紙資料からＰＤＦを作成する

機関リポジトリ新任担当者研修東日本会場
2015年2月20日
森下映理（奈良女子大学学術情報センター電子情報係）
eri@cc.nara-wu.ac.jp

本日の内容

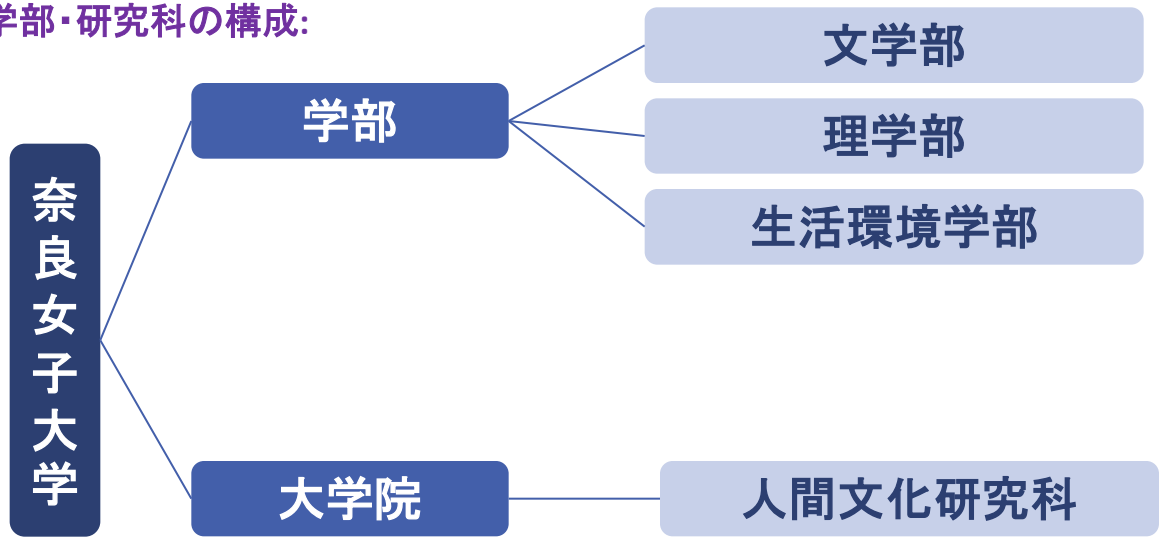
- 奈良女子大学と本学リポジトリの概要
  立ち上げ当初から現在まで

- 紙資料の電子化について

奈良女子大学について

◆大学の概要

学部・研究科の構成:



学生総数: 2,671名（学部学生2,097名、大学院学生574名）
教員総数: 212名

（平成26年5月1日現在）

学術情報センターについて

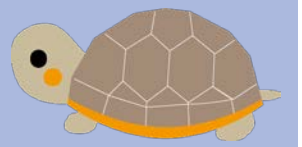
2014年4月　附属図書館と総合情報処理センターが統合

◆構成：

情報管理係・情報サービス係・電子情報係
・情報基盤係
（常勤：7名　非常勤：7名）

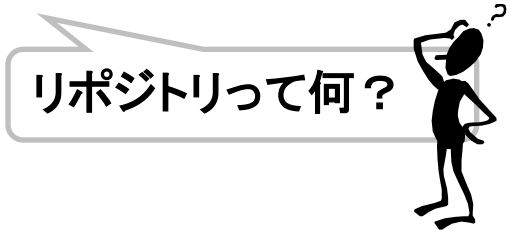
★リポジトリ担当は電子情報係
（常勤1・非常勤1）

本学リポジトリについて



平成18年

学内委員会の立ち上げ
「平成18年度次世代学術コンテンツ基盤共同構築事業：領域１」委託業務に採択される！
学内刊行物の調査・把握→公開許諾依頼
近年の刊行物のPDF化
D-Space 構築（業者に依頼）



平成19年

領域１継続中
10月リポジトリ専任非常勤職員採用
情報収集（DRFメーリングリストへの参加）
学内広報活動の推進
公開許諾依頼 （研究者総覧に登録されている論文について）

平成20年

3月　正式公開
引き続き、個別の著者に対しての公開許諾依頼→登録作業

平成21年

3月末　リポジトリ専任の非常勤退職→後任の学内予算つかず
10月　担当係長交代 （初代担当係長、学外へ）

平成22年

3月　**領域１委託業務終了**
4月　担当係長交代 （館内移動）。新課長のもと新体制に。
領域３（近畿領域）と領域２（遺跡リポジトリ）採択される

平成23年

4月　図書館長が副学長と兼任に。
平成22-24 年度委託事業（領域2）「オープンアクセス環境下における同定機能導入のための恒久識別子実証実験」の平成23年度 **実証実験に参加**。（平成24年2月D-Space バージョンアップ。3月リポジトリと研究者DBとの連携）

平成24年

3月　**領域３（近畿領域）終了。**

本学の電子データの作成について

1 印刷会社等が作成したPDF

紀要などの印刷時に作成されたPDF

CHECK!

