JDCat サロン

データインフラの最前線

横断検索から広がる史資料の魅力

渡邉要一郎（わたなべ・よういちろう）

東京大学　史料編纂所　特任研究員

東京大学史料編纂所にてメタデータの整備に携わられている渡邉要一郎さんに、史料編纂所の取組みや今後のデータアーカイブ構想をお聞かせいただきます。

東京大学人文社会系研究科アジア文化研究専攻　インド文学・インド哲学・仏教学専門分野博士課程単位取得退学ののち、2019 年度より東京大学史料編纂所特任研究員としてデータインフラ事業に従事。

――ご自身の研究についてお聞かせください。

SAT 大正新脩大藏經テキストデータベース研究会のプロジェクトのひとつとして、保有する 1 億字に及ぶ大蔵経電子テキストを TEI（Text Encoding Initiative[1]）に準拠したかたちで構造化することに関する研究を行っています。平行して、電子テキストを、国際標準に沿った形でマークアップするために必要な作業環境の開発・構築等も行っています。

――人社データインフラ事業では、どのような仕事をご担当されていますか？

東京大学史料編纂所内が保有するメタデータを、JDCat メタデータスキーマ[2]や JPCOAR スキーマ[3]に対応した形にするための整備を行っています。また、JDCat メタデータスキーマに準拠したメタデータから IIIF マニフェスト[4]を自動生成し、画像データをより容易に公開できるような仕組みを作ることも模索しています。

――ではまず、史料編纂所が扱うデータの概要についてお聞かせください。

史料編纂所は世界・日本各地に史料調査に出かけ、注目すべき史料について複写を行い、日本史研究や史料集編纂のための素材となる史料を収集しています。近年では、これらの史料データをデジタル画像として所内・閲覧室で検索・閲覧することが可能です。史料データは狭義の意味での日本史研究に限定されないような利用もなされており、収集した史料から過去の地震に関する記述が発見されるといった事例があります。

――異分野での活用を意識した取組みが既に始まっているのですね。

日本史研究に閉じた利用方法だけでは出てこなかっ

---

[1] TEI とは、主に人文学分野のテキストをデジタル形式で表現するための規格を共同で開発・維持するコンソーシアム。主な活動として、同分野のテキストを機械可読形式にエンコーディングする方法を定めたガイドラインの策定がある。

[2] JDCat メタデータスキーマとは、人文学・社会科学総合データカタログ（Japan Data Catalog for the Humanities and Social Sciences：JDCat）で採用されているメタデータスキーマ。人文学・社会科学両方の分野をカバーし、海外関連機関との連携も視野に含めた統制語彙を採用している点に特徴がある。
https://jdcat.jsps.go.jp/JDCatmetadata.html
JDCat は学振が実施する「人文学・社会科学データインフラストラクチャー構築推進事業」の成果の一部で、複数のデータアーカイブのメタデータが一括検索できる。
https://jdcat.jsps.go.jp

[3] JPCOAR スキーマは、オープンアクセスリポジトリ推進協会（JPCOAR）が策定したメタデータ規格。日本の機関リポジトリのメタデータの国際的な相互運用性を向上させ、日本の学術的成果の円滑な流通を図ることを目的としている。
https://schema.irdb.nii.ac.jp/ja

[4] IIIF（International Image Interoperability Framework）とは、デジタルアーカイブに収録されている画像を中心とするデジタル化資料を扱うための国際的な枠組み。画像データと一緒に、IIIF に対応した画像ビューアがコンテンツを扱う際に必要となる一連の情報（IIIF マニフェストと呼ばれます）を提供することで、利用者は画像データを自前の IIIF 対応ビューアに容易く取り込むことが可能になる。

た展開で、利用方法に対して先入観を含めないデータ整備の重要性を感じています。最近では他機関の保持する史資料の画像公開の希望も増え、これまで史料編纂所が扱ってこなかった歴史気象データ等を受け入れる予定です。

関連した取組みとして、日本のみならず海外からの利用を考え、データベースの英訳化・歴史用語の英訳グロッサリーを作成しています。専門用語の翻訳は研究者によって意見が分かれる難題ですが、こういった補助的なデータの整備によって多様な観点での検索が実現できるものと考えています。

――日本語の場合、検索のためにヨミを当てることもありますね。

ヨミでの検索実現は実は悩ましい問題です。昔の人の名前のヨミについて、当時の読み方として音読みと訓読みのどちらが代表的なのか確定できていない名前も多く、ローマ字に変換するとさらに問題は複雑になります。公的な情報公開に耐えられるか、まだまだ検討の余地があると感じています。

――続いて、史料編纂所のデータアーカイブのシステム面に関する取組みについてお聞かせください。

JDCat へのメタデータ提供については、各古文書一点一点のメタデータが史料編纂所内部の規格で作成されているため、このメタデータを JDCat メタデータスキーマに沿うように変換して提供しています。史料編纂所では JAIRO Cloud[5]を利用して DOI[6]を取得することを検討中ですが、例えば資料全体で DOI を付与するのか、古文書一頁毎に付与するのか、決定にあたり課題が残っているように思います。また、メタデータによっては、HuTime API（暦法の変換や暦法に基づく期間の計算を行う機能を提供する API）を用いて和暦を西暦に修正するなどの作業を行いましたが、時間軸が曖昧なもの（例えば江戸中期など）に関して、どのように処理するのかに関しては検討を要しております。

