

- 無候補の場合は，リンク先オブジェクトが未入力なのでそれを新規入力する．

このようにデータ修正を行い，編集したSGMLテキストをDBMSに返送して再度リンク先候補の検索を行う．すべてがユニークに決定できるようになったら，リンク設定実行フェーズを起動して実際にDB間リンクを設定する．

人文系のデータベースでは関連情報を見つけてリンクを設定したいことが多いので，この類似候補検索を強化するといろいろな発展が可能であろう．

5.3 目的別出力ファイルの生成

本システムは，電子図書館システムの全文検索システム用にタグ付けしたSGMLファイル（7章で説明する），CD出版やCTS出版用にタグ付けしたSGMLファイル，plain textファイルなどを出力できるようにした．もちろん，検索条件に一致したものだけを出力することも可能である．

データベースにSQL-APIで間接アクセスしなければならないRDBと違って，ODBでは応用プログラムが直接データベースにアクセスできるので，簡単なプログラムで要求に応じたファイルを作ることができた．

6. DB構築端末の実装

6.1 SGMLテキストによるDB構築端末とDBMS間通信

4.2節で述べたように，DB構築端末はデータ構造や項目をタグで表現したSGMLテキストでもってDBMSと交信する．これにより，DB構築端末側ではこのSGMLテキストをオンライン処理することも，ファイル化してオフラインで一括処理することも可能である（図5）．DB構築端末は以下の機能を持つ．

- 新規データ入力モードでは，DB構築端末は入力データの形式と値の正当性をチェックし，SGMLテキストに編集してDBMSに送る．
- 検索命令やトラバース命令をDB構築端末からDBMSに送ると，DBMSは結果をSGMLテキストに編集してDB構築端末に返送する．DB構築端末はこれをGUI表示する．
- 更新モードにてDB構築端末の表示画面に修正を施して送信すると，新データがSGMLテキストとしてDBMS本体に送られてDBが仮更新され，Commit/Abort時に本更新が行われる．

6.2 Client-Server構成によるDB構築端末の実装

DB構築端末プログラムは，日本語対応のJava開発キットを用いて実装した．複数のDBユーザがパソコンからLAN経由でUNIXサーバ上のDBMSにアクセスするので，GUIプログラムはユーザPC上で走るGUI Clientと，UNIXサーバ上で走るGUI Serverから成る構成とし，その間の通信にはJavaのSocketを用いた（図6）．

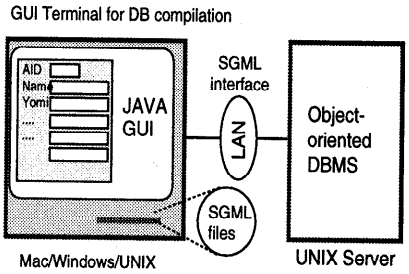

図5 DB構築端末
Fig. 5 GUI terminal for DB compilation.

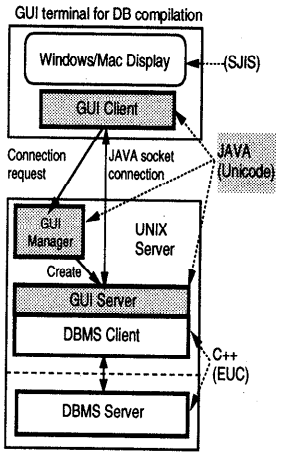

図6 DBMSとDB構築端末の機能配分
Fig. 6 DBMS and GUI terminal interworking.

ユーザのDB接続要求はGUI Managerに送られる．GUI ManagerはGUI Serverの生成とDBMS Clientの起動を行い，GUI ClinetとDBMSとを結び付けるので，GUI Clientの数だけGUI ServerとDBMS Clientが動的に生成される．

ここでDBMS（Objectivity/DB）自体もClient-Server型であり，DBMS ServerとDBMS Clientは分散配置が可能であるので，データベースがさらに大規模になったら3階層の分散システムとしても運用できる．

日本語を扱う場合，文字コードは重要な問題の1つである．本システムでは図6に示すようにUNIX部はEUC，PC部はSJIS，JAVA部はUnicodeというように3種類のコードが共存している．Javaの入出

力機能は文字コード名を指定するだけでコードの自動変換ができるので，本システムの開発においてはコード変換を意識する必要はなかった．ただし，外字を使う場合には各々の外字を定義する必要があるので，今後 Unicode 環境が整備されたら Unicode で統一した方が良いと考える．

