文化財報告書データベースにおけるテキスト可視化と地理情報

高田祐一[1] Yanase Peter[1] 武内樹治[2]
[1]奈良文化財研究所 [2]立命館大学大学院
takata-y23@nich.go.jp yanase-p7g@nich.go.jp
mikihdqt812@gmail.com

概要

文化財そのものは位置情報を保持し，関連する調査成果はテキスト媒体である．成果は報告書としてまとめられ報告書データベースには 27 億文字が登録されている．データベースは全文検索可能であるものの巨大であるため，人間可読が難しく機械で読んでいくことが求められている．テキストの頻出語や地域ごとの特徴語の算出によって可視化した．また海外でも日本の成果を位置付けていくために国際連携のためのシソーラスマッピングを推進している．さらに文化財 WebGIS の構築によって文化財位置情報 61 万件を地理データとして扱えるようになり，文化財テキストと地理を組み合わせた新たな可能性の基盤となる．

1 全国遺跡報告総覧

奈良文化財研究所（以下，奈文研）は，文化財報告書を全文電子化しインターネット上で検索・閲覧できるようにした全国遺跡報告総覧（以下，遺跡総覧[i]）を 2015 年 6 月から運営している（図１）．遺跡総覧には，全国の文化財調査機関が報告書類の書誌および全文 PDF を Web 登録する．PDF にはテキストが含まれることから，遺跡総覧では，登録データに全文検索が可能である．2023 年 1 月 13 日時点で，PDF がある書誌数 33,893 件，PDF410 万ページ，総文字数 27 億文字に対して全文検索が可能である．従来は，書庫に籠り，１ページずつ必要とする情報を探すという地味な作業であったため，劇的に資料調査が効率的かつ網羅的になった．書誌情報には，国立国会図書館の JP 番号，CiNii Books の NCID を付与し，各書誌の DB へリンクを設定しているので，実際の図書の所蔵館を調べることも可能である．遺跡総覧の利用状況として，2021 年度には 197 万件の PDF のダウンロード，9997 万件のページ閲覧数があった．

図 1 全国遺跡報告総覧トップページ

1.1 巨大データを可視化する

全文検索は至便であるもののテキスト量が巨大であるため，人間可読に適さない．そこで，機械的にテキストを可視化するために，遺跡総覧の内部には，文化財関係用語シソーラスを保持している．シソーラスに登録された文化財専門用語をもとに，テキスト解析を実施し，用語を切り出している．奈文研が独自開発した言語リソースである．専門の語彙数としては 190594 件であり，そこによみや類義語，多言語対訳を付与している．

このシソーラスを活用することで，報告書に記載されている文化財関係用語で頻出用語や地域独自の特徴語を可視化できる（図２）．報告書ごとに特徴語を算出することで，内容の類似度に応じたサジェスト機能も実装している．

報告書全体のテキストに対し，考古学関係用語の出現回数を集計し，図化した．報告書ワードマップ[ii]という機能で閲覧できる．用語をクリックすると

[i] https://sitereports.nabunken.go.jp/ja

[ii] https://sitereports.nabunken.go.jp/ja/visualization/term

頻出用語のみを検索キーとした高精度検索が可能である．検索語を入力するには事前に専門知識が必要であるが，そういった知識がなくとも選択だけで検索が可能である．また都道府県ごとに考古学関係用語の特徴語を，図化した．当該都道府県内にて頻出する用語（よく使われる用語は重要）かつ他都道府県では出現頻度が低い用語（希少用語は重要）であることを勘案するため，当該都道府県の強い特徴を示す用語を可視化できる．自然言語処理技術のベクトル空間モデルの TF（索引語頻度）と IDF（逆文書頻度）を組み合わせた TF-IDF にて算出した．

