全文テキスト化の今後の展望

2010年12月11日

国立国会図書館　中山正樹

実証実験のまとめ

- テキスト化
  - OCR文字認識の精度向上は継続的に
  - 共同校正は有効だが、大量の校正にはまだまだ
- 構造化
  - 章節項のレベルに目次を付けるだけでなく、分割して、書誌的事項を自動付与できることが必要
- 検索
  - 章節項のレベルで、関連するドキュメントとのリンクが必要
- おまけ
  - 電子書籍対応（DAISY→EPUB）により、縦書きブラウザでの閲覧が可能（Webkit、GoogleChrome等）

まとめ―テキスト化―

- フォーマット
  - OCR出力フォーマットALTOの日本語表示固有表現の標準化
  - 透明テキスト付PDF、DAISY以外のフォーマットへの対応、自動変換ツールの開発
- 共同校正・共同構造化
  - 新字・旧字対応、目次階層化レベル、柱の概念、ページ判断基準等、作業ルールの明確化

まとめ－構造化－

- 構造化機能
  - 校正結果によるOCR再学習機能
  - 構造項目ごとの特徴を踏まえた推論機能
    - 目次、本文、索引ページ
  - 文字・単語のつながり方の規則や辞書を利用した形態素分析によるチェック機能
  - JIS第3、第4水準対応
- 組織化機能
  - 章・節・項の単位での組織化

まとめ－検索－

- テキスト検索
  - 書籍単位の全文検索から、章・節・項単位での検索へ
  - 本文すべて、目次、索引の構造指定検索に加えて、見出し階層指定検索も
  - 階層化された単位での検索語出現数表示
  - 文脈検索、タグクラウド、固有名表示、引用書籍表示等の操作性のブラッシュアップ
- 読上げサービス等
  - 旧字体、旧仮名遣い対応の読上げソフト
  - 新字体での検索用データとは別に、旧字体での表示用データの用意

まとめ－おまけ－

- DAISY ≒ EPUB
  - アーカイブファイルのサフィックスを変えるだけで、EPUBとして認識
  - DAISYファイル群に含まれるCSSに縦書き指定をするだけで、縦書きブラウザで表示可能
- EPUB３、その他フォーマット対応
  - 保存用フォーマットと閲覧用フォーマットの分離が必要
  - 中間ファイルフォーマットの行方が不透明
    - 仕様決定に、当館はどのように関わっていくか？

