全文テキスト化実証実験参加協力会社との定例会（第２回）

日時：平成２２年１２月１３日（月）１０時 ～１１時
場所：国立国会図書館新館３階研修室

議事次第

１　報告
　○ 全文テキスト化実証実験の進捗について

　○ 全文テキスト化実証実験の対象資料について

　○ 全文テキスト化実証実験の評価の実施方法及び項目について

２　質疑応答

３　閉会

配布資料

- 資料１　　全文テキスト化実証実験スケジュール
- 資料２　　全文テキスト化実証実験の対象資料について
- 資料３　　全文テキスト化実証実験の評価の実施方法及び項目について

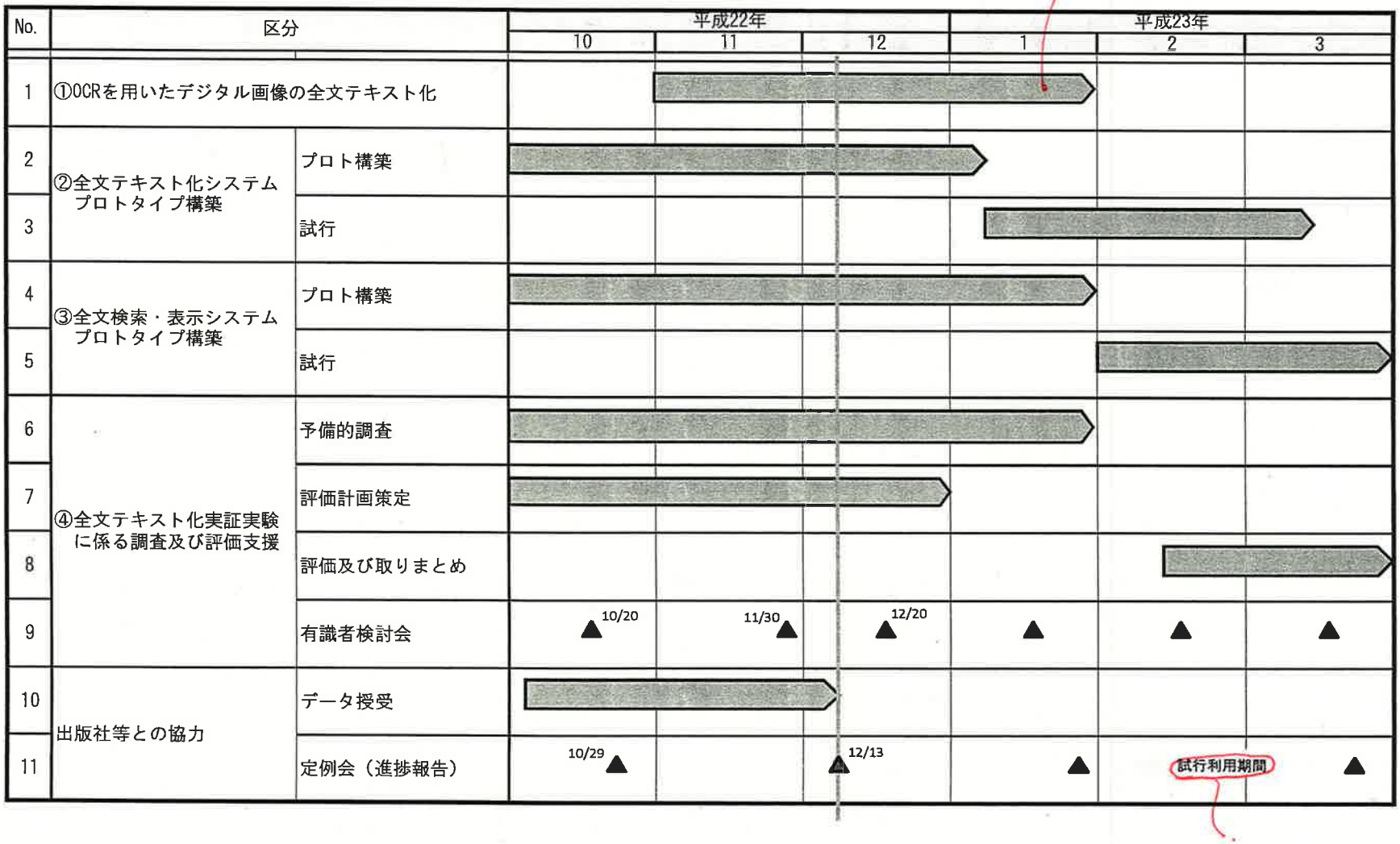
- 参考２－１　全文テキスト化システムプロトタイプ全体構成図
- 参考２－２　全文検索・表示システムプロトタイプ画面イメージ
- 参考３　　評価項目一覧表（案）