2 Microsoft Office のファイル (WORD等)

PDFプリンタで PDF化

3 紙媒体の資料（本・雑誌等）

SCAN

PDF化（画像化）

OCR

登　録（メタデータ付与）

OCRって何のこと？

Q1. OCR（Optical Character Reader)とは？

OCR（Optical Character Reader)は、光学式文字読取装置。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。

Q2.なぜOCR機能を利用して透明テキストをつけるのか？

紙媒体をスキャンしたPDFはただの画像で、文字データを持たない。OCR機能で透明テキストを付与して、検索可能にするために、OCRを行う。←論文が検索されやすくなる！

印刷会社が作成したPDFやMicrosoft Wordなどから変換したPDFは、もともと文字データがあるため、PDFではあるが、なりたちが違う。

PDFの成り立ちの違い

電子媒体由来のPDF

文字データと画像データの組み合わせ。フォント情報を有する。ファイルは、スキャンして作成したPDFより小さく、拡大縮小した場合にも文字等がつぶれず美しい。

紙媒体由来のPDF

一枚の画像。文字情報がないので、OCR機能を使って、上から透明テキストをかぶせる必要がある。

紙資料の電子化①

自力でのPDF化

裁断 → SCAN → PDF化

資料によって、作業手順が変わる

【著者等から提供された資料の形態】
- 紙 or 電子媒体
- 裁断可能 or 裁断不可（裁断後、製本？）
- カラー, 白黒, グレー？
- PDF , WORD, Power Point?

紙資料の電子化②

裁断　裁断機

電動タイプ　or　手動タイプ

手動タイプの場合は、刃が垂直に降りてくる垂直裁断型がお勧め。

刃が斜めに降りてくる遮断機型は×

SCAN　スキャナー・複合機等

フラットヘッドスキャナー　or　ADF (自動給紙)型スキャナー

★裁断しない場合もあるので、フラットヘッド型が望ましい。画質も〇

★フラットヘッドとADF両方の機能がついていたほうが楽。

複合機

★本学では複合機（レンタル）利用

紙資料の電子化③

本学ではSCANに複合機を使っています。

使用機器
SCAN : RICHO カラー複合機 MPC3003

苦労しました・・・・
★開始当初は、スキャナも買えず、ペーパーレス化のため事務局から支給されたA4用ScanSnapを使用。裁断したものしかScanできず・・・。

紙資料の電子化④

アクセスしやすいPDFを作る！

スキャン（PDF化） → OCR（透明テキスト付与）

【SCAN時の設定】

解像度高→画質◎、サイズ大。
解像度低→サイズ小。画質荒い

●解像度：400dpi程度
（詳細な画像が必要な場合は、600dpiに変更する場合も）
●裏写り防止設定
●カラーモードを使い分ける。
（カラー写真等はカラーモード。白黒写真はグレースケール。）

紙資料の電子化⑤

複合機の原稿種類設定：モノクロ写真の場合

No.	「原稿種類」の設定	ファイルサイズ
①	白黒/文字	75KB
②	白黒/文字・画像	345KB
③	白黒/文字・写真	466KB
④	白黒/写真	575KB
⑤	グレースケール	259KB

【参考写真資料】

解像度：400dpi

ファイル形式: PDF

読み取りサイズ： A5

●**「原稿種類」の設定**を**「白黒 / 文字」**（モノクロ①）にすると、ファイルサイズは小さいが、黒くつぶれてしまう。

●**「原稿種類」の設定**を**「白黒 / 文字・画像」**（モノクロ②）**「白黒/文字・写真」**（モノクロ③）**「白黒/写真」**（モノクロ④）にした場合、画質はよくなるがファイルサイズは大きくなる。

●「白黒/文字」モード以外の白黒モードよりもグレースケールの方がファイルサイズは小さい。

画像の重要度にもよるが、写真など画像がある資料はグレースケールでスキャンした方がよい。

●綴じがきつい資料などは白黒でスキャンすると、影がまっ黒に出てしまい、文字が読めない。グレースケールでスキャンすると、影がグレーになり、文字は読めるようになる。

紙資料の電子化⑤

複合機の原稿種類設定：カラー写真の場合

No.	「原稿種類」の設定	ファイルサイズ
①	白黒/文字	55KB
②	白黒/写真	945KB
③	グレースケール	230KB
④	フルカラー	274KB

【参考写真資料】

解像度：400dpi

ファイル形式: PDF

読み取りサイズ： A5

●カラー原稿の場合も同様に**「白黒/文字」**（カラー①）は黒くつぶれてしまい読めない。
●設定を**「白黒/写真」**（カラー②）にした場合、ファイルサイズが「フルカラー」の3倍以上に。
●**「グレースケール」**（カラー③）は、画像は美しいが、ファイルサイズ自体は期待するほど、**「フルカラー」**（カラー④）より小さくはならない。

