資料技３－２

既存書籍のデジタル化

2010.04.27

凸版印刷株式会社

既存書籍のデジタル化への期待

1）作者、編集者等の意図を正確に永久保存
- デザイン、レイアウト、書体、外字/異体字 etc.

2）情報提示機器の多様化への対応
- PC、モバイル機器、ゲーム機、デジタルTV、専用端末 etc.

3）デジタル化による付加価値
- マルチメディアデータとの統合、検索、再利用 etc.

テキストとしてのデジタル化?
イメージとしてのデジタル化?

	メリット	デメリット	コスト
テキスト	●デジタル化のよる付加価値を付けやすい（全文検索等） ●他の目的への再利用が可能 ●データ量が少ない	●作者・編集者の意図を完全に再現することが困難（レイアウトデザイン、文字表現等）	●OCR誤認識の修正 →時間がかかる ●テキスト校正 →負荷が高い
イメージ	●作者・編集者の意図を再現できる ●様々な情報提示機器への対応が容易（画像サイズの変更。ただし画質は劣化）	●データ量が多い - 画像圧縮が必要 →文字潰れの発生 - リフローができないことによる読みやすさの欠如	●スキャナやカメラと画像処理技術を組み合わせ、ある程度の自動化が可能 →時間短縮 →負荷が低い

