# JICSTの大規模情報検索

小野修一
（日本科学技術情報センター）

## 1. はじめに

日本科学技術情報センター（JICST）は、科学技術に関するわが国の中枢的情報機関として、科学技術情報資料の収集、処理、サービスの業務を行っている。内外から収集される一次資料は、表-1に示すごとく、雑誌だけでも8000種近くにのぼり、これらに含まれる大量の文献情報の処理（抄録や索引などの二次資料化，蓄積と検索など）は、複雑な工程をたどっている。一方、JICST本来の目的である情報サービスは、科学技術文献速報（抄録誌，索引誌）などの出版物サービス，電子計算機を利用した機械検索サービス（SDI，RS*），JICSTで蓄積された二次情報ファイル（磁気テープ）の提供，複写，翻訳，調査，閲覧サービス，その他と多様な形態をとっている。このほかマシンリーダブルな磁気テープファイルとして入手される外国製の二次資料も、JICSTの処理工程を経て利用者にサービスされている。図-1はJICSTの情報サービスの概要を示したものである。

| | |
|---|---|
| 外国雑誌 | 4950種 |
| 国内雑誌 | 3000種 |
| 技術レポート | 35000件 |
| 会議資料 | 250件 |
| 特許明細書 | 48000件 |
| 特許公報 | 40種 |

表-1 JICSTの収集資料（昭和年度）

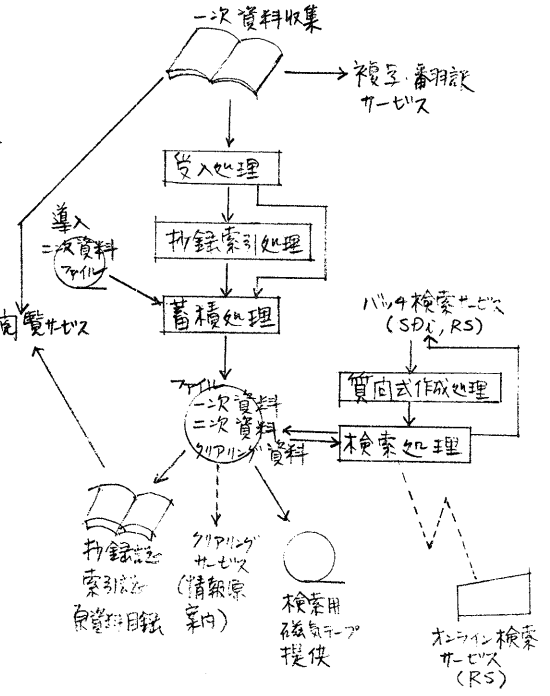


図-1 多様なJICSTのサービス

## 2. JICSTにおける計算機処理の歴史

昭和42年のFACOM230-50の導入と前後して、科学技術文献情報の収集、処理、蓄積、検索、提供の一連の流れを電算機により処理し、業務の合理化をはかることに着手した。その第1段階は、昭和43年実用化にいたった文献速報自動作成システムであり、漢字情報を含む文献情報を磁気記録化し、科学技術文献速報の編集を自動化した。このようにして蓄積された文献情報は後に多様な形態で利用者に提供されることとなった。同じ43年に完成をみた（漢字モード）及び英字カナ文字モードのIRシステムにより、S47年からSDIサービスを実施し、又蓄積磁気テープも同年度から利用者の所有する電子計算機による検索用として提供されるにいたった。

---

＊ SDI（Selective Dissemination of Information）：あらかじめ選定されたテーマに関して定期的に最新のファイルから検索を行い、回答をサービスする。 RS（Retrospective Search）：利用者の要求に応じて過去数年～10年分のファイルからそのつど検索すること。

一方文献検索に用いられる語いをコントロールする目的で、用語管理システム（DOCTOR）が建設され、S47年度には部内検討用のシソーラスを作成した。このDOCTORは、前記2つのJICSTIRシステムと蓄積段階でのインターフェイスを備えている。これらと並行してオンライン方式での会話型文献検索システムの開発がすすめられ、昭和46年にプロトタイプを完成した。（DOOR）

JICSTで蓄積されたファイルの外、外国製の蓄積ファイルを用いての検索サービスも行われている。米国国立医学図書館（NLM）で作成されるMEDLARSファイルによるSDIサービス（RESサービス）がS47年度から，化学情報に関しては数十年の歴史をもつCAS（Chemical Abstract Service）から提供されるCA condensatesファイルもSDI方式により、S48年度から実施されている。

以上は文献（二次資料）の蓄積と検索に関するシステム開発の経緯であるが、これらIRシステムの周辺の合理化にも検討が加えられ、収集される一次資料の管理を目的とする資料管理システムを昭和46年完成させた。