OCRソフトについて

本学では・・・

使用ソフト
Adobe Acrobat XI Pro と e.Typist V.12.0を併用

苦労しました・・・・
★リポジトリ開始直後は、Abobe Acrobat Standardを使用していたが、墨消し機能もなく、画像のゴミや影を消すのにも苦心。→
後にAcrobat 7.0 Professionalを購入。（博士論文公開に関連して、PDF/Aを作成するため、現在はXIにバージョンアップ。）
複数言語対応のため、e-typistを導入して、現在に至る。

OCRソフトの比較

使用ソフト:

Adobe Acrobat XI Pro・etypist v.15.0〈体験版〉

読取革命 Ver.15 （体験版）

Scan設定:

解像度:400dpi / ファイル形式: PDF

読み取りサイズ: B5 / 白黒・文字モード

複合機でスキャンしてPDF化したファイルを個々のソフトでOCR。OCR 前とのファイルサイズを比較。

解像度	OCR前	e.Typist	読取革命	Acrobat
200dpi	59KB	210KB	728KB	36KB
400dpi	150KB	223KB	701KB	62KB
600dpi	267KB	224KB	711KB	90KB

●e.Typistおよび読取革命では、もとの解像度が違っても、OCR後のファイルサイズの差があまりない。PDFから一旦画像化して作業する際に最適化され、ほぼ同一のサイズになっている。

●Acrobat Professionalでは、OCR化後に最適化され、OCR化前よりもファイルサイズが小さくなっている。

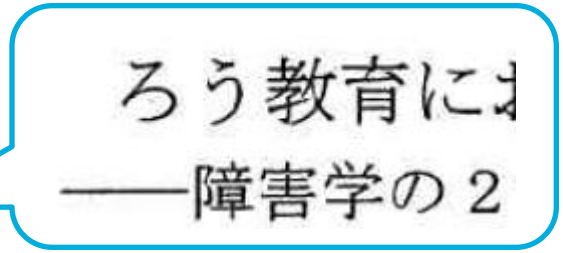
ファイルサイズの比較には、
座主果林:“ろう教育における2つの教育方法”, 奈良女子大学社会学論集, 2010, 第17号, p.243 を使用

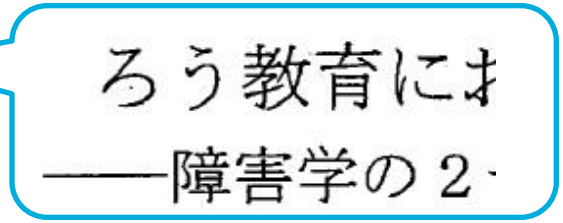
e.Typistで保存すると劣化する？

e.TypistではPDFファイルを画像に変換するため、解像度の低いPDFファイルを読み込むと画像が劣化することがある。←環境設定を操作することによって、改善が可能。
http://pac.mediadrive.jp/faq/index.php?action=artikel&cat=365&id=447&artlang=ja

が、当然、きれいにするほど、ファイルサイズは大きくなる。
元画像の精度等によって、設定を変更する必要がある。

	OCR後サイズ
デフォルト設定	223KB
「きれい」設定	2,641KB





画面で見るとデフォルトの方は、文字の後ろに影がうっすら入っているのが、わかる。印刷すると、あまり違いは目立たない。（元画像の精度や、プリンターにもよる。）

ファイルサイズの比較には、
座主果林：“ろう教育における2つの教育方法”, 奈良女子大学社会学論集, 2010, 第17号, p.243 を使用。
400dpiでスキャンしてPDF化したファイル（150KB）をOCR化した。
「きれい」設定では、環境設定の解像度を400dpiにし、認識結果の「図領域の品質」をきれい（最大値）に設定。

OCRソフトの比較
（文字認識の精度①）

英語　（661文字/ 1 page)

	誤認識文字数	正しく認識された文字の比率（%）
e.Typist v.15.0（体）	0	100
Acrobat XI Pro.	0	100
読取革命 Ver.15（体）	0	100

フランス語 (1420文字/ 1page)

	誤認識文字数	正しく認識された文字の比率（%）
e.Typist v.15.0（体）	2	99.8
Acrobat XI Pro.	1	99.9
読取革命 Ver.15（体）	×(4)	×(99.7)

日本語：横書　（1149文字/ 1 page)

	誤認識文字数	正しく認識された文字の比率（%）
e.Typist v.15.0（体）	18	98.4
Acrobat XI Pro.	18	98.4
読取革命 Ver.15（体）	6	99.4

韓国語(223文字/ 1page)

