厚生労働科学研究費補助金

特別研究事業

厚生労働科学研究成果データベースの検索機能強化に関する研究

（Ｈ２２－特別－指定－０２３）

平成 22 年度 総括・分担研究報告書

研究代表者 辻井 潤一

平成 23 (2011) 年 3 月

厚生労働科学研究費補助金

特別研究事業

厚生労働科学研究成果データベースの検索機能強化に関する研究

（Ｈ２２－特別－指定－０２３）

平成 22 年度 総括・分担研究報告書

研究代表者 辻井 潤一

平成 23 (2011) 年 3月

目　次

I. 総括研究報告

厚生労働科学研究費補助金（特別研究事業）
総括研究報告書

厚生労働科学研究成果データベースの検索機能強化に関する研究

研究代表者　辻井 潤一（東京大学 大学院 情報学環）

研究要旨

科学研究事業の社会還元に際しては、研究成果の適切なデータベース化が欠かせない。しかし、既存の厚生労働科学研究の研究成果データベースは、研究成果の全文データが収録されておらず、検索機能も制限されていた結果、国民が知りたい研究成果を効率的に検索することが困難であった。そのために、研究者への情報提供だけでなく、実用化への橋渡しや政策に対する基礎資料の提供などに際しても、適切な情報の還元がなされていない可能性が懸念されていた。

そこで、本研究は、厚生労働科学研究成果の活用促進を図るため、厚生労働科学研究成果データベースの利便性を高めるための機能向上について検討し、高度な知的検索に対応した高機能版の厚生労働科学研究成果データベースを暫定運用することを目的として開始された。

実際の研究は、「研究成果報告書を対象とした全文検索や知的検索に対応するインデックス化に関する研究」、「医学文献の知的検索に関する基礎研究」、「高機能版厚生労働科学研究成果データベースの研究開発」に分けて実施し、WIDE プロジェクトの協力を得て、高機能版の厚生労働科学研究成果データベースの暫定運用を実現した。また、OCR 技術を用いることで、1998～2002 年度の研究成果報告書を全文検索可能な形へとインデックス化すると共に、今後の研究成果報告書のデジタル化に向けた調査と提言を行った。

研究分担者

松崎 拓也	東京大学大学院情報理工学系研究科
岡崎 直観	東京大学大学院情報学環
建石 由佳	工学院大学情報学部コンピュータ科学科
奥村 貴史	国立保健医療科学院研究情報センター
中村 修	慶應義塾大学 環境情報学部 環境情報学科

研究協力者

関谷 勇司	東京大学情報基盤センター
堀場 勝広	慶應義塾大学政策メディア研究科
谷田 和章	東京大学学際情報学府
野中 真生	工学院大学情報学部
磯野 威	国立保健医療科学院研究情報センター
宮澤 博子	国立保健医療科学院研究情報センター

A．研究目的

近年、科学研究事業の社会還元に際して、研究成果の適切なデータベース化が求められている。厚生労働省においても、国民が直面する保健医療福祉分野の喫緊の課題に対する様々な施策を進めるための科学的根拠を確保するための事業として「厚生労働科学研究事業」を行っており、その研究成果報告を「厚生労働科学研究成果データベース」として集積し、公開してきた。

しかしながら、従来の厚生労働科学研究成果データベースでは、研究成果の概要に対する検索が行えるが、近年のデータベースでは一般的になりつつある全文検索や知的検索を行うことが出来なかった。また、研究成果に対して他のサイトよりリンクが貼れない、検索クエリのスペルチェッカやサジェスト機能、研究キーワードや成果報告書の類似度に基づく検索ナビゲーション機能がないなど、機能上の多くの問題があった。厚生科学審議会科学技術部会においても、昨年より成果データベースに関する議論は行われており、改善の必要性が指摘されていた。

公的資金による研究成果の社会還元に関する問題は、科学技術研究を行う他省庁においても共有認識となっている。文部科学省の予算監視・効率化特命チームにおける「研究費・プロジェクト系教育経費の効果的予算措置に関する中間報告」においても、科学技術データベース間の連携強化、研究成果情報の活用促進、政策決定に必要なエビデンスの整備などが求められており、その結果、平成24年度中に次期eRadの運用を開始するなどの具体的取り組みが始まっている。さらに、平成22年5月に策定された「知的財産推進計画2010」においても、産学連携を促進する環境整備のため公的資金による研究成果のオープンアクセスを確保するよう各省に要請されることとなっていた。

そこで、本研究では、厚生労働科学研究成果データベースの高度化に向け、① 全文検索、知的検索に対応するインデックス化の研究、② 医学文献の知的検索に関する基礎研究、③ 高機能版厚生労働科学研究成果データベースの研究開発を行うことにより、高度な知的検索に対応した高機能版の厚生労働科学研究成果データベースを暫定運用することを目的として開始された。

B．研究方法

1) 全文検索、知的検索に対応するインデックス化の研究

厚生労働科学研究では、毎年1500課題ほどの研究が実施されている。現行のGRANTSシステムには、その研究成果報告書をスキャンしたPDFファイルが年度毎に蓄積されているが、これらはスキャン画像として保存されているに過ぎないために、全文検索や高度な知的検索に用いることが出来なかった。

