北海道大学工学部研究報告
第130号（昭和61年）

Bulletin of the Faculty of Engineering,
Hokkaido University. No. 130 (1986)

学術研究における文書情報システムの構成

前田　隆*　天野　要**

（昭和60年11月20日受理）

Document Information System in Scientific Research

Takashi MAEDA and Kaname AMANO

(Received November 20, 1985)

Abstract

Scientific documents consist of bibliographic and textual information, and also include complicated information of certain figures, tables, graphs, images, etc. Future scientific information system will be connected to an integrated database of huge dimensions and be constructed as a scientific workstation with advanced functions of data processing of various forms of multi-media knowledge information. In the design of such an information system, some specific abilities are required for the processing of non-coded information of variable length, set-valued information and structural information, in addition to the usual coded elementary information.

In this paper, we discuss design considerations and implementation problems for constructing a document information system, as a prototype of such an information system, with the function of text processing as a typical example of complicated information processing using a commercial DBMS. Future directions of an advanced scientific information system are also suggested.

1．はじめに

学術的文献は，これを情報科学的な観点からみるならば，概念的な「文章」をはじめとして種々の型の情報媒体による表現が一定の順序・構造をもって配置されたものとみなすことができる。すなわち，これらはテキストを基本として，種々の図形，表，グラフ，画像等の多元的知識情報が，ある配列と順序をもった構造を形成することにより，全体としての知識情報表現を構成している。学術的研究活動を支援する，あるいはこれを効率的かつ総合的に推進するための高機能な学術情報システムは，上記のような学術文献情報をはじめとして，様々の事実情報，統計的情報等の統合された巨大データベースである（あるいはこれと結合されている）と同時に，これらの情報表現を構成している多元的知識情報の生成，分析，蓄積，検索，変換，処理，合成等の機能および操作の単純かつ容易な入出力機能を有する，学術研究用ワーク・ステーションとして描く（構成される）ことができる。

＊　工業数学講座
＊＊　大型計算機センター

このような多元的知識情報を取扱う情報システムの設計における最も基本的な問題は多元的情報のモデル化とこれに基づく統合的な処理方式の開発である[14]。すなわち，このようなシステムにおいては，通常の定型化された基本的情報に加えて少なくとも，(1)可変長の非定型的情報，(2)集合値情報，(3)関係的構造的情報等の容易かつ柔軟な表現・格納・検索・処理等の能力が要求される[2，6，9，14]。従って，既存の一般的な汎用型情報検索システム(IRS)やデータベース管理システム（DBMS）技術の単純な適用では解決できない種々の問題が提起されると考えられる［3］。

本稿では，上記の三つの特徴を有する複合的な情報の典型的な例としてテキスト情報をとりあげ，これに対する種々の処理機能をもつ文書情報処理システムについて，その設計上の考察および実現への問題が論じられる。第2章では，特にテキスト情報を基本とする文献情報とその既存のDBMSによるシステム設計が取扱われる。第3章では，北海道大学大型計算機センターにおけるデータベース作成支援システム［1］の下で，商用DBMS (ADABAS) を用いた文献情報システムの実現方式が記述される。第4章では，多元的知識情報を取扱う統合的学術情報システムの設計に対する方法論，ユーザ・インターフェース等に関する今後の課題，方向性等についてまとめられる［12］。

2．学術研究における文書情報

学術的な研究環境においては，メモやレポート，論文，本などの種々の形式の文書情報が存在する。このような情報環境の重要な源泉の一つとして文書情報データベースを想定するとき，次の二つの側面で利用価値が考えられる。すなわち，一つは当然ながら情報手段としてであり，もう一つは研究対象としてである。我々は，次章で具体的にのべるように，情報学における文献情報システムを構想する際，上記の二つの面で有効であるようなシステム設計を考察することにする。そして，これをさらに拡張･発展させることにより，「はじめに」で述べたような，多元的知識情報を含む統合的学情報システムの構想とその設計へと進展させることを意図している。

