

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 齊藤 進¹, 庄司順一¹, 中村 敬²
恒次欽也³, 中沢明紀⁴

【要約】

厚生省心身障害研究の情報活用方法としてのデータベース化の可能性と実用性について検討し、電子データを使用したサンプル版 CD-ROM を作成した。

昨年度の報告書の記載内容をフロッピーディスク（電子データ）で提出してもらい、分析した結果、電子データは使用機種、OS、ソフトも多彩で提出の統一化は難しく、また二段組が災いしてテキストデータへ変換することが困難であることから、テキストデータ形式で提出してもらうことが無難であることが明らかになった。データベースについては検索機能の能力でその選定する必要があり、適当なソフト選定が課題となる。従って現状では、イメージデータとテキストデータを組み合わせる方法がコスト面、実用面で適当であるが、今後のパソコン環境の進歩を考慮して電子データ（テキスト）提出の義務化を推進していくことが必要である。

【見出し語】 心身障害研究報告書 データベース 電子データ テキスト形式

【目的】

近年、パソコンの利用環境は急速に整いつつあり、パソコンを利用した研究・調査情報の提供が求められている。

本研究では、厚生省心身障害研究の成果を多くの研究者や行政担当者等が有効活用を図るために、研究報告書の電子化、データベース化を進めるにあたって、その方法と課題を明らかにする目的で実施した。

【方法】

「保健・医療・福祉にかかわる医療資源の有効活用に関する研究」班（平成7年度 小宮班）の各班員の研究報告書原稿を電子データ（フロッピー）で提出を依頼し、電子化状況を分析検討した。あわせて、この電子データをもとにデータベースサンプル版（CD-ROM）を作成し、この作成過程に発生した問題点を析出した。

1 日本総合愛育研究所

2 独協医科大学越谷病院

3 愛知教育大学

4 神奈川県藤沢保健所

【結果】

1. 報告書原稿の電子化状況

提出された各研究協力者による42の報告書原稿データの電子化状況をみると、使用機種については、図1のとおり、Windowsパソコンが45.2%でもっとも多く、次にMacintoshパソコン33.3%、ワープロ専用機19.0%、不明2.4%で、パソコン使用者が多かった。パソコンだけを考えた場合、一般的な使用機種の割合に比べ、Macintosh使用割合が高いことが特徴であった¹⁾。

使用ソフトは表1のとおり、一太郎が33.3%で一番多く、次でMS Word 14.3%、EG Word 9.5%であった。一太郎ではバージョン3~6までが使用さ

れており、その割合は、Ver.3：21.4%、Ver.4：14.3%、Ver.5：35.7%、Ver.6：28.6%と多様であった。またワープロ専用機はキャノワード、オアシス、ルポ、カシオワード、書院と多彩であった。

フロッピーの種類については、使用機種と同じであったが、MS-DOS用フロッピーではNECフォーマットとIBMフォーマットが混在していた。

データベース作成の素材として、報告書原稿のフロッピーでの提出は、複数の使用機種、使用ソフトの多様性から、現時点では提出を統一することは難しく、各機種、ソフトで変換出力が可能なMS-DOSテキスト形式によることが一番標準的であるといえる。

