大規模なリファレーションデータベースの開発
ー参照データの作成と流通についてー

○浅井　勇夫*

Development of large scale database of cited literature with special attention to relationship between authors and any other ones citing them: Input and distribution of reference data.

ASAI Isao*

リンク情報を持たない既存の書誌データベースは，フルマッチ検索に弱点を持つ。一方，リンク情報を持つ新しいリファレーションデータベースは，検索や知識生成に有用である。この論文は，参照データの入力に関して APTS コード，作成時間，記憶容量，そしてコストを扱う。さらに，参照データの流通に関して CD-ROM，フロッピーディスク，そしてパソコン通信を考察する。ユーザ側でデータ処理する検索システムを提案する。〔著者抄録〕

Since bibliographic information databases available do not contain any interrelated data which might be useful in literature search or knowledge formation, it is too hard to understand the relationship between authors and any other ones citing them. The author introduces a new term "referation" which gives a lucid explanation of that relationship, and describes the APTS code system, machine time, memory capapcity and money needed for feeding such data into a computer. He also examines such tools as CD-ROMs, floppy disks and PC-based telecommunication for facilitating the distribution of them, and proposes a retrieval system for processing them on user's terminals. 〔Author Abs.〕

* 大阪府立大学工学部経営工学科（〒 593 堺市学園町 1-1）Tel. 0722(52)1161
* University of Osaka Prefecture, Dept. of Industrial Engineering (1-1, Gakuen-machi, Sakai, 593)

1. はじめに

文献データベース（DB）の世界にも大きな情報技術の波が押し寄せている。この10年間でパソコンは驚くほど高性能・低価格化した。CPU の高速化，記憶装置の大容量化，グラフィックスの高度利用などは現在も進行中である。それに伴ない，現存するシステムの多くは新しい視点からの見直しが迫られている。そして，情報検索や情報検索システムも新しい情報技術への対応が求められている。

ここでは，最初に文献 DB の将来を展望し，リファレーション DB を位置づける。次に，DB の入力問題について考察し，参照データの入力について述べる。最後に，DB の流通問題について考察する。

2. 文献 DB の展望とリファレーション DB

研究者の文献収集手段として，(1)購入雑誌をブラウジングして直接収集，(2) DB を検索して原文の収集，(3)参照文献から収集，(4)コミュニケーションによる収集，などがある。(2)に関して抄録型 DB のオンライン検索が発展してきた。最近，抄録型 DB の作成に時間がかかることから，書誌型 DB が作成されるようになった。

文献 DB の種類とその構成を図１に示した。情報分野の研究は検索に関するものが多く，特にフルマッチ検索の短所を回避するためのものが多い。文献 DB 自体の研究はあまりなされていないが，最近，全文型 DB や関連型 DB が注目されている。オンライン情報検索システムを開発したときに比べて利用者の情報環境は大きく変化しており，利用者の情報環境に適したシステムの開発が求められている。いずれにしても，検索を目的とした DB の必要性は高く，課題は多い。

全文型 DB	テキスト＋式＋図表
書誌型 DB	著者＋標題＋出典
抄録型 DB	書誌型＋抄録
関連型 DB	書誌型＋参照

図１　文献 DB の種類と構成

最近，全文型 DB が普及してきた。まだテキスト部分だけを DB にしたものが多いが，グラフィックスが扱えるウィンドウズが普及すれば式や図表が DB 化され，完全なものになる。そして，マルチメディアの進展とともに，文献はテキストや静止画だけでなく，著者の音声による紹介や実験結果などの動画表現など新しい形に発展する可能性がある。

冊子体の雑誌の購入は CD-ROM 版の全文型 DB の購入へと発展する。これは冊子体の抄録誌から電子化したオンライン検索に移行した場合と同様な現象である。CD-ROM 版の全文型は原文に直接アクセスできるため，研究者に大きな影響を与える。さらに書誌型 DB や抄録型 DB にも大きな影響を与えることが予想される。

文献 DB を構成する文献はバラバラに存在しており，文献間の関連は全く考慮されていない。文献間のリンク情報に関して，筆者は参照文献データから得られるリファレーション（参照文献＋文献＋被参照文献）という新しい概念を提案し，リファレーション DB の研究を行ってきた[1~4]。図２は，文献428が９編の文献とリンクしている例を示す。そして，リファレーション属性を持つ文献 DB をリファレーション DB と呼ぶ。