全国的な専門用語の使い方と都道府県ごとの地域性を可視化した．

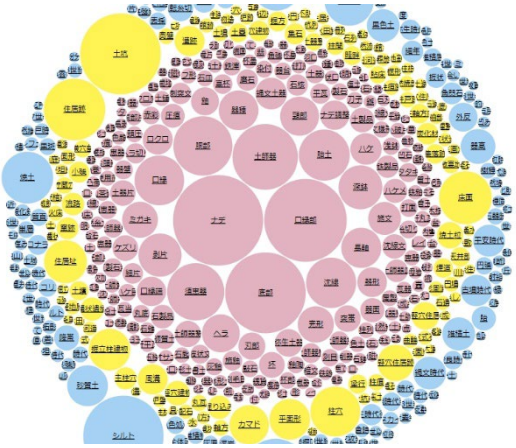

図２　報告書ワードマップ（頻出用語俯瞰図）

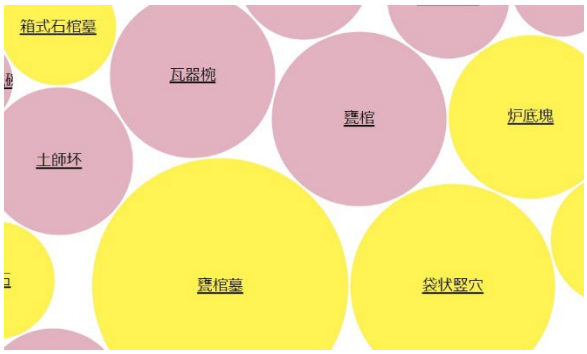

図３　報告書特徴語ワードマップ-福岡県

2 国際連携のための自然言語処理

日本の調査成果を海外からアクセスできるようにするため，多言語化対応を推進している．奈文研は ARIADNEplus という国際プロジェクトに参加している．ARIADNEplus とその前身であった ARIADNE は，欧州委員会の資金を受けて 2013 年から 2022 年までの間に実施された考古学研究プロジェクトであった．その主な目的は，言語の壁と国境を乗り越え，ヨーロッパの諸国の考古学データを整理し，その目録を公開することであった．ARIADNE は本来ヨーロッパ内のプロジェクトであったが，最終的にヨーロッパ圏外の国からの参加機関も関わるようになった．2022 年 12 月の時点で公開カタログには約 350 万件ものデータセットが含まれていた．ARIADNEplus の名義で行われたプロジェクト自体が 2022 年に完了したが，公開カタログの維持と更新はこれからも予定されている．

ARIADNEplus のカタログは「時間」，「空間」と「モノ」で検索ができる．日本のデータをこの国際的なカタログに統合するためには，奈文研には①データのクレンジング，②データベースのスキーマの変換，③データマッピングの三つを行う必要があった．その中で，日本の考古学用語の英語へのマッピングがとりわけ困難であった．ARIADNE は言語を問わない横断検索を実現するために，それぞれの言語・