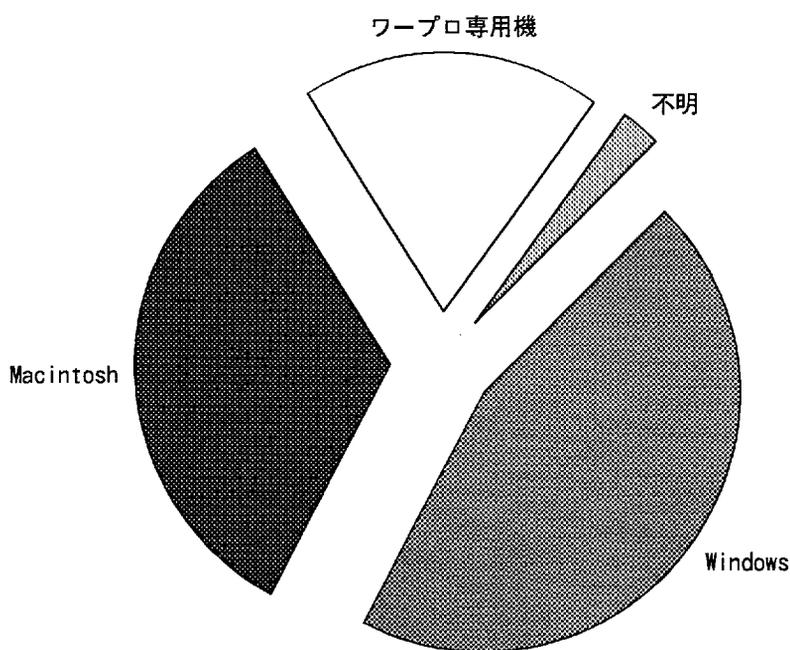


図1 使用機種

表1 使用ソフト(N=42)

ソフト名	実数	%
一太郎	14	33.3
MS-Word	6	14.3
EG Word	4	9.5
マックライトII	2	4.8
キャノワード	2	4.8
オアシス	2	4.8
東芝ルポ	2	4.8
Mac Word	1	2.4
Page Maker	1	2.4
テキスト	1	2.4
カシオワード	1	2.4
シャープ書院	1	2.4
不明	5	11.9

表2 フロッピーの種類(N=42)

フォーマット	実数	%
NEC(MS-DOS 1.2)	14	33.3
IBM(MS-DOS 1.44)	5	11.9
Mac	14	33.3
ワープロ専用	8	19.0
不明	1	2.4

データコンバート作業

前述の電子データ(フロッピー)を市販のテキストコンバータソフトおよび手持ちのワープロソフト(Win/Mac)を使用してテキストまたはMS-Wordファイルへの変換を試みた結果、次のことが明らかになった。

- ① フロッピーに使用機種やソフト、ファイル名が記載されていないことが少なくなく、機種やソフトの確認が一つ一つ必要となるため、作業能率が著しく低下する。ただし、これは提出時に記載条件の義務を付せば解決する。
- ② 一部にワープロソフトで作成されたファイル

に、ソフトで二段組みしたものでなく、スペース等を工夫して見かけ上二段組みにしたものがあった。このような場合、テキスト変換を行うとスペースを使用した段組みに変換されたり、半角、全角の関係で位置がずれる場合があった。

- ③ ワープロ専用機作成の電子データをテキスト変換する場合、段落末に改行コードが入っていないため、段落がなくなり、すべての文が繋がってしまうことがある。ワープロファイルに改行コードが入っている場合は問題はないが、一般的にワープロ専用機の場合、改行コードを入力しなくても作成可能な機種があるためである。
- ④ 罫線を使用して作成された表の場合、MS-Wordファイルへの変換で、表組みがずれるなど一部に不具合がみられた。

以上から、ワープロ専用機の文書ファイルコンバートにおける大きな問題点は、二段組みに関係することであった。段間スペースになったものをテキストデータにすることは、時間的、労力的にかなりの困難な作業であることが明らかになった。

データベース化の試み

前年度に報告した、心身障害研究報告書をデータベース化する3つの方法は次のとおりであった。

- ① 全文テキスト(図表はイメージ)データで入力し、検索は全文テキスト検索によるデータベースにする。
- ② 全文テキスト(図表はイメージ)データで入力し、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にする。
- ③ 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にする。

このうち、過去の報告書については、印刷された報告書をイメージ(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、データベース化する 3 番目の方法が適当であると考えられ、そのサンプル版 (CD-ROM) を作成した。これについては、今年度、別途実際の使用感や実用度を検討した (中沢らの今年度報告を参照)。

