ディジタルドキュメント (8)

高久雅生

2015年6月4日（木）3・4時限

1

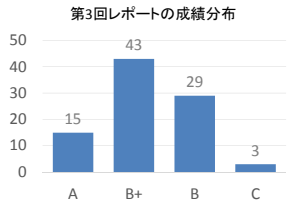
本日のお品書き

- （第3回レポートの返却・講評）
- （前回の振り返り）
- （前回講義での質問・要望への回答：１件）
- オフィス文書とデジタルドキュメント
  - ✓ワークフロー
  - ✓処理モデル
  - ✓文書フォーマット：OOXMLとODF
    - 標準化（独占規格から国際標準へ）
    - パッケージ形式による複合オブジェクトの埋め込み
- デジタル化テキスト
  - ✓大規模デジタル化
  - ✓テキストと文字，デジタル翻刻
  - ✓ドキュメントフォーマットTEI

2

第3回レポート講評

- 成績は全体にとても良かった。
  - ✓内容的にもよく書けている人が多い印象
- 選択した電子書籍の種別
  - ✓小説 58
  - ✓一般書 16
  - ✓コミック 13
  - ✓論文 2
  - ✓その他 1



3

第3回レポート講評 (2)

- 電子書籍プラットフォーム，提供元
  - ✓青空文庫　44
  - ✓（その他：iOSアプリ）　8
  - ✓Kindle 7
  - ✓iBooks　7
  - ✓Comico　3
  - ✓Digital Book Japan　3
  - ✓Book Live!　2

4

（前回の振り返り）

- テキスト処理とドキュメント形式
  - ✓ドキュメント形式（フォーマット）とは
  - ✓テキストフォーマットとは
    - 文字コード
  - ✓電子メールのドキュメント形式
  - ✓軽量テキストフォーマット
  - ✓ページ記述型フォーマット

5

質問へのお答え (1)

- 形式と形態の違いは何でしょうか？

「形式」は、前回の講義で扱ったような、文書形式に対応します。つまり、対象とする文書がどのように符号化されているかを示す用語です。
一方で、「形態」はより一般的な概念として、その文書の様態、様子を総合的に表現するように用いています（少なくとも本授業では）。

6

オフィス文書のフォーマット

7

オフィス文書の具体例

- 会議資料
- 議事録
- プレゼンテーション資料
- 広報資料
- 案内、お知らせ
- 会計書類：見積書、発注書、納品書、請求書
- 出張願、出張依頼、出張復命書
- 報告書
- 企画書
- 説明書、マニュアル
- 業務メモ
- 経歴書、業績書
- 連絡文書
- レターカバー

8

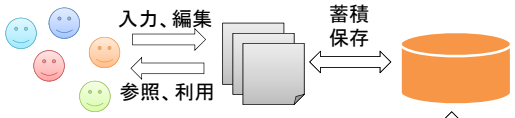
オフィス文書とデジタルドキュメント

- オフィス文書はデジタルドキュメントの王道
- 文書のサイクルが出現
  - 作成
  - 編集
  - 承認
  - 管理
  - 配信・提供
  - 保存管理
- 関わるひとが多いことも特徴
  - ドキュメントの作成者
  - 共同入力者・編集者
  - 承認・決裁者
  - 利用者
  - 管理者
  - 2次的な利用者
- 紙の文書との併用も

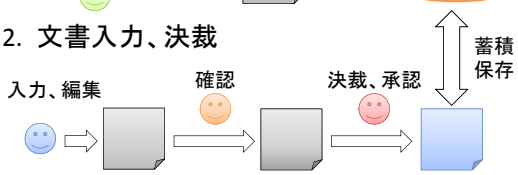
9

オフィス文書に関わる処理の類型

1. 文書共有、保存



2. 文書入力、決裁



10

オフィス文書の特徴

- 複数人による情報共有が前提とされている
  - ✓他者による引継ぎや検査
- 情報の保存
  - ✓日付（保存期間）、担当者など
- 定型形式
- 紙文書としての処理フロー
- 版管理機能

11

統合オフィスソフトウェアの事例
（オフィススイート; Office suite）

名称	文書	表計算	プレゼンテーション	画像処理
Microsoft Office	Word	Excel	Powerpoint	
LibreOffice	Writer	Calc	Impress	Draw
Google Docs	Docs	Spreadsheet, Form	Presentation	
iWorks	Pages	Numbers	Keynote	
JUST Suite	一太郎	三四郎		花子
Lotus (IBM)	WordPro	1-2-3		

