

国立国会図書館 御中

全文テキスト化実証実験に係る  
調査及び評価支援等作業  
実証実験報告書

平成 23 年 3 月

 株式会社 三菱総合研究所

# 全文テキスト化実証実験報告書

平成 23 年 3 月

# 目次

1	はじめに.....	1
1. 1	背景・目的.....	1
1. 2	概要.....	1
1. 3	実施体制・スケジュール.....	3
2	実証実験の実施.....	4
2. 1	実証実験における評価事項.....	4
2. 2	実証実験のために構築したシステムプロトタイプ.....	7
2. 3	実証実験の対象書籍.....	21
3	テキストデータ作成に関する実証実験に関する評価.....	23
3. 1	テキスト化システムの構築の評価.....	23
3. 1. 1	日本語対応の OCR 出力フォーマットの評価.....	23
3. 1. 2	構造化メタデータの評価.....	25
3. 1. 3	OCR の再学習機能の評価.....	26
3. 1. 4	出力フォーマットの評価.....	28
3. 1. 5	出版社データから OCR 出力フォーマットへの変換の評価.....	31
3. 2	テキスト化システムを用いた作業の効率化、高度化の評価.....	33
3. 2. 1	レイアウト校正作業の効率化、高度化の評価.....	33
3. 2. 2	共同校正機能による効率化、高度化の評価.....	34
3. 2. 3	OCR の再学習機能による効率化、高度化の評価.....	37
3. 2. 4	共同構造化機能による効率化、高度化の評価.....	38
3. 2. 5	構造情報推論機能による効率化、高度化の評価.....	42
3. 2. 6	読上げ順序編集機能による効率化、高度化の評価.....	47
3. 3	テキストデータ作成にかかる作業時間の評価.....	49
4	全文テキストデータの検索・表示に関する実証実験に関する評価.....	51
4. 1	検索画面における機能の評価.....	51
4. 1. 1	キーワード検索（構造指定検索機能・難易度検索機能）の評価.....	52
4. 1. 2	自然文検索の評価.....	52
4. 1. 3	サジェスチョンの評価.....	53
4. 2	検索結果一覧画面における機能の評価.....	55
4. 2. 1	ランキングの評価.....	55
4. 2. 2	スニペットの評価.....	56
4. 2. 3	連想検索の評価.....	58
4. 3	書誌詳細表示画面の評価.....	60

4. 3. 1	目次の評価 .....	60
4. 3. 2	文脈検索の評価 .....	61
4. 3. 3	固有名表示の評価 .....	62
4. 3. 4	タグクラウドの評価 .....	63
4. 4	本文表示画面の評価 .....	64
4. 4. 1	書籍本文のテキスト表示の評価 .....	64
4. 4. 2	目次・本文リンクの評価 .....	65
4. 4. 3	検索語出現数表示の評価 .....	66
4. 4. 4	検索語ハイライトの評価 .....	67
4. 5	ページ構成の評価 .....	69
4. 6	視覚障がい者等向けの読上げサービス等の評価 .....	71
4. 6. 1	全文テキストデータの読上げの評価 .....	71
4. 6. 2	全文テキストデータの品質の評価 .....	71
4. 6. 3	OCR 認識率が読上げサービスに与える影響の評価 .....	73
4. 6. 4	DAISY ファイルの読上げの評価 .....	75
4. 7	全文テキストデータのインデキシング処理時間の評価 .....	76
4. 8	文字コード対応の評価 .....	77
5	実証実験の成果と課題 .....	79
5. 1	テキストデータ作成に関する実証実験の成果と課題 .....	79
5. 2	全文テキストデータの検索・表示に関する実証実験の成果と課題 .....	83

## 1 はじめに

本章では、全文テキスト化実証実験の背景と目的、概要、実施体制およびスケジュールを示す。

### 1. 1 背景・目的

平成 22 年 1 月に施行された改正著作権法により、国立国会図書館を含む図書館等において、視覚障がい者等のための著作物の複製および自動公衆送信を、著作権者の許諾なしに行えるようになった。これに対応するため、国立国会図書館は、所蔵資料のデジタル化画像データから全文テキストデータを作成し、読上げソフトウェア等を用いて提供することを基本方針とした。

一方、総務省、文部科学省、経済産業省が「デジタル・ネットワーク社会における出版物の利活用の推進に関する懇談会」を平成 22 年 3 月に設置し、同 6 月に報告書を取りまとめた<sup>1</sup>。この報告書では、「全文テキスト検索については、様々な課題が存在するため、国立国会図書館と出版物のつくり手等との連携による実証実験等を通じて、課題解決について検討を進めることが適当である」とされている。

これらの状況を受けて、国立国会図書館は、全文テキストデータの作成、および検索・表示に関する技術的課題の評価を目的とした、全文テキスト化実証実験を実施した。

### 1. 2 概要

本実証実験では、所蔵資料のデジタル化画像データ等から、OCR<sup>2</sup>で文字を読み取り、結果を校正し、目次・見出し文字などの情報から構造化<sup>3</sup>を行う作業を支援する「全文テキスト化システムプロトタイプ」（以下、テキスト化システム）と、全文テキストデータを蓄積し、目的の書籍を検索し、結果を表示する「全文検索・表示システムプロトタイプ」（以下、検索表示システム）を構築した。

これらのシステムプロトタイプを用いて、テキストデータ作成に関する実証実験と、全文テキストデータの検索・表示に関する実証実験を実施した。各実証実験の概要は以下のとおりある。