問い合わせ先
国立国会図書館総務部企画課
全文テキスト化実証実験担当
TEL：03-3506-5297（直通）
E-mail：digi-jimu@ndl.go.jp

・API
・.indd.
・転送形態

資料1

全文テキスト化実証実験スケジュール



資料2

平成 22 年 12 月 13 日
国 立 国 会 図 書 館

全文テキスト検索実証実験の対象資料について

1 対象資料

区分	対象	数量	作業	成果物
① OCR を用いた画像の全文テキスト化	デジタル化画像データ	2 万冊	OCR 処理	テキスト PDF
② 全文テキスト化システム	デジタル化画像データ	40 冊※1	OCR＋校正＋構造化	PDF XHTML
	出版社提供データ	600 冊※2	構造化	XHTML
③ 全文検索・表示システム	①＋②の成果物	2 万 640 冊	全文検索	(XHTML)
			表示	(テキスト) (PDF)

※1：約 4 冊は全ページ作業対象とし、残りは構造種別（図表、段組み、ルビ等）に応じて、一部分を作業対象とする
※2：約 1 冊は全ページ作業対象とし、残りは自動構造化機能により作業を実施する

2 画像データの概要

時代	形式	数量 ①2 万冊 内訳	数量 ②40 冊 内訳	対象資料例	特徴
明治～昭和戦前	2 値／モノクロ	19,790 冊	30 冊	(1) オセロー（坪内逍遥訳） (2) 初等代数学（千本福隆編） (3) 四書白文 (4) 和仏法律学校講義録　雑報 (5) 飲水要論（石塚左玄著） (6) エスペラント語速成教書 (7) 政治学研究（C. ビーアド著） (8) 岩崎弥太郎（山路愛山著）	(1) 縦中横 (2) 数式 (3) 漢文 (4) 割注 (5) 漢字と仮名サイズが混在 (6) 左横書き、日英字混在 (7) 傍点 (8) ルビ
昭和戦後	カラー	210 冊	10 冊	(1) 戦国人名事典（阿部猛等編） (2) マルクス=エンゲルス全集	(1) 多段組み (2) 索引

※　ただし、数量、内訳数及び対象資料例については、変更の可能性もあり

3 出版社提供データの概要

出版社提供データ（21 社）の概要は次のとおりです（平成 22 年 12 月 13 日時点）。

区分	タイトル	冊数	ジャンル（冊数） 一般	ジャンル（冊数） 学術	出版年（冊数） 2000 年以降	出版年（冊数） 1999 年以前	フォーマット（冊数） PDF	TXT	.book	XMDF	他
図書	249	316	301	15	192	124	138	94	20	60	4
雑誌	6	33	33	0	33	0	22	0	0	0	11
総計	255	349	334	15	225	124	160	94	20	60	15

※ .book は TTX 含む。また、XMDF は素材データ付き。
※ フォーマットの「他」の内訳は、indd 3 冊、XML 12 冊

資料3

平成22年12月13日
国立国会図書館

全文テキスト化実証実験の評価の実施方法及び項目について

1 評価の概要

(ア) 評価の目的

国立国会図書館が本年度実施する全文テキスト化実証実験において構築する全文テキスト化システムプロトタイプ及び全文検索・表示システムプロトタイプを用いて、技術的課題の検証・評価を実施する。

(イ) 評価の方法

プロトタイプシステムの設計情報に基づく技術評価、定量データに基づく評価、プロトタイプシステムを利用して、使用性などに関する定性的評価を行う。

(ウ) 評価の項目

全文テキスト化、校正・構造化、検索、表示等について、(イ)評価の方法で示した内容の評価を行う。

(参考3:「評価項目一覧表(案)」参照)

