

電子メール長期保存の課題： ePADDDの有用性と問題点

堀内暢行（国文学研究資料館研究部）

橋本陽（京都大学大学文書館）

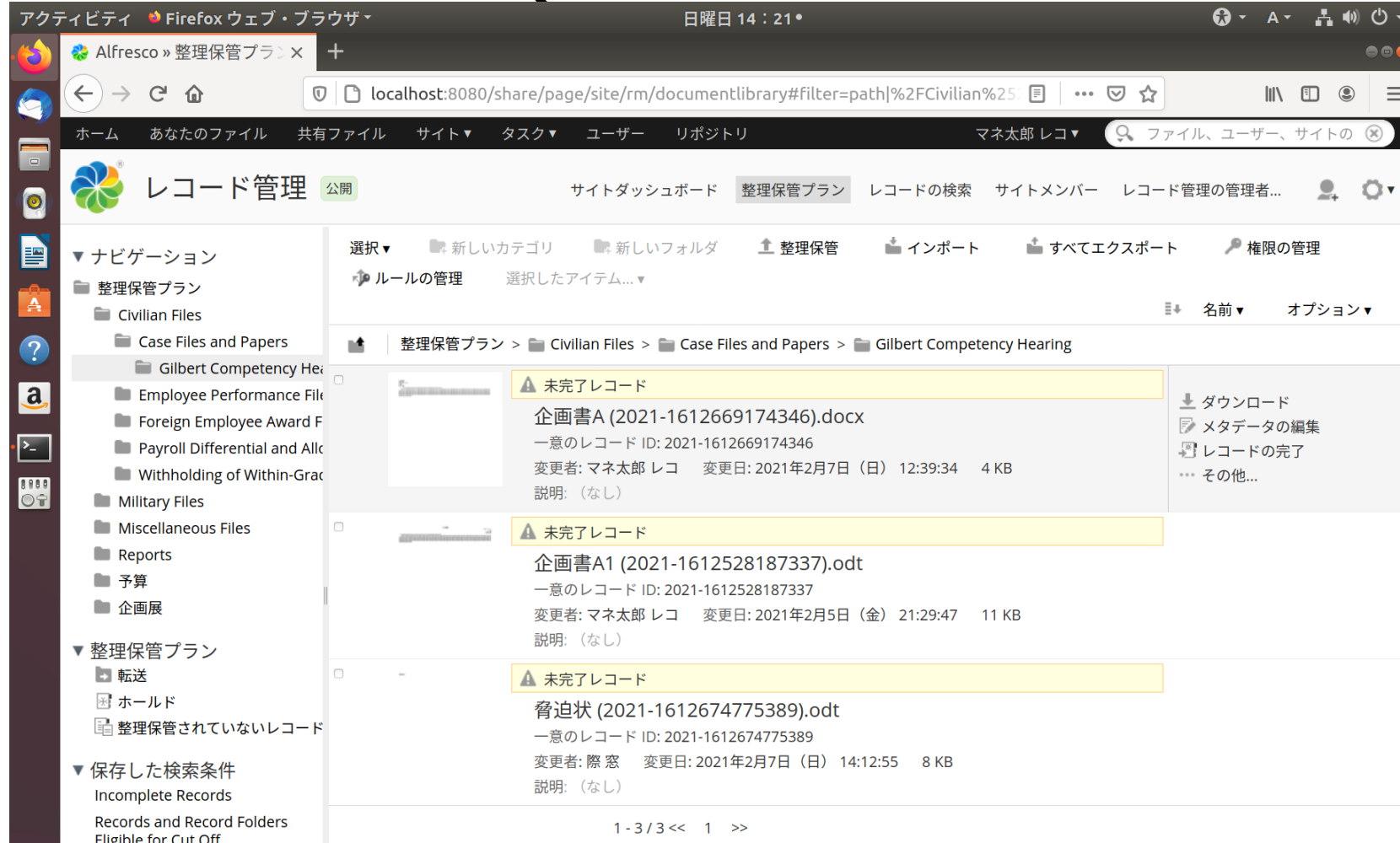
1. はじめに：本報告のテーマ

- デジタル・アーカイブ：MLAを包摂
- 本報告はA（＝アーカイブズ）の視点
- アーカイブズ：記録の集まり；長期保存記録の保有機関
- 記録とは
 - 法的義務を遂行または業務を実行する中で、組織または個人が、証拠および資産として作成、受領し保管する情報（ISO 15489, 2016）
- 電子メールは記録の一種
- 電子メールの長期保存に関する課題と対策を考える