そこで、本研究分担においては、研究成果報告書に対して全文検索や知的検索が可能となるように、既存のPDFデータをOCR(光学文字読み取り)処理したうえで知的検索に必要なインデックス化を行うと共に、今後の研究成果報告書の効率的なデジタル化、インデックス化のための戦略提言を行うことを計画した。

文献のOCR処理については、様々な研究がなされまた商品化も進んでいるが、機械的な識字には限界がある。そこで、識字率を上げるため、読み取り対象の特徴を踏まえたチューニングを行うとともに、対象分野の辞書を用いるなどした読み取り結果の自動的な補正と、最終的な人力による目視確認を試みた(II-1章)。

ただし、これらの処理には相応の費用が掛かるために、保存されている全ての研究成果報告書を本研究において処理することは困難である。そこで、過去の研究成果報告を効率的にデジタル化するために、過去の研究者に自発的なファイルの提供を依頼し、必要データの回収を試みた。研究者からのデータ回収は、既に研究実施から年数が経過していることからも回収率に限界があるうえ、提出されたデータの正当性についての検証も必要となる。そこで、今回の回収実験による統計をもとに、今後のデータ回収手法についての検討と、今後のOCR精度の向上や低廉化の予測に基づいた厚生労働科学研究成果報告書データベース全体のインデックス化戦略について検討した(II-2章)。

2) 医学文献の知的検索に関する基礎研究

次に、研究成果報告書の有効利用に向けて、膨大な文献から必要な文献を効率的に検索するための技術について基礎研究を進めた。たとえば、救急外来に関する文献を検索するのであれば、単に「救急外来」というキーワードで検索をすれば良いが、「薬物過量摂取により救急外来を受診した患者の数」を調べたい際、その表そのものを検索、表示してくれるような技術は、文献調査の効率を大幅に高めるだろう。

そこでまず、研究成果報告書に含まれる表や図の見出しも含めて全文検索に対応させるために、OCRの精度向上に向けた検討を進めた。具体的には、OCRによって画像から獲得したテキストには、あらかじめ電子化されていたテキストと異なり、OCRの文字認識エラーや、二段組み等の文書レイアウトに起因するエラー等が含まれる。そこで、OCRの認識結果に含まれる単語表記の誤りを訂正し、シソーラスなどの既存の言語資源とOCRテキストとを結びつける手法について検討した(II-3章)。

また、研究成果報告書に対する知的検索に際しては、前節に示したように、OCR化されたテキストからキーワードや類似キーワードを自動抽出し、効率的なインデックスを作成する必要がある。そうした情報の効率的な抽出に必要な自然言語処理の基礎研究を行った。

とりわけ、医学文献からの効率的な情報抽出のためには、抽出すべき情報の手本(コーパス)を計算機に教えることで自動抽出の精度を上げる手法が一般的である。そのためには、コーパスの質、量ならびに抽出したデータの整理が重要となるために、その手本の作成作業(アノテーション)を効率的に行うための手法と疾患知識ベースの構築に関する研究を行った(II-4章)。

3) 高機能版厚生労働科学研究成果データベースの研究開発

最後に、高機能版の厚生労働科学研究成果データベースを仮設運用するための研究を行った。その際、事業として運用を行っている現行の厚生労働科学研究成果データベースとの独立性を保つために、現行システム上のデータベースを複製した上で、高機能版の研究成果データベースを低コストに暫定運用しうる手法について検討を行った(図1)。

まず、全文検索、知的検索に対応する研究成果データベースシステムを短い期間で開発するため、1週間毎に部分的な機能を開発し、短い間隔でデモを行いユーザーからのフィードバックを元に開発を進めていく反復型開発を行った。

また、開発した高機能版のデータベースシステムを、現行の研究成果データベースと統合するまでに暫定的なサービス提供が行えるよう、日本のインターネット研究コンソーシアムであるWIDEプロジェクト