- これらのバランスを考慮すると、現状ではイメージが主流
  → 今後の技術に期待

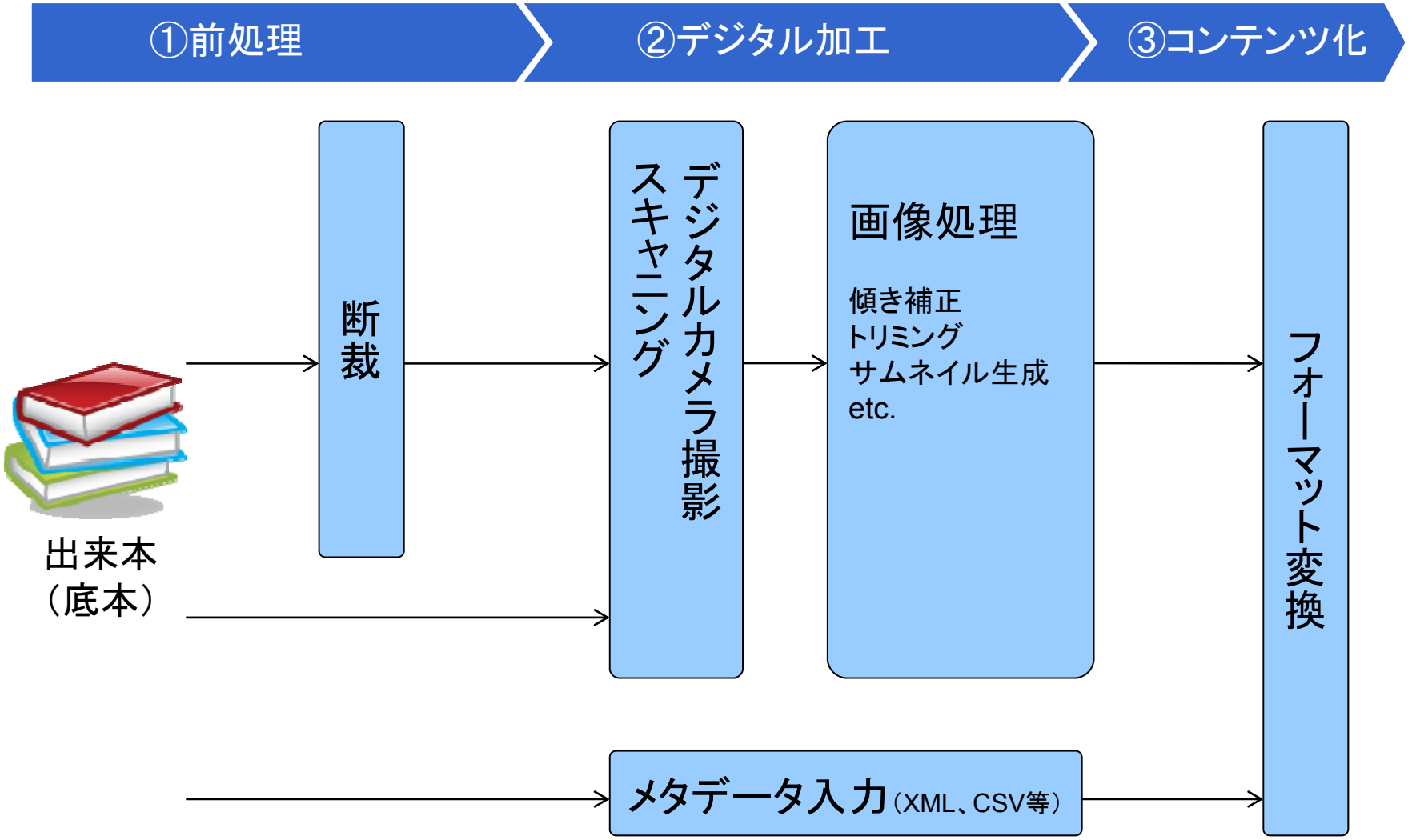
既存書籍のデジタル化への取り組み（1）

- イメージ
  - 国立国会図書館
    - JPEG 2000
  - 国立公文書館
    - JPEG 2000
  - Google社（米国）
    - Google Book Search
    - 一部PDF、EPUBでダウンロード可能
    - 約1200万冊（2010年4月現在）
  - Internet Archive（米国）
    - スキャンされたパブリックドメイン書籍。一部はfull-text化
    - 約2400万冊（2010年4月現在）

既存書籍のデジタル化への取り組み(2)

- テキスト
  - 紀伊國屋書店
    - NetLibrary(米国)への提供
    - PDF + テキスト
    - 約1500冊以上(和書、2010年4月現在)
  - 青空文庫
    - ボランティアによる底本のデジタル化
    - 約8900点(2010年4月現在)
  - プロジェクト・グーテンベルク(米国)
    - 青空文庫の海外版(1971年～)
    - 約3万5000点(2010年4月現在)

デジタル化の工程（既存書籍）



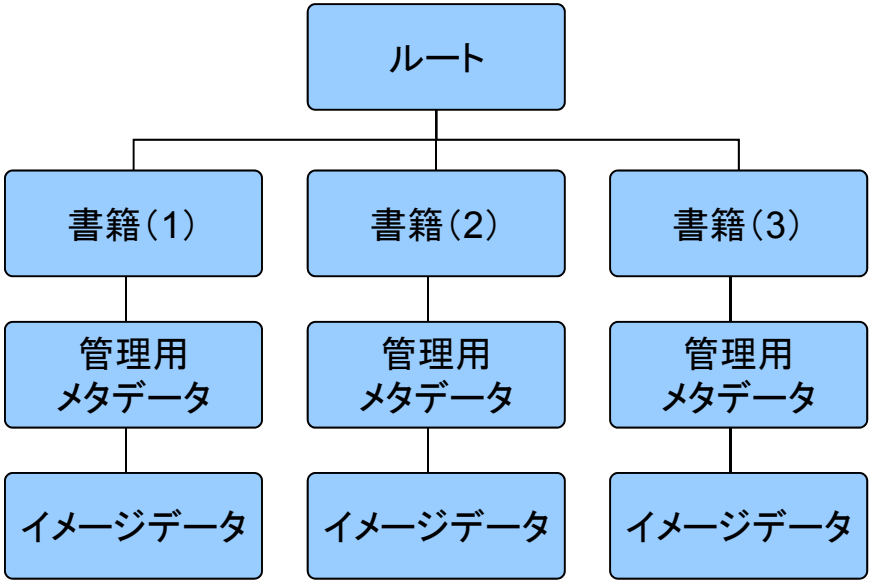
イメージデータの取り扱い（1）

- フォーマットは用途に応じて複数用意
  - 保存用
    例1）TIFF（可逆圧縮）
    → アプリケーション非依存、印刷品質に耐えうる
    例2）JPEG（可逆/非可逆圧縮）
  - サムネイル用
    例）TIFF、PDF、JPEG etc.
    → 保存用データを再利用して作成
- 解像度
  - 保存用：OCRの利用を考慮し400dpi程度
  - サムネイル用：システム要件に応じて都度検討
- 撮影条件
  - 外光の影響を受けない環境下で撮影する
  - 照度ムラ等はソフトウェアで補正