12

オフィスソフトウェアが取り扱うドキュメントフォーマット

- それぞれの文書ソフトウェア毎に独自フォーマットが用いられることが多い。
- 1990年代後半からは、Windowsとともに、Microsoft Office製品が市場シェアを伸ばし、それらのソフトウェアがオフィス文書コンテンツの代表格となってきた。
- 2000年代以降、オープンフォーマットの重要性が高まり、標準化された文書ソフトウェアが伸張している。
  - ✓OpenOffice.org (2000-2011)
  - ✓→ドキュメントフォーマット ODF (Open Document Format)
- また、別の動きとしてオンライン上での文書ソフトウェアの提供の動きもある。
  - ✓Google Docs

13

Microsoft Officeのドキュメントフォーマット

- 長年の間、Microsoft Officeは独自形式（バイナリデータ）を用いていた
  - ✓Microsoft Officeの初期バージョン以来、MS Office 2004まで
  - ✓MS Office 2004以前の環境におけるドキュメントフォーマットは一社独占による文書形式 (proprietary)
    - 拡張子：.doc, .xls, .ppt
- 仕様の詳細が明らかでなかったため、MS社以外によるドキュメント利用が困難
  - ✓別のOS・機種における閲覧可能性が低い
  - ✓検索エンジンによるテキスト検索などでの応用
  - ✓社内文書の管理、検索においても問題
- 公的部門における調達への準拠（グローバル化への対応）

14

オフィス文書における ドキュメントフォーマットの標準化 (ODF)

- 2000年にOpenOffice.orgがオープンソースソフトウェアとして公開されると、MS Officeの代替としての期待が高まった。同時に、データ交換のためにも、オープンフォーマットへの期待、ドキュメントフォーマットの互換性への期待も高まった。
- 2002-2005年にかけて議論され標準化された。
- Open Document Format (ODF)
  - ✓国際標準ISO/IEC 26300:2006

15

オフィス文書における ドキュメントフォーマットの標準化 (OOXML)

- ODFの動きに呼応して、Microsoft側も、Office 2007リリースにあわせ、新しいドキュメントフォーマットを提案、標準化
- Office Open XML (OpenXML, OOXML) の提案
- 2006年：ヨーロッパの標準化団体ECMAにおける標準化
  - ✓ECMA-376
- 2008年：国際標準化機構ISOにおける標準化
  - ✓ISO/IEC 29500

16

OOXML と ODF

- 国際標準化に向けた競争の激化
- 両者のドキュメントフォーマットは基本的には、かなりの部分で類似のもの
  - ✓（しいて言えば、OOXMLの方がより複雑：Microsoft Officeが持つ機能全体に対応）
- 3つの主要な文書種別をカバー
  - ✓ワードプロセッシング、スプレッドシート、プレゼンテーション
  - ✓それぞれ Word/Writer, Excel/Calc, Powerpoint/Impress に対応
  - ✓ODFフォーマットの拡張子：.odt, .ods, .odp
  - ✓OOXML新フォーマットの拡張子：.docx, .xlsx, .pptx
    - WordPressingML, SpreadsheetML, PresenationML
- 複合オブジェクトを許す汎用形式
  - ✓パッケージ形式（Zip）
  - ✓XML（コンテンツ+メタデータ）
  - ✓マルチメディアファイル

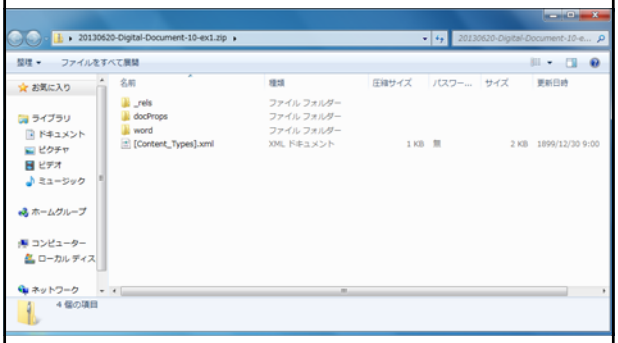
17

OOXMLにおけるパッケージ形式

- Zip形式によるアーカイブ化＋圧縮
  - ✓XML形式以外のファイル（画像等）の埋め込み、展開
  - ✓テキスト形式であるXML文書の圧縮によるファイルサイズ縮小に貢献
- 任意のWord文書等の拡張子を .zip に変更してZipファイルとして展開、閲覧してみる
  - ✓[Content_Types].xml
  - ✓_rels/.rels
    - *word*/document.xml , …
    - *docProps*/core.xml , …