1. ISO 15489の定める要件

- 記録のもつ特質
 - 真正性：記録が真；作成（送信）者が真；作成（送信）時間が真
 - 信頼性：内容が業務を表しているか
 - 完全性：改ざんされていないか
 - 利用可能性：検索して発見できるか
- 記録のためのメタデータ
 - 記録の作成、管理、利用を可能とする情報
 - 記録 = 内容 + メタデータ
- 記録の補足（作成した情報を管理システムへ取り込むこと）
 - a. 一意の識別番号（Unique Identifier、UI）の付与
 - b. メタデータの補足と作成
 - c. 他の記録、作成者、業務との関連性の作成

2. EDRMS (画面はアルフレスコ)



- * 記録のためのメタデータ（例：xmlファイル）が付与され続ける
- * デジタル・フォレンジックスの観点からは、オープンソースであることが望ましい

2. 夢破れたEDRMS

- 21世紀初頭は、夢のレコマネ・ツール
- 2010年になると…
 - 他の製品が台頭、市場の競争力を失う
 - 万能薬ではないと認識される
- 明らかになった弱点
 - 複雑、コスト (Lappin, 2010)
 - 使うのは人間→文化・慣習
 - 特に電子メールは補足されない

2. アーカイブズにとっての問題

- 保管の連鎖（Custodial Chain）
EDRMS（ISO 15489準拠）



アーカイブズの長期保存システム（ISO 14721準拠）

- EDRMSで捕捉できないと記録として信用できない
- 電子メールを記録として長期保存できる枠組みの必要性
 - メール・アカウント単位を対象
 - Capstone
 - ソフトウェアの開発→ePADD

3.1 ePADDDの基本的特徴

- Stanford Univ. Lib.がOpen Source Software (OSS) として開発・公開
→ユーザーは無料で利用 | 改変が可能
- 対応OS
Windows・MacOS・Linux (Ubuntu16.04)
→汎用性が高く、ユーザーを限定しない
- MBOX形式で出力可能
→アーカイブズ学 (ISO 14721) に基づく保存に対応
- Lexicon、NLPを実装
→大量のメールデータの整理作業を軽減