さらに今年度は、昨年度とは異なるソフトウェア (検索機能が充実) を使用したイメージ中心 (一部 MS Word ファイル) のデータベースおよび全文テキストデータを用いたデータベースの 2 種類の CD-ROM サンプル版を作成し、各 1 枚を厚生省に提出した。

その作成プロセスと、使用したソフトウェアは次のとおりであった。

(1) イメージデータベース

<作成プロセス>

図 2 のとおりである。

<使用ソフト: Alchemy>

Alchemy は、さまざまな OA ソフトウェアで作成されたファイルを、データベースとして CD-ROM に格納するソフトウェアで、データベースを構築する Alchemy Build と、構築されたデータベースを検索表示する Alchemy Search で構成され、その特徴は次のとおりである。

- ① データファイルを階層化された「バーチャル」フォルダに格納する。このフォルダ構造は CD-ROM の目次として、ファイル検索に役立つ。
- ② ドキュメント情報 (索引カード) がそれぞれのファイルに生成され、これによりキーワード検索が可能になる。
- ③ 異なるファイル同士 (テキストファイルとイメージファイル等) を関連づけられる。
- ④ MS Word、MS Excel およびテキストファイルは、テキストデータの内容も索引化できる。
- ⑤ 作成された CD-ROM には、検索・参照するためのエンジンソフトである Alchemy Search

も格納されるので、Alchemy のインストールされていないパソコンでも、検索・参照が可能である。

- ⑥ 専用ビューアソフトをインストールしなくても、テキスト、表計算、画像ファイル等を表示、印刷できる。
- ⑦ イメージファイルでは非常に多くのフォーマットをサポートしている。
- ⑧ ファイルに関連づけられた外部アプリケーションを起動できる。

(2) テキストデータベース

<作成プロセス>

図 3 のとおりである。

<使用ソフト: DynaText>

Dyna Text は ISO8879 SGML (Standard Generalized Markup Language) テキストをそのまま受け付けることができる電子出版システムである。前年度に報告した SGML の特徴をいかした大量文書のフルテキストサーチが可能である。

2. データベース作成にあたって

ここでの課題として次のことが明らかとなった。

- ① 提出報告書原稿の文書構造の標準化が必要である。
- ② 変換作業の効率化、低コスト化のためには、使用ソフトウェア等の統一が必要である。
- ③ マルチプラットフォーム (Windows と Macintosh に対応) で使用できる検索ソフトと表示ソフトの選定が必要である。
- ④ 検索機能の能力をどの程度にするかを決定する必要がある。
- ⑤ 提供サービス方法の選定が必要である。

