資源としての文化
第10回

福田一史

大阪国際工科専門職大学

https://scrapbox.io/fukudakz/資源としての文化

図. 講義ウェブサイトリンク（QRコード）
※LMSの資料のページにもリンクがあります

授業計画

回次	タイトル
1	ガイダンス・文化資源の定義
2-3	文化資源研究の系譜と基礎概念
4-6	デジタルアーカイブ
7-9	メタデータ
10-12	文化資源データの分析
13-14	事例研究
15	総括

コンテンツ

1. メタデータの検索と活用（振り返り）
2. テキストマイニングによる文化資源データ分析

メタデータの検索と活用

前回の課題と復習

前回の課題：データ検索演習（３）

- これまでに得た知識や、サンプルクエリも参考にして、あなた独自のジャパンサーチのSPARQLのクエリを作ってみよう。
  - サンプルクエリは、ジャパンサーチのクエリ入力欄右（画面構成によっては下段）にリストとして表示されています。
  - 作成したクエリについて、**1)クエリ結果が生成されたページのURL**と、**2) そのクエリの説明（どのような機能を持つか）**を、コミュニケーションノートの「作成したクエリ（URL）」と「作成したクエリの説明」から投稿してください。

Wikidata Query Service

- https://query.wikidata.org/
- LODのデータサービスであるWikidataの**SPARQLエンドポイント**
- クエリ・ヘルパーが設定されており、直接SPARQLのクエリを書かずとも、半自動でクエリを生成してくれる。複雑なクエリはかけないが、基本的検索は十分だし、エラーも少ないので役立つ。
  - 「フィルター」はWHEREのパターン指定、「表示」はOPTIONALで特定の属性を表示する。

分野特化 vs 分野横断

- WikidataとJapan Searchはそれぞれ別のデータ構造を持つデータセット
  - Wikidataは百科事典で扱われる事実や概念を対象とする
  - Japan Searchは図書館や博物館などが構築するデジタルアーカイブを対象とする
- それぞれのサービスで、それぞれの目的に沿って、データモデルやメタデータスキーマが設計される
  - 分野特化：リッチな構造と検索機能・その他のデータセットとの接続性が低い
  - 分野横断：プアな構造と検索機能・抽象的／総合的・分野特化データセットのハブになる
- 何にでも使えるデータ構造・データセットは存在しない。ただし、ID（URI）を用いて、接続していくことでデータの相互交換と総合的検索／識別／分析が可能となる。

まとめ

- RDFは**主語**と**目的語**およびそれらの関連を示す**述語**という3つの要素（トリプル）からなる、グラフで表現される抽象構文のモデル。
- RDFは**URI**を用いたウェブにおけるグローバルなデータ交換枠組みであり、Linked Open Dataの基盤技術として用いられている。
- RDFで記述されたデータは論理的な関係性が明示され、URIを通じて直接的なアクセスを提供可能である。さらに、その問い合わせ言語である**SPARQL**を用いて**リッチな検索やデータ取得**が可能となる。

テキストマイニング による文化資源データ分析

With KH Coder

テキストマイニング

- 計量テキスト分析
- テキスト（文章・文字データ・文献など）を対象とするデータマイニング。
- 以降、Scrapboxコンテンツを参照のこと
  - テキストマイニング
  - KH Coder

図. 夏目漱石「こころ」の抽出語リスト
http://khcoder.net/ スクリーンショット集より

データ分析演習（準備編１）

- 分析のテーマを決める。
  - 自分が興味を抱く／取組中のテーマやトピックに関連する、デジタル形式でテキストを取得可能な（a.k.a. コピペできる）データを設定する。

データ分析演習（準備編２）

- 分析対象とする**日本語テキストデータ**を収集し、分析用ファイルを作る。
  - 文字数基準：10,000字以上、上限なし（PCのスペックによる）
  - 入力可能ファイル形式：**テキストファイル**（.txt）、CSVファイル（.csv）、MS Excelファイル（.xslx）
    - この講義ではテキストファイル形式の作成を推奨します
  - 分析しやすいように段落や属性などの外部変数を記録する
    - 「話」や「章」などの見出しで文書を分割する
    - 3～10件に分割するイメージだと分析しやすい
    - KH Coderを使って分析するためには、どのようにデータを準備すればよいですか？

テキストファイルの保存と見出しの付け方

- テキストファイルとは、文字データだけで構成されたファイル。「メモ帳」や「サクラエディタ」「VS Code」「EmEditor」などのアプリケーションで作成できる。
- KH Coderでは、**見出し**を特定の形式で記録することで、章や節とその内容を構造化することが可能
  - 右は見出しによる構造化の事例（本文）

```
<h1>一　午後の授業</h1>
「ではみなさん、さういふふうに川だと云はれたり、乳の流れたあとだと云はれたりしてゐた、このぼんやりと白いものが何かご承知ですか。」……

<h1>二　活版所</h1>
　ジヨバンニが學校の門を出るとき、同じ組の七八人は家へ歸らずカムパネルラをまん中にして校庭の隅の櫻の木のところに集まつてゐました。…（以下続く）…

<h1>三　家</h1>
　ジヨバンニが勢よく歸つて來たのは、ある裏町の小さな家でした。…
```

ファイルの作成と提出

- 本日、演習用ファイルを提出してください。
  - 演習用はScrapboxからダウンロードできます。
- 今回は、書けるところまで書いてから、提出してください。
  - **準備編１**（テーマ設定）か、**準備編２**（分析ファイルの出所と特徴）まで。残りは次回以降の講義で分析してから書きます。保存しておいてください。
- 講義中に、**準備編２**まで進められなかった場合は、**来週**の講義までにファイルを作成してください（**!!!宿題!!!**）。