以下では，学術的文献のデータベース化とその高度利用を意図する情報システムの構想から，統合的学術情報システムの必要性およびそのようなシステムの設計への指針を考察することにする。

学術的研究における文献をはじめとする様々な文書 — メモ，計算，引用，手紙，ノート，描図，原稿等 — の容易な入力，処理，編集，合成，出力が可能であるためには，定型的なデータと共に，非定型的なデータの統合的な管理と処理が要請され[17]，また公共的および個人的の両レベルでのデータベースの利用が可能であることが望まれる。このようなシステムを利用することにより，学術的研究において，次のような効果が期待できる。

＊ 集積された知識情報の効果的利用，
＊ 知識情報の系統的な獲得と配布，
＊ 整理された知識データの観察による創造活動への刺激の獲得，
＊ ある種の知的活動や操作への支援，他。

以下では，このようなシステムの構成に関連するいくつかの要請および問題について，上記の二つの面からそれぞれ考察することにする。

2．1 情報手段としての文書情報 学術的文献の情報は一般にその分野の研究動向を適切に把握するために，また関連する個々の事実や方法，技術等について具体的な知識をうるために，

学術的研究活動において定期的にあるいは必要に応じて利用されている。個々の内容的な事柄は当然であるが，その他に例えばある論文の著者や研究機関およびその住所等の情報もその都度必要となることが多い。これまで，書誌情報を中心とする文献データ・サービスは各分野において実現され，一般化してきており，それぞれ一定の役割を果している。しかし，このようなシステムは，しばしば指摘されているように，必ずしも充分な満足度をもって受入れられている訳ではない。特にテキスト部分など非定型的情報の扱いが困難であり，一言で言えば，これらの情報の有効利用のためにはシステムの一層の高度化，高機能化が求められているのである。

一般的に，利用者サイドからみたデータベースへの要請項目として，

(1) 網羅性（守備範囲をカバーし，もれがない等），
(2) 完全性（データを構成する属性等に対する完備性等），
(3) 柔軟性（任意のデータへの容易かつ迅速なアクセス），
(4) 操作性（データの処理・編集・変換等が容易），
(5) 高度性（知識データの処理・編集等の機能），

等が考えられるが，現状での不満はとくに(3)，(4)に関連している。これはそれ自身の不十分さと同時に(2)に関連するデータ表現，あるいは(5)を考慮すればデータの知識としての情報表現の問題に行きつく性質のものである。すなわち，文献データベースの高度化・高機能化は，システムの操作性を高め，テキスト・データをはじめとする各種データの知識としての情報処理機能をもつ一つの統合された情報システムへの発展をめざすものである。そして，その基本は種々の機能を具体化する柔軟なユーザ・インターフェースであり，これを有効に実現するための知識情報としての文献データの情報表現の問題であると考えられる。

2．2 研究対象としての文書情報 学術的文献情報の取扱いを研究対象として眺めるとき，様々の観点からのアプローチが可能であることは明らかである。しかし，ここでは文献データベースあるいは情報システムの構成，特にその学術的研究活動における高度利用という観点に限定して考える。前節で議論したことと関連させると，大きく分けて次の五点が特に重要な項目として挙げられる：

(1) 文献内容の適切な情報表現 —— 文献（テキスト）情報のデータモデル，
(2) 内容検索のための手法 —— 内容向き索引の利用等，
(3) データベース作成の方法 —— 高度の DBMS 技術の利用・拡張など，
(4) ユーザ・インターフェースの改良，
(5) 汎用データベースと結合した個人向き文書管理システムの構成法。