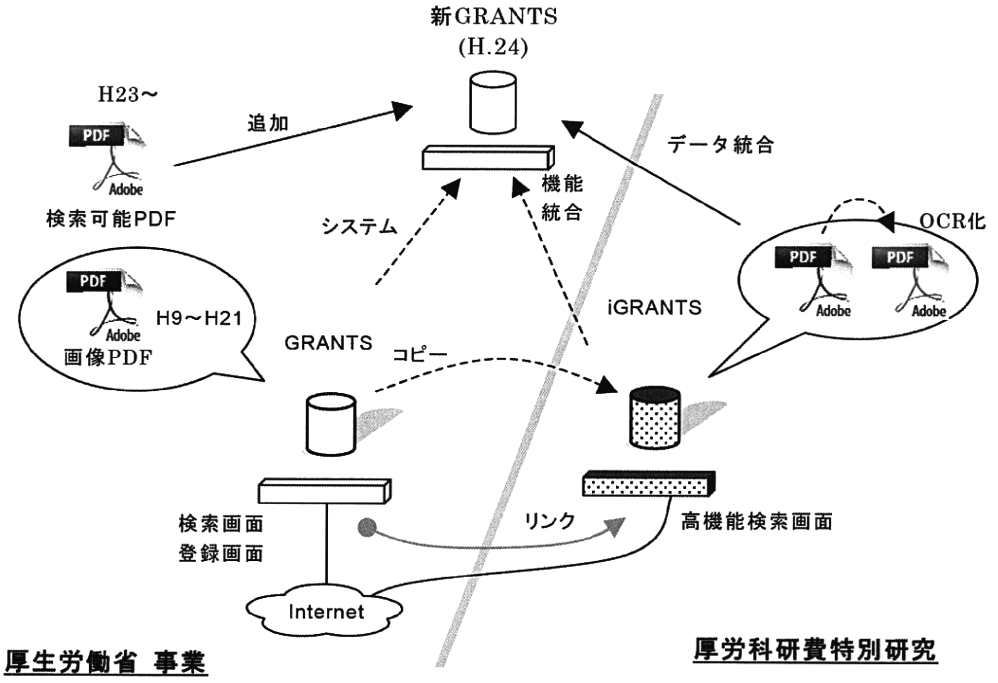

図1 既存の研究成果報告データベースとシステム構成

の協力の下、WIDEプロジェクトの運用する研究用クラウドシステム上にシステム構築を行った(II-5章)。

最後に、現行の厚労科研成果データベース本体を訪問するユーザーが、今回開発した高機能版研究成果データベースを違和感なく利用できるよう、現行データベースを運用する国立保健医療科学院研究情報センターとのコーディネーションを行った。また、今後、現行データベースを更改する際に、今回開発したシステムや技術を統合しうるよう、必要な技術的調整を行った。

（倫理面への配慮）

本研究は、既発表の研究成果報告書を対象とした研究であり、個人情報や実験動物等、倫理的問題が生じうるデータを扱わない。

C．研究結果

本研究により、厚生労働科学研究成果データベースに登録されている研究成果報告書の1998年から2002年度までのOCR処理を行い、フルテキスト化を行うとともに、知的検索に対応するインデックスを作成することが出来た。これにより、既存の研究成果報告書の2次利用における利便が格段に向上し、資料としての付加価値が増大した(参考資料III-1)。

また、全文検索機能に加え、検索クエリサジェスト機能、類似文献提示機能等の知的検索機能を備えた次世代版厚生労働科学研究データベースのプロトタイプを開発し、仮設運用を開始することが出来た。このデータベースには、フルテキスト化はされていないものの2003年から2009年までの研究成果報告書と要約が収録されているため、現行の厚労科研費研究成果データベースの

検索機能を代替することが可能となった(参考資料 III-2、III-3)。

また、本研究で試みた OCR の精度向上研究により、既存の研究成果報告書のフルテキスト化をより効率的に行えるようになった。さらに、過去の研究代表者からの研究成果報告書を直接回収するための基礎情報が得られたことに加えて、今後、厚生労働科学研究における研究成果報告をファイルにて提出するに際したノウハウの集積を行うことが出来た(参考資料 III-4、III-5)。これらにより、膨大な分量に上る既存の研究成果報告に加え、今後、蓄積されていく電子化された研究成果報告を効率的にデータベース化していくための戦略を定めることが可能となった。

さらに、今回実施した医学文献の知的検索に関する基礎研究により、医学文献の自然言語処理に関する新たな知見が得られるとともに、研究に必要なコーパスを整備するためのアノテーションツールを開発することが出来た(参考資料 III-6)。また、公的な疾患知識ベースの構築に向けた戦略の整理を行うことが出来た(参考資料 III-7)。これらは、厚生労働科学研究の研究成果を広く国民へと還元し医学研究の支援に生かせるだけでなく、医療用情報システムへと応用することにより、医療従事者の負担軽減などの波及効果が期待される。

D. 考察

研究成果のデータベース化は、公費によって行われる科学研究の成果を社会へと還元していく上で欠かすことが出来ない作業である。とりわけ、専門的な検索技能を有さない一般国民が簡便に利用できる検索システムは、研究助成の説明責任を果たすうえで少なくない意義を有する。実際、インターネットの検索エンジンは、図書館の蔵書検索システムよりもはるかに広く利用されていると考えられ、検索の効率化が知識の普及に果たす役割は少なくない。また、厚生労働科学研究は、わが国における厚生労働行政との結びつきが強く、施策に必要な各種統計や学術的エビデンスが効率的に検索、引用できる体制を整えることは、既存の研究成果の価値を高めるだろう。

一方、本研究においては、研究予算上の制約により、過去の研究成果報告書の全文テキスト化が1998年から2002年の範囲までしか行うことが出来なかった。これは、仮設運用するデータベースにおいては、当該年度の範囲においてしか全文を対象とした検索を行えないことを意味している。一方で、高度検索に対応するインデックス化は、1998年から2009年までの範囲をカバーしているため、既存の研究成果データベースの検索機能を代替しうる情報量を備えている。

本来、既存資料のデータ化は、研究ではなくデジタルライブラリの整備事業として施行すべき内容でもあるために、今後は、本研究が提言する既存資料の効率的なデータベース化を進め、より網羅性の高いデータベースの構築と維持を行うことが望ましい。