2 参加協力会社への依頼事項

(ア) プロトタイプシステムの試用及び評価の実施

現在開発中のプロトタイプシステムの試行利用及び評価の実施

(イ) 評価の内容

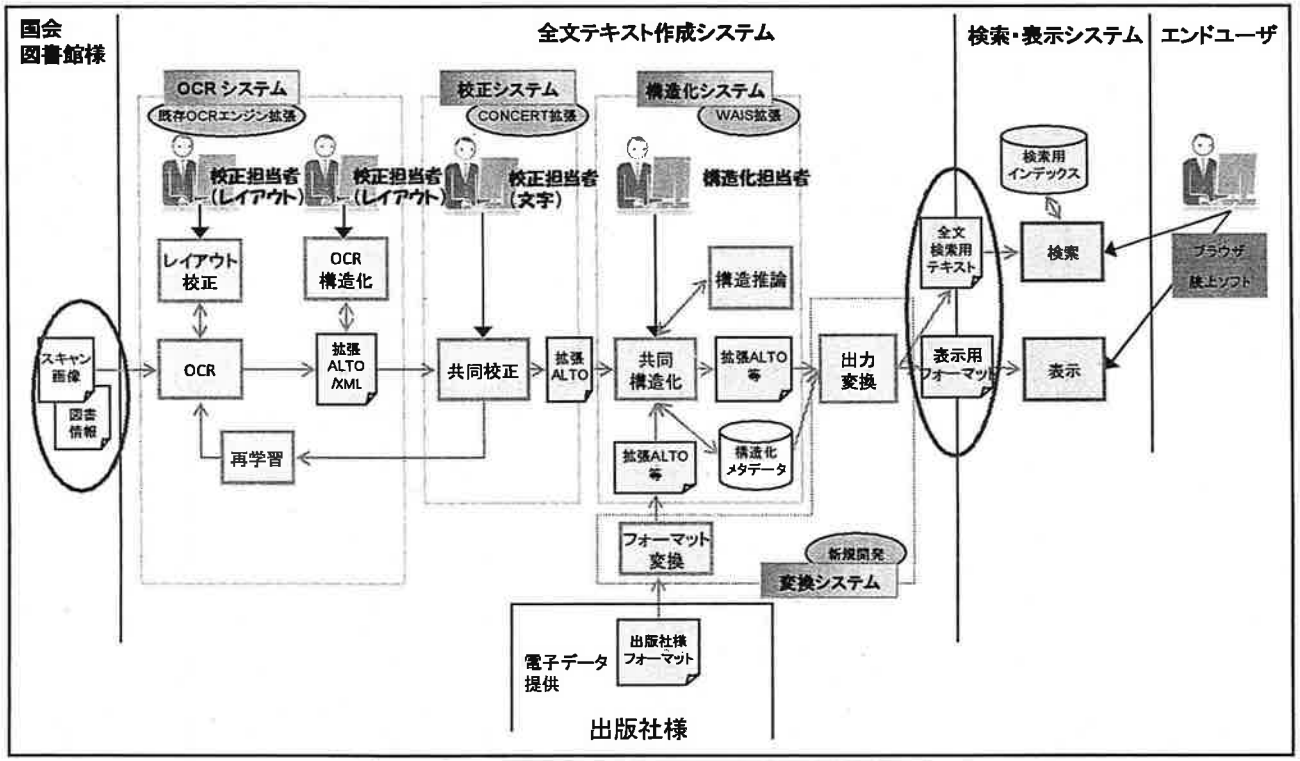
- サーチャビリティ(情報の探しやすさ)
  検索機能の有効性の検証
  書誌に加え、全文テキストを検索対象とすることの検証
- 検索結果の表示の有効性検証
  テキストの表示(Web ブラウザ)
  イメージ形式の表示
- API機能の利用

3 試行・評価の実施について

- 日 程 平成23年2月(数日間に渡り実施の予定です。)
- 会 場 国立国会図書館東京本館内(東京都千代田区永田町1−10−1)

