JDCat サロン

データインフラの最前線

デジタル人文学最前線の一翼を担って

宮川 創（みやがわ・そう）

国立国語研究所　研究系　テニュアトラック助教

国立国語研究所で言語資源のデジタル化及びデジタルコーパス開発に携わられている宮川創さんに、デジタル人文学の観点からお話をお聞かせいただきます。

ゲッティンゲン大学エジプト学コプト学専修博士課程修了。ドイツ研究振興協会（DFG）特別研究領域研究員（2015 年～2020 年）、関西大学アジア・オープン・リサーチセンターPD（2020 年～2021 年）京都大学大学院文学研究科助教（2021 年～2022 年）を経て、2022 年より現職。

――ご自身の研究についてお聞かせください。

古代エジプト語の最終段階であるコプト語の資料や、同じくエジプト・スーダンの古ヌビア語資料のデジタル化及びデジタルコーパス[1]開発を行っています。コプト語は 17 世紀頃に母語話者がいなくなったとされますが、19 世紀末頃から、典礼言語[2]として残っていた情報を手掛かりに言語復興の機運が高まり、現在でも言語教育が続けられています。こういった消滅危機にある言語を収集し、デジタルコーパスとして整備しながら、自然言語処理やコーパス言語学などの手法を用いてエジプト語史を紐解いています。最近は、このような手法を、消滅危機にある日本の諸言語、例えば、アイヌ語や琉球諸語に適応する研究も行っています。

――国立国語研究所では、どのような仕事を担当されていますか？

国立国語研究所（国語研）では、消滅危機にある日本の言語や方言の文献資料・音声資料・動画資料のデジタル化及びデジタルコーパス開発に関わっています。現在は、幕末～明治期における日琉諸語の聖書翻訳を用いたパラレルコーパスを開発し、テキストは TEI XML[3]形式 、琉球語訳など一部の画像は IIIF[4]形式 でデータを整備しています。その他、国語研『沖縄語辞典』の多言語版ウェブアプリ化や、「日本の危機言語データベース」管理、文献や音声やテキストコーパスなど言語データを中心とした国立国語研究所デジタルアーカイブ（NINDA）構築などにも関わっています。

――ではまず、デジタル人文学で扱うデータについてお聞かせください。

デジタル人文学が扱うデータは、デジタル化された文献だけではなく、3D 画像、音声データ、美術品の画像データなど多岐に渡ります。データの作成・取得先も様々で、既にある程度整理されている博物館・図書館・文書館・美術館などに留まらず、遺跡やフィールドワークなどから入手する場合もあります。

――順番にお伺いします。まず、文献のデータはどのように作成されるのでしょうか。

文献から得られるデータは、写真画像データと翻刻[5]されたテキストデータの 2 種類があります。画像デ

---

[1] 自然言語で記述された文章や発話記録を大規模に収集し、構造化したもの。言語学の研究で主に用いられる。

[2] 教典や宗教行為のみで用いられるようになった言語のこと。

[3] TEI (Text Encoding Initiative)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。同コンソーシアムによって策定された国際標準として、TEI XML 形式が知られている。

[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る）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。

[5] 古典籍や古文書などに記載された文字を解読し標準的な文字

ータの作成に関しては、研究者の研究に資するため、または一般の方々に文化遺産にもっと触れてもらおうといった動機から、主に図書館・博物館・文書館が行っています。作成方法には自前でやる場合と外注する場合がありますが、自前で「ちゃんと」やる場合は、画素数の高いカメラ、書籍の撮影台、専用の画像ソフト等が必要になります。さらに、高品質なデータを取得するためには、撮影台を黒いカーテンで覆って照明器具やカメラを両側から設置する、ページを押さえるアクリル板などを準備する、といった環境も重要になります。一方で、これらの機材を準備できないフィールドの写真や、震災画像などの場合は、研究者個人のカメラ、時にはスマホのカメラなどで撮影されています。そのため、文献の画像データもスマホで十分ではないかという考え方もあり、これに対応した装置開発も進められています。例えば、インスブルック大学のREAD COOP チームでは、スマホで写本撮影をする装置”ScanTent”と、深層学習を使って写本や手稿の手書き文字を機械翻刻する”Transkribus”というソフトウェアを開発しています。

――遺跡やフィールドワークでは、どのようなデータを入手されるのでしょうか。

考古学や遺跡を扱う歴史学などの分野では、現地で得られる許可申請の関係上、研究対象を調べることが可能な時間が限られています。そのため、後日詳細な分析を行えるよう、現地の状況を再現できる 3D データ、ポリゴンデータ等が必要になります。私が関わったエジプトのピラミッド発掘プロジェクトでは、現地政府による許諾のもと、ドローンを用いて外観の写真を撮影したり、３D 計測したりしていました。また、言語学の分野であれば、対象となる言語の話者と交渉し、談話や読み上げを録音する形で音声データを入手します。さらに、ジェスチャーが重要になる言語では、談話を録画するケースもあります。どのケースでも、関係者との信頼関係の醸成や粘り強い交渉が求められます。

――多種多様なデータを扱っているのですね。これらのデータは、どのように活用されているのでしょうか。

例えば、私の恩師が関わっているプロジェクトの一つに、IIIF に準拠した動画ファイルへのアノテーション機能開発があります。このプロジェクトは、IIIF に対応した代表的なビューワーである Mirador[6]を対象に、動画上で IIIF に準拠した音声データ、字幕データを表示できるようにすることが目的ですが、この機能が実装された場合、言語学的な注釈を含む字幕を動画上で表示できるようになります。また、人文学以外では、対象地域の経済・政治の歴史研究に活用されたり、考古学的な観点から進化人類学に活用されているような事例もあるようです。さらに、最近では Cultural Japan が提供するセルフミュージアム機能[7]のように、一般の方でも楽しめるアプリ開発にも活用されています。