今後の展望

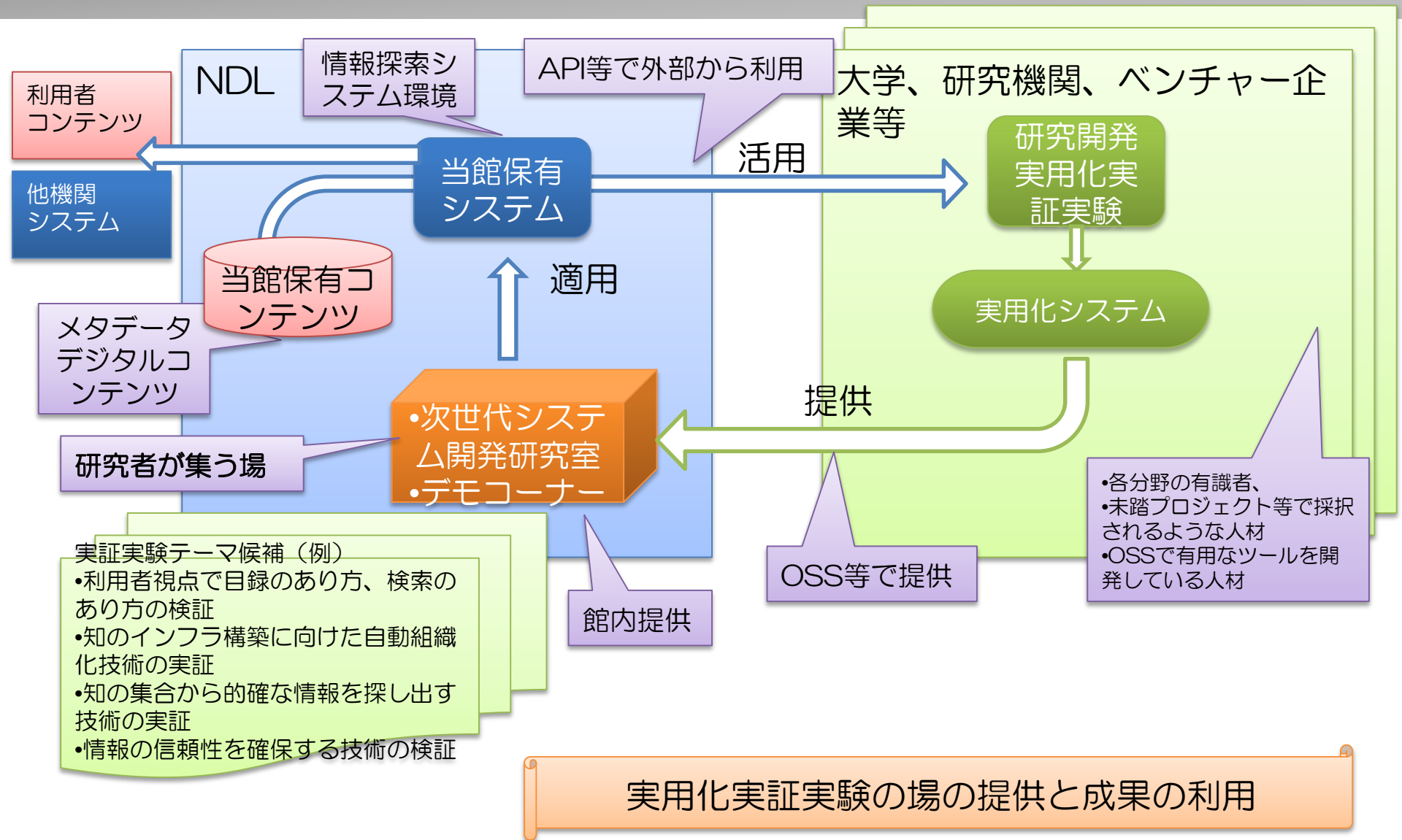
- 館内外の動き
  - 知識インフラの構築を目指す
  - 保存のためのデジタル化（館内閲覧）、視覚障害者へのアクセシビリティ（読上げ、配信等）
  - 文化庁検討会議中間報告の実現
- 直近で、目指すところ
  - 全文デジタル化
    - 検索及び部分表示に応えられるレベル
    - 障害者の視聴に耐えれるレベル
  - 知識として利用できるレベルでの構造化
  - 知識としての検索サービス

実現を加速させるために

- NDL内体制及び環境
  - 次世代システム開発研究室、非常勤調査員の採用、有識者会議
  - 実証実験環境（NDLラボ（仮称））の設置
- 資金（想定）
  - 関係省庁の研究開発予算で当館と共同研究、当館予算要求
- 実施（想定）
  - 電子出版環境整備事業(新ICT利活用サービス創出支援事業)【総務省】の成果の活用
    - 電子書籍交換フォーマット、EPUB、近刊情報書誌情報、共通メタデータ
    - 音声読み上げ対応電子出版制作ガイドライン、画像情報からのテキスト抽出アクセシビリティガイドライン
  - NICTの研究開発への協力→共同研究（今年度から）
    - 旧字体/新字体変換、言い換え/含意/矛盾関係アノテーション、図書館蔵書を対象にした質問応答システム
  - JST、NIIの予算で構築するシステム、サービス
    - J-GLOBALとNDLSearchでの引用文献リンク、検索サービスでのJSTシソーラス辞書の活用
    - 文献の関連リンク：ジャパンリンクセンター構想