3. JICSTIRシステムの特徴

JICSTIRシステムを支えている原則が3つあり、これは将来とも変ることはないと思われる。

① Man-Machine Interaction

IRシステムは、完全自動化システムではなく、人間と機械との間で各々が得意とする作業の相互分担によって成立するものである。抄録作成や索引作成作業、シソーラス作成作業におけるキーワードの選定、キーワード間の関係づけなどの自動化は技術的に難点が多く、実用レベルのサービスを行うJICSTでは、当分の間採用されることはないだろう。

② 科学技術の全分野を統一的に扱う。

総合情報センターとしてのJICSTは、科学技術の全分野にわたる情報について共通の扱いをする。現在JICSTで蓄積される文献に用いられる語いのコントロールを全分野について統一的に行う努力がなされている。又導入ファイルとJICSTファイルを含む、ファイル間の統一（対象分野の重複からくる記事の重複削除，キーワードの統一的扱いなど）も今後の課題として残されている。

③ 情報の提供形態は多様である。

利用者の情報要求は多様であり、したがって提供形態も多様である。サービスの多様さは前述した通りである。情報処理工程の機械化は、蓄積工程，検索工程のスピードアップなど、それなりの効果をあげ、今後の情報量の増加を考えればその傾向はますます強まるだろう。そこで機械処理を中核とした多様性を考えると、現システムは、段階的業務別に作成されたプロセスオリエンテイドなものであり、この多様さに耐えられる順応性に乏しい。近年よくいわれているデータベースの考えが必要だ。つまりデータベースを核として種々のアプリケーションは考えられるべきであろう。

このほかJICSTIRシステムの特徴としてあげられるものを下記に示す。

• 漢字情報の処理が大きなウェイトをもっている。

JICSTで収録される文献には、日本語の抄録とキーワードが付加される。このため、文献情報の蓄積（入力からファイルの作成まで）及び抄録誌，索引誌の編集枠組，版下作成は一貫して漢字モードで処理されている。従って入力には漢字テレタイプライタを出力には漢字プリンタを用いる。

・大量データの処理である。

JICSTが現在対象としている3つのファイル、JICST理工学，MEDLARS，CACの各々の年間収録数は，40万，23万，36万件に達しており、合計で100万件にせまっている。今後もますます増加する。1件は1文献（1記事）に相当する。各記事は文献の書誌事項，主題項目等、その文献を表わす二次資料からなる。