従来の検索は検索キーにマッチした文献だけを検索するが，リファレーション検索は文献間のリンクの度合い（関連性）が強いものを検索する。そのため，検索キーにマッチしていない文献も検索可能になり，検索結果は関連性の高い順に表示できる。さらに，専門分野の情報構造であるキーワード間，著者間，そして出典間の関係をマッピングすることも可能である。DB はリンク情報を持つと知的検索や知識構造の把握が可能になる。

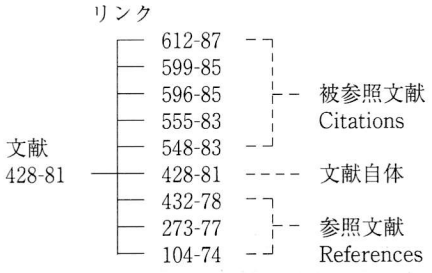

図２　文献428のリファレーション

3. DBの入力問題と参照データの入力

3.1 DBの入力

DBの作成は時間，コスト，労働力を必要とする。例えば，抄録型DBの作成は資料の収集，選択，抄録の作成，データの入力などのステップを経る。景気後退でDBの作成にブレーキがかかり，長い間継続してきたDB作成ルールも変更せざるを得ないかも知れない。

最近の情報処理は，POS（Point of sales）に見られるように情報を発生時点で入力し，即時に利用する。全文型DB作成の場合，資料を収集してテキストを再入力し直すことは経済的に成り立たなくなる。出版社から写植データを収集し再編集できれば全文型DBは短時間に作成できる。DBの作成はPOP（Point of publications）が目標となる。

全文型DBの作成は資料の収集からでなく写植データの収集から始まる。全文の中に書誌情報や抄録が含まれるため，全文型DBから副次的に書誌型・抄録型DBを作成できるようになる。DBの入力は大変な作業であり，できるだけ自動化してコストダウンを図らなければならない。

3.2 オープン化

初期のリファレーション研究で使用した参照データは特定分野に関連した文献のみで構成した[1]。この方法は参照文献の中から分野に関連する文献を選択する作業を必要とした。しかし，データ入力は簡単になり，リファレーションの作成やその操作も簡単になり，リファレーションの有効性を確かめることができた。

以前に作成したクローズドなDBは対象とする分野しか利用できない。そこで入力の対象を参照文献リストにあるすべての文献に拡大した。このオープンな方法は選択の負荷を免れる代わりにデータ量は莫大になる。

参照文献の書誌項目を入力したDBからはリファレーションの作成は難しい。なぜなら，ある文献がいろいろな文献の参照文献の中に点在している場合，文献を同定するためのコンピュータ処理プログラムの開発が難しくなるからである。文献の基本的書誌項目だけで200バイトくらいあり，不完全なデータやエラーが混入しやすくなる。この文献同定問題を解決するために導入したのが次に述べる文献コードである。

3.3 APTSコード

1992年の研究集会では文献同定のためにAPTSコードを提案した[4]。この文献コードは，支援システムがなくても，誰でも，どこででも，簡単なルールで文献をコード化できる。この方法を使うと参照データの作成は比較的短時間に，安いコストで作成できる。

APTSコードは，著者，発行年，標題，そして出典をそれぞれ4桁で表し，合計16桁のコードで文献を表す(図3)。そして，各構成要素の英文の頭文字からAPTSコードと名付けた。

□□□□+□□□□+□□□□+□□□□			
著者	発行年	標題	出典
Author	Publication Year	Title	Source

図3 16桁のAPTSコードの構成

雑誌論文は複数のDBに登録される。例えば，Information Processing & Managementはその目次に索引した20以上のDB名をリストしている。すなわち，雑誌論文は20以上のDBに重複して登録されていることを示す。やがて文献の同定が問題となり，文献のコード化は必要になる。DBが文献コードを介して互いにリンクし，さらに参照文献もリンクするようになれば，互いに孤立しているDBの利用価値は高くなる。

3.4 文献台帳

文献DBは抄録型DBを中心に発展しており，主要なDBは1960～1970年代から蓄積を開始している。参照文献の中にはそれより以前の年代のものがある。また，参照文献の記述が標準化されていないため，発行年の位置は著者の直後か，一番最後に記述される。さらに，出典のページ数が記載していないとか，副題を省略したり，あるいは間違って記述した参照文献が存在する。

このような場合に，文献の基本的書誌項目だけから構成される身軽な書誌型DBがあれば，文献の確認作業ははかどる。古い年代の文献を含む全

文献を網羅するような文献台帳の必要性を痛感した。図4は文献台帳の構成例を示した。CD-ROM 1枚で200万編の文献を採録可能である。

コード	FOXE-1993-DLIO-4441
著　者	Fox, Edward. A.　Lunin, Lois. F.
発行年	1993.
標　題	Digital Libraries: Introduction and Overview.
出　典	Journal of the American Society for Information Science, 44(8), 441-445.