また、本研究から派生し構築された厚生労働科学研究成果データベースは、あくまで研究名目の仮設運用という位置づけにある。今後、このプロトタイプの運用経験を蓄積すると共に、実際の利用者からのフィードバックを収集することにより、来るべき事業としての研究成果報告書データベースの更改に向けた基礎資料とすることが望まれる。

E. 結語

本研究では、「研究成果報告書を対象とした全文検索や知的検索に対応するインデックス化を試みる研究」、「医学文献の知的

検索に関する基礎研究」、「高機能版厚生労働科学研究成果データベースの研究開発」を実施し、高機能版の厚生労働科学研究成果データベースのプロトタイプを WIDE プロジェクトのクラウドサービス上に構築した。また、OCR 技術を用いることで、過去の研究成果報告書の 1998～2002 年分を全文検索可能な形へとインデックス化すると共に、今後の研究成果報告書のデジタル化に向けた調査と提言を行った。

研究成果データーベースの全文テキスト化に必要なコストとしては、今回のコストモデルを用いると、未だ全文テキスト化が行われていない2003年から2009年度の処理のために 9,000～13,000 万円が必要となることが明らかとなった。また、今後の研究成果報告書の電子化に要するコストは、各年度で 990 万円(回収率 50%)～1,700 万円(回収率 10%)程度と見込まれた。これらの費用は、電子ファイルでの提出を義務化した際に生じうる費用と比して 1/10 程度に抑えられており、今後、電子ファイルでの提出率が 100%になるまでは、ファイル回収と OCR を効率的に組み合わせた戦略が効率的であることが示された。ただし、以上の費用推計は電子ファイルの回収率とOCR 処理の単価に依存しており、今後、ファイル回収率を上げるとともに、OCR そのものの単価を下げる工夫により、さらにコストを下げられる可能性がある。

今後、WIDE プロジェクトの協力により研究用の仮設データベースを暫定運用していくことで、研究開発した技術の検証と、厚生労働科学研究成果データベースの改良に向けたフィードバックや運用データの収集が可能となる。これにより、厚生労働科学研究の研究成果を広く国民に還元していくための基本資料が得られるとともに、厚生労働科学分野の文献に対する知的検索技術の高度化と医療用自然言語処理技術の発展が期待される。

F．研究発表

1．論文発表

なし

2．学会発表

野中真生, 奥村貴史, 建石由佳, 谷田和章, 辻井潤一,「疾患プロファイル作成のための症状名抽出」,言語処理学会第 17 回年次大会(NLP2011)予稿集, pp. 643-646, 2011 年 3 月.

宇佐美佑, Han-Cheol Cho, 岡崎直観, 辻井潤一. 自動構築した大規模訓練データを用いた固有名抽出. 言語処理学会第17 回年次大会(NLP2011), pp. 782—785, 豊橋技術科学大学(愛知県), 2011 年3 月.

Han-Cheol Cho, Naoaki Okazaki, Jun'ichi Tsujii, Token Boundaries or Named Entity Boundaries. 言語処理学会第17 回年次大会 (NLP2011), pp. 778-781, 豊橋技術科学大学, 2011 年3 月.

関谷 勇司,「大学間クラウド環境の実現に向けて」, 私立大学情報教育協会 大学教育と情報, Vol. 19, No. 4, pp. 2-4, 2011 年 3 月.

G．知的財産権の出願・登録状況

1．特許取得

なし

2．実用新案登録

なし

3．その他

開発した XMLinemarker(アノテーションツール)を、無償ソフトウェアとして公開準備中。

Ⅱ. 分担研究報告

厚生労働科学研究費補助金（特別研究事業）
分担研究報告書

全文検索、知的検索に対応するインデックス化の研究

研究分担者　松崎　拓也（東京大学大学院情報理工学系研究科）

研究協力者　宮澤　博子（国立保健医療科学院研究情報センター）
磯野　威　（国立保健医療科学院研究情報センター）

研究要旨

厚生労働科学研究データベース（GRANTS）システムには、厚生労働科学研究成果報告書が年度ごとに蓄積されているが、現行システムでは報告書冊子からのスキャン画像PDFとしてこのデータを保存しているため、報告書本文を対象とする全文検索や高度な知的検索に対応することは不可能であった。

本分担研究では、報告書本文データを対象とする全文検索、さらに類似文書の自動検出などの知的検索機能を実現するために、既存の画像PDFデータにOCR（光学文字読み取り）処理を施すことでテキスト情報を抽出した。これにより、報告書全文を対象としたインデックス化を行うことが可能になった。 さらに、OCR処理によって顕在化する科学研究成果データベースにおける著作権法上の課題について考察した。

A．研究目的

本分担研究は、研究報告書本文に対する全文検索および高度な知的検索に対応した高機能版の厚生労働科学研究成果データベースを暫定運用するための第一段階として、現行システムで画像PDFとして保存されている報告書データの一部にOCR処理を施し、テキスト情報を抽出することを目的とする。さらに、今後の研究報告書を電子データ化していくうえで必要となる著作権上の課題について考察する。

