心身障害研究報告書のデータベース化に関する研究

分担研究 保健福祉情報の整備と活用に関する研究

研究協力者 斉藤 進[1] 庄司 順一[1] 中村 敬[2]
恒次 欽也[3] 中沢 明紀[4]

要約：

母子保健・医療の臨床や行政施策の充実に貢献している心身障害研究の情報をより有効に活用する方法として、その報告書のデータベース化の可能性と実用性について、既存資料や関係者からの聞き取り調査データをもとに検討した。 その結果、1.報告書の配布状況から、データベースとしては全文を入れたものが有効である 2.データベース作成にあたっては、実施時期を基点に将来と過去との二方向別々に検討する 3.将来部分は全文テキストのデータベースが理想である 4.過去の報告書はイメージデータ中心のデータベースが適切である 5.データベース情報の配布方法としては CD-ROM が適当であることがわかった。 イメージデータを利用したデータベースのサンプル版作成を試みた。

見出し語： 心身障害研究報告書 データベース CD-ROM SGML

【目的】

母子保健・医療の臨床や行政施策の充実に貢献している心身障害研究の情報をより有効に活用する方法として、その報告書のデータベース化の可能性と実用性について検討し、具体的作成方法とそのためのシステムについて明らかにする目的で本研究を実施する。

【研究方法】

AiKEN-CD（日本総合愛育研究所発行の母子保健・児童福祉文献データベース）作成の経緯資料やデータベース検索ソフト等の既存資料、および関係者からの聞き取り調査データをもとに、研究協力者で検討し、具体的なデータベース作成の可能性を明らかにした。

1 日本総合愛育研究所
2 東京都母子保健サービスセンター
3 愛知教育大学
4 神奈川県藤沢保健所

【結果】

1. データベースとして必要な項目と内容

データベースの利用は、検索がやさしく、手軽で、早くでき、おおまかな内容が見れると便利である。したがって、最低限、一般的な書誌項目（題名、著者、所属、報告書名、掲載ページ、発行年、キーワード）プラス要約がデータ化されていることが望ましい。

AiKEN-CD に収録されている心身障害研究データベースは、分担研究者までの報告論文についての書誌情報である。問題点として、分担研究までの報告事項書誌ではおおまかすぎて実用的とは言えない。また、内容をまったく見ることができないことから、書誌プラス要約は必須であることを確認できた。収録範囲は最小の報告論文までデータベースに収録することが理想である。

次に、報告書の配布は限定されているので、検索してもその報告書自身の所蔵先を探し、報告内容そのものを見ることはむずかしい。したがってデータベース化にあたってはデータベース中に報告書の内容すべてが何らかの形で収録されている必要がある。

研究報告書の内容すべてがデータベース化されることを前提に実際の方法を検討することにした。

2. 項目データの電子化の方法

実際的にデータベースを作成するにあたって、項目内容の電子化方法を、1.テキストデータ（文字コード情報） 2.画像データ の二面から検討した。結果は、表１の通りである。

本文等の文字主体の部分については、テキストデータで入力し、図表部分についてはイメージデータで入力する方法が良いと考えられる。

既存ソフトの利用、例えば文字情報は「一太郎」「ワード」、図表は「エクセル」「ロータス」等を使用することも検討したが、データベースの性格上、長期間にわたって蓄積され利用されることから、特定のソフトになるべく依存しないデータ形式で電子化する方が有効であると考えられる。

将来は、世界標準化機構の規約（ISO 8879）、日本工業規格（JIS X 4151）で定められている「文書記述言語 SGML」の使用が望ましい。SGML は特定のソフトに依存しないデータ形式で文字、画像両方を扱うことができ、文書データベースの作成保守や電子出版に適したものである。

表１　情報の電子化方法の比較

	ファイルサイズ	入力の容易性	検索可能性	利便性
テキストデータ	小さい	作業量多い	自然語検索可	引用等が簡単
画像データ	大きい	作業比較的簡単	検索工夫が必要	写真として使用

現在、SGML が注目されているのは CALS（ Continuous Acquisition and Life-cycle Support）においてである。CALS（カルス）は米国国防総省における各種調達業務の合理化計画としてスタートし、現在は一般官庁や民間企業間の取り引きにも適用されてきており、製造業を含めた産業活動全般をカバーする情報化システムである。SGML はこの CALS の提出文書の規格として採用されている。航空機関係のマニュアルをはじめ、製造関係の製品仕様書や各種の技術文書は SGML 規格が使用されている。SGML を使用し標準化した文書データは、依存ソフトがなく、テキストデータなので、全文データベースの作成が容易で、有効性が高い。学術情報センターの「学術情報センター紀要」は SGML を使用して作成されており、「学術論文データベース」等へも応用されている [1]。実際には、SGML 化の入力、変換ソフトや全文検索ソフトの開発等の問題が残されており、今回のデータベース作成にあたっての導入は難しいことがわかった。しかし、将来必ず移行する必要性があると思われる。

3. 検索方法と検索ソフト

検索方法は、1.項目内容すべてを検索、2.特定の項目のみを検索の 2 種類が考えられる。また、検索方式として 1.自然語検索（文中の文字を検索させる方法）、2.検索用インデックスを使用した検索（検索用キーワードを作成して検索効率を図ったものや統制語によるシソーラスを使用したもの）を検討した。

検索用にインデックスを付与し、シソーラスによる検索を取り入れるためには、シソーラス用の統制語を選定、決定する必要があり、また統制語をキーワードとして改めて報告論文に付与する作業が必要となる。このため検索ソフトの問題はないが、シソーラス作成の困難性が予測された。