ここで，(1)は文献における書誌事項等の定型的な情報以外に内容に関する非定型的な概念的意味情報をどのように情報表現するかという問題である。単純なキーワード表現以上の，テキスト情報をベースとする方法が基本である。(2)は(1)に基づいてこれらに効率的にアクセスしたり処理したりするための手段であり，従来の転置索引法などが有力ではあるが，情報表現の形式や蓄積文献量とのかねあいで，より適切かつ有効な手法が必要とされる。(3)は(1)，(2)に基づいて，最適かつ有効なシステムを作成することが基本であるが，それ自身が特別の目的でない場合には既存の DBMS あるいは IRS を必要に応じてその一部を変更ないし拡張することにより実現することもしばしば有効である[2，7，15]。このような制約のもとで，文献をはじめとする文書情報の外部表現，内部表現を含むシステム設計が基本的な問題となる。上記のすべてが関連する(4)は学術的研究における様々の形式の文書情報の取り扱いを含むシステムの在り方に関する考察に基づ

くモデルの研究によって，必要とされる諸機能の自然な操作性のよい実現方式が問題である。近年の高解像度のビットマップ・デイスプレイ装置を用いてのマルチ・ウインドやマウス等を利用するインターフェースの拡張も大いに参考になるであろう[8]。しかし，これらの技術は，従来のものと比較すれば格段に有効ではあるが，それ自身その操作性等の点においてまだまだ改良の余地があり，これを一新させるような方式が必要であると考えられる。以上の点を踏まえて統合化されたシステムとして(5)の性格が特徴づけられる。

2．3 設計上の考慮 これまで述べてきたように，学術的研究環境における文書情報システムの二つの側面と機能は，そのようなシステム設計上での重要な指導原理としての役割をも果す。すなわち，情報手段としての側面は，システムの公共的なデータベースへのアクセス機能に対応しており，また研究対象としての側面は個別的なデータ処理機能に対応していると見なすことができる。従来，これらの機能はそれぞれ文献データベースシステムおよび文書処理システムとして別々に開発研究されてきた。高度の学術情報システムは，少なくともこれらの二つのタイプの文献・文書情報システムが一つに統合されたシステムとして，特に多元的データに対する操作(処理）におけるより高度の機能を備えたシステムとして構成されるべきである。このようなシステムは，単純に言えば，汎用データベースと結合した専用型（個人向き）文書管理システムとして設計・実現することができよう。

3．文書情報システムの一実現

上記の論点を踏まえて，以下では，小規模ではあるが複合的・非定型的情報の例としてテキスト情報を取扱う文献情報システムの実現の方法論について議論する。この実現にあたっては，既存の商用DBMS（ADABAS）を利用することにした[19]。それはシステム作成の負担をできるだけ軽減するため本学大型計算機センターでサービスしているデータベース作成支援システムを使用すること，およびここでの主眼である「テキスト処理機能」および「よりよいユーザ・インターフェース」の作成のためにサブシステムであるNATURAL言語が有用であると考えたからである。しかし，既存のDBMSに共通する問題点として，一般にレコード長やフィールド長に関する制限等，テキスト情報のような非定型的データには基本的に不都合な性格が存在している。また，用意されているデータ操作言語，検索言語等も通常は基本的には定型的なデータ構造を前提としており，同様にそのままでは我々の目的にはそぐわない。

我々は次の三点をシステム設計の基本方針として採用した。

(1) 文献情報の適切な分割……基本的には，各文献を単位とするレコード構成において生ずる冗長性の排除が目的であるが，(2)との関連で，分割された個別ファイルの有意義な利用をも意図している。

(2) データの意味の保存……形式的な正規形への変換などによる機械的な冗長性の導入を排除すると共にこれにともなう各レコードの意味の破壊やデータの分散を避けること，(3)との関連で利用者にとって自然な取扱いの単位を追求する。

(3) 柔軟なアクセス機能……視覚（スクリーン）情報の活用が重要であるが，コマンド型の言語による場合でもより柔軟かつ有効な問合わせ言語の具備および集合とテキスト処理機能の充実を基本におく。