B．研究方法

現行のGRANTSシステムに蓄積されている報告書データの内、1998年度から2002年度までの報告書、計4634件、374,113ページに対してOCR処理を施した。このデータ量は予算の制約下で処理できた最大の量であり、現行GRANTSシステムに研究開始時点で蓄積されていた1998年度から2009年度までの報告書データ全体の約16%にあたる。参考資料1に、今回OCR処理の対象としたもの含め、現行GRANTSシステムに蓄積されている報告書データについての統計値を示す。

OCR処理によって抽出したテキストには目視による校正を施し、最終的にはインデクシング用のテクストファイル、および、画像PDFにテキスト情報を付加した透明テキストPDFとして保存した。目視チェックによる校正作業後の精度として、文字あたり精度99.9%を仕様とした。さらに、OCR業者からの納品後、研究グループ側でOCR精度についての抜き出し調査を行った。

その他、OCR処理に際して、抽出したテキストの検索システムでの利用のために、以下の2点の処理を依頼した。

- テキストブロック右端の印刷上の改行はＴＡＢコードに、また文の終了に伴う改行には改行コードを入れる
- 報告書中の図表は、タイトル(キャプション)のみを残し、中身を削除する

前者は、文の切れ目を印刷上の改行から区別するためである。また、後者は、標準的なOCR処理では図・表内のテキストを正確に認識することは難しく、図・表の認識結果がノイズとして出力に残ることを避けるためである。図・表のタイトルはテキストとして抽出されるため、検索の対象となる。

C．研究結果

上記374,113ページの報告書画像データ全てに対しOCR処理を施した。これに要した費用は約2500万円、ページあたり約66.8円であった。OCR業者からの納品後、研究グループ内で精度についての抜き取り調査(各年度から2報告書づつ、計10報告書)を行い、1報告書(99.7%)を除き、それ以外の全ての報告書で99.9%の文字認識精度であることを確認した。

D．考察

D-1) 研究報告書データ内の著作権譲渡済み著作物

冊子体の研究報告書には、そのデジタル化(複製権)およびインターネットによる一般公開(公衆送信権)が、出版社・学会など著作権者の権利に対する侵害であると見なされうる雑誌発表論文別刷や書籍からの抜粋などが含まれる。今回OCR処理の対象とした1998年度から2002年度の研究報告書を含む、2003年度以前の報告書については、現行GRANTSシステムへの画像データ蓄積の時点で、上記に該当する著作権譲渡ずみの既発表論文などが目視により取り除かれ、書誌事項のみが表記されているため、問題はないと考えられる。

しかし、2004年度以降の報告書については、規定類の見直し(「厚生労働科学研究費補助金事務処理要領」「厚生労働科学研究費補助金取扱細則」)により、研究者の公開承認を明記して冊子体提出のまま画像PDF化・公開されることが原則化されたため、画像データ内に著作権譲渡ずみの論文などが含まれている。

国内外の学術雑誌、学会誌へ論文を投稿した場合、著作者は当該論文の著作権を投稿先の出版者へ譲渡する場合が多く、その場合の著作権者は出版者となる。この結果、当該著作物の著作者の権利は制限を受けることとなる。よって、たとえ報告書データベースの運営主体が著作者である研究者からの公開承認を得ていても、「複製権」「公衆送信権」などが報告書データベースにおいて侵害されているとの指摘を著者(研究者)が出版者等(著作権者)から受ける可能性がある。この問題は、画像データを利用する現行GRANTSシステム、全文テキストを対象とする新GRANTSシステムともに共通の問題である。

D-2) OCR処理および全文検索機能による著作権に関わる問題の顕在化

今後、2004年度以降の報告書についてもOCR処理によるテキスト化を進め、新GRANTSシステムへと追加して行くとき、新システムの全文検索機能により、研究成果報告書に収載されている発表論文別刷、パンフレット、著書の一部または全部、な

どを含めた全てのテキストが検索対象領域となる。特に問題となるのは著作権者の許諾を得ていない著作物が公開データベースの中に存在しうること、また、新GRANTSシステムの全文検索機能、あるいは外部のサーチエンジンからの検索によって、それらのデータの発見が容易になることである。

たとえば「著者名」「雑誌名」などで外部のサーチエンジンから検索を行った場合、ヒットした著作物の著作権者による WEB ページに並んで、新 GRANTS システム内のページも検索結果として表示される、といったことが生じ得る。

D-3) 研究成果還元への要請と著作権に関わる問題との対立

一方、「厚生労働科学研究費補助金事務処理要領」等により、研究代表者による研究報告書の公開承認は義務付けられているが、「補助金により印刷物を作成した場合」「補助金により成果を雑誌等に掲載した場合」は研究報告書に盛り込むことを促しており、研究者が助成期間中に発表した研究成果は適時、公表されそのまま研究報告書へ掲載することを認めている。

研究成果の国民と行政課題への迅速な還元が最優先される厚生労働科学研究費補助金事業においては、研究者による報告書全文に対する公開承認のもとで報告書全体のデータを公開している現在の状況はやむをえないものと考えられる。

また、報告書データベースにおける全文検索機能、および引用機能向上のために個別検索結果に個別 URL を付与する機能は、上記のように著作権との関連において問題となる可能性がある一方で、報告書データベースの有用性を大きく高め、研究成果の社会還元・研究期間後の有効活用に対する効果が非常に大きい。