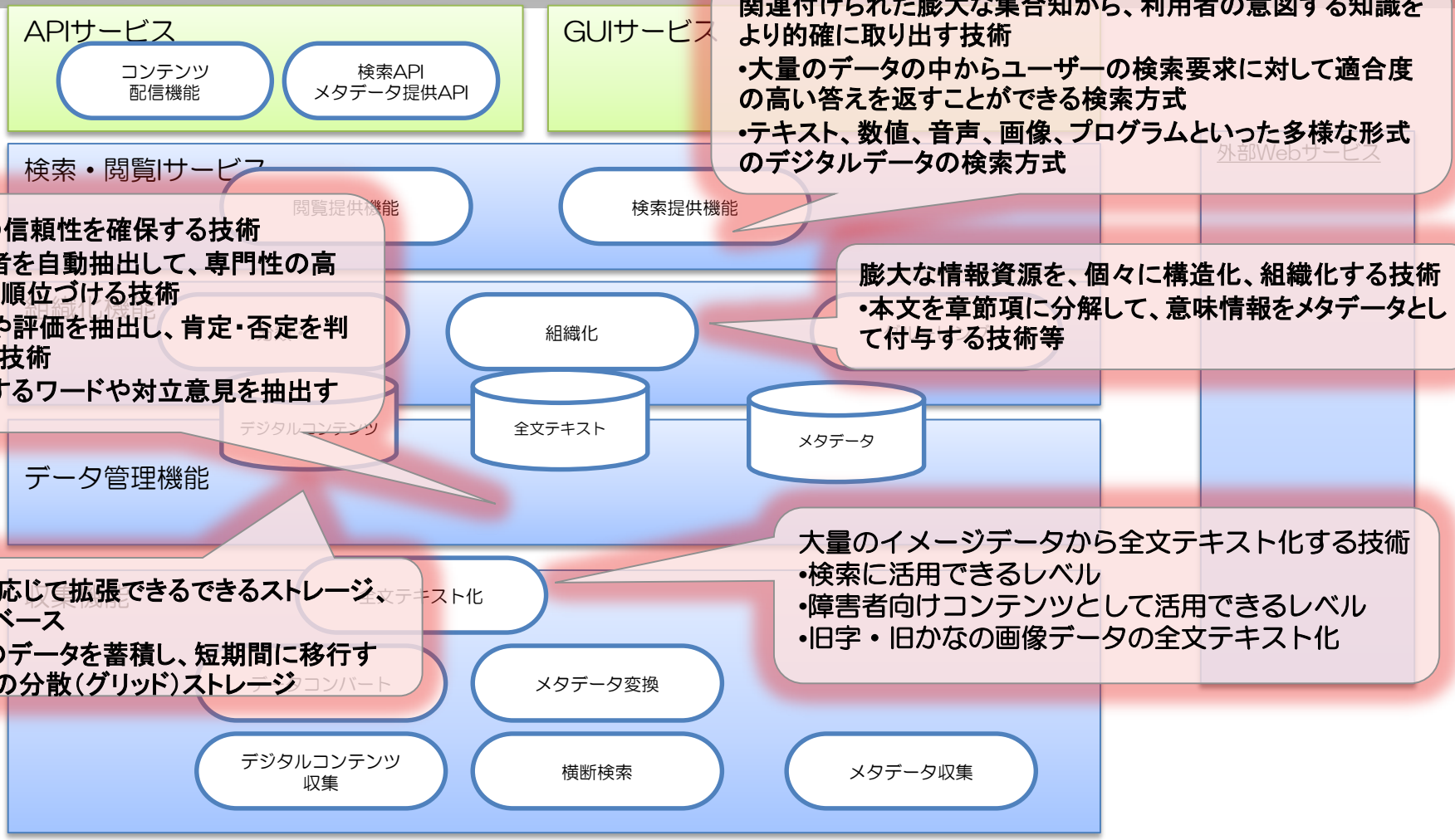
NDLラボ（仮称）の設置

次世代サービスの研究開発と実用化を促進するために



活用したい研究開発成果（例）

全文テキスト化実証実験の延長として



終わりに

- デジタルアーカイブとして、
  - 情報資源を将来に亘って保存するだけでなく
  - 当館が所蔵している情報資源は、利用しやすく
  - 他機関が保有している情報資源は、たどり着けやすく
- 知識インフラとして、
  - 情報資源を、新たな知識の創造に活用できるように
- 高度化する社会的なニーズに応えるために、
  - 一つの機関の情報資源、技術だけでは実現できない
  - 関係機関と連携して、研究開発成果を適用していく。