3．1 文献データとファイル

学術的文献の情報構造を一意に定めることは不可能であるが，ここでは学術的雑誌等に発表された論文を直接的な対象とするので，通常は次のような項目（属性）により，これを特徴づけることができよう：

文献－0：＝［文献コード，標題T，著者M（著者名，住所），書誌（雑誌名，巻数，号数，発行年，頁），分類M，鍵語M，抄録T（抄録文M），参考文献M，言語，所蔵M］，

ここで，T，M，（ ）はそれぞれテキスト情報，多値情報および複合的情報という属性を持つことを表わしている。

このような文献情報の枠組みに対して，実際の文献データを写像することは基本的には難しい問題ではないと考えられる。しかし，良く知られているように，他の文献との間で各々の属性値が同一である場合も多く含まれている。典型的なものは書誌情報，分類，鍵語等であり，また著者，参考文献等においてもしばしば同一の値が繰返して生起する。これらの値は相対的には少ない要素からなるものである。従って，ファイル編成における冗長性をできるだけ少なくするためには，このような性質を踏まえたファイル設計が望ましい。

上で述べたシステム設計の基本方針に基づき，原型システムとしての文献データベースにおけるファイルは次の7個に分割することとした。

1．文献：＝［文献コード，標題T，著者M（第n著者，姓，名），書誌（雑誌コード，巻数，号数，発行年，頁），分類M，鍵語M，参考文献コードM，言語］，
2．著者：＝［著者コード，姓，名，住所T］，
3．雑誌：＝［雑誌コード，雑誌名，出版社T，所蔵コードM］，
4．抄録：＝［雑誌コード，文章M（文番号，文T）］，
5．用語：＝［鍵語コード，鍵語，関連語M（関連型，鍵語）］，
6．所蔵：＝［所蔵コード，住所T］，
7．参考文献：＝［参考文献コード，文献コード，参考文献名T］。

3．2 文献情報データベース

文献情報の資源は，現在，情報学の分野における以下の三つの雑誌であり，近い将来，人工知能学における主要な雑誌等の文献を併合する予定である［19］：

(a) Information Processing and Management （以前は，Information Storage and Retrieval）：Pergamon Press,
(b) Journal of The American Society for Information Science : ASIS
(c) The Journal of Documentation : Aslib.

最初に，予備的な第0版として本学大型計算機センターで提供しているDBMS ADASASを用いて，3．1節の「文献－0」の形式（但し，参考文献なし）で約720件の小さな文献集合を対象とする原型データベースを構築している。これは引続いて必要な加工処理を行ない，上記の六つのファイルに分割・変換されることになる。

ここでの主要な関心は，当然ながら「文献」および「抄録」の二つのファイルに対してである。

これらのファイルの「標題」「文章」の属性をもつテキストデータは非定型・不定長データであり，ADABASのレコード長およびフィールド長の制限の下で，ある固定的な長さに分割し，多値属性のフィールドを利用して格納された。

3．3 テキスト処理

多元的知識情報を取扱う高度の文書情報システムにおいてはテキスト処理の豊富な能力を具備することが望ましいことは言うまでもない。一般に，テキスト情報に対するデータ操作には三つの異なるレベルが存在する：語あるいは文字列のレベル，文字列の列すなわち文のレベル，および文の列すなわち文章のレベル。

従来，この中で第一のレベルは文字通りワード・プロセッサにおいて取扱われており，第二のレベルはある種の制限の下でテキスト・エディタ［13］あるいは小数ではあるがデータベース・システム［16］に組込んでいるものがあり，また第三のレベルにおけるデータ操作は現在幾つかの領域において興味が持たれはじめてきている段階である［2］。