図4　文献台帳の構成例

3.5　参照データの作成時間

APTS コードを使用した参照データの作成例を図5に示す。著者名，発行年，標題，そして出典などを入力する場合に比べて，容量は約8分の1になる。そして文献コードのマッチングだけで文献を同定できるためリファレーション処理は簡単になる。

雑　誌	## JASIS-44-6-2
文　献	LINA-1993-CAGC-4322
参照1	KOTJ-1974-CAOP-5242
参照2	MEDH-1971-SGCE-BHUP
参照3	STNM-1985-RMUS-6130

図5　APTS コードを使った参照データ例

参照データは雑誌の号数ごとに作成する。まず1冊の雑誌に掲載された全文献の参照文献リストを収集し，用紙にコード化する。1時間約100編のコード化が可能である。それからパソコンなどでデータを入力する。参照データの作成はコード化，データ入力，そしてチェックなどを含めて1編当たり約1分かかる。1雑誌当たりの文献数を20，1文献当たりの参照文献数を30編とすれば，1雑誌分の参照データは600編となり，その作成時間は約10時間である。

3.6　参照データの記憶容量とコスト

APTS コードは，1文献を改行を含めて21バイトで表す。1冊の雑誌が600編の参照文献を持つ場合，1冊当たりの参照データは約13K バイトになる。このテキストファイルを圧縮すると約40%になり，5Kバイトのバイナリーファイルになる。

1巻6冊とすると雑誌1年分の参照データは60時間の労働力を必要とし，テキストファイルで80K バイト，圧縮ファイルで30K バイト作成される。そのコストを8万円とすると，参照データ1K バイト当たり1,000円である。

3.7　参照データの作成者

参照データは誰が作成するのだろうか。おそらく2とおりの方法が考えられる。第1は，既存の抄録作成システムに参照データの作成を追加する場合である。1文献当たり500～1,000円のコスト増になる。昨今の経済状況から追加システムの実現は難しい。第2は，多数の人に雑誌の号ごとに参照データの作成を依頼する場合である。誰に依頼するのか，データの信頼性，データ収集システムなど，まだ課題は多い。いずれにしても重複して入力することだけは避けなければならない。

4.　DB の流通問題と参照データの流通

4.1　DB 普及のカギ

発展の要因として，性能や品質などのほかに量の占めるウェイトが高い。ハードの世界では半導体の DRAM は1社当たり月産数百万個生産する。ソフトの世界ではゲームのヒット作は100万本以上，パソコン用のウィンドウズは全世界で1,000万本以上販売する。ウィンドウズは2万円で販売されているが，その開発費は恐らく100億円以上である。逆に言えば，われわれは100億円のものを2万円で買っていることになる。DB もソフトと似た世界であり，量の重要性は高い。

DB の流通にはその内容だけでなく利用ソフトの果たす役割は大きい。現在，数百万台のパソコンが普及しているが，オンライン検索で得られたデータを有効に活用できるソフトは少ない。大規模な情報検索システムのソフトの改良は時間もコストもかかる。そのためユーザの情報環境とのギャップは広がる一方である。ユーザが利用できるソフトの開発が望まれる[5]。

4.2　参照データ流通の指針

リファレーション DB を処理する大規模なシステムの開発は当面考慮しない。大規模な情報セン

ターでDBを処理し，その結果をユーザに渡す代わりにユーザが必要とするデータを渡す。すなわち，書誌データや参照データを雑誌単位，巻単位，あるいは号単位に配布する流通システムを構築する。

そして，データ処理はエンドユーザ自身が行う。参照データからリファレーションの作成，リファレーション検索，知識構造の生成などのソフト開発は誰でも参加できるようにする。そこに競争原理が働き，良い利用技術が生まれる。DBの作成者，ソフト開発者，そしてユーザの三者が共に利益を享受できるような仕組みが望まれる。