7. 電子図書館用の検索閲覧サービスシステムの実装

前述のように，電子図書館システムにおける DB 検索閲覧サービスシステムは，DBMS が作成した SGML ファイルを全文検索エンジンで検索し，結果をユーザに返送する．全文検索エンジンでは，高速検索を可能にするためにあらかじめバッチ作業でインデックス情報を作っておく必要がある．全文検索エンジンは，分かち書きや形態素解析に基づいてインデックス情報を生成するものが多いが，古文の場合は辞書が整備されていないため形態素解析はうまく動作しない．したがって，分かち書きや形態素解析に依存しないインデックス方式が必要である．また文書構造を指定した検索を可能とするために SGML 対応が望まれる．この両条件から全文検索エンジンとしては OpenText[7] を使用した．

前述のように，電子図書館サービスの環境には WWW 環境を利用しているので，WWW サーバが全文検索エンジンを制御する．

7.1 電子図書館サービス用の SGML ファイル

全文検索エンジン OpenText を使うには，まず (1) SGML ファイル（SGML タグ付けしたテキストファイル）を用意し，次いで (2) OpenText のインデックス作成機能を用いて，SGML ファイルからインデックスファイルを作っておく（バッチ作業）必要がある．後は，OpenText に検索コマンドを与えると，OpenText はインデックスファイルを用いて該当個所を高速に探し当て，その内容を SGML ファイルから出力したりする．RDB の SQL が検索式の中にフィールド名などを指定できるように，OpenText の検索コマンドは，検索個所や文脈などを SGML タグを用いて指定することができる．

この全文検索用 SGML ファイルは，オブジェクト指向データベースの内容から以下のように作成する．

- 本データベースは，著者 DB，著作 DB および書誌 DB から成るので，各々に対応した 3 個の SGML ファイル（著者 SGML，著作 SGML，書誌 SGML）を作成する．
- 本 DBMS 内では，“著者 obj” とその著者の “著作 obj” がアソシエーション機能でつながれているように，関連オブジェクトは著者 DB，著作 DB，書誌 DB といった DB の枠組みを越えて双方向リンクを組んでいる．SGML ファイル上では，これらを後述の手法で SGML ファイル間リンクに表現しなおす．
- 図 3 に示すように，たとえば著者オブジェクトは，複数の別称オブジェクトなどを抱えている（複合オブジェクト）．電子図書館サービスでは，DBMS 内のオブジェクト単位に表示するよりも，別称は著者情報内に入れ子で表示したほうが読者に便利である．したがって，このような DBMS 内の複合オブジェクトは，SGML ファイルでは入れ子表示に変換する．

```
<!--- SGMLテキスト例 --->
<auth>
<aid>1234</>
<name>紫式部</>
<yomi>むらさきしきぶ</>
.....
<work>
 <worklink wid=8765>源氏物語</>
 <worklink wid=9876>紫式部日記</>
 .....
</work>
 .....
 .....
</auth>
```

図 7 電子図書館用 SGML テキストでのハイパーリンク
Fig. 7 Hyperlink in SGML text for Digital Library.

著者ファイルの SGML テキスト例を図 7 に示す．このように SGML ファイル間のリンクは，単純にリンク先情報の ID 番号を用いて，たとえばリンク先が著作情報でその ID（wid）が 8765 の源氏物語だとしたら，“<worklink wid=8765>源氏物語</>” と SGML 表現している．

電子図書館サービスのユーザが，リンク先情報へのトラバースを指令すると，全文検索エンジンはこの ID を用いて検索するが，インデックスファイルが作られているので高速に検索が完了する．

DBMS から著者・著作・書誌の各 SGML ファイルを作り，これに検索用インデックス情報を生成する作業は，1 つのシェルコマンドで行える．SGML ファイルの合計規模は約 400 MB であるが，SGML ファイル作成の所用時間は約 30 分，インデックス情報生成の所用時間は 4 時間強であった（Sparc Server 1000）．

7.2 処理の流れと CGI

電子図書館システムでの処理は，図 8 に示すように汎用 WWW サーバ，CGI（Common Gateway In-

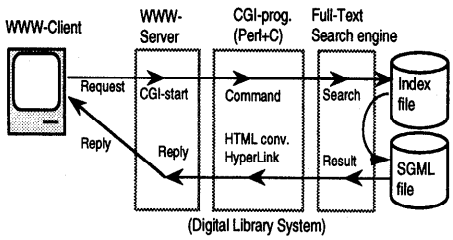

図8 全文検索エンジンを使った電子図書館サービス
Fig. 8 Digital library service using Fulltext search engine.

terface program)[8]と全文検索エンジンによって行われる.