――非常に幅広い活用の道があるのですね。どのようにデータを整備されているのでしょうか。

データ整備の際は、可能な限り国際標準となっている規格や形式に合わせるようにしています。画像データであれば国際的な規格である IIIF、テキストデータであれば TEI XML 形式、メタデータであれば Dublin Core[8]などが基準になりますが、これらの国際標準に準拠しておくことで、プロジェクトが終わっても他のプロジェクトに簡単に引き継ぎができたり、他のプロジェクトや研究者に活用してもらったりすることができます。データの持続可能性という観点から、おすすめです。

もう一つ、再利用可能なライセンスを付与する点が挙げられます。国際標準となるクリエイティブ・コモンズ・ライセンスを前提に、自由な再利用・二次利用が可能な CC BY や、一切の制約がない CC0 で公開するのが理想ですが、著者の意向を尊重すべきケースもあります。そのままの形で二次利用可能としなければならない場合は CC BY-ND、商用利用不可にしなければ

にして書き写す作業のこと。現代的な文脈では、コンピュータに打ち込むことにより、デジタルテキスト化する作業も含む。

[6] https://projectmirador.org/

[7] https://self-museum.cultural.jp/

[8] ダブリン・コア・メタデータ・イニシアチブ（DCMI）が制定した、主に検索のためのメタデータを記述する標準仕様。

ならない場合は CC BY-NC で公開することになります。

――再利用性の高いデータ整備を進めるにあたり、課題はどこになるでしょうか。

データ整備の課題は、やはり、その資源の所有者のマインドセットの改革だと思います。もちろん、長い年月をかけて、多大な労力を持ってその資源を保全・管理していただいていることに対して、社会は感謝と敬意を払わなければなりませんが、やはり文化遺産はみんなのものであると言う感覚で、デジタル技術を活用してその文化遺産を社会とシェアしていただくことが理想的です。

――続いて、デジタル人文学を支える人材育成についてお聞かせください。

前述のように、デジタル人文学では多種多様なデータを扱うための技術が必要になります。私が以前に研究活動をしていたドイツでは、デジタル人文学の専修を持つ大学や研究所が複数設置されており、ゲッティンゲン大学やライプツィヒ大学などでは博士号も取得できます。大学院生の出自も人文学と情報学が半々で、もともと橋渡し的なキャリアを持つ人材も在籍しており、大学院生が働きながら技術を確立していくルートが出来ています。また、図書館のデジタル部門に進んだりプログラマーになったり、といったキャリアパスが示されていることも着目すべき点に思われます。

――専門職として明確に位置付けられているのですね。日本の状況はどうでしょうか。

最近では東京大学人文情報学（UTDH）の活動を始めとして、九州大学でデジタル人文学の大学院が設立されたり、岡山大学で機関横断プロジェクトが始まったりと、人材育成に繋がる動きが見られます。一方で、やはりまだまだ講習会等を通じたアウトリーチが必要です。前述したような国際標準となる TEI や IIIF などの規格・形式や、リンクトオープンデータなどへの理解、そして文化遺産や文化資源のオープン化の社会的促進に向けた活動が重要になるものと思います。

――今後、デジタル人文学のプラットフォームとして、どのようなものが必要と思われますか。

デジタル人文学のプラットフォームとしては、現在のところ Drupal、Omeka、WordPress、Concrete 5 など様々な CMS（Contents Management System）が用いられています。どの CMS を使っても、あるいは自前のプラットフォームでも良いのですが、データやメタデータが国際標準に沿って公開されることが重要です。そのためには、例えば方言などの辞書を誰でも簡単に、ウェブで見やすい形で公開できるとともに、データも TEI XML 形式などでリポジトリに追加できるようなシステムが作れればと思います。加えて、API（Application Programming Interface）なども整備しておくことで、他のプロジェクトや利用者がデータをどんどん二次利用、活用していけるようにするのが理想です。専門家だけが使えるプログラムやデータセットを作るのではなく、誰もが使えるインクルーシブなデジタル技術、プラットフォームを整備していきたいと思っています。

――最後に、人文学・社会科学データインフラ事業について、期待やご意見をお聞かせください。

データ活用の観点からは、JDCat 分析ツール[9]が Google Colaboratory[10]並みの使いやすさになり、RStudio でできる全てのことが JDCat[11]上でもできたら素晴らしいです。人文学系ではデータ可視化に関する需要が大きく、R Shiny などが良く用いられるほか、最近ではウェブ上でテキストマイニング結果の可視化ができるツール（例：https://voyant-tools.org/）

---

[9] JDCat 分析ツールとは、統計ソフトをインストールしたりデータを手元にダウンロードしたりすることなしに R や Python のプログラムを作成・実行しデータを分析できるツール。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。

[10] https://colab.research.google.com

[11] JDCat とは、Japan Data Catalog for the Humanities and Social Sciences の略。人文学・社会科学総合データカタログ。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。

などが話題になっています。GUI ベースでデータの可視化ができ、自分のウェブページで公開出来るような仕組みができたら本当に素晴らしいと思います。

（座談会開催：令和４年１１月２日／聞き手：南山泰之）