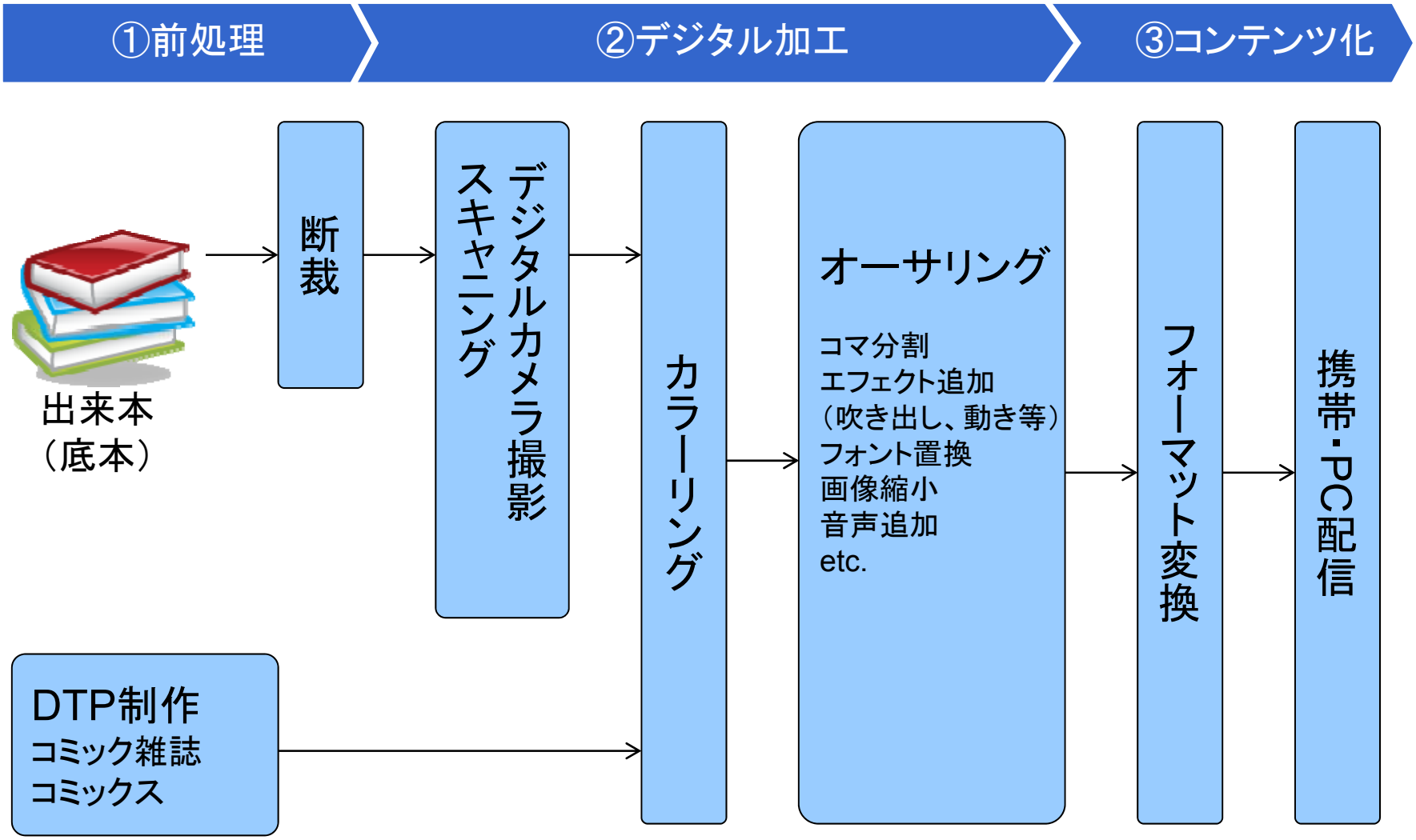
イメージデータの取り扱い(2)

- データ管理
  - 管理ルールの定義
    - データの管理単位(ディレクトリ・データ構成 etc.)
    - 管理用メタデータ
    - 保管形態(メディア/サーバ) etc.
  - ソフトウェアの利用