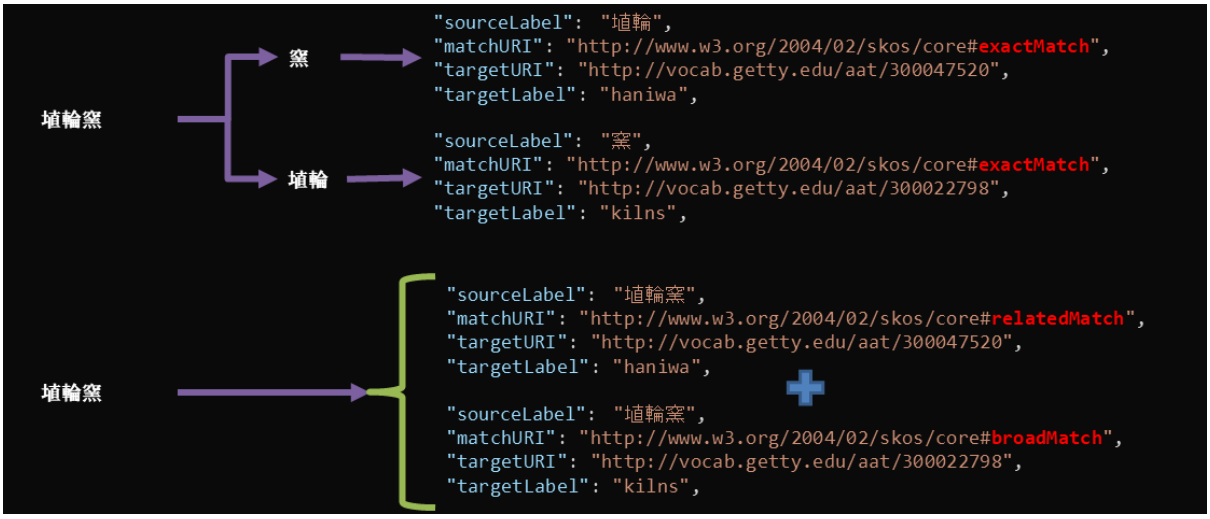

図４　マッピング作業の流れ

分野のシソーラスを米国の Getty 財団が運営している芸術関係用語のシソーラスである Art & Architecture Thesaurus にマッピングさせた．ただ，ご存知の通り，日本ではそもそも芸術・文化財に関する整備されたシソーラスが存在していない．そのうえ，日本考古学の用語は，日本の特有な文化・歴史に由来するものも多いが，それより独自の道を歩んできた日本考古学そのもの文化・伝統に由来する表現が多い．

これらの用語を AAT にマッピングするためにはまず遺跡総覧から抽出した用語をいったん手動でマッピングした．ただ，それだとマッピングの再現性が低いので，次は手動で作ったマッピングを分析し，シンプルな自然言語処理に基づいてもう一度機械的にマッピングを作成した．これによってデータのばらつきがなくなり，再現性も高まった．この処理ができるためには，ARIADNEplus が本来想定していた 1 対 1 のマッピングではなく，1 対多の採用が前提となっていた．また，マッピング作業においてはマッピング・プロパティーも明記する必要があった．そのため，複合名詞に含まれているそれぞれの名詞に対して，その位置に基づいて自動でマッピング・プロパティーを付与した（図４）．

3 文化財総覧 WebGIS

奈文研では，全国の文化財に関する約 61 万件のデータを文化財総覧 WebGIS[iii]（図５）として公開しており，地域・場所に根差した文化財を地図上で調べることができる．これまで報告書などでは位置情報を把握しづらいという課題があったが，この WebGIS によって位置情報を頼りに文化財情報を検索することが可能になった．

3.1 WebGIS に登録されているデータ

ここでは，当 WebGIS に登録されているデータを紹介する．

まず，背景地図として，地理院地図（標準地図や単色地図など），空中写真，色別標高図，傾斜量図などが備わっているのに加えて，奈文研空中写真や平城宮跡の遺構図や地形図も含まれている．兵庫県・静岡県・岐阜県については CS 立体図も追加している．また，土砂災害警戒区域や地すべり危険箇所などのハザードマップも背景地図として表示することができる（図６）．

文化財情報として，文化庁や国土交通省が公開している国・都道府県指定文化財情報や，奈文研作成データ（遺跡データベース，平城宮・跡に関するデータ，遺跡地図データ），地方公共団体が公開している遺跡範囲等のデータなどが登録されている．