・文章情報である。

単なるデータでなく意味内容を伴う言語情報である。

・ほとんどが可変長データである。

各文献の掲載された資料名、文献の標題，著者名，著者の所属機関名，キーワード，抄録などすべて可変長の文字データである。

## 4. JICSTIRシステムの現状

前述のようにJICSTIRシステムは、文献速報自動作成システムに上のせする形で、段階的、業務別に作成された。現行の各システムの関連図を図-2に示す。

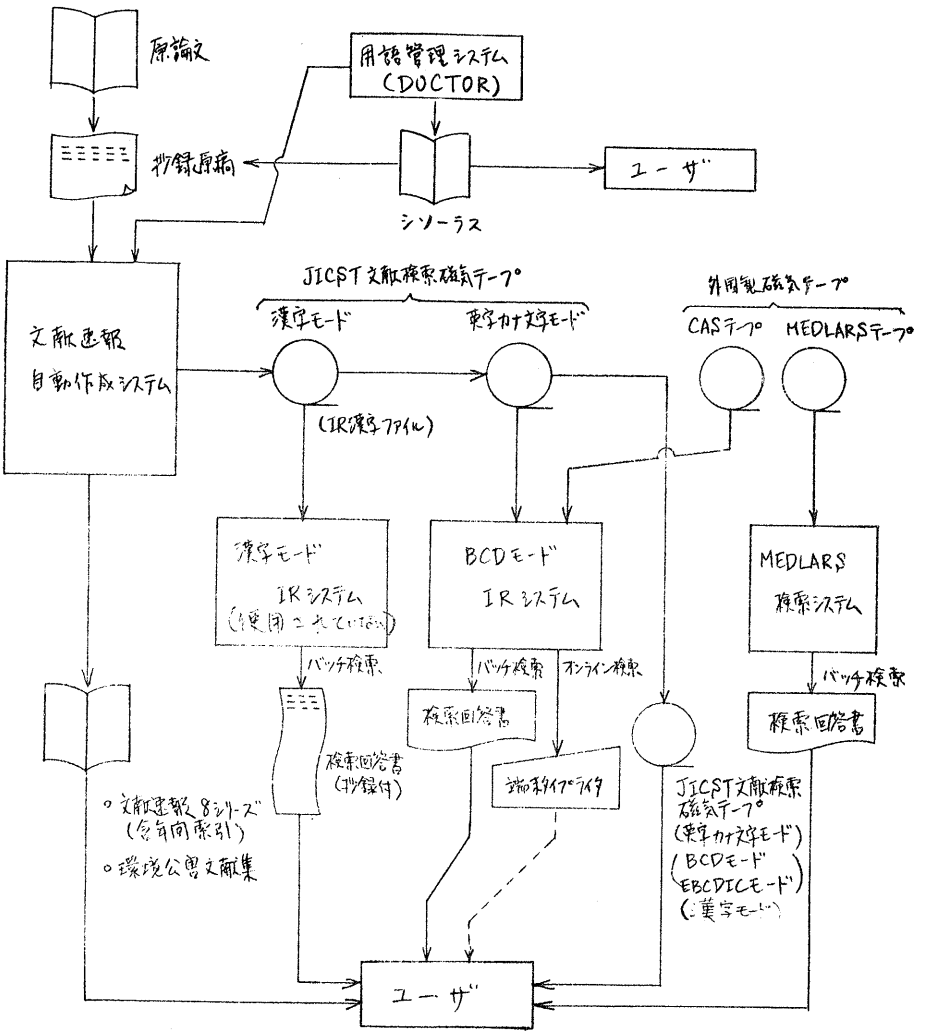


図-2
JICST各種
システム関連図

現行システムでは、蓄積された文献や用語はすべて磁気テープに保存され、磁気テープ上でメンテナンスされ、分割編集される。（オンラインプロトタイプのDOORシステムのみが唯一のディスクシステムである）ファイル形式は固定長で、可変長データを取り扱うために、セグメント方式なるものを採用している。（図-3 参照）

## 4.1 文献速報自動作成システム

収録された文献（二次資料）を磁気テープに記録保存することと、これらの蓄積磁気テープから文献速報とその索引をプリントアウトし、オフセット印刷用版下を作成することが目的である。システムフローを図-4に示す。