1. システム概要
他のプロトタイプとのインタフェース

- 当プロトタイプにおける外部システムとのインターフェースは以下の2つ（下図赤丸部分）となることを想定している。
  - （1）国立国会図書館及び出版社から提供されるデータ（画像ファイル）
    - CD/DVD 等の媒体又はHDD等の補助記憶装置で受領し当プロトタイプに投入
  - （2）全文検索・表示システムプロトタイプに提供されるデータ（全文検索用テキスト、電子書籍フォーマット）
    - CD/DVD等の媒体又はHDD等の補助記憶装置での授受、若しくはメール送付で提供



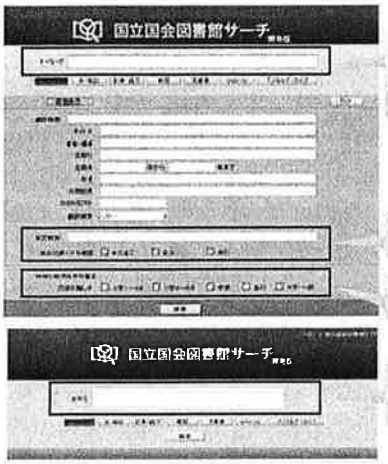
当プロトタイプの全体構成図

© 2010 IBM Corporation

（参考資料２－２）　　【取扱注意】

2-2　検索画面

HITACHI Inspire the Next



サジェスチョン
検索式の入力時に検索語を補完

構造指定検索
全文テキスト中の検索する範囲を指定

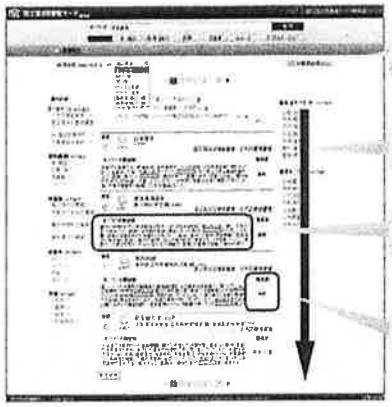
難易度検索
書籍の難易度の選択が可能

自然文検索
任意の文章を検索条件として検索を行い、文章の内容が似ている書籍を取得

7

2-3　検索結果一覧画面

HITACHI Inspire the Next



もしかして検索　もしかして：図書館
ヒット件数が0件の場合、検索語の綴りの誤りをチェックして、正しい検索語を表示

ランキング
書誌と全文のデータ量を考慮してスコア付けし、検索語に関連の強い順に資料を表示

スニペット
本文中の検索語前後の文章を表示

難易度表示
本テキストから本文の難易度を判定した結果を表示

8

2-4　検索結果一覧画面

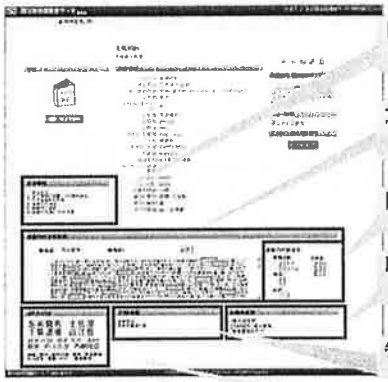
HITACHI Inspire the Next

連想検索
資料を選択して似ている資料を検索

9

2-5　書誌詳細画面

HITACHI Inspire the Next



目次表示
全文テキストまたは情報探索プロトタイプのメタデータから抽出して表示

文脈検索
任意の検索語を入力して、どのような文脈で検索語が使われているか表示

固有名表示
本文によく登場する人物名や地名を表示

内容ベースのレコメンド
参照した資料の書誌IDを元に連想検索エンジンを使用して類似資料を検索し、お薦め資料を表示

参考文献リンク
本文中に記載されている文献を書誌情報として表示

タグクラウド
本文中の特徴的なワードを表示

10

2-6　本文表示画面

HITACHI Inspire the Next

目次・本文リンク
目次から、該当する本文箇所へ表示を移動

検索語可視化
検索語の出現回数を可視化

本文検索
本文中を任意のワードで検索

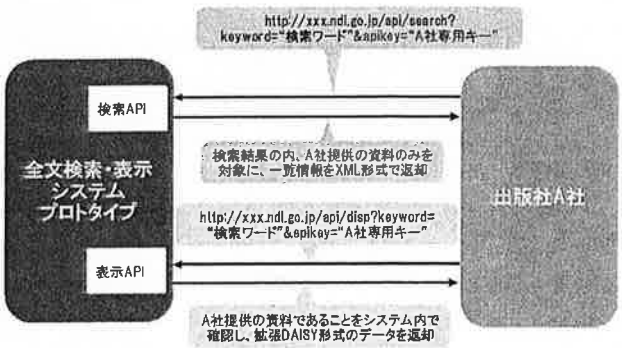
ハイライト表示
検索語をハイライト表示

11

2-7　出版社向け検索・表示API

