2L1-1

地域研究のための資料メタデータ編纂システムの構築

Developing a Metadata Compilation Platform for Area Studies Research Materials

亀田 尭宙 *1
KAMEDA Akihiro

*1京都大学 東南アジア地域研究研究所
Center for Southeast Asian Studies, Kyoto University

Various kinds of research materials are shared and used in area studies. To search, aggregate, arrange, and publish a list of materials, we developed a metadata compilation system. This system suggest additional information based on original metadata, too.

1. はじめに

近年、研究データの共有は、研究活動を促進するための重要なテーマとして取り組まれている．例えば、京都大学理学研究科附属地磁気世界資料解析センターが開催しているオープンサイエンスデータ推進ワークショップは 2015 年からはじまり現在までに四回開催され、研究データの共有に関わる議論が行われている．代表的な研究データリポジトリの一つである figshare*1 は 2012 年以降 200 万以上の研究データを蓄積し共有しており、ダウンロード数も 750 万を超える．生命科学を中心として急激に成長している Linked Open Data (以下、LOD) も新たな研究データ共有の一つの形と考えてよいだろう [Abele 17].

一方、研究プロジェクト単位、機関単位でのウェブへのデータベース公開は以前より行われてきており、日本の代表的な研究ファンドである日本学術振興会の科学研究費情勢事業では、少なくとも 1999 年に日本学術振興会へ審査・交付業務が移管された時点から研究成果公開促進費としてデータベースの作成・公開について支援してきている *2.

2017 年 1 月から統合によって京都大学東南アジア地域研究研究所となった旧京都大学地域研究研究所（以下、旧地域研）は共同利用・共同研究拠点として「地域研究情報資源の統合と共有化」を目標の一つとして掲げ、地域研究における研究資料のデータベース化を支援してきた．地域研究資源共有化データベース *3 は、各機関に分散した地域研究関連データベースの統合検索を目指している．本システムにより、旧地域研の１７のデータベースと外部の３４のデータベースの統合検索を可能にしている．また、その旧地域研での１７のデータベースは My データベースシステムというデータベース構築システムで構築されており、csv 形式での表構造のデータのアップロードと検索に関わる幾つかの設定を行うことで、通常の語彙検索に加えて、地図を利用した空間検索とタイムラインを利用した時間検索を含めた検索のインタフェースと、それに対応する検索の API を提供するデータベースシステムになっている．

本稿では、そのように多数の研究データがウェブ上に公開されてきている現状を踏まえ、個々の研究者の研究目的のもとで文献を含めた地域研究資料群としてまとめ上げることを支援するシステムについて紹介する．具体的には、各データベースでの検索結果を JSON の形で URL を付与して保存し、それらを結合したり、アイテムを絞ったり、コメントを加えたり、カテゴリのタグを加えたり、項目値を外部の LOD の URI に紐づけたり (このタスクは Entity Linking と呼ばれる) といったことを可能にした (図 1).

以下、システムの詳細について述べていく．

2. 資料メタデータ編纂支援システム

2.1 API からの返答を JSON として書き出す

My データベースシステムの API は SRU (Search/Retrieval via URL) *4 という図書館のデータベースに使われる標準に従って作られており、URL のクエリ文字列に検索クエリを埋め込むことで JSON での返答を受け取ることができる．従って、その API に対して検索を行い結果を受け取って HTML にレンダリングする検索ページに、検索結果の保存の機能を付け加えることでまず初めの JSON を生成することは可能になる．

2.2 JSON の加工

オープンソースの JSON エディタ *5 に前述の JSON を読み込むことで、その編集を可能にしている．また、機能拡張として、入力した文字列とスキーマに基づく外部 LOD の URI の推薦と分類のためのタグの推薦を行うようにした (図 2) ．URI の推薦は rdfs:label のリテラルを前方一致検索しており、分類のためのタグはデータ群の中で中程度の文書頻度 (Document Frequency) を持つものを推薦するようにしている．

このエディタは JSON Schema *6 を用いてスキーマを定めることができ、その際に Entity Linking 先の候補も絞り込むことで適切な推薦を可能にしている．

2.3 LOD としての再発信

JSON のデータを JSON スキーマの情報を元に RDF として RDF ストアに保存する機能を備えている．しかし、この変換は試験的であり、さまざまな地域研究資料に対して実際に適

連絡先: 亀田 尭宙，京都大学 東南アジア地域研究研究所，〒 606-8501 京都市左京区吉田下阿達町 46，kameda @ cseas.kyoto-u.ac.jp

∗1 https://figshare.com/. 各種の数字は 2017 年 3 月 15 日現在のもの．figshare のサイト上における情報.