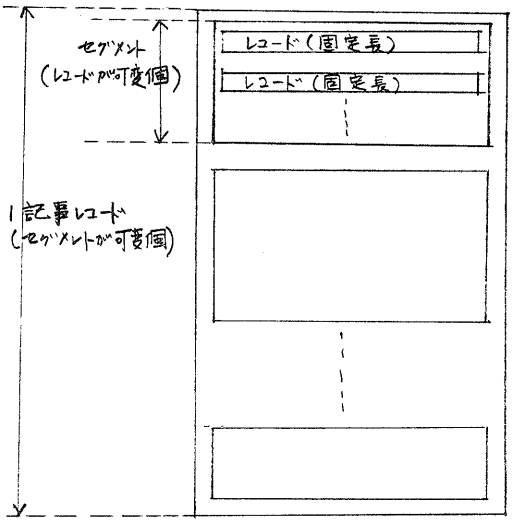


図-3 セグメント方式

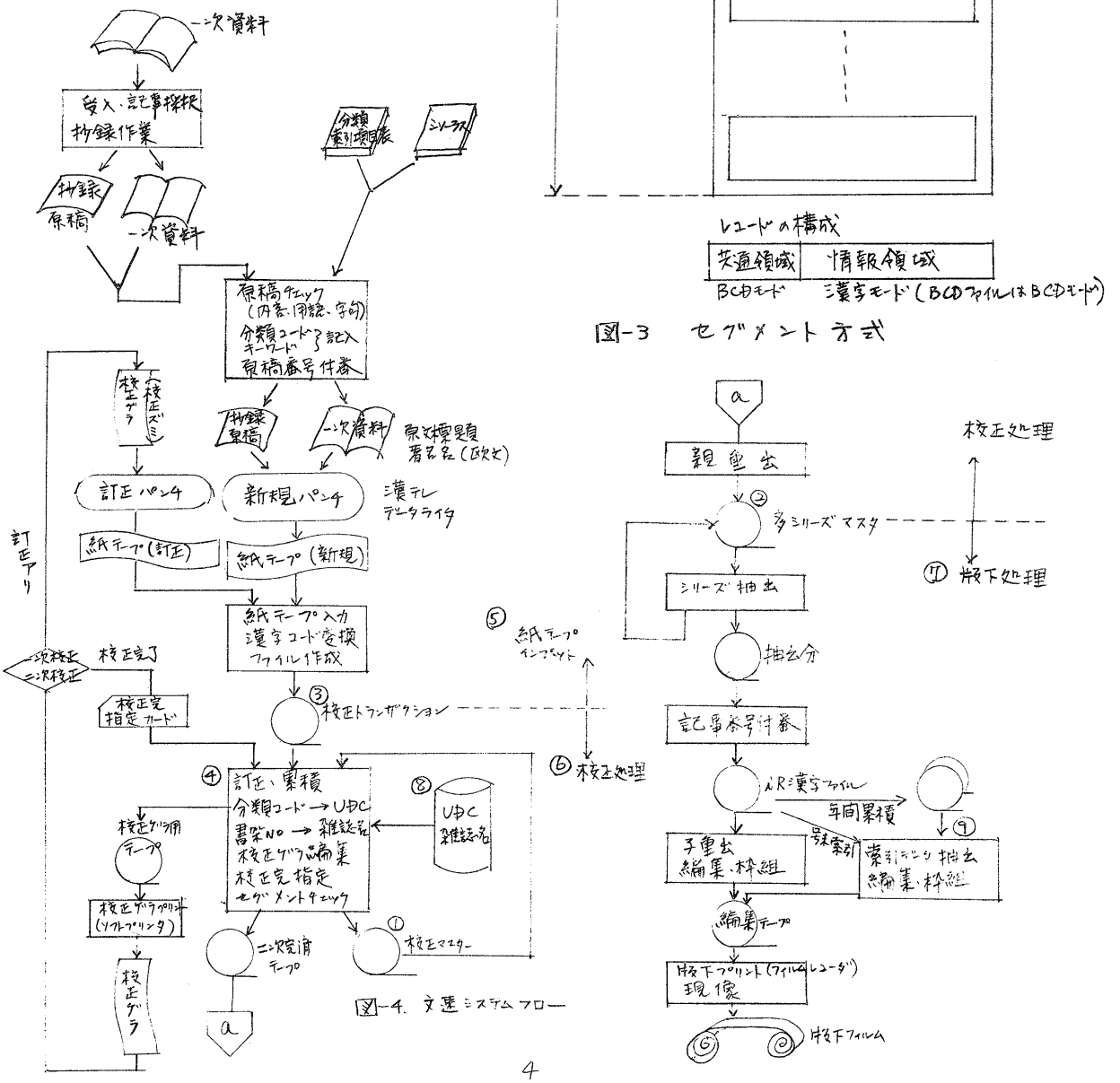


図-4. 文速システムフロー

① 校正マスターは、校正未完の文献データのプールで、常時約40,000件がプールされている。これは800 bpi, 2400フィートの磁気テープ7本分に相当する。

② 多シリーズマスタは、校正完了の文献データで、関連するシリーズに複製されたものを含むプールで、25000～30,000件がプールされ、磁気テープ4～5本分に相当する。

③ 校正トランザクションは、新規、訂正いづれのデータも含み、約13,000件である。

④ 抄録校正の処理時間は約2時間で、このシステムの中では、最大の処理時間を要している。

⑤ 紙テープインプットは新規、訂正とりまぜて毎日1ロット（1300～1500件）の処理が行われる。計算機処理時間は、約1.2時間である。

⑥ 校正処理は1日おきに行われ、④の処理を含めて約5時間を要する。

⑦ 校下処理は1日おきに行われ、処理時間は約2.5時間である。文献速報の2シリーズ分の同時処理が可。

⑧ UDCと雑誌名は各々分類コードと書架NOをキーとして、テーブルから自動的に付加されるが、雑誌名についてはテーブルメンテナンスがあり、漢字モードで週1回，英字カナ文字モードで月1回，約0.5時間の処理が行われている。

⑨ 年間索引処理は、年1回行われ、47年度実績で60時間の処理を行っている。

以上の処理時間はすべて漢字プリンタでのプリント時間を含まない。
このシステムで蓄積された情報は、JR漢字ファイルとして保存され、JICST で収集した理工学文献に関するサービスの情報源となる。

4.2 検索システム（漢字モード，英字カナ文字モード）

① 漢字モードで蓄積された文献ファイルに、漢字モードの質問をぶつけて漢字モードの回答を出すのが、漢字モードの検索システムである。

② 漢字モードのJR漢字ファイルを英字カナ文字モードに変換したファイルに、英字カナ文字モードで質問し、回答をうるのが、英字カナ文字モードの検索システムである。

JICST ファイルに対する検索では質問入力と探索ファイルは英字カナ文字モードで、回答は日本語抄録文付きの漢字モードで行っている。これは漢字モードの質問の入力校正が繁雑で時間がかかり、かつ漢字パンチの熟練者を要するなど、実用的でないためである。SDIシステムのフローを図-5に示す。質問は検索タグ（表-2参照）の論理結合（ブーリアン）で表わされ、リスト展開される。ファイルはシリアルで逐次に探索される。現行（FACOM230-50での処理）では、単位探索時間が8.3ms/記事質問（探索プログラムはアセンブラ）となり、JICST 理工学文献の場合、探索時間だけで、10時間/月に達している。今後質問数の増加、対象ファイルの増加は必至でありSDIに要するマシンタイムは増加の一途をたどると考えられる。