HITACHI Inspire the Next

全文検索・表示システムプロトタイプで保有する出版社提供データを検索し、取得するAPIを出版社向けに提供します。



12

評価項目一覧表（案）

参考3

実施項目	評価の目的	具体的な評価指標の例
A. テキストデータ作成に関する実証実験		
(1)標準フォーマットによる全文テキスト化システムの構築		
a. 日本語対応のOCR出力標準フォーマットの検討・定義・検証		
日本語対応のOCR出力標準フォーマット	日本語対応のOCR出力標準フォーマットについて、以下の観点から検証・確認を行う。 ・汎用性の確保 ・旧字・新字、外字の扱い、文字コード等への配慮 ・日本語表示固有の表現（図表、ルビ、縦書き等）、数式等の特殊表現の配慮	以下の項目に対して、定性的に5段階で評価。 ・将来的にオープンな標準化団体に提出可能な程度の質的水準を確保しているか ・ALTOなど海外のデファクト標準のフォーマットとの互換性にも十分に配慮しているか ・旧字・新字、外字の扱いを十分に考慮しているか ・日本語表示固有の表現への配慮が行われているか
b. 全文テキストデータを構造化するために付与するメタデータの検討・定義・検証		
構造化のためのメタデータ	全文テキストを構造化するメタデータの構造と利用方法について、検証・確認を行う。	以下の項目に対して、定性的に5段階で評価。 ・メタデータ化による構造情報保持方式の妥当性 ・メタデータの構造と利用方法
c. OCR、共同校正機能、共同構造化機能の連携システムの構築、将来の実用化時の要件検討		
OCR学習用辞書の整備	・OCR学習用辞書について検証・確認を行う。	・識別率向上のために適切な、文字認識用辞書が作成されているか。 ・作成された辞書は、他のOCR製品の辞書との互換性があるか。
d. 中間フォーマットから全文テキスト化された書籍を、利用者に見やすい形で表示するための当館指定のフォーマットへの出力システムの構築、将来の実用化時の要件検討		
資料を構造化する上での分類方法（標題紙、目次等）	資料の要素分解に関わる機能について、検証・確認を行う。	標題紙、目次など資料の構成要素を適切に保持する方法を確認する。
レイアウト認識	OCRの読み取り精度を高めるために必要となるレイアウト認識について、検証・確認を行う。	資料の要素分解にレイアウト情報を利用することで、しみや変色、ごみ等のノイズの影響の除去を含め、テキストの識別率が向上することを確認する。
本文補足文字除去	OCR結果の利用で必要となる本文補足文字除去技術について、検証・確認を行う。	本文補足文字の属性（ルビ、ノンブル、欄外記載文字、註への参照番号、等）ごとの認識と取り扱いに関する技術的検討結果
e. 出版社から提供される電子書籍データや版下データを中間フォーマットに変換するシステムの構築、将来の実用化時の要件検討		
中間フォーマットへの変換	出版社から提供されるデータの中間フォーマットへの変換技術について、検証・確認を行う。	中間ファイルフォーマットとして、どれだけ汎用性・網羅性があるかを定性的に5段階で評価