検索方法は、自然語検索を中心にすると良く、インデックスを使用する場合、コンピュータソフトを利用したキーワードの切り出しによるインデックス付与程度にする方が簡便で、検索効率が良いと思われる。この辺については、データベースに加工するオーサリングソフトの性能でカバーできる場合もあるので、可能な限り使用を検討する必要があると考えられた。

実際のソフトについては、現在 2～3 のソフト（新日本製鐵製 NSEARCH、NEC 製 infoSHOT、DynaText 等）の情報を収集中であるが、全文を検索する方法は、データ量、検索時間、コンピュータの性能等との兼ね合いが課題である。現状では、書誌事項と要約部分の自然語、インデックス検索を中心に作成した方が良いだろう。

4.データベースの作成

データベース作成にあたっては、実施時期を基点に将来と過去との二方向別々に検討する必要がある。大部分を電子データとして収集できる可能性を持つ将来の報告書と、印刷物としてしか入手できない過去の報告書では、同一の方法でデータベース化することは難しい。

過去の報告書全文をテキストデータとして入力するには、OCR（光学式読取り）等のシステム、スキャナーで読込み、OCR ソフトで文字に

変換するシステムと、タイピングによる入力が検討された。タイピングによる全文入力は莫大な費用が必要である。また、OCR方式は検討中であるが、完全性への不安が懸念される。

このことから、次の3つの作成方法が考えられる。

- 全文テキスト（図表はイメージ）データで入力、検索は全文テキスト検索によるデータベース
- 全文テキスト（図表はイメージ）データで入力、検索は書誌、キーワード、要約項目の検索によるデータベース
- 本文イメージデータ、検索用に書誌、キーワード、要約をテキスト入力し、検索するデータベース

過去の報告書については、データ全部をイメージ（画像）ファイルとして入力し、検索に必要な書誌事項のみをタイピング入力してデータベース化する3番目の方法が良いと考えられた。実際の使用感や実用度を検討するために、この方式でサンプル版の作成を試みた。

将来部分は全文テキストのデータベースが理想である。入力コストや作業性から検討すると、報告書原稿のデータを電子データで収集できると実現の可能性が高いが、現実にはワープロ、パソコン等の機種の違いや使用ソフトの違いなどをどうクリアするかが課題である。この問題については、今年度の報告書の一部を実際に報告書原稿を電子データ（フロッピー）として収集し、検討する予定である。

5.配布メディア

AiKEN-CDの作成配布経緯の資料から、情報やデータの提供方法は、制作コスト、使用コスト、配布コストからCD-ROMが適当である[2]ことがわかった。現在、販売されているノート型以上のパソコンでは、ほとんどCD-ROMドライブが装着されており、個人ユース上は問題はないと考えられる。

6.サンプル版の試作

平成6年度の一報告書を使用し、サンプル版データベース（CD-R版）を作成した。これには電子出版用に開発されたCD-ROMソフト「経葉」を試用した。このソフトは、書物として存在する書籍を廉価に配布、保存することを目的に開発されている。本文イメージ部分は、マウスを使用して書物を読むのと同じ感覚で使用できる。イメージデータのメリットを生かし、拡大•縮小が自由にでき印刷も可能である[3]。検索用に書誌事項とキーワードをテキスト入力した検索体系を付加して作成した。

テキスト入力データは、報告書の内容に忠実に入力したため、報告者名等において異字や外字に関係する課題が発生した。予測どおり、報告者名や固有名詞の表記法など、今後のデータベース化にあたっての重要な課題であることがわかった。

【考察】

今年度のサンプル版データベースは、一冊の報告書から作成したものであるため、検索のスピードや操作性など利用者の立場に立った十分な検討は実施できなかったが、過去の印刷され

た報告書のデータベース化の方法としては、細部の検討課題はあるものの基本的に有効であると考えられる。

報告書全体を収録したデータベースの作成は、利用性、保存性、活用性などは従来の印刷物以上に向上するだろう。しかし、有用性は十分であっても、今後の課題として検討を要することは、パソコンなしでの利用はできない点についての対策である。作成機関を情報センター化し、継続的にデータベースを作成すると同時に、オンラインやFAXサービス、代行検索、印刷サービスを実施することが望まれる。

将来、学会論文、研究報告書等はSGMLでの提出、交換が一般化され、印刷出版、データベース化図られるだろう。SGML形式に変換するソフトの開発、発売が待たれる。報告書の提出様式決定にあたって、今後はSGMLを視野においた報告書様式や文章構造を定義してゆくことが必要である。

現時点では、ソフトやデータ形式は別として、報告内容の電子データ（フロッピー）提出の義務化を最優先したい。次に報告書の構成や様式等について検討することが課題である。

次年度は、サンプル版の評価と今年度報告書の電子データを使用しての全文テキスト入力データベースの作成を試みる予定である。

【文献】

1. 根岸正光：SGML普及への展望，SGMLの活用（根岸正光・石塚英弘共編）：P159-164，1994，オーム社
2. 斉藤進 他：母子保健・児童福祉分野におけるデータバンク事業の現状と課題，日本総合愛育研究所紀要第31集：P91-101，1995，日本総合愛育研究所
3. イメージCD-ROM「経葉」参考資料：1995，経葉社

要約:

母子保健・医療の臨床や行政施策の充実に貢献している心身障害研究の情報をより有効に活用する方法として、その報告書のデータベース化の可能性と実用性について、既存資料や関係者からの聞き取り調査データをもとに検討した。　その結果、1.報告書の配布状況から、データベースとしては全文を入れたものが有効である　2.データベース作成にあたっては、実施時期を基点に将来と過去との二方向別々に検討する　3.将来部分は全文テキストのデータベースが理想である　4.過去の報告書はイメージデータ中心のデータベースが適切である　5.データベース情報の配布方法としては CD-ROM が適当であることがわかった。　イメージデータを利用したデータベースのサンプル版作成を試みた。