Vol.2019-CH-119 No.12
2019/2/16

東アジア絵葉書データベースのシステム設計

亀田尭宙[†1], 貴志俊彦[†1], 原正一郎[†1]

概要: 京都大学東南アジア地域研究研究所では, 戦前戦中に発行された東アジアの絵葉書をデータベースとして整理・公開している．これまで国際連携のために， Linked Open Data や International Image Interoperability Framework に対応した公開を進めてきた．また，それぞれの弱点である，ドメイン研究者によるデータの簡便な登録と更新や応答の早い検索 API について，当研究所が構築してきた My データベースや Elasticsearch との連携によって補っている．本稿では，データの具体的な形式やシステム間の連携について詳述し，活用に至るまでの課題について議論する．

1. 背景

複数のデータベースを横断して検索・活用したいという要求は各学術分野で生じている．人文科学とコンピュータ研究会と関連の深い分野では，人間文化研究機構によって提供されている「人間文化研究機構統合検索システム nihuINT （nihu INTegrated retrieval system, 以下, nihuINT）」[a] や東南アジア地域研究研究所（以下，東南地域研）によって提供されている「地域研究資源共有化データベース」[b] がある．また，生命科学分野においては，ライフサイエンス統合データベースセンターが行った統合データベースプロジェクトの成果として，生命科学分野の論文の横断検索から蛋白質の横断検索まで多様な統合検索が提供されている[c]．また，分野横断的な情報リソースを扱う図書館関係では, 国立情報学研究所が提供する「CiNii Books」 [d] は全国 1,300 以上の大学図書館などが所蔵する図書，雑誌，古典籍などの学術資料約 1,150 万件の情報をさがすことができるサービスを提供している[e].

幅広い種類のリソースを横断的に検索する場合には，タイトルや著者といったリソースに共通の要素に関するメタデータ構造を策定し，そのメタデータ構造に合わせたメタデータを各参加機関に提供してもらう，もしくは各参加機関の持っているメタデータ構造を共通のメタデータ構造にマッピングするという手段が取られる．例えば，欧州委員会が運営するデジタルアーカイブの統合検索ポータル Europeana の場合，Europeana Data Model （EDM）[f] を策定し，そのデータモデルに各機関のデータモデルをマッピングするためのマッピングガイドラインを提供することで，質を担保した統合検索を可能にしている．

本論文では，東南地域研がラファイエット・カレッジ，ハーバード・イェンチン図書館と共同で構築している東アジア絵葉書データベース（以下，絵葉書 DB）のシステム設計について紹介し，同様の人文系画像資料共有のためのシステム設計の参考となる知見を提供したい．

2. 東アジア絵葉書のデータ

第二次世界大戦の戦前・戦中に発行された東アジアの絵葉書は各大学やコレクターの間で分散しているのが現状である．そのうちのいくつかはデジタル化されオンラインで検索できるようになっている[g][h].

絵葉書のデータであるのでもちろん画像が重要な役割を果たすが，それとともにタイトルや写真の内容，発行された時期などのメタデータも，その絵葉書を用いた研究にとって重要な役割を果たす.

表 1 はラファイエット・カレッジ側の絵葉書のメタデータの一例である（一部データを省略している）．項目数が多く詳細に構造化されている．また，OCM [i] のサブジェクトカテゴリについては 1 つの絵葉書に対して複数が割り当てられるようになっており，先方でのデータのエクスポートの際にそれはセミコロン区切りで表現されているなど，一部，表形式では構造を扱いきれない部分がある．また，日付について，可能性を点ではなく範囲で示していたり，絵葉書の日付と写真の日付を別々に扱っていたりするため，日付だけで複数項目存在する．多くの情報は英語で書かれている.

表 2 は東南地域研側の絵葉書のメタデータの一例である．構造はシンプルで，タイトルや場所といった情報は共通しているものの，時期は範囲ではなく点での表現になっているなどラファイエット・カレッジ側のメタデータ項目と 1 対 1 に対応するわけではない.

絵葉書の画像データについては，ラファイエット・カレッジは JPEG2000 で，東南地域研は普通の JPEG で持っている．対象とする画像の印刷解像度が高いわけでもなく，サイズも葉書サイズであるので，一枚当たりの画像サイズについて大きいものが求められるわけではない.