これらの課題を検討した結果、文書構造の標準化については別途様式を検討し、報告した (中沢らの今年度報告を参照)。

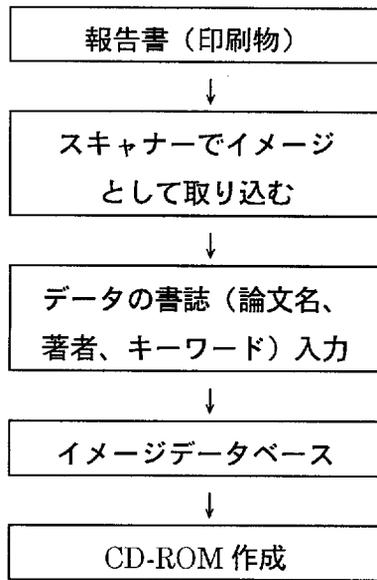


図2 イメージデータベース作成プロセス

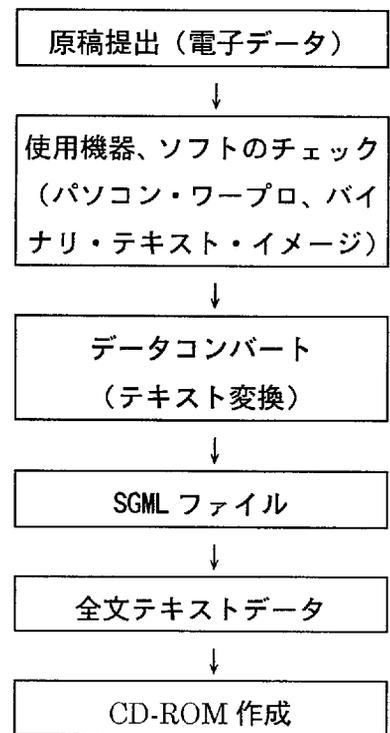


図3 テキストデータベース作成プロセス

提出電子データの統一については、テキスト形式による提出（図表はイメージ）を基本とし、標準的ワープロソフトを選定すべきである。これはインターネットのホームページの普及により、変換ソフトやワープロソフトが高機能化し、SGML ファイルに変換可能なものが増加する傾向にあるためである。

③、④の点について、検索機能を重視しない場合は電子出版ソフトがマルチプラットフォームで利用しやすいと考えられる。今回、最近マルチメディアタイトルによく使用されているマルチプラットフォームな出版ソフト（エキスパンドブック）は、試みることはできなかった。

【考察】

データベースサンプル版を作成し、その過程で明らかになったことから、次のことが確認できるだろう。

原稿の電子データ提出については使用機種、OS、ソフトも多彩で統一は難しく、テキストデータ形式での提出が無難である。これと併行して、インターネットの普及でHTML（SGMLの一形式）ファイル化が可能なワープロやコンバータソフトが開発されていることから、一定のワープロソフトを何種類か指定することも検討する必要がある。

報告書の様式として、現在二段組みが使用されているが、ワープロ専用機での二段組はテキストへの変換が難しいことと、画像（イメージ）データベ-

スとしてディスプレイに表示する場合は読みにくいことから、一段組みを基本とすることも検討すべきである。

現状では、イメージとテキストを組み合わせる方法がコスト面、実用面でもっとも適当と考えられるが、今後のパソコン環境の進歩を考慮して、現時点から電子データ（テキスト形式等）の提出を義務化していくことが必要である。

データベース化ソフトは、マルチプラットフォーム対応で、かつ検索機能が強力であることが必要である。また、データベースを CD-ROM として配布をする場合は、使用料や継続性を考慮してソフトを選択することである。

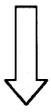
今後の課題は、HTML（インターネットのホーム

ページ用）ファイル化の検討や OCR（光学式読み込み）方式の採用、パソコン環境の進歩にあわせた情報提供方法の検討（インターネット、電子ブック）である。

来年度は、今年度のサンプル版の評価とあわせて全研究班報告書のイメージデータベース（CD-ROM 版）を作成し、その実用性や有用性についても検討したい。

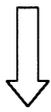
【参考文献】

1. 齊藤進他：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(2)，平成7年度日本総合愛育研究所紀要 第32集，p107--114,1996



検索用テキスト OCR(光学的文字認識)ソフト使用

論文の一部ですが、認識率の関係で誤字が含まれる場合があります



【要約】

厚生省心身障害研究の情報活用方法としてのデータベース化の可能性と実用性について検討し、電子データを使用したサンプル版 CD-ROM を作成した。

昨年度の報告書の記載内容をフロッピーディスク(電子データ)で提出してもらい、分析した結果、電子データは使用機種、OS、ソフトも多彩で提出の統一化は難しく、よた二段組が災いしてテキストデータへ変換することが困難であることから、テキストデータ形式で提出してもらうことが無難であることが明らかになった。データベースについては検索機能の能力でその選定する必要があり、適当なソフト選定が課題となる。従って現状では、イメージデータとテキストデータを組み合わせる方法がコスト面、実用面で適当であるが、今後のパソコン環境の進歩を考慮して電子データ(テキスト)提出の義務化を推進していくことが必要である。