18

OOXMLパッケージの展開例



19

OOXMLパッケージの展開例 _rels/.rels

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<Relationships
xmlns="http://schemas.openxmlformats.org/package/2006/relationships
">
 <Relationship Id="rId3"
Type="http://schemas.openxmlformats.org/officeDocument/2006/relati
onships/extended-properties" Target="docProps/app.xml"/>
 <Relationship Id="rId2"
Type="http://schemas.openxmlformats.org/package/2006/relationships/
metadata/core-properties" Target="docProps/core.xml"/>
 <Relationship Id="rId1"
Type="http://schemas.openxmlformats.org/officeDocument/2006/relati
onships/officeDocument" Target="word/document.xml"/>
</Relationships>
```

20

OOXMLパッケージの展開例 word/document.xml

21

デジタル化テキスト（デジタル翻刻）

22

文書のデジタル化

- 紙の資料をデジタル化する
  - ✓画像として撮る
    - 例：近デジ，ScanSnap
  - ✓テキスト（文字）としてデータ化
    - 例：青空文庫，Project Gutenberg
- 前者と後者をつなぐ
  - ✓OCR（Optical Character Recognition）
  - ✓大規模デジタル化（Mass-digitization）
- 利点と欠点
  - ✓画像：撮影の手間は低いが、全文検索などはできない
  - ✓テキスト：人的コスト（手間、労力がかかる）

23

デジタル化の例：NDLデジコレ

http://dl.ndl.go.jp/info:ndljp/pid/959448/40

デジタル化テキスト

- 青空文庫に代表されるように、過去に出版された書籍などを電子化し、テキストを作成する動きは数多くある
  - ✓青空文庫，Project Gutenberg
  - ✓Amazon「なか見！検索」，Googleブックス，HathiTrust，（NDLデジタルコレクション）
  - ✓Internet Archive
- 2000年代を通じて、大規模デジタル化（Mass-digitization）はひとつのトレンドとなった
- ここでは、紙の時代から蓄積されてきた文書の電子化（デジタル化）とそのテキスト化を考えてみたい

25

デジタル化テキストの利点

【文書の多様な活用】

- 端末やブラウザの多様化：PC，スマートフォン，タブレット，ブラウザ，アプリ
- 全文検索
- 再利用・活用
  - ✓第3者による再販
  - ✓書き下し
  - ✓音読，オーディオ，点字
- 学術研究：コーパス研究の進展
  - ✓古来から、様々な古典籍に出現する用語や概念の関係をまとめることは人文科学における重要な仕事であった
  - ✓Google n-gram
    - https://books.google.com/ngrams
    - http://www.ted.com/talks/what_we_learned_from_5_million_books TEDxBoston, 2011-07

26

一言でデジタル化といっても…

- レベルは色々
  - ✓文字
  - ✓テキスト（文字列の集まり）
  - ✓版組
- どこからがデジタルドキュメント?
  - ✓画像で撮る（スキャンデータ）

27

テキストとは?

- 先に授業で触れたように、文字として認識し、符号化した集まり
- しかし、そもそも「文字」とは何だろうか?
  - ✓一種の文字「概念」を意識する必要がある

28

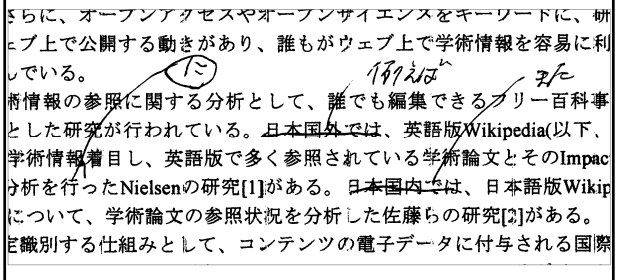
文字とテキスト (1)

- どの文字はどの文字か?

出典：小林龍生．ユニコード戦記：文字符号の国際標準化バトル．東京電機大学出版局，2011, p.37.「図2.1 ホネホネ問題」より

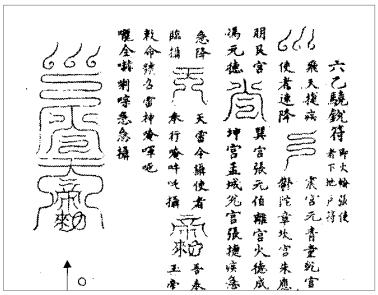
文字とテキスト (2)

- （未刊行のとある原稿にて…）



30

文字とテキスト (3)



出典：林宏美, 宇陀則彦, 松本浩一, 二階堂善弘. 道教資料「道法會元」の電子化実験. 情報知識学会誌, 2001, Vol.11, No.4, p.36-45.「図4」より一部改変

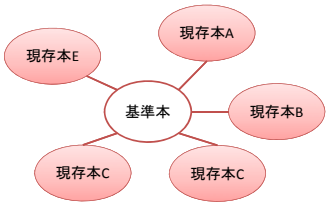
文字とテキスト (4)