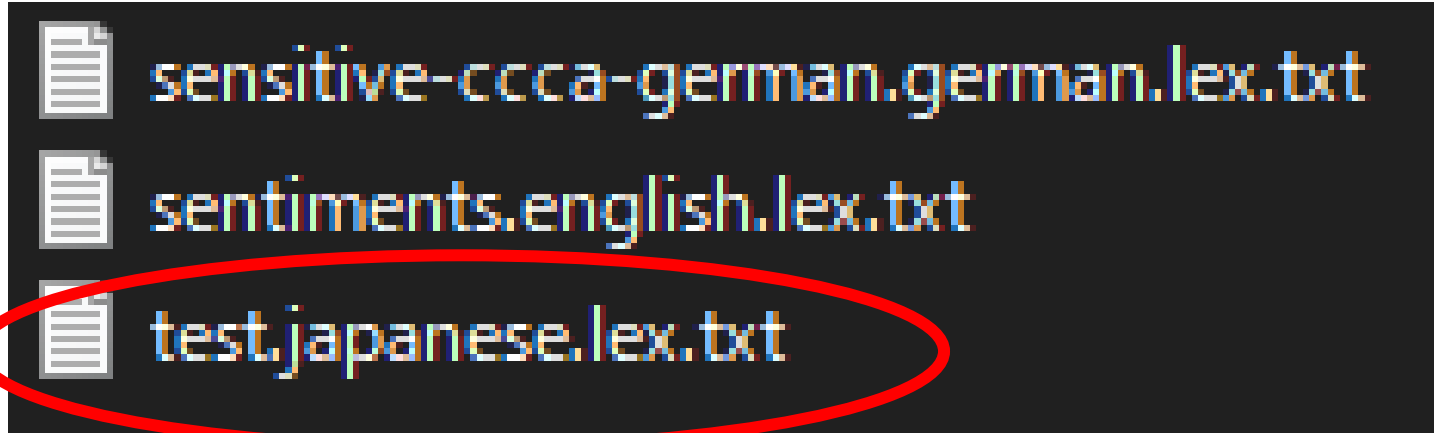
3.2 各種モジュール

- 評価選別(Appraisal)：保存処置を行う前に寄贈者・アーキビストとともにデータの内容を確認し、評価選別作業を行うことができる。
- 整理(Processing)：メールデータを整理し、アーカイブズに必要な記述を付すことができる。
- 検索(Discovery)：制限下のユーザーがメールの事項群などを検索することができる。
- 提供(Delivery)：閲覧室などでユーザーがメールコレクション情報にアクセスすることができる。

4.1 日本語化—Lexicon

epadd-appraisal> user> data> lexicons
内に”test.japasene.~”を作成

ここでは、テスト用に左の語彙を作成し保存。



test.japanese.lex.txt - メモ帳

ファイル(F) 編集(E) 書式(O) 表示(V) ヘルプ(H)

家族:妻|父|母|姉|姪|
職場:大学|国士舘|国文研|国文学研究資料館|
仕事:調査|史料館|資料館|アーカイブ|航空
プライベート:銀行|F X|”on line store”|beams|保険|

4.1 日本語化—Lexicon

Appraisal | Lexicons

test-collection ⓘ

⊕ | ⓘ | Search:

Lexicon	Number of Categories ▼
persona.environmental.artist.projects.stanford	50
sentiments	42
sensitive	9
persona.academic.administrator.sensitive.duke	8
sensitive-cca-german	8
persona.composer.nypl	6
persona.faculty.uci	4
persona.writer.theater.nypl	4
sensitive-cca-french	4
test	4
persona.author.princeton	3
persona.microbiologist.uluc	3
persona.journalist.activist.politics.and.travel.ucb	2
regex	2

Showing 1 to 14 of 14 entries

sensitive-cca-french

test

persona.author.princeton

作成した“test”ディレクトリが
反映されていることを確認

4.1 日本語化ーLexicon_testモジュール

Appraisal | Lexicon Hits: test

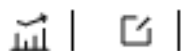


test-collection

test.japanese.lex.txt スモ帳

ファイル(F) 編集(E) 書式(O) 表示(V) ヘルプ(H)

家族:妻|父|母|姉|姪|
職場:大学|国士舘|国文研|国文学研究資料館|
仕事:調査|史料館|資料館|アーカイブ|航空
プライベート:銀行|FX|"on line store"|beams|保険|



Search.