③ 外国製蓄積ファイルの探索

MEDLARS ファイルと CA condensates がすでに導入され、JRでサービスされていることは前述の通りである。

英文カナ文字モード
検索可能項目
- 資料種類
- 記事区分
- 資料番号
- 発行国
- 発行年
- 巻
- 整理番号
- 使用言語
- 著者
- 所属機関
- 分類コード
- UDC
- キーワード
- 欧文標題
- 和文標題

表-2

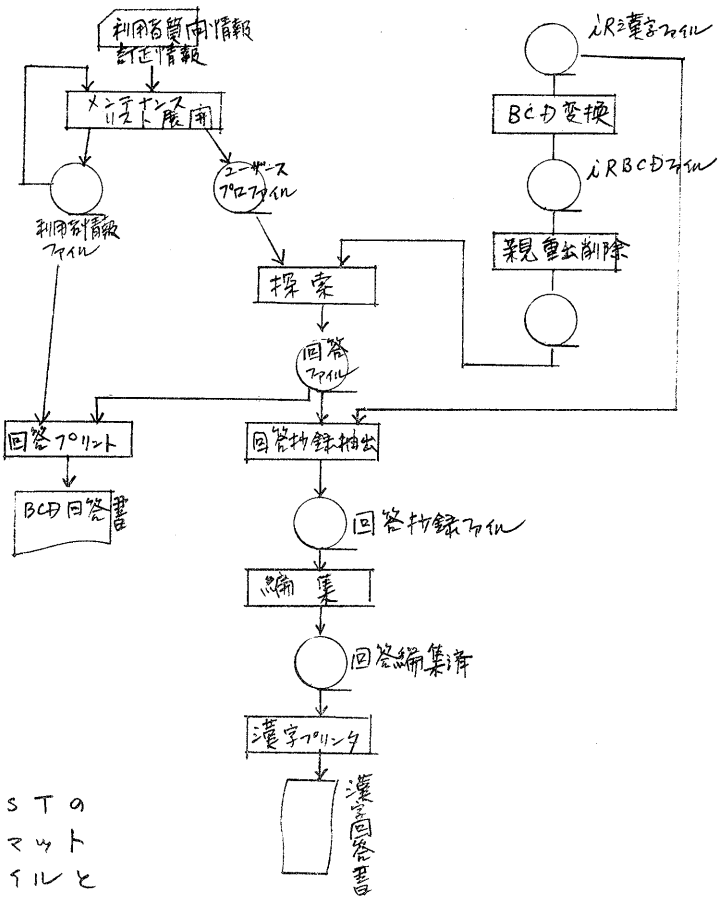


図-5　SDIシステムフロー

・CACファイルはJICSTのJICST BCDファイルのフォーマットに変換され，JICSTファイルと同じプログラムで検索される。但しファイルの内容とくにキーワードの表現形式、主題分類体系などにちがいがあるため、質問の内容が異なり、検索は別立てで行われている。

・MEDLARSファイルは、MeSHと称する詳いファイルから、質問に検索語を付加するなど、特異なシステムであるため、JICST，CACとは別の検索システムを用意している。

④ DOORシステム

会話型のオンラインリアルタイムシステムを志向して設計されたDOORシステムは、検索タグのうち、主題を表わすキーワードと分類コードについてインバーティドファイルをもつ、ディスクファイルシステムである。質問は会話方式により、最終的には、文献集合の論理結合で表現される。現在JICST理工学文献の約半年分20万件を蓄積ファイルとしてもち、SDIの質問式作成等に用いている。ファイルメンテナンスはファイルの再構成で行われ、

・SDIに用いられる親重出削除後のJICST BCDファイルを、検索タグと書誌部分に分割。（20万件で約5時間）

・ローディング（20万件で約1時間）

の手順で行われる。ファイルはユーザに常駐されるエントリーテーブルからインデックステーブル，タグテーブルをへて書誌ファイルに至る階層構造で、その間はアドレスでポイントされる。レスポンスタイムは平均2～3秒で、

部分マッチは、前方マッチのみで中間、後方マッチはできない。通信回線を介して入ってくるメッセージ　　の処理ルーチンはもたず、ジャーナルもとっていない。

4.3　用語管理（DOCTOR）システム

JICSTIRシステムで用いるシソーラスの編成作業を自動化したものである。特定のIRシステムに依存せず、インターフェイスを通じてあらゆるIRシステムと接続できるように設計されている。又すべての既存のシソーラスファイルを統合することも原理的には可能である。現状ではJICST理工学文献ファイルを対象とするIRシステムで使用される用語を漢字モードで登録し、それらの用語間の関係（同義、階層、関連）を定義している。この用語ファイルを用いて、蓄積文献ファイルに対し、ふりがなの付加、上位語の付加などの蓄積キーワード調整を行う。又、索引作成者や探索者の用語の選択に供するため、冊子体のシソーラスも出版される。47年度に作成されたシソーラスにはJICST理工学文献の全分野を統一的に扱う用語約32000語が登録された。用語管理システムのフローを図-6に示す。