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

a) http://kyoyusvr.rekihaku.ac.jp/GlobalFinder/cgi/Start.exe
b) http://mydatabase.jp/GlobalFinder/cgi/Start.exe
c) http://lifesciencedb.jp/
d) https://ci.nii.ac.jp/books/
e) 2017 年版 CiNii パンフレットより
f) http://pro.europeana.eu/page/edm-documentation
g) 東南アジア地域研究研究所 戦前期東アジア絵はがきデータベース
http://app.cias.kyoto-u.ac.jp/infolib/meta_pub/G0000022PPC
h) East Asia Image Collection, Lafayette College
https://dss.lafayette.edu/collections/east-asia-image-collection/
i) Topics Covered (OCM Subjects) | Human Relations Area Files
http://hraf.yale.edu/resources/reference/outline-of-cultural-materials/

3. 同テーマ異構造データベース間統合

テーマが多様なデータベースの統合検索を扱う場合は，前述の Europeana のように幅広い項目をカバーする共通のメタデータスキーマを用意するか，逆に nihuINT のように検索に使うのは限られたメタデータスキーマに絞り，表示の際に詳細なメタデータを示すかといった方法が取られる[1].

しかし，ある固定のテーマでデータベースを統合する場合には，ただリソースを発見したいというだけではなく，研究の分析などに利するために求められるため，より詳細なデータを手がかりとして提供してほしいという要求が生じる．一方で，研究それぞれの研究者はそれぞれの興味に基づいてデータベースを構築しているため，メタデータの構造は前述のようにそれぞれのデータベースで大きく異なる．

こういった同テーマ異構造データベース間の統合に関して，Linked Data が技術として適切であると考え，Linked Data 化を行った．

Linked Data [j] は RDF（Resource Description Framework）のデータモデルを用いて書かれたデータをウェブ上で共有し，つなげる技術である．The Linking Open Data cloud diagram [k] で示されているように多様なデータが Linked Data として公開され，つながっている．

例えば，博物館情報に関する Linked Data である LOD.AC プロジェクト[2][l]は，各博物館のデータを収集し共通のオントロジにマッピングすることでデータベースの統合を実現している．また，生物情報の追加 [3] や絶滅危惧種情報の追加 [4] など各種情報をスキーマのデザインとともに追加している．

絵葉書 DB では，対応する項目がある場合，例えば，「タイトル」と「title_japanese」をそれぞれのデータベースが述語として持っていた時に，それら２つが一般的な語彙である dcterms:title のサブプロパティであると記すことで，タイトルに対する検索が横断的に可能になる．これは，それぞれのデ

表 1　絵葉書のデータ例（ラファイエット・カレッジ）

title_english	[ob0021] [Yueyanglou Pavilion]
title_chinese	[岳阳楼]
title_japanese	岳陽楼
subject_ocm	340 STRUCTURES;341 ARCHITECTURE;530 ARTS …
description_text_japanese	古島松之助筆
description_ethnicity	Chinese;Japanese
coverage_location_country	China
coverage_location	Hunan Province
format_medium	Picture postcard
description_indicia	陸軍需品本廠;軍事郵便;済閲検;[printed in ora …
creator_maker	Kojima Matsunosuke
creator_company	Rikugun Juhin Honshō (Main Workshop for …
relation_seealso	[oa0044]
contributor	Li Guo
date_original	
date_artifact_upper	1941/7/7
date_artifact_lower	1937/7/7
date_image_upper	
date_image_lower	
date_search	
identifier_dmrecord	4926
relation_ispartof	East Asia Image Collection;Postcard Albums; …
format_digital	Master TIF image captured at 4000 pixels across …
publisher_digital	Special Collections & College Archives, Skillman …
rights_digital	This image is posted publicly for non-profit edu…
creator_digital	The East Asia Image Collection is a joint project ,,,
project_name	pa-omitsu02
item_number	21
object_url	https://digital.lafayette.edu/collections/eastasia/…
object_url_front_jpeg	https://digital.lafayette.edu/collections/eastasia/…
object_url_back_jpeg	https://digital.lafayette.edu/collections/eastasia/…

表 2　絵葉書のデータ例（東南地域研）

場所	東京
タイトル	（大正博覧会第一会場）正門
袋のタイトル	上野不忍池畔の美観
作者	
出版者	松本幸盛堂（東京市神田）
時期	1914/3/
ファイル名_頭文字	EX
備考	「大正3年3月東京大正博覧会」のスタンプあり。
フォルダ名	J040906
画像	http://app.cias.kyoto-u.ac.jp/infolib/www/data…
サムネイル	http://app.cias.kyoto-u.ac.jp/infolib/www/data…

j) オープンであるという条件も付随して Linked Open Data として言及されることも多いが，ここでは本稿では個々のリソースに関するライセンスなどの議論を行わないため Linked Data で統一する．

k) Linking Open Data cloud diagram 2018, by John P. McCrae, Max Schmachtenberg, Christian Bizer, Anja Jentzsch and Richard Cyganiak. http://lod-cloud.net/