現在，多くは第一のレベルでの文書の編集（Editing）と書式化（Formatting）の機能をもつテキスト処理システムとして実現されている［5，13］。これらのシステムにおけるテキスト処理は基本的操作として，対象テキストにおける特定の語あるいは文字列等が存在する位置の「探索」，それらの「削除」，テキストの特定の位置への文字列の「挿入」，あるテキスト（の一部）の別の位置などへの「複写」と「移動」，ある文字列の他の文字列による「置換」等を含んでいる。

これらの基本的な操作を上述の第二および第三のレベルに拡張することは，とくに文脈的なテキスト処理およびその種々の文書処理への応用にとって極めて重要である。従って，以下のような方向への拡張が強く望まれる。

(1) 構文指向……多くのテキスト・エディタで現在使用可能な操作の，構文の特徴を利用する拡張，

(2) 文法論型……対象テキストの言語における文法的情報を用いる方向での拡張，例えば，品詞，文型，テキスト構造等の情報の利用，

(3) 意味論型……意味的情報の利用，例えば，語の意味，語や語句間の種々の意味的関係等の利用，

(4) 語用論型……語用論的情報の利用，例えば，慣用句や慣用的表現の利用等。

これらの中で，特に(2)，(3)，(4)の各方向に対しては，言語学的な知識と共に，対象領域における知識としての辞書あるいは「知識ベース」のようなものが必要とされるであろう。これらの拡張の部分的な実現であっても，広い応用に対して高度のテキスト処理の能力をもたらすことになる。我々の予備的なシステムにおいてはこの中で，特に(2)と(3)の方向において拡張の試みがなされており，これはいわゆる機能的テキスト構造分析法に関連するデータ操作法として重要なものである。［10，11］。

3．4 ユーザ・インターフェース

よりよいユーザ・インターフェースはその情報システムの最も重要な構成要素の一つであり，有用かつ柔軟なユーザ・インターフェースを持たないシステムはその価値を半減させるといっても過言ではない。高度の入力装置をもち，柔軟かつ操作性のよいユーザ・インターフェースに関する種々の技術は近年とくにいわゆるオフィス・オートメーションの分野で発展してきている［8］。これらのうちの幾つかは，伝統的な情報検索システムの分野においても，そのユーザ・インターフェースの一部として取入れられてきている［4］。これらのシステムは，ビットマップ・ディスプレイとマウス等の位置指示子を使い，画面上のマルチ・ウインドにおいてコマンドの代りにそれを像化したアイコンをさし示す操作等により，種々の文書の検索・編集・加工処理を行うという一種の「電子机」（Electronic Desktop）として機能させようとするものである［8］。

よりよいユーザ・インターフェースに関する基本的観点は，結局，その情報システムの有用性と柔軟性という二つの概念の実現ということである，と考えられる。

ここで，有用性は次のような機能を可能とすることにより実現されるであろう：

a．要素的な部分だけでなく，複合的なデータへのアクセスおよび操作，

b．種々の型のデータ集合の保存と引出し，

c．必要に応じて利用過程の適当な段階への後戻り，

d．並列かつ多重の利用，等。

また，柔軟性は次のようなことを要求する：

e．利用過程やデータの視覚化，

f．各操作および使用装置の容易かつ単純な操作，

g．ユーザとシステムとのコミュニケーションの相互主導的方式，等。

構築中のシステムでは，これらのうち，有用性ではa．の点，柔軟性ではe．の点について，ADABASのユーティリティおよび通常のキーボードと表示装置を用いて，実現が試みられている。

4．おわりに

高度の学術情報システムの設計に関して考察してきた。学術的研究環境におけるこのようなシステムの構成に対する二つの重要な指導原理として，システムの情報手段としての側面および研究対象としての側面が取上げられた。これらはそれぞれこのような文書情報システムの汎用的なデータベースへのアクセス機能および文書処理機能として把握することが可能であり，これらの機能をもつ原型システムの実現の試みが提案された。この原型システムは現在，本学大型計算機センターで既存のDBMSを用いて開発･構築中である。このシステムではとくにテキスト処理能力とユーザ・インターフェースの改良に重点がおかれている。