CGIは以下の処理を行う専用のプログラムで, Perl[9]とCで記述した. サイズは各々400行と1400行程度である.

- 入力情報を分析し, 検索コマンドを編集して全文検索エンジンOpenTextを起動する.
- OpenTextが返したSGMLテキストに対し, WWWサービスのためのHTMLタグの追加, SGML → HTMLタグ変換, ヒットした検索語を着色表示するタグ付けなどを行う.
- WWW上でハイパーリンクを実現する. 前節のようにSGMLテキスト上ではハイパーリンクは相手オブジェクトのID情報として表現されている. これを該当オブジェクトを検索するCGIコマンドのリンクに変換する. たとえば, "`<worklink wid=8765>`源氏物語`</>`" を "`<a href="../cgi-bin/search?wid=8765">`源氏物語`</>`" という形式のHTMLタグに変換する. これは『searchという名前のCGIプログラムをパラメータwid=8765で実行せよ』を意味し, ユーザがブラウザ上でこのリンクをクリックすると, wid=8765のオブジェクトが検索され表示される.

8. 関連研究

本システムは, 情報間の関連の表現に重きを置いた込み入った内容を持つ人文系DBへのODBの活用, 全文検索エンジンとSGMLテキストを活用した電子図書館サービス, SGMLテキストによるシステム間の結合に特徴がある.

通常の目録DBは目録レコードの集合であり, レコード間の関係も複雑でないのでRDBの特性にあっており, 大部分はRDBによって実現されている. これに対し, 本システムは著者DB, 著作DB, 書誌DBといった自己完結的なDBの集合で, DB内およびDB間の関連するオブジェクトが相互にリンクを張り合っていて自在にトラバースできるもので, ODBを用いてこそ簡単に実現できたといえる.

電子図書館システムの研究は各地で行われているが[2),10),11)], 本報告のように古典籍を対象として著者・著作・書誌・原本画像などの間を自在にトラバースして検索閲覧することをねらったシステムは見当たらない(原本画像DBは別途検討が進んでおり, 将来結合する予定である). オブジェクト指向DBに記録された込み入った内容をSGMLテキストに変換し, それを全文検索エンジンを用いて電子図書館サービスする点も独特である. 全文検索エンジンによる非常に簡便で高速な検索は, 電子図書館ユーザである国文学者に好評である.

フルテキスト閲覧サービスに関しては, SGMLテキストと全文検索エンジンの組合せによるサービスが今後が広まると思われる. なお, 図面を持つ科学技術論文の閲覧サービスに関してはページイメージが主流といえる.

9. 評価

9.1 ODBの有効性

元情報が持っているデータ間の関係をダイレクトにオブジェクト間のリンクとして表現でき, 直観的に目的データにアクセスできるので, 人文系DBのように関連情報間のつながり表現を重視する場合には効果的である. 特にアソシエーション機能は適切にオブジェクト関係の表現とデータ管理を可能とする. たとえば, 国文学者にRDBのテーブル図を理解してもらうのは必ずしも容易ではないが, 図3のようなオブジェクト関連図ならただちに理解してくれた.

現在のDB内容はメインフレーム上のRDBから移植したものである(約400 MB)が, RDBよりも格段に使いやすくなった. また, SQL-APIを経由せずに応用プログラムから直接DBにアクセスできることによるプログラム開発の容易化も大きい.

本DBシステムの永続オブジェクトは, クラス数が19, フィールド合計が220である. 本体プログラムは既存データ移植や電子図書館用機能まで含めて約18K行(C++)であり, 1.5人年程度で開発できた.

なお, 人文系DBにはRDBで適切に表現できるものも多くあり, 問題内容に応じてRDBとODBを使い分ければよい.