D-4) 著作権に関わる問題に対する方策

今後、全文電子テキスト化された報告書を検索対象とする新 GRANTS システムの上で、上記の問題に対応するには以下の方策が考えられる。

A案）公開前にあらかじめ対象となりうる著作物を特定し、当該著作物の著作権者から許諾を得る。対象となりうる著作物の特定については、全文検索機能を利用し、例えば雑誌名などによる検索を一括で行う、といった方法も考えられる。著作権者による許諾を得られなかった著作物は公開対象から外す。この作業は現行 GRANTS で公開しているすべての報告書の見直し作業となる。併せて、今後、研究代表者が誤って、著作権者から未許諾の著作物が入らぬよう担当課によるチェックを厳しく行う必要があろう。

B案) 研究者から(透明テキストつき)PDF 提出を義務化するにあたり、研究成果報告および公開のルールを見直し、あらかじめ「公開方式の原則」を定める。こうした方法としては、米国 National Institute of Health (NIH)が「パブリック・アクセス・ポリシー」を 2008 年 4 月より公表し、研究費助成の成果を国民に広く公開している事例がある。NIH は、研究成果の公表に先立ち、研究者、学会、出版者などと検討を重ね、著作権の処理についてのルールについて結論を得ている。現在、NIH の研究助成による研究成果論文は、Naitonal Library of Medicine (NLM)の PubMED Central より一般に公開されている。その利用状況は 2009 年度では月間、2,140 万件にも達し、利用の社会還元が進んでいる。

その主要な目的は「アーカイブの保存」「科学の推進」および「患者家族を含むユーザのアクセス性の向上」としている。

今後、研究成果の社会的資源としての有効利用の上から、関連データベース(Medline、医学中央委雑誌など)との相互リンクを進めていくとすれば、B 案の「公開方式の原則」を定める方法が妥当であろう。併せて、論文別刷り、著書、パンフレットの一部など著作権が研究者にある内に著作権の処理が出来る制度を作ることが望ましいと考えられる。

E. 結論

1998 年度から 2002 年度までの厚生労働科学研究成果報告書、総計 374,113 ページの画像データに対し OCR 処理を施し、99.9%以上の文字あたり精度でテキストデータ化、および画像 PDF へのテキストデータ付加(透明テキスト PDF)を行った。併せて、全文テキストデータを対象する高機能な新報告書データベースシステムにおいて特に問題となり得る著作権に関する問題について考察を行い、今後の方策への提言を行った。

F. 研究発表

なし

G. 知的財産権の出願・登録状況

なし

H. 参考文献

平成 19 年度厚生労働科学研究費補助金総括・分担研究報告書「科学研究費計画書の作成支援システムに関する研究」(研究代表者：土井徹), 2007 年 3 月, pp.1-4.
野添篤毅, 磯野威, 緒方裕光, 「研究成果の公開システム：米国 NIH パブリック・アクセス・ポリシーとデジタル・アーカイブ PubMed Central」, 平成 22 年度厚生労働科学研究費補助金「循環器疾患・糖尿病等生活習慣病対策総合研究事業【健康日本 21 の中間評価、糖尿病等の「今後の生活習慣病対策の推進について(中間取りまとめ)」を踏まえた今後の生活習慣病対策のためのエビデンス構築に関する研究】総括・分担研究報告書, 2011 年 3 月, pp.60-65.

厚生労働科学研究費補助金（特別研究事業）
分担研究報告書

既存の研究成果報告書の効率的なインデックス化

研究分担者　奥村 貴史（国立保健医療科学院 研究情報センター）
研究協力者　谷田 和章（東京大学学際情報学府）

研究要旨

厚生労働科学研究成果データベースの検索機能強化に向けて、全文データの効率的な確保手段が求められている。そのために、本研究分担では、研究代表者に依頼することにより研究成果報告書を電子ファイルとして直接収集する手法について検討を行い、今後、研究成果報告書の電子化に求められるコストの推定を試みた。

1998 年度の研究代表者を対象としたファイル回収実験では、回収率は全体の 1.4%に留まり、実際の利用に足るファイルの回収率は 0.57%に過ぎなかった。また、回収されたファイルには意図したものと異なるファイルも多く含まれていた。今後、研究代表者からの電子ファイルによる研究成果報告書の回収率を上げていくためには、研究成果報告書の作成方法に関する調査を行いつつ、回収方法そのものを改善しながらファイル回収を進めていく必要性がある。この試みは、今後、研究成果報告書の提出自体を電子化していく上でも基礎的な資料となることが期待される。

研究成果データーベースの全文テキスト化に必要なコストとしては、今回のコストモデルを用いると、未だ全文テキスト化が行われていない 2003 年から 2009 年度の処理のために 9,000～13,000 万円が必要となることが明らかとなった。また、今後の研究成果報告書の電子化に要するコストは、各年度で 1,700 万円(回収率 10%)～990 万円(回収率 50%)程度と見込まれ、電子ファイルでの提出を義務化した際に生じうる費用と比して 1/10 程度に抑えられることが明らかとなった。