Lexicon category	Messages ▼
プライベート	1136
職場	805
仕事	533
家族	8

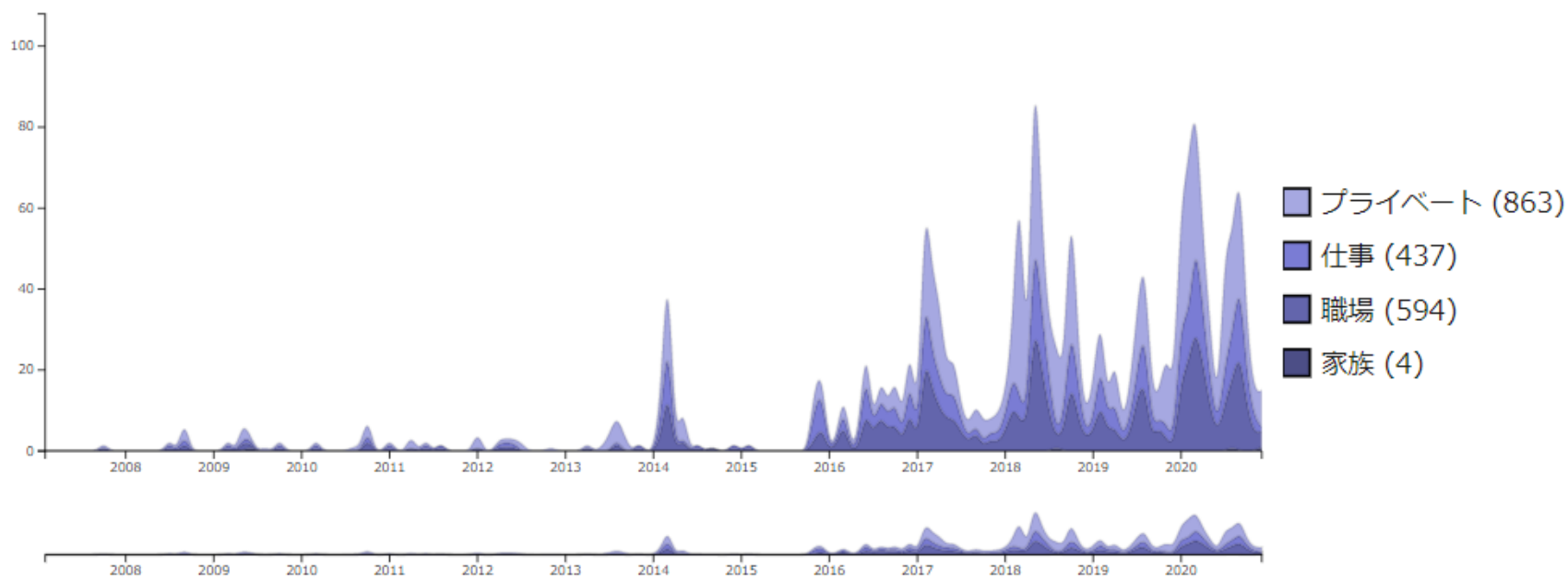
Showing 1 to 4 of 4 entries

4.1 日本語化ーLexicon_testモジュール

Appraisal | Lexicon Graph: test



test-collection



4.2 日本語化ーNLP

他の言語でも同じように作ると簡単に対応できる

src> java> edu> stanford> muse> ner> dictionary> EnglishDictionary.java

```
* A language related util class, created so as to be able to easily expand to other languages by just creating and using a similar dictionary for the language*/
public class EnglishDictionary {
    private static final String adverbsFile = "dictionaries/en-pure-adv.txt";
    private static final String adjFile = "dictionaries/en-pure-adj.txt";
    private static final String verbsFile = "dictionaries/en-pure-verbs.txt";
    private static final String prepFile = "dictionaries/en-prepositions.txt";
    private static final String pronounFile = "dictionaries/en-pronouns.txt";
    private static final String abbFile = "dictionaries/en-abbreviations.txt";
    private static final String tagDictFile = "dictionaries/ANC-written-count.txt";
    private static final String fullDictFile = "dict.words.full";
    private static final String dictStatsFile = "stats.txt";
    private static final String commonNamesFile = "firstNames.txt";
}
```

NLPの日本語化

- NER(Named Entity Recognition)モデルにDbpediaのインスタンスファイルを実装。
- 追加学習＝日本語対応させるためには学習用データファイル（例えばDbpedia Japanese）を準備し、NER_READM_md.docxに書かれている手順でHOME> epadd-settings> SeqModel_ser.gzに追記することで可能。
～> ner> SequenceModel.javaで追加学習内容が反映される

！！BUT！！

日本語対応NLPの結果を反映させるためには、関連プログラムを書き換える必要があることから容易には追加できない。（例：**EnglishDictionary.java**

5. おわりに

- 本報告のまとめ
 - 電子メールのアカウントに着目した保存対策
 - ePADDの利用を検討
- 今後の課題
 - ePADDの日本語化と運用における課題
 - 日本語対応に向けて関連箇所の洗い出しとプログラムの変更
 - メール所蔵者に合わせたLexicon項目毎のファイル作成
 - 電子メール以外の記録、作成者、業務との関連性が捕捉できず、ISO 15489の要件を満足できない