DBpedia Japanese

Hiroki UEMATSU（上松大輝）
総合研究大学院大学 / 国立情報学研究所
hiroki_u@nii.ac.jp

DBpediaとは

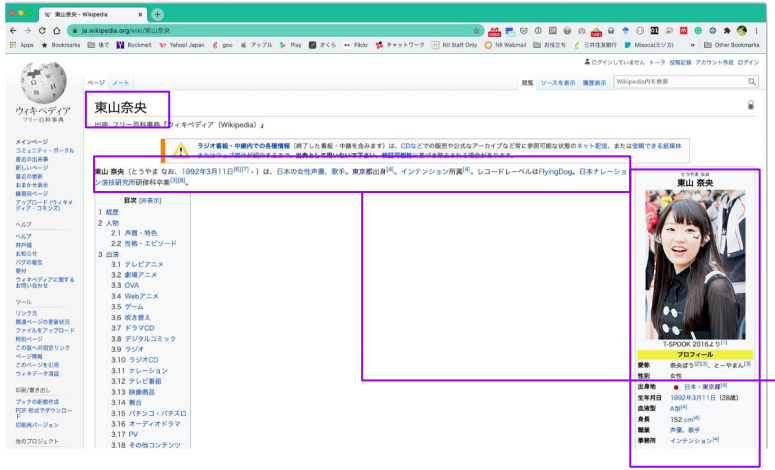
DBpedia Project

- 目的
  - Wikipediaから構造化データセットを抽出
  - フリーの知識ベースを構築し、公開

- 本家DBpediaでは英語版Wikipediaのデータを公開
  - https://www.dbpedia.org/
- 各言語のWikipediaのデータはチャプターと呼ばれる有志が作成して公開
  - 日本語版：http://ja.dbpedia.org/
  - スペイン語版：https://es.dbpedia.org/
  - フランス語版：http://fr.dbpedia.org/
  - etc

WikipediaとDBpedia

- Wikipediaの情報を構造化したDBpedia



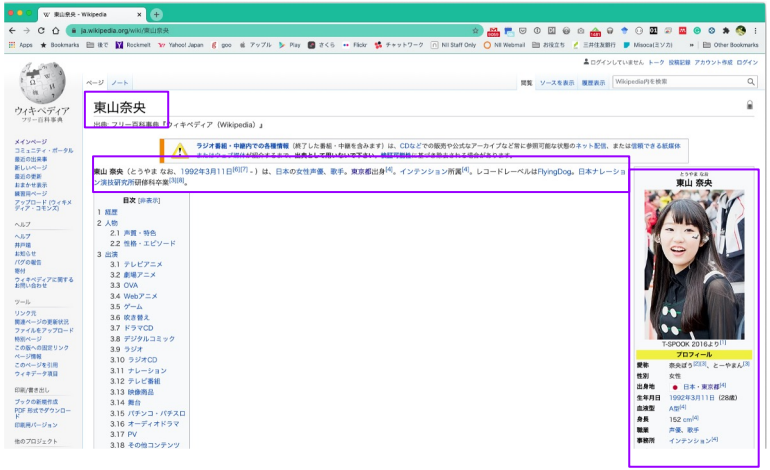
項目名	内容
名前	東山奈央
職業	女性声優、歌手
出身	東京都
…	…

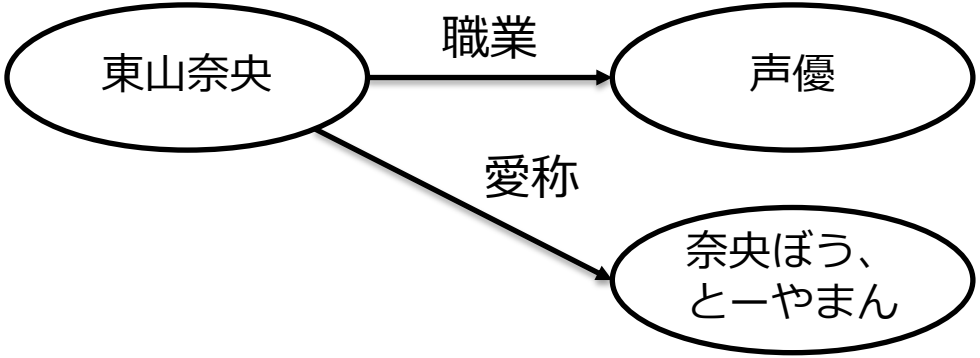
記述されている内容を解釈する必要がある

東山 奈央（とうやま なお、1992年3月11日[6][7] - ）は、日本の女性声優、歌手。東京都出身[4]。インテンション所属[4]。レコードレーベルはFlyingDog。日本ナレーション演技研究所研修科卒業[3][8]。

DBpediaが扱うデータ

- Wikipediaの各記事の関係性（リンク）
  - タイトル、概要、インフォボックスから抽出





- 構造化された知識
  - RDF（Resource Description Framework）による記述
  - Linked Dataに沿ったRDFを用いたデータモデル