したがって、今後、電子ファイルでの提出率が 100%になるまでは、ファイル回収と OCR を効率的に組み合わせた戦略が合理的である。ただし、費用は電子ファイルの回収率と OCR 処理の単価に依存しており、今後、ファイル回収率を上げるとともに、OCR そのものの単価を下げる工夫により、さらにコストを抑えられる可能性がある。

A．研究目的

厚生労働科学研究研究成果データベース GRANTS には、現在、1998 年度から 2009 年度の研究成果情報が蓄積されている。これらの情報は、研究課題のタイトルや成果の要約に関する「課題データベース」と、それぞれの研究代表者に関する情報が含まれる「研究者データベース」に加えて、それぞれの研究報告書をスキャンした「研究報告書ファイル」に大別される。この研究計画書ファイルは、20 ページ毎に PDF

(Portable Document Format)化されファイルのコピーを禁じるなどのセキュリティ処理が施された状態で保存されているが、基本的にスキャンされた画像データであり、全文検索に用いることは出来ない。そのために、厚生労働科学研究成果データベースの検索機能強化に向けて、全文データの効率的な生成が求められていた。

そこで、今年度、本研究の「全文検索、知的検索に対応するインデックス化の研究」において、OCR(Optical Character Reader)処理によりテキスト情報を機械的に読み取ることを試みた。しかし、機械的な OCR そのものは低コストに行うことが可能だが、機械処理による文字読み取りには一定確率で誤読が生じてしまう。そこで、OCR 処理した文章から誤読を取り除く作業が必要となるが、現在の技術では人手による目視確認を欠かすことが出来ず、全文テキスト化の費用を押し上げる要因となっていた。たとえば、今年度の研究では、1998～2002 年度の 374,113 ページの全文テキスト化に 2,500 万円が掛かっており、これは現在データベースが保有している総ページ数の 16%に過ぎない。さらに、目視確認を行ったとしても、目視を人力で行う以上一定確率でのミスの混入は避けられず、100%の精度を期待することは出来ない。

こうした問題を回避するためには、紙媒体で提出された報告書を読み取るのではなく、研究代表者側に依頼して提出された報告書の元となった電子ファイルそのものを入手する方法がある。これにより、報告書の読み取り費用を抑え、また、文字読み取りに伴うミスの混入を避けられる可能性がある。一方で、研究成果データベースには 10 年以上の昔の研究も少なくなく、何割程度の研究班が電子的に報告書を取りまとめていたかは未知数である。さらに、研究者はキャリアのうえで所属を変えていくことが一般的であり、研究成果データベース上で把握している所属が変更されている場合、ファイル提出の依頼そのものに大きな困難が生じることになる。

そこで、本研究分担では、今後、過去の研究報告書を効率的に全文テキスト化するための基礎資料とするために、研究代表者への依頼よるファイル回収がどの程度現実的であるかを検証する。そのために、まず、研究成果データベースより過去に報告書を提出した研究代表者のメールアドレスのリストを復元した。また、インターネット上にファイル回収のためのシステムを構築し、研究者より効率的にファイル収集を行うための体勢を整えた。その上で、研究代表者に対して、報告書の元となった電子ファイルの提出を電子メールにて依頼した。その後、回収した報告書ファイルを分析し、今後の研究報告書の全文テキスト化を行う手段としてのファイル回収について考察を行った。

B．研究方法

B-1) 連絡先メールアドレスの復元

研究代表者への電子メール送付に際しては、まず、連絡先メールアドレスを用意する必要がある。しかしながら、GRANTS の研究成果報告書を含む「課題データベース」には、研究課題、研究年、研究者・組織、研究費など、研究課題にまつわる情報は含まれているものの、電子メールアドレスは含まれていない。研究者の固有の情報は、研究者データベース側のみに含まれており、たとえば、研究者名、研究者 ID、電子メールアドレス、性別、機関名、卒業大学、卒業年などの情報が含まれている。一般的なデータベースの構成では、研究者 ID によりこれらの情報がシステム上で「紐付け」られているが、GRANTS の過去データではこの紐付けが完全ではなかった。そこ

でまず、「課題データベース」に含まれる研究者を「研究者データベース」から探し出し、回収年度の対象となる研究代表者の電子メールアドレスを取得、整理する必要がある。

B-2) ファイル回収システムの構築

次に、ファイルの回収作業を行うファイルアップローダシステムを準備した。メールによるファイルの提出依頼からファイルの回収を自動化するために、依頼メールに「以下をクリックして、手持ちの PDF をアップロードして下さい」といった形式でアップローダへのリンクを含め、そのリンクに研究課題 ID を埋め込むことで、研究成果報告書の電子ファイルを効率的に管理できるよう配慮した。また、収集したファイルの管理を円滑化するために、web による管理システムを用意した。

B-3) ファイル提出依頼

ファイルの提出依頼そのものについても検討が求められた。たとえば、提出されるファイルの形式を整えるために、PDF のバージョンを指定すると、新しいソフトウェアを用いている研究者にも、古いソフトウェアを用いている研究者にも不便が掛かることになる。とりわけ今年度の回収実験においては、10 年以上昔の研究報告書に対する PDF 提出を依頼することになるため、古いバージョンから新しいバージョンまでさまざまな形式でのファイルが集まる懸念がある。一方で、旧バージョンのファイルに対しては機械的な変換も可能であることから、今回は、特にバージョン等の細かな形式を指定せず、報告書の電子ファイルを指定した URL よりアップロードするよう依頼する説明文を作成し、電子メールにて依頼状を一斉送信した。