実施項目	評価の目的	具体的な評価指標の例
(2)共同作業支援システム（共同校正機能及び共同構造化機能）によるコスト削減等の効果の検証		
a. 共同校正機能による校正作業の効率化、高度化の検証		
校正	共同校正機能による校正作業の効率化、高度化の検証、ツールの使いやすさ、館外体制の共同作業について、検証・確認を行う。	システムが出力する電子ファイルに対して、どれだけ修正が必要かを、単位ページあたりの修正率（全文字数に対する修正文字の割合）で評価 各ページごとに ①全文字数 ②修正した文字数 なお、対象文書の特性（原資料の作成時期、テキスト領域のレイアウト、枠線の有無、ルビの有無、活字の形状、活字と背景文書のコントラスト、等）についても必要に応じて考慮する。
		校正作業を実施し、その簡便性を定性的に5段階で評価 定性的な指標を構成する要素の例 ・校正作業画面のレイアウト、文字サイズ ・校正対象文書の閲覧容易性 ・文章修正の容易性 ・修正対象文字に関する修正候補の表示、等
		単位ページあたりの構成に要する作業時間を測定 各ページごとに ①手作業での修正作業に要した時間
		・館外体制(障がい者団体等)による共同校正作業は可能なシステムとなっているか。 ・館外体制を有効とする工夫がなされているか。
b. 共同構造化機能による構造化作業の効率化、高度化の検証		
テキストの構造化作業	テキストの構造化作業の効率化と高度化について、検証・確認を行う。	区切り位置（見出し位置、章、ページ、段落、行、等）の挿入作業を実施し、その簡便性を定性的に5段階で評価 簡便性を表す指標を構成する要素の例 ・区切り位置確認・修正画面のレイアウト（それぞれの区切り位置について） ・修正対象文書の閲覧容易性（それぞれの区切り位置について） ・区切り位置修正作業の容易性（それぞれの区切り位置について）、等
		構造化項目（段組み・多段構成、コンテンツ区切り（標題／目次／本文／索引）、ページ番号、柱、等）に関する取り扱いに関する技術的検討結果
		単位ページあたりの区切り位置挿入に要する作業時間を測定 元テキストでの１ページあたりのページ・段落・行区切りについて、修正を行うのに要した時間

実施項目	評価の目的	具体的な評価指標の例
(3)半自動校正システム、半自動構造化システムによるコスト削減等の効果の検証		
a. OCRの再学習システムの構築と、それによる校正作業の効率化、高度化の検証		
「(2) 共同作業支援システム（共同校正機能及び共同構造化機能）によるコスト削減等の効果の検証」におけるOCRの再学習システムの有効性	OCR再学習機能の有効性について、検証・確認を行う。	継続的に全文テキスト化を実施する中で、都度識字率を算出し、その推移を評価 全文テキスト化作業実施日ごとに ①全文字数 ②誤認識した文字数
		OCRの継続的／網羅的なかけなおしにおいて以下を考慮してコスト削減を図れるか評価。 ・複数作業者間で学習機能に反映すること ・異なるソフト間(バージョン間)で辞書を共有していくこと ・以上の時間コストへの影響
b. 構造情報を半自動的に付加する機能と、それによる構造化作業の効率化、高度化の検証		
「(2) 共同作業支援システム（共同校正機能及び共同構造化機能）によるコスト削減等の効果の検証」における、構造推論機能の有効性	構造推論機能の有効性について、検証・確認を行う。	システムが提示した構造情報の正しさを、文書あたりの採用率（提示した構造情報のうち、作業者が採用した割合）で評価 なお、対象文書の特性（原資料の作成時期、テキスト領域のレイアウト、枠線の有無、ルビの有無、活字の形状、活字と背景文書のコントラスト、等）についても必要に応じて考慮する。
目次、索引などの構造化への利用の検討	目次、索引などによる構造化の効率化と高度化について、検証・確認を行う。	目次、索引などによる構造化への利用法を聞き取り、将来的な可能性を検証する。
c. 読み上げ順序の視覚化システムの構築と、構造化作業の効率化、高度化の検証		
「読み上げ順序の視覚化システム」の構造化作業の効率化、高度化への寄与を評価	「読み上げ順序の視覚化システム」の構造化作業の効率化、高度化について、検証・確認を行う。	システムが出力した「読み上げ順序」に対して、どれだけ修正が必要かを、単位ページあたりの修正の有無で評価 なお、対象文書の特性（原資料の作成時期、テキスト領域のレイアウト、枠線の有無、ルビの有無、活字の形状、活字と背景文書のコントラスト、等）についても必要に応じて考慮する。
		・「読み上げ順序の視覚化システム」の機能 ・構造化作業の効率化、高度化への寄与
全文テキスト化全体の所要時間、処理容易性、対象文献種類ごとの難易度とコスト	全文テキスト化全体に必要な技術、文献種類による難易度について、総合的に検証・確認を行う。	全文テキスト化に関する一連のプロセスを担当者が実施し、その簡便性を定性的に5段階で評価
		単位ページあたりの全文テキスト化作業の所要時間を測定
		プロトタイプの操作結果から得られる、資料の特性ごとの全文テキスト化の難易度に対する5段階評価、および操作結果から試算される全文テキスト化に要するコスト 【パラメータ】 ・OCR対象の資料の作成時期、使用文字種、レイアウト