	誤認識文字数	正しく認識された文字の比率（%）
e.Typist v.15.0（体）	14	93.7
Acrobat XI Pro.	49	78.0
読取革命 Ver.15（体）	×	×

OCRソフトの比較
（文字認識の精度②）

日英混合　（731文字/ 1 page)

	誤認識文字数	正しく認識された文字の比率（%）	スペース抜け
e.Typist v.15.0（体）	1	99.8	78/78
Acrobat XI Pro.	10	98.6	14/78
読取革命 Ver.15（体）	6 (12)	99.1(98.3)	0/78

日本語：縦多段　（646文字/ 1 page)

	誤認識文字数	正しく認識された文字の比率（%）
e.Typist v.15.0（体）	2	99.6
Acrobat XI Pro.	10	98.4
読取革命 Ver.15（体）	2	99.6

・句読点などを除いた文字数で計算。
・テストデータは400dpiで読み込み。
・読取革命は日英のみ対応だが、仏語を読み込んだところアクセント記号が読みとれないだけでかなり正確に読み込んだ。
・日英混合の読取革命の（）内の数字は全角半角間違いを含む。
・韓国語は紙媒体が古く、認識されにくかった模様。
・e-typistは日英混合対応となっているが、文字認識はするものの、英文のスペースが全て抜けていた。

英語：外国文学研究（奈良女子大学）, Vol.28, p.85
仏語：外国文学研究（奈良女子大学）, Vol.28, p.113
日本語縦：前述の座主論文　を利用
韓国語：10日間のハングル（JICC出版局）, p.117
日英混合：外国文学研究（奈良女子大学）, Vol.31, p.3
日本語縦多段：叙説（奈良女子大学）Vol.41, p.24

ソフトの長所・短所
（Adobe Acrobat XI Pro)

長所：

- 認識時間が短い。
- 墨消し機能あり。
  （画像や文字を墨消し可能なので、プライバシーに関係するような画像や文字等を消すことができる。Standardにはなし。）
- Wordやエクセルなどのソフトに対応。
- 出来上がりサイズが小さい
- PDF/Aの生成が可能
- PDFプリンターに対応。
- 中国語・韓国語などのアジア言語に対応

短所：

- 複数言語が混在する場合、対応できない。
- 透明テキストの確認方法・書き変え方法が煩雑。

ソフトの長所・短所
（e.Typist v.15.0：体験版）

長所：

- 文字認識率が高く、専門用語の単語登録も多い。
- 多言語対応（58ヶ国語）しており、他のソフトにはない言語もOCR化可能。
- 複数の言語が混在する文もOCR化可能。
- 認識範囲の指定、透明テキストの確認、修正が可能。
- 画像編集機能あり。（トリミング、消しゴム機能、直線描画等）
- Abobe Acrobatでエラーが起きてOCR化できないPDFでもOCR化できる場合がある。

短所：

- PDFファイルを読み込むと、いったん1Pずつの画像として認識し、それぞれ分割して処理を行うため、作業効率が悪く、認識時間も若干長い。
- デフォルトでPDFを処理した場合、文字の後ろに影が入る場合がある。（改善は可能だが、ファイルサイズが大きくなる。）

ソフトの長所・短所
(読取革命Ver.15：体験版)

長所：

- 単語辞書が充実。文字認識率が高い。状態の悪い原稿にも強い。
- 認識範囲の指定、透明テキストの確認、修正が可能。
- 画像編集機能あり。(トリミング、消しゴム機能、直線描画等)
- Abobe AcrobatでエラーがおきてOCR化できないPDFでもOCR化できる場合がある。

短所：

- PDFファイルを読み込むと、いったん1Pずつの画像として認識し、それぞれ分割して処理を行うため、作業効率が悪く、認識時間も若干長い。
- 日英の二言語のみの対応。
- 生成するPDFのファイルサイズが大きい

OCRソフトについて

- AcrobatプリンタやOfficeソフトとの連携、セキュリティの設定、PDF/A対応等のことを考えると、Acrobat Proがおすすめ。
- Acrobat Standardは墨消し機能がなく不便。Professionalを。
- 多言語資料のPDF化や画像処理を頻繁に行うのであれば、e-typistが便利。
- 読取革命は状態の悪い原稿に強く、文字認識力が高い。（日本語、英語対応のみ）
- 広告を見ているだけではわからないこともある。体験版やネットの口コミ情報、他大学の状況も要調査。

PDF化のまとめ

- 自力でのPDF化は時間がかかるだけでなく、ファイルサイズも大きくなる。新規に出版される紀要等は事前に出版団体に依頼。印刷業者にPDFも納品してもらうこと。
- 自力でPDF化する場合は利用者のことを考えたPDFを作成。（重すぎない、粗すぎない。外部から検索されやすいように、透明テキストを！）