l) http://lod.ac/

ータの詳細度が異なっている場合により有効になる．例えば既に述べた時間情報について，おおよその時期が知りたい場合は，時間にかかわる多くの述語を１つの述語のサブプロパティとみなして検索することになり，細かく時間が知りたいときは，適切な少数の語彙だけを選択してまとめあげるほうが有効であろう．rdfs:subPropertyOf などスキーマに相当するような関係を SPARQL 検索の対象となるグラフに適宜入れることで，こういった検索の粒度を変えることができる．例えば，RDF データベースの Virtuoso では，”define input:inference <スキーマの URL>” というプラグマを指定することによって推論に使うオントロジを組み込むことができる[m].

実装にあたっては，まず各々の機関のメタデータを csv ファイルの形で用意し，その中には絵葉書の画像の URL を記述しておく．次に，csv ファイルの方を，Virtuoso に扱いやすいインタフェースを被せた InfoLib-LOD [n]に読み込ませ，データセット間の比較に基づいて定義した述語を割り当てることで Linked Data 化した．

Linked Data 化は，既に存在する知識構造を再利用することで，外部の知識と関連付けることにつながる．例えば，絵葉書の写真の撮られた時期は，そこに写っている事物や袋の情報などから，絞り込んでいくことができる．暦などの時間の構造を Linked Data として表現し，曖昧な時間の知識の定式化に取り組んでいる体系として HuTime があり[5]，今後，時間表現を HuTime につなげていきたいと考えている．

4. 絵葉書統合検索の実装

一方で，こういった詳細で自由度の高い検索を行う前に，まずキーワード検索ベースで画像を検索・比較したいといったニーズが絵葉書資料を用いる研究者から挙げられた．SPARQL はリテラルの一部の文字列で検索するといったことが苦手で一般的にスピードが遅く，全文検索エンジンの方が自然言語による検索は得意である．そこで，RDF ストアに入れている表形式のデータを json に変換して全文検索エンジンである Elasticsearch と同期することで，キーワードによる画像検索を実現した．その際は構造を問わず，自然言語で書かれている箇所はすべて検索の対象とすることで，複数のデータベースをまず統合で検索してブラウジングするということを可能にした[o].

5. IIIF 対応

また，近年，International Image Interoperability Framework (IIIF)を用いた人文系の画像データベースの構築が盛んである[5][6][7].

東南地域研の絵葉書データも Image Server に Digilib[p], ビューワーに Mirador[q], Presentation API は前述の csv をスクリプトで変換したものを用いて構築することで IIIF での画像提供を始めた．Presentation API は Linked Data のシリアライゼーション形式の１つである JSON-LD で表現されているため，当初は既に構築した Linked Data をリクエストごとに SPARQL を介して IIIF Presentation API の形式に変換する実装にしていた．こういったアプリケーションをすぐに作れることも Linked Data の強みである．ただし，現在は，直接 csv と画像のセットから Presentation API 用の json ファイルをスクリプトで生成するようにし，rdfs:seeAlso 語彙で IIIF のメタデータと Linked Data の URL を関連付けるように変更している．

(1) IIIF のデータには画像のサイズ情報を含める必要があり，それを Linked Data の方にも含める手間が画像サイズの変更などを通して面倒になった
(2) 今後 IIIF の提供を別組織に依頼することを検討中であり，サービス同士を直列につなげると Linked Data 側の変更が IIIF のメタデータ生成に影響し，IIIF の機能維持のために別組織とのコミュニケーションが必要になってしまうこと

この２つが変更の主な理由である．

連携先のハーバード・イェンチン図書館は既に IIIF 化を行っており，ラファイエット・カレッジも対応を予定しているため，今後３機関の絵葉書を横断した画像のアノテーションや比較がより一層便利になると考えている．

6. 様々なプレゼンテーションへの対応

サイト上では，時間軸上の表現（図１）や地図上の表現など絵葉書の特性を生かしたプレゼンテーションへの対応に取り組み始めている[r]．また，それらの表現や IIIF での表示だけでは，ストーリーを読み取ることは困難であるため，書籍からの引用文や研究者の考察などと組み合わせてストーリーを語るためのインタフェースを用意した（図２）．右側は HTML プレゼンテーションになっており，縦にスクロールすることでスライドをめくることができ[s]，それに応じて左側の画面が IIIF 画像や地図の画面に切り替わるようになっている．