4.4.　資料管理システム

文献の蓄積と検索に関するシステムとは、多少異質は、目録作成システムであるが、IRシステムと密接な関係を有する。つまり文献ファイルに蓄積される文献の二次資料は、このシステムの管理対象である1次資料から收録される。現状ではIRシステムとは全く別個に、JICSTで收集している逐次刊行物、会議資料，モノグラフなどの書誌事項（資料名，発行国等）を、英字カナ文字モードで入力し磁気テープに記録し、これをメンテナンスし、各種所蔵目録をプリントしている。

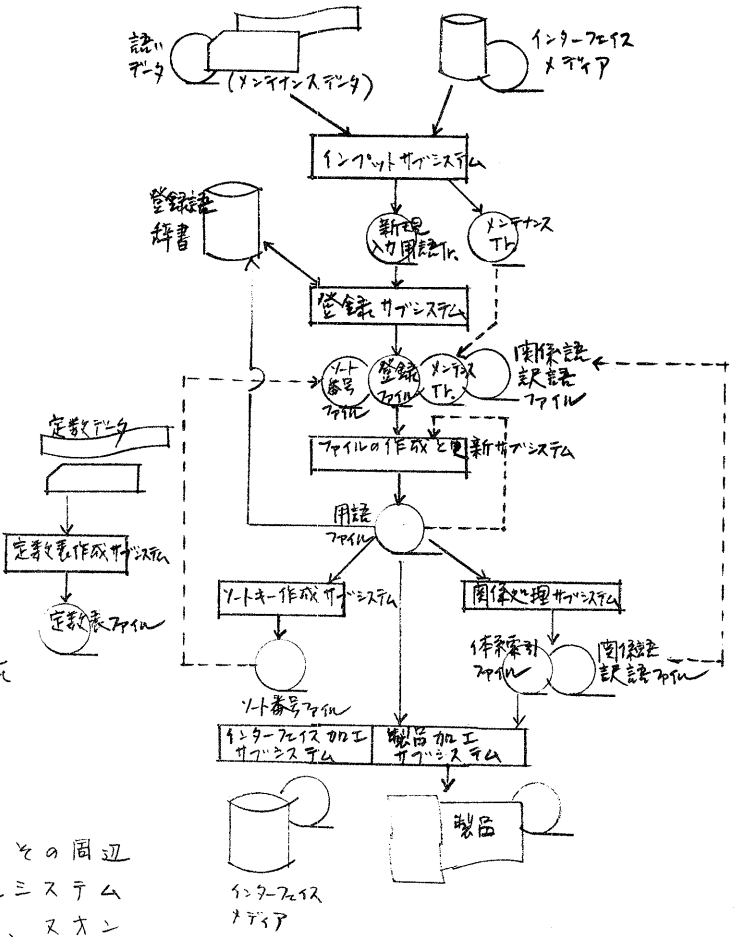


図-6　用語管理（DOCTOR）システムフロー

以上JICSTIRシステムとその周辺について現状を述べて来た。現システムは後述するような問題点を含み、又オンラインの会話型検索システム実用化の要求もあるため、JICSTでは近く予定されているマシンのレベルアップを機に、システム再設計に向けてスタートした。

## 5. システム再設計のための必要条件

### 5.1 システムの統合

前述の如く、JICST iRシステムとその周辺は、段階的に業種別に行われトータルシステム的なアプローチが完全でなかったため、システム間のインターフェイスが悪いとか、データや機能の重複があるなど、現行システムを整理統合する必要が生じて来た。問題点をいくつかあげてみると、

①ファイルのもち方がプロセスオリエンティドなものとなり、ファイル間のデータの重複や不一致がみられる。例えば文献ファイル（iR漢字ファイル，iR BCDファイル）と資料管理側で、資料名の表記法に差異がみられる。その上、資料名データの入力とメンテナンスを両方で行っている。これはどちらかに統一。

②システム間のインターフェイスについて

iRシステムと資料管理システムとの間に何らインターフェイスがないために①で述べたような矛盾が生じている。一方すでにインターフェイスをもっているものでも、もっと早くできる処理が、インターフェイスのタイミングが遅すぎるために遅れているケースがある。たとえば、検索用のファイルは、文選システムの校正完の文献データで充分で、現在のように抄録記用の記事番号付番を待つ必要はない。又用語管理システムと文選システムとのインターフェイスである、蓄積キーワード調整は、校正処理段階で行える。