9.2 ODBと標準化

RDBに対しODBの標準化の遅れを懸念する意見も聞かれるので, Objectivity/DB用プログラムの一

部を別のODBであるObjectStore[12]用プログラムに書き換える実験を行った．その結果，プログラム開発がRDBよりも格段に容易なので，異機種ODBへのプログラム書き換えも比較的容易であることを確認できた．データ自体はSGML化テキストファイルとして簡単に移植できる．

9.3 DB構築端末プログラムのJAVAによる実装

日本語対応のJava開発キットを用いて実装した．豊富なクラスライブラリーのおかげで，GUI表示，DB修正のための表示内容の編集機能，DBMSとの通信などを含めて約3K行で実現できた．同一バイナリコードでWindowsおよびUnixが走り，日本語の入出力もできるので，JAVAの採用は効果的であった．

9.4 電子図書館としての使い心地

本電子図書館サービスのユーザは，WWWブラウザから著者DB・著作DB・書誌DBのいずれかを指定し検索語を入力するだけで，それを含むオブジェクトが表示され，またリンクボタンをクリックするだけで関連情報に自在にトラバースでき，非常に簡便に必要な情報にアクセスできる．検索も短時間で処理できる．たとえば，書名に“源氏”を含む作品を検索すると，約14秒で1111件が表示される（内訳：OpenText：1秒，CGI：5秒，NW転送：9秒）．CGIに時間がかかっているが，これはPerl記述部のコンパイル時間と実行時間なので，C言語などで書き直せば高速化できる．

DB構築端末からODBMSを直接検索した場合は，複雑な検索式を入力すれば大変込み入った情報検索も1回の検索で行える，数値演算をともなう検索が可能であるなど，より強力な検索が行える（なお，本DBでは数値演算対象のデータはほとんど含まない）．文字列の検索速度に関しては，完全一致と前方一致検索はOpenText同等に高速であるが，中間一致検索は数分かかることもある．

10. おわりに

人文系データベースといっても多様であるが，その中で情報間の複雑な相互関係や込み入ったデータ構造を扱わなければならない場合には，ODBの待つ融通性は非常に有利である．また，この内容を電子図書館サービスとして使いやすい検索閲覧サービスを提供するには，全文検索方式が効果的で，実際にユーザの評判も良い．大多数のODBMSは高速全文検索機能を含んでいないが，ODBMSと全文検索エンジンはSGMLファイルを介させることで簡単に協調動作させることができる．

今後の課題は，原本画像や本刻テキストとのリンク，複数図書館にまたがった分散サービスの実現，ネットワークを利用したデータベース内容の分散協調開発環境の実現があげられる．

本システムの電子図書館サービスは，書誌情報（SGMLテキストで162 MB）に関しては著作権や原本所蔵者の公開許諾などの課題があるため，外部には著作情報（同224 MB）と著者情報（同38 MB）のみを`http://www.nijl.ac.jp/infocenter.html`にて公開している．

参考文献

1) 八村広三朗：人文科学とデータベース，情報処理，Vol.38, No.5 (1997).
2) 長尾ほか：電子図書館Ariadneの開発，情報管理，Vol.38, No.3 (1995).
3) ACM: Digital Libraries, *Comm. ACM*, Vol.38, No.4 (1995).
4) Bryan, M.: *SGML*入門，アスキー (1991).
5) Objectivity Inc.: *Objectivity/DB*説明, http://www.ogis-ri.co.jp/otc/products/objectivity/technical.html
6) JavaSoft Inc.: *JAVA online documents*, http://www.javasoft.com/
7) OpenText社：*OpenText*説明書, http://www.opentext.com/
8) The WWW Consortium: *HTTP online documents*, http://www.w3.org/
9) Warry, L., et al.: *Programming Perl*, O'Reilly and Associates (1993).
10) ディジタル図書館学会：ディジタル図書館ネットワークホームページ, http://www.DL.ulis.ac.jp/
11) 安達：学術情報センターのディジタル図書館プロジェクト，情報処理，Vol.37, No.9 (1996).
12) 石塚：オブジェクト指向データベース，アスキー (1996).

(平成10年8月13日受付)
(平成11年1月 8日採録)

丸山 勝巳（正会員）

1970年東京大学大学院工学系研究科電子工学専攻修士課程修了．同年電電公社(現NTT)入社．1995年国文学研究資料館教授．1998年学術情報センター教授．高水準言語，最適化コンパイラ，実時間システム，並行オブジェクト指向，分散OS，電子図書館，ODB等の研究に従事．1982年電電公社総裁表彰．1993年本学会論文賞．1997年電気通信普及財団賞．工学博士．