その他、一般に IIIF を導入するにあたり、困難のひとつはサーバーを立てたり、IIIF マニフェストを生成したりする技術的なハードルの高さがあるように思われます。そのため、JDCat メタデータスキーマ対応のメタデータから自動的に IIIF マニフェストを生成でき、GitHub pages[7]を用いて比較的簡単に資料を公開できるような仕組みを考えています。

――史料編纂所のデータアーカイブを持続的に管理していくためには、今後どのような仕組みが必要になるのでしょうか。

汎用的で利用価値の高いアーカイブを構築するためには、安定した規格のもと業務を進められる体制をつくることが大切であると考えます。各機関独自の創意ももちろん重要なことであると思いますが、ともすれば秘伝のようになってしまったデータは、読み解くことが次第に困難になり、分からない箇所は誰に聞いていいかもわからず、継承に困難をきたしてゆくことが想定されます。

――安定した規格の存在、あり方がポイントになりそうですね。

方向性としては 2 つあると考えており、人文系の延長線上に作成するか、あらかじめ一定程度標準化された規格を作っておくか、のどちらかになりそうです。JDCat のように、人文系の人間からすると何だか分からないけれども今後必要になる規格があれば、利用者から使ってみようという意思が生まれてくるはずですので、そのフィードバックをどう確保するのか、とい

---

[5] JAIRO Cloud とは、オープンアクセスリポジトリ推進協会（JPCOAR）と国立情報学研究所が共同運用する、クラウド型の機関リポジトリ環境提供サービス。

[6] DOI (Digital Object Identifier)はデジタルネットワーク上で、コンテンツへのアクセスを管理するために用いられる国際的な識別子。https://doi.org/ に続けて DOI をブラウザに入力することで、自動的にコンテンツの所在情報（URL）に変換されるサービス名称でもあり、登録機関が DOI に紐づく URL のメンテナンスを行うことで、利用者からの恒久的なアクセスが実現される。

[7] GitHub Pages は、GitHub のリポジトリから HTML、CSS、および JavaScript ファイル を直接取得し、任意でビルドプロセスを通じてファイルを実行し、ウェブサイトを公開できる静的なサイトホスティングサービス。

う体制が大事になるのではないでしょうか。

そのうえで、私も人文学出身の人間ですが、特に技術的に強いバックグラウンドが無い人間であっても、少し検索すれば利用方法が分かり、データアーカイブに携われるようであるべきだと考えております。人文学のデータは人文学の人が責任を持って扱っていくべきで、できる限り資料に近い人間が直接にデータを扱った方が、伝言ゲームのような情報伝達の齟齬が起こらないのではと考えます。

――技術的な知識と人文学の知識の両者を持つ橋渡し的人材への期待、ということでしょうか。

自ら開発する程度の知識までは求められないかと思いますが、最低限、どういうイメージでデータを作るか、技術畑の人と会話が出来るレベルになれると良いと思っています。個人的には、最近はプログラミング教育も始まっているので、ある程度自然に人材が確保できるかもしれないと楽観的に見ています。一方で、そのために、研究者としてデータアーカイブに携わる人間を評価する仕組みも必要であるように感じます。

――本事業では、社会科学系データアーカイブとの垣根を超えたデータ利活用が期待されています。JDCat の公開、活用によって、どのような利活用が期待できるでしょうか。

人文系では、機関ごとに優れたデータアーカイブはあるのですが、それを横断する仕組みないし発想は薄いものであったように思われます。その一方で、社会科学では研究手法の統一性を前提にした横断検索が企図されているように見受けられ、従ってそれを軸として各データを横並びにすることが可能になると思われます。各データアーカイブに、実際に横断性があるか否かは別として、仮にそれがあったとして、どのように比較の表の上に研究資料を落とし込むか、という発想自体が人文系にとって面白い視点を与えてくれるのではないかと考えます。

――横断検索という考え方を通じて、人文学の研究手法自体にも影響が及ぶかもしれないということですね。

現在のところ、JDCat で仮想された横断性というものが、カタログの名前空間と統制語彙に他ならないと思われます。さらに言えば、研究手法が同じでも、結果の解釈は人によって異なります。解釈を成立させるための可視化、言説間のネットワークなども作れると面白いかもしれません。

――最後に、今後データアーカイブはどのような役割を果たしていくことが期待されるか、お考えをお聞かせください。

本事業は、従来通りの実証的な研究の基盤となるのみならず、JDCat のように横断的な規格が有効にはたらきはじめることによって、各研究者にとって狭義の専門分野以外の関連性が必然的に目に入る、可視化されるようになるものと思われます。したがって、史料収集の役割だけではなく、発見・再発見のツールとしての役割を果たすことになるのではないでしょうか。各研究者が、思っていたほど重要そうではないデータから新しい発見がある、ということへの期待と、社会科学系との相互作用性を自覚しながらツールを活用できる姿が望ましいように思います。

そして、そのためには、充分なデータのみならず、先入観によって用途を限定しない設計が求められるように感じられます。横のつながりを実現する手段とともに、プラットフォームとして機能する場となれば素晴らしいのではないかと考えております。

（座談会開催：令和３年８月５日／聞き手：南山泰之）