③従来の業務別構成にとらわれず、機能別に整理されたシステムが望ましい。現在、ファイルメンテナンスやプリントルーチンなどは、機能的に類似したものも多い。これらは何らかの方法で汎用化できる。

④ファイルのもち方とも関連するが、JICST作成のファイルのみでなく、外部導入ファイルも含めた統一的扱いを必要とする。これについては後述する。

⑤オーソリティデータの一括管理

現在iRシステムで用いられる語いのコントロールはJICSTファイルについては総合的に行われている。未だ完全ではないが、入力の標準化、冊子体検索バッチ検索には今後とも有効であると思われる。又会話型のオンライン検索では検索の補助的なツールになるであろう。これらのほか、資料名，分類表，機関名なども、ある奥拠に基づいた標準化が必要であるが、これらの追加，更新，削除は、一本化して行うべきである。これらのオーソリティデータは理想的には外部導入ファイルも含めた総合的管理が必要であるが、実用性をふまえた堅実な努力が続けられるであろう。

### 5.2 ディスクファイルの採用。

現在RSはMEDLARSファイルについて一部サービスされているが、今後JICSTファイルを主体とした、MEDLARS，CACファイルを含む5～10年分の蓄積ファイルに対するそ及検索サービスが要求されるだろう。定期的に質問をまとめて行うバッチ方式と、オンライン端末からの会話方式の両方がある。これらの大量データを含むファイルからの検索は、

① 従来のMTベースではオペレータの負担も大きい。

② 大量のデータの中からランダムに1部のデータを引き出すRSでは、検索の速度だけを考えても、ランダムアクセスデバイス、なかでもディスクの使用が必然である。

③ 会話型のオンラインリアルタイムのファイルアクセスがある。

などの理由でディスクファイルの採用は確定的である。現在JICSTファイルは年間約40万件，MEDLARS，CACファイルは各々年間約23万件，36万件に達しているが、各ソースファイルの大むねの大きさを表-3に示す。

SDI用のファイルはJICSTファイルは校正完了直後のもの、外部導入ファイルは入手直後のものと考えられる。

| | SDI用(現在) | RS用(49年度から蓄積した場合の5年分) |
|---|---|---|
| JICST | 30000件/月 | 230万件(1035MB) |
| MEDLARS | 20,000件/月 | 133万件(931MB) |
| CAC | 7000~8000件/週 | 209万件(1150MB) |

表-3 各文献ファイルの大きさ

*RSファイルのメンテナンス

大量データをシリアルに逐次サーチしていくことは現実的にばかげたことで、現在の技術では、検索タグについてインバーティドファイルをもつことになる。RSファイルへの新しい文献データの追加は、たとえばSDIの周期に合わせて行われ、削除はもっと長い周期で行われるかも知れない。RSファイルは極力コンパクトにする必要があり、可変長フォーマットでの記録もさることながら、追加文献は、重要な文献のみにしぼる圧縮が必要である。又収録データエレメントも検索タグと必要最小限の回答項目に限定するのが妥当であろう。削除はインプット年月日とか、雑誌の発行年とか、その文献の使用頻度などをKeyとして行われる。このような文献の追加や削除は、基本的には、ファイル内のランダムな位置に発生すると考えられ、スペース効率と文献へのアクセス効率のいづれを補いうる手法が必要となる。又インバーティドファイルは、文献の追加や削除に伴って、該当文献の記事番号又はアドレスをKeyとしてメンテナンスされる。

*可変長ファイルの採用 固定長方式で可変長データを取り扱うのに便利なセグメント方式も、スペース効率の点で可変長方式のそれにはほど遠い。まして大量データはコンパクトにもつのが経済的であろう。処理が容易であるか否かも、数レコードにまたがる可変長データの多い文献データの処理では、一概にすぐれているとは言えない。

5.3 会話型のオンライン検索への要求

オンライン ~~リアルタイム~~ システムであるからといってRシステムであることに変りはないが、エンドユーザが直接ファイルと対話しながら、連想作用を伴う試行錯誤検索を行えるという判定がある。JICSTでは、実用的な1つのマンマシンインタラクションとして、オンラインリアルタイムで、会話型検索を行うために実状に即した段階的アプローチをしている。究極的な目標は下記の通りである。