実施項目	評価の目的	具体的な評価指標の例
B. テキストデータ検索・表示に関する実証実験		
(1) 視覚障がい者等向けの読上げサービス等に関する課題改善（アクセシビリティの向上）の検証		
a. 視覚障がい者等向けの読上げサービス等のアクセシビリティの検証		
視覚障がい者向けの読上げサービス	視覚障がい者向けの読上げサービスの妥当性の検証・確認を行う。	読み上げサービスを実施し、その利便性や読み上げの妥当性について定性的に5段階で評価 定性的な指標を構成する要素の例 ・読み上げサービス起動ボタン等画面上のレイアウト ・ショートカットキーへの割り当て可能性 ・音量調整ボタン等の操作性 ・読み上げソフトの発声、等
ビューアのダウンロード	ビューアのダウンロードの妥当性の確認の検証・確認を行う。	ビューアのダウンロードの操作容易性、処理時間
全文テキストデータの閲覧	全文テキストデータの閲覧の妥当性の検証・確認を行う。	全文テキストデータの閲覧の操作容易性、処理時間 【パラメータ】 非ダウンロード方式、ダウンロード方式
b. 視覚障がい者等向けの読上げサービス等のためのフォーマットへの適切な変換の検証		
視覚障がい者向けの読上げサービス（再掲） にて適切な変換がなされているかの検証	読み上げサービスのためのOCRの妥当性の検証・確認を行う。	読み上げサービスを実施し、読み上げの精度について定性的に5段階で評価して、文字認識率のレベルなどが与える影響を評価
(2) 全文テキストデータの検索・表示に関する課題改善（サーチャビリティの向上）の検証		
a. 検索結果一覧表示における次の全文テキストデータの活用方法の検討		
検索結果一覧表示における全文テキストの活用		
スニペット	スニペット表示技術の妥当性、有効性の検証・確認を行う。	スニペットの作成方法に関するロジックについて、その汎用性を定性的に5段階で評価 定性的な指標を構成する要素の例 ・スニペットで表示される内容の妥当性、わかりやすさ ・スニペットのレイアウト(表示位置、行数など)
ページ構成の妥当性	検索結果一覧表示におけるページ構成の妥当性、有効性の検証・確認を行う。	検索結果一覧表示におけるページ構成の妥当性、有効性を定性的に5段階で評価 定性的な指標を構成する要素の例 ・1ページ当たりの件数の妥当性、有効性 ・結果が複数ページなる場合の全数の表示や、前ページ、次ページへの移動しやすさ
文字コード	検索において特殊な扱いが必要とされる文字コードの取り扱いの妥当性の検証・確認を行う。	新旧いずれの字体でも検索可能であるか、またJIS漢字コード非対応文字や外字に対して提供される検索方法の利便性や操作性について定性的に5段階で評価 定性的な指標を構成する要素の例 ・旧字体テキストに関する検索容易性 ・検索文字列の指定方法の柔軟性、等
目次	・目次表示の妥当性と有効性の検証・確認を行う。	目次情報の表示が、どれだけ適切なものかを定性的に5段階で評価 定性的な指標を構成する要素の例 ・目次情報の表示レイアウト ・目次情報を表示することでの検索結果に対する付加価値性向上、等