∗2 https://www.nii.ac.jp/sparc/event/2012/pdf/20120725_2.pdf 日本学術振興会 研究事業部: 科学研究費補助金（研究成果公開促進費）学術定期刊行物の現状について．2011 年 10 月 7 日.

∗3 http://app.cias.kyoto-u.ac.jp/GlobalFinder-lg/cgi/Start.exe

∗4 http://www.loc.gov/standards/sru/

∗5 https://github.com/jdorn/json-editor Jeremy Dorn: JSON Schema Based Editor

∗6 http://json-schema.org/

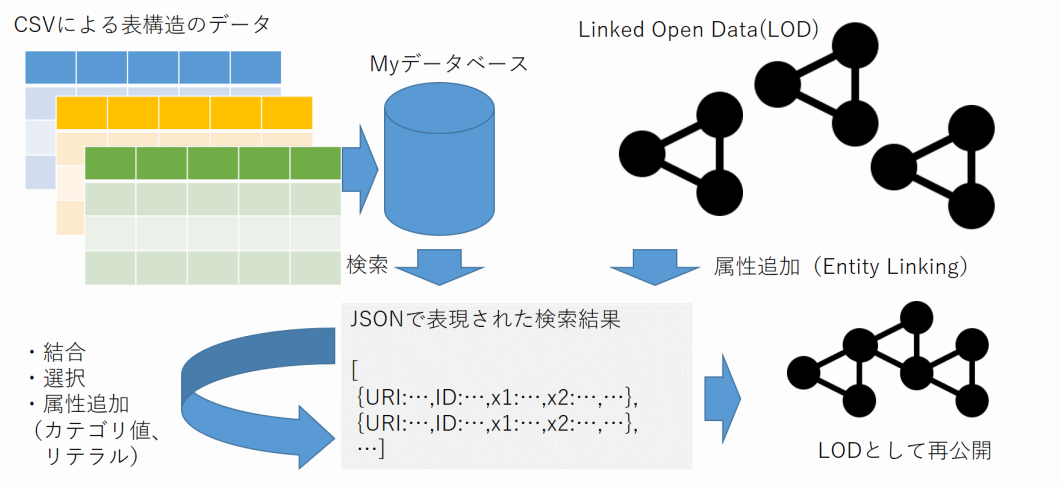

図 1: システムの概要

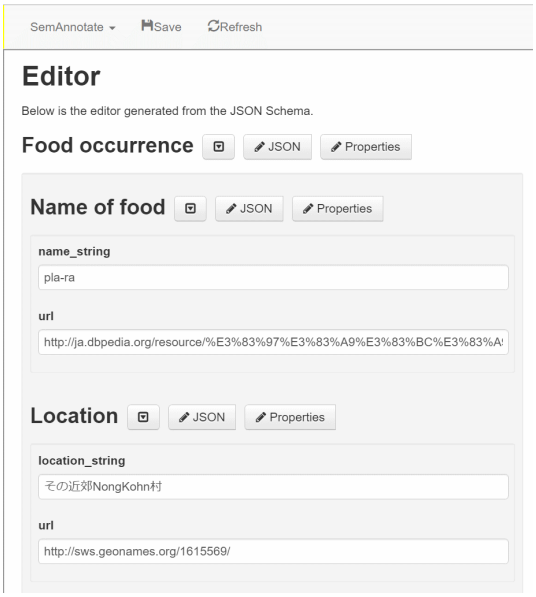

図 2: JSON の加工

用しつつ、適切な変換を模索したいと考えている．例えば、現在、JSON で基本的なデータ構造である配列には順序の意味があるが、RDF で順序付きの配列を表現するのは煩雑であるため順序の無い複数のグラフとして表現し直すなど、利用者の利便性を考えたヒューリスティックスが含まれている．

3. おわりに

地域研究で使われる資料は、メタデータのつけられた絵葉書からテキスト化された雑誌本文に至るまで多様であり、学際的であるゆえ関連文献も複数の学会やドメインに分散して存在する．これらの情報を個々の研究者のニーズに応じて検索、集約し、整理し、再発信するためのメタデータ編纂システムについて紹介した．類似のシステムとしては、画像データの共有の標準となっている IIIF で表現された画像アイテムをコレクションとして編纂するシステム [Kitamoto 16] があげられる。このようなシステムの評価手法は定まっていないが、今後の運用を通して、ユーザからのフィードバックや編纂されたメタデータの質などを評価したいと考えている．

参考文献

[Abele 17] Andrejs Abele, John P. McCrae, Paul Buitelaar, Anja Jentzsch and Richard Cyganiak: Linking Open Data cloud diagram (2017) `http://lod-cloud.net`

[Kitamoto 16] 北本 朝展, 山本 和明: 人文学データのオープン化を開拓する超学際的データプラットフォームの構築 (2016), 人文科学とコンピュータシンポジウム じんもんこん 2016, pp. 117-124