---

<sup>1</sup> 「デジタル・ネットワーク社会における出版物の利活用の推進に関する懇談会報告」(2010年 6 月 28 日デジタル・ネットワーク社会における出版物の利活用の推進に関する懇談会) [http://www.soumu.go.jp/main\\_content/000075191.pdf](http://www.soumu.go.jp/main_content/000075191.pdf)

<sup>2</sup> Optical Character Reader の略。光学式文字読取装置。手書き文字や印字された文字を光学的に読み取り、前もって記憶されたパターンとの照合により文字を特定し、文字データを入力する装置のこと。

<sup>3</sup> 標題紙・目次・本文・索引・奥付、本文の章・節・項など、書籍が持つ構造をコンピュータが認識できるようにテキストにタグ付け等の処理を行うこと。

#### ○テキストデータ作成に関する実証実験

全文テキストデータの整備を進めるために解決すべき技術的課題を、プロトタイプを構築して検証した。具体的には、校正作業の結果を OCR の文字認識に反映させるために必要な OCR エンジンの再学習機能の可能性、OCR 出力フォーマットの在り方、校正・構造化作業のシステムによる共同作業化（複数人による作業の共同化）の可能性・有効性、OCR により読み取ったデータの構造化作業の自動化による効率化の可能性・有効性、また、構造化するために付与するメタデータや読上げサービスに必要な読上げ順序情報の付与の在り方等について、検証した。

#### ○全文テキストデータの検索・表示に関する実証実験

従来の図書館システムが検索対象としてきた書誌情報に加えて、本文の全文テキストを検索対象にすることによる検索サービスの高度化について、情報の探しやすさ（サーチャビリティ）の向上の観点から、プロトタイプを構築して検証した。具体的には、構造化された全文テキストデータとその書誌情報を用いて、検索結果一覧表示における全文テキストデータの活用の可能性、ナビゲーション・リコメンドなど高度な検索機能への全文テキストデータの活用の可能性、全文テキストを検索対象とする場合の検索結果のランキングの在り方、本文の表示方法（ハイライト・目次と本文のリンク）の在り方等について、検証を行った。また、構造化された全文テキストを用いた視覚障がい者等向け読上げサービス等の有効性および高度化の可能性（アクセシビリティ）の検証を行った。

全文テキスト化するデータとしては、国立国会図書館の所蔵資料に加えて、出版社より書籍の電子データの提供を受けた。

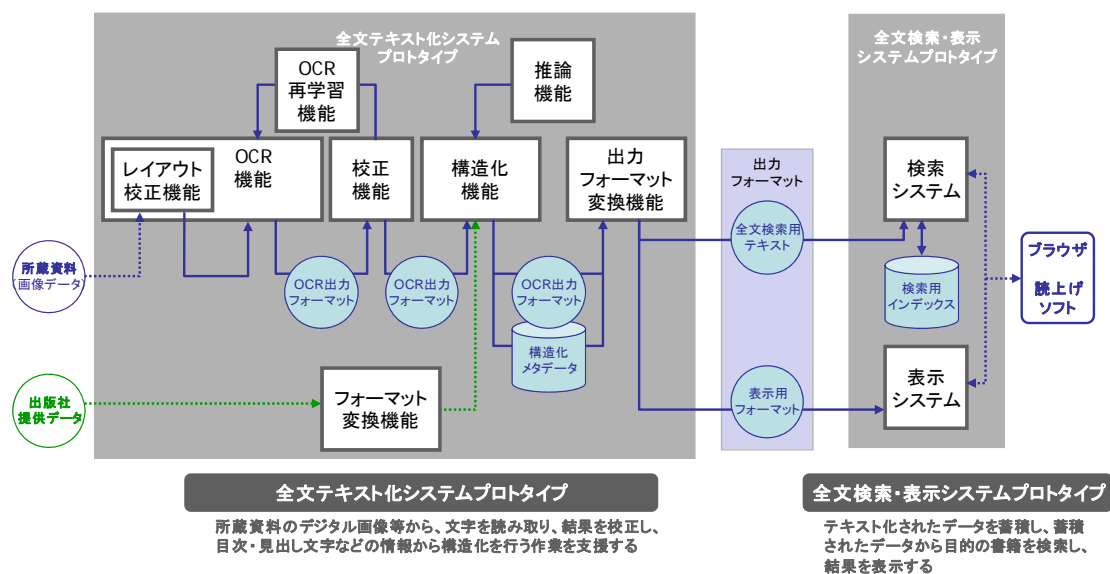


図 1-1 全文テキスト化実証実験の全体像

### 1. 3 実施体制・スケジュール

本実証実験の実施にあたり、全文テキスト化システムプロトタイプの構築は日本アイ・ビー・エム株式会社が、全文検索・表示システムプロトタイプの構築は株式会社日立製作所がそれぞれ担当した。また、実証実験の実施に関連した調査、および各システムプロトタイプを用いて行う実証実験の評価・とりまとめ支援は株式会社三菱総合研究所が担当した。なお、評価に際しては、有識者の助言を得て実施した。

本実証実験のスケジュールは以下のとおりである。

- ・平成 22 年 10 月 ～ 平成 23 年 2 月： プロトタイプ構築
- ・平成 23 年 2 月 ～ 平成 23 年 3 月： プロトタイプを用いた試行利用と評価