本論文で議論したような高度の学術情報システムおよび学術研究用ワークステーションに関する研究・開発は今後益々増加していくであろう。この領域で提起される多くの問題は広範かつ多くの研究分野，特に計算言語学，人工知能，計算機およびソフトウェア科学等に関連している。これらの諸分野に関連する最も重要かつ基本的な問題は，いわゆる多元的知識情報の基本部分をなすテキスト情報の処理手法の拡張・開発，その基礎でもありかつそれ自身としてのテキスト情報のモデリングおよび概念的・意味的情報表現形式の工夫，およびこれらの多種多様の情報を取扱う高度かつ拡張された意味でのデータベース管理システムの構成等であると考えられる。本論文ではこうした基本的課題についての導入部分のみが取扱われたにすぎず，残された課題は多い。

参考文献

[1] K. Amano and T. Maeda : Database Management in Research Environment, *Proc. of 3rd EFISS Symposium,* Roskilde, Denmark, 1985.

[2] F. Bancilhon and P. Richard : Managing Texts and Facts in a Mixed Database Environment, in "New Application of Databases" (G. Gardarin et al. eds.), Academic Press, 1984.

[3] R. G. Crawford : The Relational Model in Information Retrieval, *J. ASIS,* 32 (1), 1981.

[4] H. P. Frei and J.-F. Jauslin : Graphical Representation of Information and Services : A User-Oriented Interface, *Inf. Tech. Research & Development,* 2 (1), 1983.

[5] R. Furuta, et al.: Document Formatting Systems : Survey, Concepts, and Issues, *Computing Surveys*, 14 (3), 1982.
[6] G. Gardarin : Towards the Fifth Generation of Data Management System, in "New Application of Databases" (G. Gardarin et al. eds.), Academic Press, 1984.
[7] R. L. Haskin and R. A. Lorie : On Extending the Functions of a Relational Database System, *Proc. ACM SIGMOD Conf.* 1982.
[8] A. Lee, et al. : User Interface Design, in "Office Automation : Concepts and Tools", (ed. D. Tsichritzis), Springer-Verlag, 1985.
[9] V. Lum, et al. : Design of an Integrated DBMS to Support Advanced Applications, *Proc. Intl. Conf. Found. Data Organization,* Kyoto, 1985.
[10] T. Maeda, et al. : An Automatic Method for Extracting Significant Phrases in Scientific or Technical Documents, *Inform. Process. & Management,* 16 (3), 1980.
[11] T. Maeda : An Approach toward Functional Text Structure Analysis of Scientific and Technical Documents, ibid., 17 (6), 1981.
[12] T. Maeda and K. Amano : Document Information System in Scientific Research, *Proc. of 3rd EFISS Symposium,* Roskilde, Denmark, 1985.
[13] N. Meyrowitz and A. van Dam : Interactive Editing Systems : Part I, II, *Computing Surveys,* 14 (3), 1982.
[14] F. Rabitti : A Model for Multimedea Documents, in "Office Automation : Concepts and Tools", (ed. D. Tsichritzis), Springer-Verlag, 1985.
[15] H.-J., Schek, et al. : Data Structure for an Integrated DataBase Management and Information Retrieval System. *Proc. VLDB-8,* 1982.
[16] M. Stonebraker, et al. : Document Processing in a Relational Database System, *ACM Trans. on Office Information System,* 1 (2), 1983.
[17] M. Togashi : Databox, RIMS Kokyuroku, 461, June 1982.
[18] C. Woo, et al. : Document Management Systems, in "Office Automation : Concepts and Tools", (ed. D. Tsichritzis), Springer-Verlag, 1985.
[19] 前田 隆：SOARによる情報学文献情報データベースAIRISの構築，北海道大学大型計算機センター・ニュース，Vol. 7, No.4, 1985.