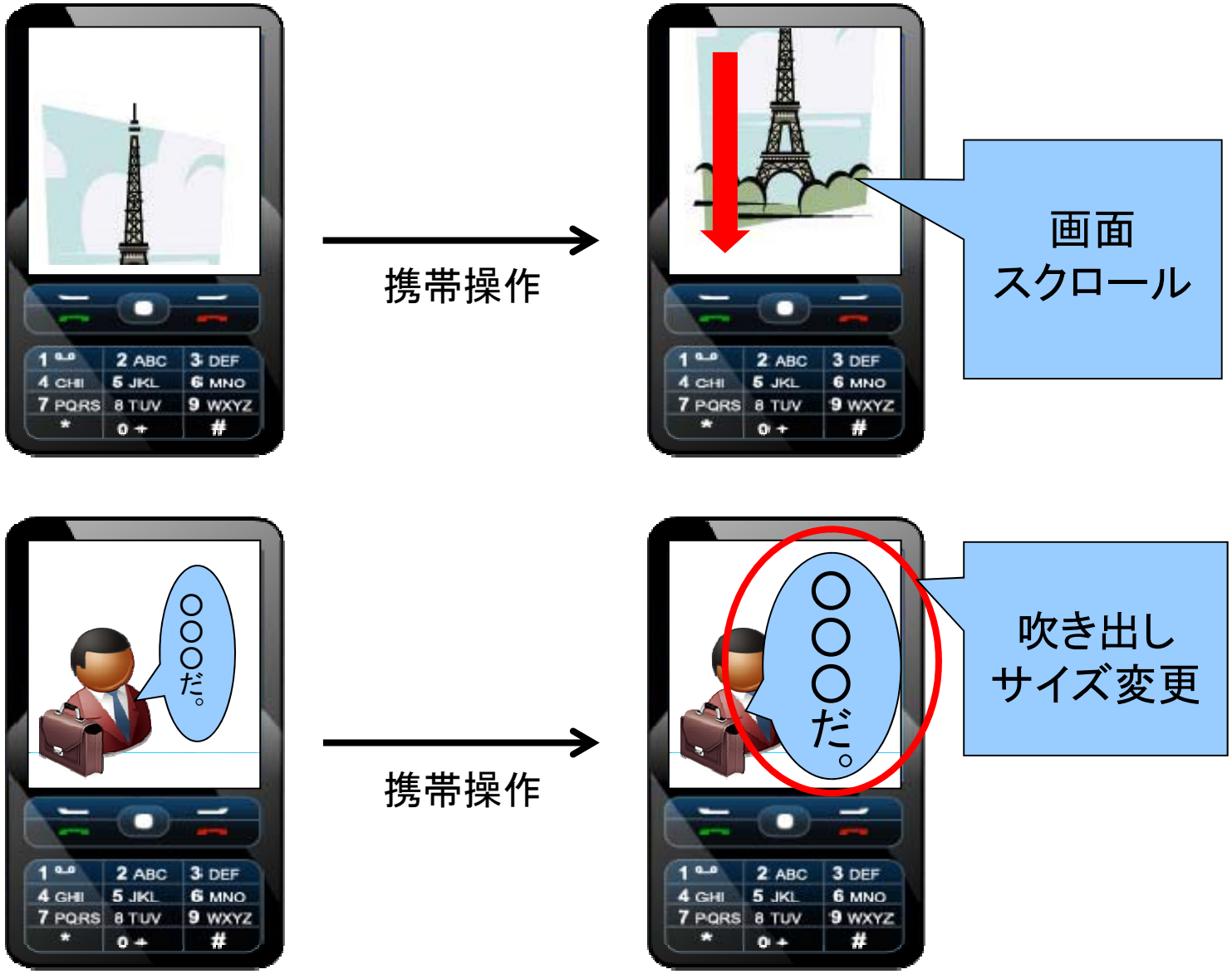
    例) ファイル命名規則に則り自動採番、管理用ディレクトリへ自動配置



デジタル化の工程（コミックス）



コミックスの例



品質管理

- ソフトウェアによる自動補正
  - イメージデータのトリミング

- 目視による検査
  - 既存書籍との比較
    - 可読であるか
    - 色調・明るさ・コントラスト等が再現されているか
    - イメージデータが傾いていないかどうか
    - 欠損・汚損等が発生していないか

    →1イメージごと行う

課題と今後への期待(1)

- 膨大な量のデジタル化は時間・コストとの戦い
  - いかに管理システムに落とせるか
    - ワークフローのルール決め
  - 各工程をどれだけ自動化できるか
    - 入力処理、画像処理、品質管理、etc.

- 目的に応じたデジタル化の品質設計

技術の進歩により、デジタルデータは現物に近づくが、
決してイコールにはならない(別な付加価値は付くが)。

課題と今後への期待(2)

- イメージからテキスト化
  - OCRの利用
  - 外字/異体字の取り扱い

- 色を正しく保存し、正しく伝える環境
  - 書籍のジャンルによっては色の保証は必須