- 活字
  - ✓言語、時代、地域による差異
  - ✓字体、字形、フォントなどの差異
- 手稿
  - ✓文字単位を区分することそのものが難しい
    - そもそも読める人が限られる場合も多い
- 元テキストの配置や構成
- 文字であるか否か判別しがたいもの
- バージョン間の差異

32

テキスト研究における翻刻

- 異版、バージョン間の参照の研究
  - ✓校合
- 書誌学、人文学、歴史学、思想史など



出典：安形麻理.デジタル書物学事始め：グーテンベルク聖書とその周辺. 勉誠出版, 2010, p.163.「図4-2 基準本との校合の手順」より改変

TEI (Text Encoding Initiative)

- 1987年に設立
- テキストのデジタル化のためのマークアップ手法の確立、標準化（ガイドライン）
  - ✓人文学、言語学、計算機科学の各領域から
  - ✓デジタル人文学（Digital Humanities）
- もともとSGMLに基づく応用
- 最新のガイドライン: TEI P5
  - ✓XMLに基づく応用
  - ✓様々な文書種別に対応する要素集合

34

TEIガイドラインの構成 (TEI P5)

1. The TEI Infrastructure
2. The TEI Header
3. Elements Available in All TEI Documents
4. Default Text Structure
5. Characters, Glyphs, and Writing Modes
6. Verse
7. Performance Texts
8. Transcriptions of Speech
9. Dictionaries
10. Manuscript Description
11. Representation of Primary Sources
12. Critical Apparatus
13. Names, Dates, People, and Places
14. Tables, Formula, Graphics and Notated Music
15. Language Corpora
16. Linking, Segmentation, and Alignment
17. Simple Analytic Mechanisms
18. Feature Structures
19. Graphs, Networks, and Trees
20. Non-hierarchical Structures
21. Certainty, Precision, and Responsibility
22. Documentation Elements
23. Using the TEI

35

デジタル文書の編集作業

- 紙の文書に対して行っていたような編集作業をデジタル文書に対しても行えるだろうか
  - ✓意外と大変…
- 普通のワードプロセッシング環境を考えてみる
  - ✓紙面情報上の特定のテキスト列に対する編集操作を記録し、改変を指示するマークアップを付与
    - データ表現レベル：該当テキストの（論理構造上または物理構造上の）位置を特定し、そこにタグを付与する
  - ✓適切な位置特定のための画面表示をリアルタイムに更新しながら、編集結果を保持
  - ✓編集操作の改変操作の履歴も保持

36

ドキュメントフォーマットの切り口（再掲）(1)

- テキスト (text) vs バイナリー (binary)
  - ビットデータ
  - 文字コードによる解釈
  - 外字
- フォーマットの指定・識別・判別
- シンプルコンテンツ vs 複合メディア
  - 埋め込みコンテンツ
  - ハイパーリンク
- メタデータ
  - 埋め込みメタデータ
  - 外部メタデータ記述
- 文書レイアウト
  - ページ概念
- 文書内の書式要素
  - 見栄え / スタイル
  - フォント
- ファイル vs ストリーム
  - データの保存・蓄積と配信

37

ドキュメントフォーマットの切り口（再掲）(2)

- オープンフォーマット
  - 移植可能性 / ソフトウェア独立性
  - Free / proprietary
- 標準化
  - デファクト標準とデジュール標準 (“de facto” vs “de jure”)
- 文書フォーマットのバージョン
- フォーマット変換
  - プレインテキスト → HTML
  - LaTeX → PDF
- 圧縮
  - 可逆 vs 非可逆
- セキュリティ
  - パスワード
  - 電子署名
- 長期保存
- デジタルフォレンジック

38

まとめ

- オフィス文書とデジタルドキュメント
  - ✓ワークフロー
  - ✓処理モデル
  - ✓文書フォーマット：OOXMLとODF
    - 標準化（独占規格から国際標準へ）
    - パッケージ形式による複合オブジェクトの埋め込み
    - 文書内容にあわせた文書要素
- デジタル化テキスト
  - ✓大規模デジタル化
  - ✓テキストと文字，デジタル翻刻
  - ✓ドキュメントフォーマットTEI

39