4.3 パソコン通信の世界

現在，パソコン通信の通信速度は2,400ボーで，通信効率を85%とすると，1秒間で約200バイト送受信できる。最近，9,600ボーの通信も行われるようになってきた。パソコン通信は電子メール，電子掲示板，OSL，そしてDBのアクセスなど利用分野は広い。

SIG情報をオートパイロットを使って15分くらいモニターすると約100Kバイトの情報が得られる。また，100Kバイトのソフトは8分くらいである。パソコン通信の世界では数百Kバイトのデータのダウンロードは日常的な出来事である。

参照データの圧縮ファイルの送受信時間を試算すると雑誌1冊分は約25秒，雑誌1年分は2分30秒である。9,600ボーではその4分の1である。

4.4 特定分野の参照データの収集

特定分野を定義することは難しい。理論的な分野，陳腐化の激しい科学技術の分野，あるいは人文的な分野などのほかに，目的により収集する文献の広がりや量は異なる。

多くの場合，情報は特定の雑誌に集中している。参照データとして，8雑誌5年分とか4雑誌10年分を収集して分析すれば，従来と異なった情報を得ることができる。この場合，合計40巻分になり，文献数は約5千，参照データは15万である。参照データの容量は3.2Mバイト，圧縮データは1.2Mバイトで，フロッピー1枚分である。

4.5 参照データの流通

電子媒体の流通は磁気テープからオンラインへ移行して以来，長い時間が経過している。現在，フロッピー時代を経ないで，CD-ROM時代に移行しつつある。CD-ROMの読み取り装置は5万円前後であり，書き込み装置は100万円以下になった。テキスト系のCD-ROMには読み切れないほどの情報が詰まっており，すでに情報過多の時代が始まっている。

今後の情報流通形態は，蓄積した多量のデータはCD-ROM，少量のデータはフロッピー，最新情報はオンラインになる。参照データの場合，ユーザが希望する雑誌とそのバックボリュームに関するデータをCD-ROM，フロッピー，そしてパソコン通信で流通させることになる。

5. おわりに

ダウンサイジングが進行しているが，データを集中的に集めてオンライン検索サービスする従来のスタイルは変化するかも知れない。DBの作成者は素材のデータを作成し，大量に配布し，作成コストを短期間に回収する。そして，データの料理はユーザの好みに任せる。そうなればソフト開発は小規模ですみ，情報環境の変化やユーザの利用形態の変化に十分対応できるようになる。

抄録型DBや全文型DBには文献間の関連性情報が含まれていない。ここで述べたAPTSコードによる参照データの作成は，関連情報であるリファレーションの作成を容易にし，関連性検索の開発に導く一歩となる。

参照文献

1) 浅井勇夫．パソコンによる引用文献データベースの開発．第21回情報科学技術研究集会発表論文集．日本科学技術情報センター．東京，1985，p.21-31.

2) Asai,I. Referation Database for Document Information Analysis. Proc. 50th Meeting of the American Society for Information Science. 24, 1987, p.1-5.

3) 浅井勇夫．大規模なリファレーションDBの開発：文献同定のためのAPTSコードの導入．第29回情報科学技術研究集会発表論文集．

日本科学技術情報センター．東京，1993，p. 273-278.
4) 浅井勇夫．リファレーションデータベースによる知的検索の可能性．情報の科学と技術．43(10)，945-950(1993)
5) 浅井勇夫．検索ログファイルの有効利用．第28回情報科学技術研究集会発表論文集．日本科学技術情報センター．東京，1992，p. 1-5.

質　疑　応　答

質問　木戸口隆夫（石川県職業訓練短期大学）
DBMSを作成しているのか。

回答　作っていない。汎用ソフトを使用している。しかし，リンクやマッピング機能は現状のソフトでは網羅していないため，新たに作成する必要がある。

質問　鈴木陽市（浜松インターナショナル㈱）
情報のリンク化によって，検索時の抽出数が膨大にならないか。

回答　設定したリンク数を基に関連性の高い順に検索・抽出を行うため，大量の抽出を防げると思う。

第30回
情報科学技術研究集会
発表論文集
INFORUM '93

JICST 日本科学技術情報センター
The Japan Information Center of Science and Technology