実施項目	評価の目的	具体的な評価指標の例
索引	・検索時の索引情報の活用可能性の検証・確認を行う。	索引情報の利用が、どれだけ適切なものかを定性的に5段階で評価 定性的な指標を構成する要素の例 ・索引情報の表示レイアウト ・索引情報を表示することでの検索結果に対する付加価値性向上 ・索引情報が提供されないことによるデメリット、等
ハイライト表示	ハイライト表示の妥当性、有効性の検証・確認を行う。	検索結果に対して表示される検索語のハイライト表示の見やすさについて定性的に5段階で評価 定性的な指標を構成する要素の例 ・ハイライト表示部分の視認性 ・ハイライト表示部分の妥当性 ・ハイライト表示部分の十分性（漏れがないか）、等
総合的な使いやすさ	検索結果一覧表示の総合的な使いやすさの検証・確認を行う。	検索機能の操作性や利便性について定性的に5段階で評価 ・検索画面のレイアウト ・検索キーの指定方法 ・検索キー以外の検索用付加情報の指定方法 ・検索速度 ・検索結果表示画面のレイアウト ・絞り込み検索に関する操作性、等
b. 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方法の検討		
高度な検索機能への全文テキストの活用 （ナビゲーション、リコメンド等）	ナビゲーション、リコメンドの有効性の検証・確認を行う。	検索結果と合わせて表示される関連語句の妥当性について定性的に5段階で評価 定性的な指標を構成する要素の例 ・関連語句の表示レイアウト ・関連語句の表示順序 ・表示された関連語句と検索内容の関係が類推できるか、等
c. 全文テキスト化情報を含めた場合の適切な検索結果のランキング、表示方法の検討		
ランキングへの全文テキストの活用	ランキングなどの有効性の検証・確認を行う。	ランキングの方法、表示方法 ランキングの表示の有効性を定性的に5段階で評価 定性的な指標を構成する要素の例 ・検索結果に関する重み付けの妥当性 ・ランキング情報による検索結果に対する付加価値向上性、等
d. デジタル化資料のイメージ形式の表示結果の検証		
テキストの表示（Webブラウザ）		
ハイライト表示	テキストの表示に関わる機能の妥当性、有効性の検証・確認を行う。	検索結果に対して表示される検索語のハイライト表示の見やすさについて定性的に5段階で評価 ・ハイライト表示部分の視認性 ・ハイライト表示部分の妥当性 ・ハイライト表示部分の十分性（漏れがないか）、等
目次から本文へのリンク		リンク機能を操作し、適切なリンク先が表示されるかについて定性的に5段階で評価 目次を選択したときに、当該ページが「正しく表示され」ない比率を評価。 なお、「正しく表示される」の定義を別途行う必要あり

実施項目	評価の目的	具体的な評価指標の例
デジタル資料の表示		
ページめくり表示	デジタル資料の表示に関わる機能の妥当性、有効性の検証・確認を行う。	ページめくり表示を操作し、その利便性や操作の妥当性について定性的に5段階で評価 定性的な指標を構成する要素の例 ・ページめくり表示機能の操作性 ・ページめくり表示機能の応答性（画面切り替えの速度）、等
ハイライト表示		検索結果に対して表示される検索語のハイライト表示の見やすさについて定性的に5段階で評価 ・ハイライト表示部分の視認性 ・ハイライト表示部分の妥当性 ・ハイライト表示部分の十分性（漏れがないか）、等
許諾権限管理	許諾権限管理方法とその表示法などの妥当性の検証・確認を行う。	検索結果の表示内容が許諾権限情報を反映した内容になっているかを定性的に5段階で評価 ・著作権切れ ・著作権有効期間内文書 ・文化庁裁定文書 ・実験協力企業（出版社、印刷会社）提供電子書籍データ等
e. 全文検索を行うために必要となるテキストの認識精度の検証		
全文テキストのインデキシング	テキストのインデキシングの妥当性の検証・確認を行う。	対象全文テキストのインデキシングに要する所要時間を測定 ①対象全文テキストのインデキシング所要時間（投入ページ数との関係）
全文テキストの高度活用（クラスタリングなど）	全文テキストの高度活用の有効性の検証・確認を行う。	全文テキストの高度活用に関する機能の有効性を、実際に操作を行い定性的に5段階で評価