① 主要枢要を結ぶ全国的ネットワークで端末の数は約100と想定される。
② キーボード付きの文字(EBCDIC)表示装置とプリンタをもつこと。
③ 応答時間は平均数秒程度。
④ 検索式はキーワード，分類コード，著者，資料番号，記事番号，記事区分等のAND，OR，NOTによる論理結合。完全一致、部分一致を可とすること。
⑤ 前回質問式に対する新質問式のAND，OR，NOTによる追加ができること。又前回質問式に対するキーワードの追加、削除、修正ができること。
⑥ シソーラスルックアップにより、端末から入力したキーワードの関係語がディスプレイされること。

⑦部分一致したキーワード群の記事件数付きリスト，記事の書誌事項とキーワード，インプットデータのエラーメッセージ等がディスプレイされること。
⑧検索された文献のうち、複写，翻訳が必要なものは、端末から記事番号の指定により、複写又は翻訳のターゲットをセンタープリンタに出力できること。
⑨SDIはバッチで行うが、SDI用のプロファイルの作成及び修正を端末から行い、検索式をプロファイルに登録できること。
⑩ハードウェア，ソフトウェアの障害からうける被害を最小限にとどめるため、検索途中結果の保存，チェックポイント・リスタート等の処置が必要。
⑪文献情報へのニーズが今後どのように増加するかは、不確定要因が多く、一概に言えないが、6年後にはSDIで3000プロファイル/年，RSで10,000件/年と予想される。

## 5.4 多種ファイルを含み多様なサービスにたえるデータベース

①情報発生量の飛躍的増加と国際的な情報の流通という点から、MEDLARS，CACに続いて今後も外部蓄積ファイルの導入が考えられる。情報支援用二次情報ファイルのフォーマットについては国際的標準化の動きもあるが、現実にはほとんどが、内容、形式ともに異っている。このようなファイルの多重利用に際して現システムは汎用性にとぼしい。(CACはJICSTファイルのフォーマットに変換されているが、キーワードなどの内容には差異がある。) 表-4に各ソースファイルの情報源と収録分野を示す。3つのファイルは索引内容にちがいがあり、内容の統一には難点が多いが、少くともファイルフォーマットの統一は可能で、したがって検索プログラムを共用できる。なおファイル間の分野の重なりからくる記事の重複は、記事を一義的にidentifyできるKeyが見い出せれば、どちらかを削除することにより避けることができる。

②機械検索サービスをはじめ、出版物，クリアリングサービスなど多様な形態のサービスに適用できる項目を含み、ファイル体系をもつ必要がある。又将来のライブラリオートメーションに備えて、複写閲覧、資料の受入れ管理なども考慮に入れておく必要があろう。

| | 情報源 | 収録分野 |
|---|---|---|
| JICST | 全世界50余か国の理工学領域の学協会誌，雑誌，技術レポート，会議報文集 | 基礎、応用物理，化学<br>電気工学，機械工学<br>金属工学，地学，鉱山学<br>土木・建築学，経営管理工学<br>原子力工学 |
| MEDLARS | わが国の127種を含む世界の約2,200種の雑誌 | 薬学，生物学，獣医学<br>農学，環境科学 |
| CAC | 雑誌12,000種<br>26ヶ国の化学部門特許<br>会議録，技術レポート | 生化学，有機化学<br>高分子，応用化学<br>化学工業，物理化学<br>分析化学 |

表-4 各ソースファイルの情報源と収録分野

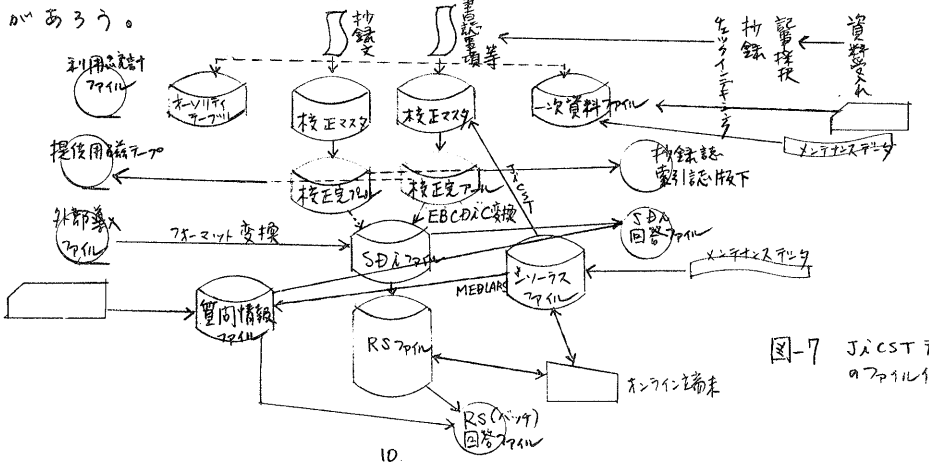


図-7 JICSTデータベースのファイル体系