C. 研究結果

GRANTS システム上には、1998 年度から 2003 年度の研究課題として 8,078 件が登録されていた。また、email アドレスを含む研究者データとして、13,365 件が登録されていた。今回の試みでは、まず、課題データベースに含まれる研究課題からそれぞれの課題における研究代表者のメールアドレスを抽出した。参考資料 III-4 に示す操作の結果、8,078 件中、姓名の一致が見られたのが 5,624 件(69.6%)あり、さらに、組織名、ならびにインターネットに公開されている研究者情報を用いたマッチングの結果、4,458 件(55.2%)の研究代表者メールアドレスを対応付けることができた。また、姓名は一致したものの、組織名や研究者情報での確認が取れなかった研究者が 1,166 件(14.4%)存在した。

次に、復元したメールアドレスから 1998 年度の登録済み報告書 697 件より、復元することが出来た 493 件のメールアドレスを利用して、研究代表者宛にファイル提出を依頼した(参考資料 III-5)。その結果、ファイルのアップロード依頼に対して、10 研究課題、24 件のファイルを回収することができた。これは、697 件のうち 1.4%に相当する。回収したデータの一覧を表 1 に示す。

Case 1 は、提出されていた報告書と回収したファイルの内容が完全に一致していた。提出されたファイルは、全文テキストが透明化されているもので、理想的な状態であった。Case 2 は、提出されていた紙媒体の報告書自体が 5 ページと短いもので、提出されていた報告書と回収したファイルとが完全に一致した。研究代表者は、今回、一太郎形式と PDF 形式をアップロードし、PDF ファイルには全文テキストが含まれていた。Case 3 は、既提出の 85 ページの報告書に対して、対応する全文テキストが

Case	判定	GRANTS		提出ファイル				備考
		ページ	サイズ	種別	内容	ページ	サイズ	
1	○	23	875K	PDF	研究報告書	23	1964K	
2	○	5	175K	PDF	総括研究報告書	5	216K	
				JTD			75K	一太郎ファイル
3	△	85	5115K	PDF	分担研究報告書	60	838K	分担研究報告
				PDF	添付資料	12	9987K	
				PDF	添付資料	12	9987K	重複アップロード
				PDF	分担研究報告書	6	288K	分担研究報告
				PDF	分担研究報告書	6	348K	分担研究報告
				PDF	総括研究報告書	4	260K	総括研究報告書
4		10	227K	PDF	総括研究報告書	3	139K	提出ファイルが総括研究報告書のみで、分担報告書が欠落。提出ファイルは、研究報告書とレイアウトも異なっており、厚労省書式でない。
				PDF	総括研究報告書	3	139K	
5		235	6238K	PDF	総括研究報告書	8	356K	フォントが筆文字。
				DOC	総括研究報告書	8	37K	
				PDF	総合研究報告書	9	388K	
				DOC	総合研究報告書	8	37K	
				DOC	抄録	4	26K	
				PDF	抄録	4	276K	
6	×	24	1020K	PDF	総括研究報告書	7	543K	6ページはスキャンPDFで、最終ページに検索用に透明テキスト化PDFが添付されている。
7	×	30	1247K	PDF	研究報告書	30	2785K	提出報告書と完全に同一なスキャンファイル。
8	△	-	-	PDF	総括分担研究報告書	30	442K	フォントが筆文字。
				DOCX	総括分担研究報告書	30	69K	
9		-	-	PDF	研究報告書	3	135K	
				DOC	研究報告書	3	8K	
10		-	-	DOC	研究報告書	41	261K	提出ファイルは、厚労省提出書式でなく、他の研究報告書と大きく異なる。

表 1 回収ファイル一覧

含まれる PDF が分割してアップロードされた。Case 4、5 は、提出されていた研究報告書のうち、総括研究報告書のみが PDF ないし Doc(Microsoft Word)形式でアップロードされた。したがって、データとして不完全であり、このまま GRANTS システムに登録することが困難なファイルであった。Case 6、7 は、提出された報告書をスキャンした PDF ファイルの提出であった。これらもまた、GRANTS システムに登録することが困難なファイルである。Case 8、9、10 は、GRANTS システムに報告書が登録されていない研究課題であった。GRANTS システムの報告書データは、国会図書館や厚生労働省図書館への提出後に残部を国立保健医療科学院に提出して頂く形となっており、Case 8、9、10 の場合は、未提出の研究者が今回の依頼にご協力頂いた形になる。これらのうち、Case 8 は利用可能な形式のファイルであったが、Case 9 は、3 ページと極めて短いうえ総括研究報告書の記載がなく、Case 10 は推奨されている報告書レイアウトと極めて異なる報告書であった。

結果的に、GRANTS に収録しうるデータとして、4 課題の報告書ファイルの回収を行うことが出来た。これは、1998 年度の登録済み報告書 697 件に対して 0.57%を占めるに留まり、極めて少ない回収率であった。

D. 考察

D-1) ファイル回収率の低値と改善策