Linked Open Data

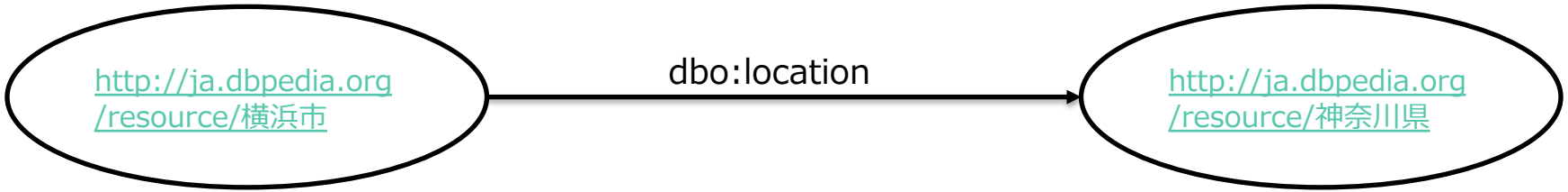
- オープンなLinked Data
  - Web上のデータをリンクさせる取り組み
    - 関係性を示したリンクを作成
- Linked Dataの原則
  - あらゆるデータの識別子としてURIを使用する。
  - 識別子には（URNや他のスキームではなく）HTTP URIを使用し、参照やアクセスを可能にする。
  - URIにアクセスされた際には有用な情報を標準的なフォーマット（RDFなど）で提供する。
  - データには他の情報源における関連情報へのリンクを含め、ウェブ上の情報発見を支援する。

参考：Linked Open Data（Wikipedia）, 2020.03.06

Linked Dataの構造

- RDFを用いて記述
- <主語><述語><目的語>の3つ組で簡単に記述

<http://ja.dbpedia.org/resource/横浜市> dbo:location <http://ja.dbpedia.org/resource/神奈川県>



DBpediaでの検索のコツ

- リソース名は同じ
  - WikipediaのURLとDBpediaのURLは似ている
    - https://ja.wikipedia.org/wiki/東山奈央
    - http://ja.dbpedia.org/resource/東山奈央
- インフォボックスの中身が構造化
  - 同じジャンルのリソースは同じデータを持つ（ことが多い）
    - Wikipediaではジャンルごとにインフォボックスの構造が決められている
- Wikipediaでリンクになっている情報は、dbo:wikiPageWikiLink にまとまっている（ことが多い）
  - 逆の構造の場合もあり
    - ?s dbo:wikiPageWikiLink dbpedia-ja:東山奈央
- まずは検索したいもののページを見て、使えそうな項目を探す
    - http://ja.dbpedia.org/page/東山奈央