7. おわりに

本論文では、東アジアの絵葉書史料という同じテーマを有しながら異なるデータ構造を持ったデータの統合検索を実現するための課題とその解決策について述べた．

今後，絵葉書 DB 活用のためには，連携先を増やしてデータを増やすとともに，新たな検索や情報提示のインタフ

m) 16. RDF Data Access and Data Management
http://docs.openlinksw.com/virtuoso/rdfsparqlrule.html
n) https://service.infocom.co.jp/das/product/infolib/lod.html
o) http://asian-postcards.mydatabase.jp/search201807/
p) http://digilib.sourceforge.net/
q) http://projectmirador.org/
r) http://asian-postcards.mydatabase.jp
s) https://revealjs.com/

ェースの構築と改善を繰り返していく必要があると考えている．

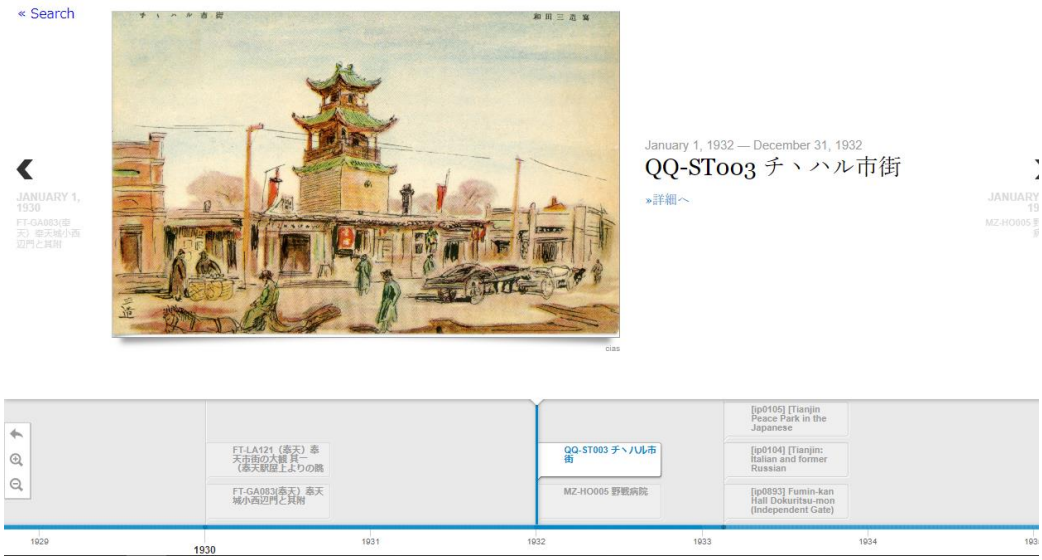

図１　時間軸上に整理した表現

図２　絵葉書の文脈を提示するインタフェース

参考文献

[1] 山田 太造，山本 泰則，古瀬 蔵，安達 文夫: 人文科学データベース統合検索のためのメタデータとその応用．じんもんこん 2012 論文集，no. 7，p. 71-78，2012.

[2] 嘉村哲郎，加藤文彦，大向一輝，武田英明，高橋徹，上田洋: LOD.AC: Linked Open Data によるミュージアム情報の結合，第 3 回知識共有コミュニティワークショップ，情報社会学会，2010.

[3] 武田英明，南佳孝，加藤文彦，大向一輝，新井紀子，神保宇嗣，伊藤元己，小林悟志，川本祥子：生物情報基盤構築のための生物種データの Linked Open Data 化の試み，人工知能学会全国大会（第 26 回）論文集，No. 3，C2-OS-13b-3，2012.

[4] Akihiro Kameda, Fumihiro Kato, Utsugi Jinbo, Ikki Ohmukai, Hideaki Takeda: Integrate Japanese Red List into LOD of Species, PNC Annual Conference and Joint Meetings 2013, 2013.

[5]関野 樹: Linked Data におけるあいまいな時間の記述，じんもんこん 2018 論文集，pp. 303-308，2018.

[6]橋場 天紀，三原 鉄也，永森 光晴，杉本 重雄: マンガの内容と構造のメタデータ記述を利用した IIIF に基づく検索・閲覧環境の構築，研究報告人文科学とコンピュータ（CH），2018-CH-116, vol.12, pp.1-5, 2018.

[7]永崎 研宣，下田 正弘, Muller A. Charles，蓑輪 顕量: 横断型デジタル学術基盤を目指して―SAT2018 の構築を通じて―，研究報告人文科学とコンピュータ（CH），2018-CH-117, vol.1, pp.1-7, 2018.

[8]吉賀 夏子，只木 進一，伊藤 昭弘: 小城藩日記データベースの構築，研究報告人文科学とコンピュータ（CH），2018-CH-117, vol.3, pp.1-7, 2018.