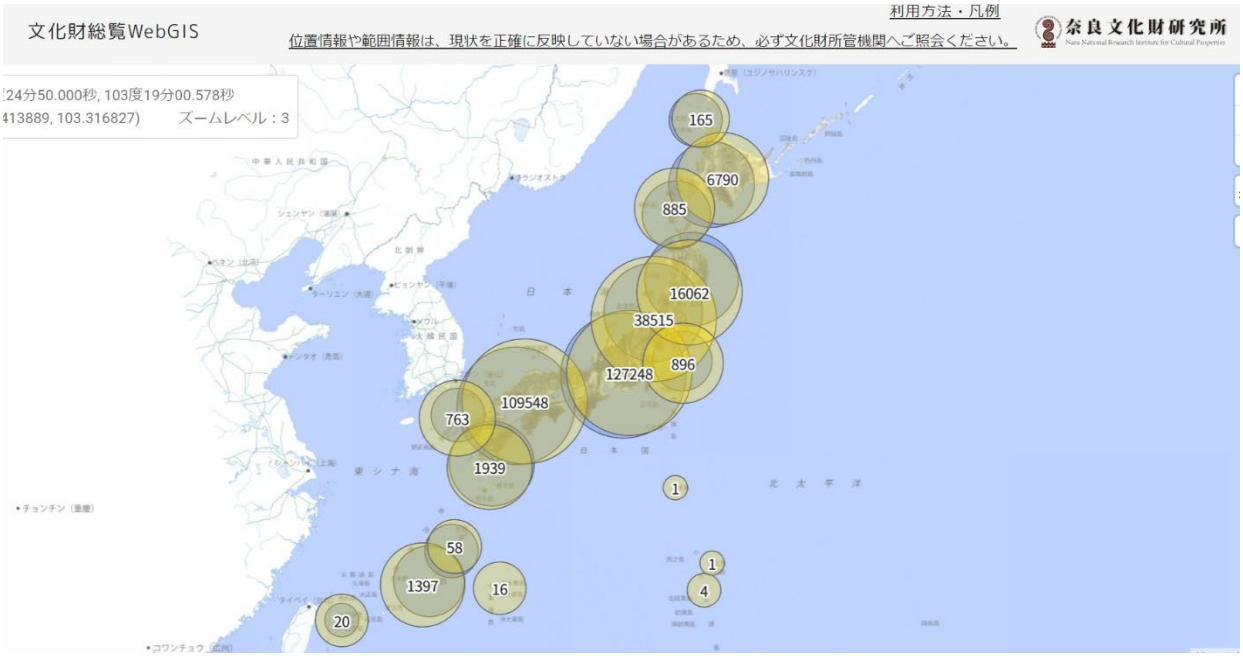

図５　文化財総覧 WebGIS

[iii] https://heritagemap.nabunken.go.jp/

3.2 WebGIS の機能

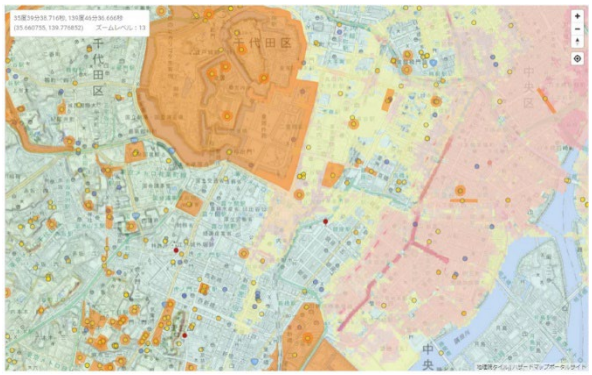

図６　ハザードマップと文化財分布（東京駅周辺）

当 WebGIS では,対象文化財のポイントやポリゴンをクリックすることで，対象文化財の名称や種別などの文化財情報を知ることができるとともに，その文化財の報告書が遺跡総覧で電子公開されているものであれば，当該報告書のページへ遷移し，報告書を閲覧することができる．WebGIS の機能として「都道府県」「種別」「時代」のそれぞれの項目で地図上に表示する対象文化財を絞ることができ，ある地域の古代の集落遺跡のみを表示することなどが可能である．「フリーワード検索」として，文化財名や遺構・遺物の名称などからも検索をかけることができる．背景地図との重ね合わせでは，文化財情報とハザードマップを重ねることで文化財の災害リスクなどを予測することができ文化財防災の観点でも利用可能である．そして，平城宮跡で出土した木簡や墨書土器について，木簡に書かれた内容で検索できるなど,平城宮の調査研究として役立つ機能もある．以上のような機能を用いて，利用者は地域の文化財の再発見や地域学習，学術研究の基盤として活用することができる．

3.3 文化財地理情報と自然言語処理

ここでは今後の文化財総覧 WebGIS の発展可能性について，自然言語処理の応用という点を中心に述べる．

地図上で検索した遺跡の情報をさらに知りたい場合には，報告書があれば閲覧できるものの，必ずしも報告書は万人に対して読みやすいようには書かれていない．そのため，数百字程度で報告書内容の要約を生成し表示する機能などもあれば様々な利用者にとって情報を得やすいシステムになると思われる．さらに，発掘調査報告書では，調査成果などをもとに対象遺跡の付近や関連のある遺跡との関わりが記される場合が多い．報告書で言及された遺跡名を抽出することができれば，ある遺跡を調べた際にその報告書内で言及された他の遺跡を GIS 上に表示し，調べた遺跡と関連する遺跡へと利用者のデータ閲覧の流れを作ることに繋げられ，より深く文化財情報を活用する機会を提供できる可能性がある．さらに，関わりの深い遺跡であれば報告書内で言及されることが多いと仮定すると，その遺跡間の関わりを報告書内の遺跡名の出現頻度を自然言語処理の手法によって算出し，地図上でその関わり度合が表示されることで，誰もが容易に遺跡同士の関連性を知ることができると思われる．