DBpedia Japaneseの仕組み

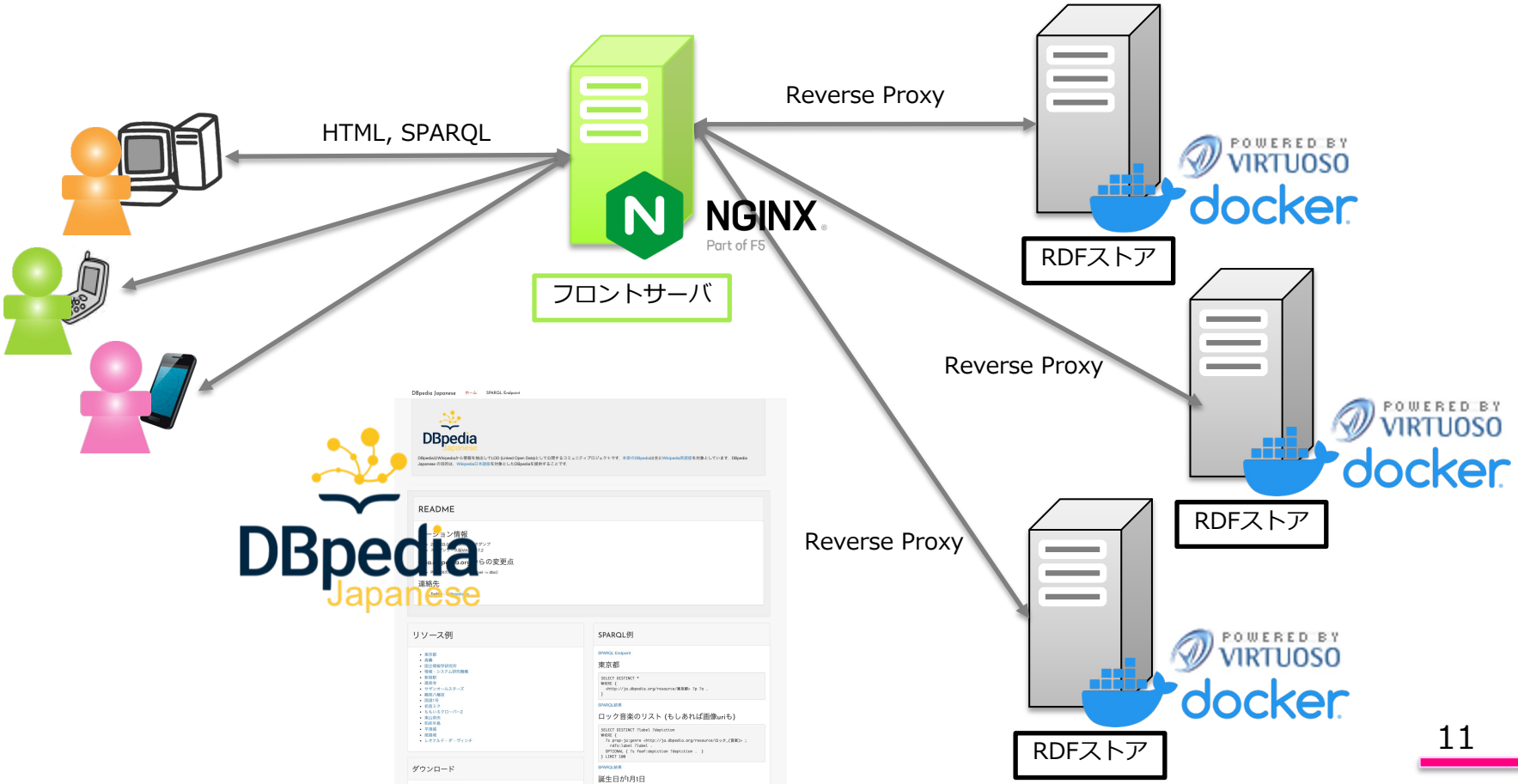
現行版

- https://ja.dbpedia.org/

- 日本語版Wikipediaのデータを使用
  - 2022.03版のExtractionデータを使って公開
  - Docker + databusを用いたChapter（世界初）
  - ~~前バージョンの最終更新は2016.04~~

- 初版は2012年公開
  - 日本語版独自でWikipediaのデータを変換
    - Extraction Frameworkを改修して利用

システム構成


HTML, SPARQL
Reverse Proxy
NGINX
Part of F5
フロントサーバ
RDFストア
Reverse Proxy
RDFストア
Reverse Proxy
RDFストア
POWERED BY VIRTUOSO
docker
DBpedia
Japanese
README
リソース例
SPARQL例
ダウンロード

DBpediaクローンの起動

ウィキペディア
フリー百科事典

DBpedia Information Extraction Framework

```
######
#     #   ##   #####   ##   #####  #    #  ####
#     #  #  #    #    #  #  #    # #    # #
#     # #    #   #   #    # #####  #    #  ####
#     # ######   #   ###### #    # #    #      #
#     # #    #   #   #    # #    # #    # #    #
######  #    #   #   #    # #####   ####   ####
```

作成したデータを保存、公開
グラフごとにコレクションを作成

https://github.com/dbpedia/virtuoso-sparql-endpoint-quickstart

POWERED BY VIRTUOSO
docker

Virtuoso（RDFストア）を
ベースにしたコンテナを利用

Docker版DBpedia

- VirtuosoベースのRDFストア
  - Github:Dockerized-DBpedia
- Store, Load, Downloadの3構成
- config.envを設定して起動
  - パスワード、ポートなど
  - DataのCollection URI
    - https://databus.dbpedia.org/hiroki_u/collections/dbpedia-ja_latest
  - virtuosoのdataフォルダ
- Dataフォルダに保存されたトリプルをインポート
  - DownloadがDatabusのCollection URIからトリプルを取得
  - Dataフォルダにトリプルを置いても自動でインポート
    - 拡張子に注意（nt nq owl rdf trig ttl xml gz bz2）

起動の手順

- DATABUSにてCollectionを作成
  - または、インポートしたいトリプルを用意
- Docker版DBpediaをclone
  - `git clone git@github.com:dbpedia/virtuoso-sparql-endpoint-quickstart.git`
- envファイルを修正
  - パスワード、Collection URI、ポート、メモリなど
    - 過去のCollecition
      - https://databus.dbpedia.org/hiroki_u/collections/dbpedia_ja-202105
      - https://databus.dbpedia.org/hiroki_u/collections/dbpedia_ja-202203
- Loadコンテナを作成
  - `cd ./dbpedia-loader`
  - `docker build -t dbpedia-virtuoso-loader .`
- Store, Download, Loadを起動
  - `docker-compose up`
  - 時間がかかるので注意
    - Storeのみ起動している状態になったら完了
    - Download, Loadは処理が終了したら自動でdown

まとめ

- DBpedia Japaneseの紹介
- Linked Open Dataの（簡単な説明）
- DBpedia Japaneseのシステム構成と構築手順
- ~~SPARQLを用いたDBpediaへの問い合わせ~~

- DBpedia Japanese では
  - 不具合、改善点等の報告お待ちしてます

twitter：@dbpedia_ja