JDCat サロン

データインフラの最前線

JDCat 開発を通じて

朝岡誠（あさおか・まこと）

国立情報学研究所 オープンサイエンス基盤研究センター
特任助教

国立情報学研究所 オープンサイエンス基盤研究センターにて人社データインフラ事業に携わられている朝岡誠さんに、本事業の取り組みとデータカタログについてお聞かせいただきます。

東北大学文学研究科人間科学専攻　行動科学専門分野博士課程単位取得退学ののち、東京大学社会科学研究所特任研究員(2009 年～2014 年)、立教大学社会情報教育研究センター助教（2014 年～2019 年）を経て、2019 年より現職。

――ご自身の研究についてお聞かせください。

リポジトリを用いた柔軟な研究データ公開の方法について研究しています。一部の研究データには、プライバシー情報など機密性の高い情報が含まれていることがあるため、すべての研究データをオープンにすることはできません。そこで、海外の研究データのアクセス制限の方法について研究し、国内におけるオープンアンドクローズ戦略に沿った研究データ公開を実現するための機能開発を行っております。

――人社データインフラ事業では、どのような仕事を担当されていますか？

人社データインフラ事業では、JDCat[1]と JDCat 分析ツール[2]の開発進捗管理を行うほか、拠点機関とのメタデータ連携における技術的サポートを行っています。また、今回の事業では、大阪商業大学、慶應義塾大学、一橋大学に WEKO3[3]を利用していただいているため、これらの拠点機関のメタデータ作成サポートも行っています。

――ではまず、人社データインフラ事業が対象とするデータについてお聞かせください。

人社データインフラ事業では、名称の通り社会科学のデータと人文学のデータの両方を扱っています。社会科学データは数値化された CSV ファイルや統計データといった計量的なデータが中心となっており、複数のデータを組み合わせて研究対象のメカニズムを解明することに重点をおいているように見受けられます。一方で、人文学データはテキスト、画像、音、映像と様々な形式で描写されており、ある研究対象が持つ情

---

[1] 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。

[2] 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。

[3] WEKO3 は、国立情報学研究所が開発を進める、研究者が公開すると判断した研究データや関連の資料を公開するためのデータ公開基盤。論文も含めたデータを登録、公開するために必要な機能を有している。
https://rcos.nii.ac.jp/service/weko3/

報を異なる角度から読み解くための記述に重点をおいているように思われます。例えば、研究対象となる過去の日記から、著者の人間関係を知るだけではなく、食事から当時の生態系を推察したり、天気や災害の記録を読み解いたり、といったように、多様な使われ方が想定されています。

――これまでのインタビューでも、社会科学データと人文学データの違いは随所に見られました。この点、データ整備の方向性においても違いが出てきそうですね。

社会科学データの場合、同じルールでデータの計測方法が記述され、比較可能になっていることが求められるため、専門家による語彙の統制やメタデータのクリーニングを含む人的なデータキュレーションが特に重視されます。対して、人文学データはいろいろな解釈の余地が必要となるために、語彙の統制はなかなか難しいようです。一方で、出来るだけ豊富な情報が記述される観点からは、機械学習などによって解析し、植物、動物、天候のデータが含まれるかどうかをチェックして、データの情報をリッチにするといった情報学的な手法と相性が良さそうです。

――共通して実施できる部分はあるのでしょうか。

どちらの分野においても、一次的にはデータの保有者がキュレーションを行う必要があるため、分野別に専門的なノウハウが必要となります。一方、人文学と社会科学はどちらも人の作品、思想、行動を対象とする学問であるため、データに含まれるプライバシー情報への対応や著作権処理といった問題は共通しています。また、時間情報、空間情報は共通のルールで整備することが可能だと思います。もっとも、データの保有者がこういった問題に単独で当たることは難しく、分野横断的な視点を持ったデータキュレータによるサポートが必要になると思われます。

――人社データインフラ事業で開発中の機能についてお聞かせください。

人社データインフラ事業では、JDCat、JDCat 分析ツールに加えて、公開されている研究データをシームレスで利用できるよう両者を連携させる機能開発を進めました。一般に、公開されたデータを利用するためには、データの利用申請を行い、データをダウンロードし、ダウンロードしたデータをローカルな分析環境にアップロードする、といった手順を踏むことになりますが、大変工数がかかります。オンライン上でこれらの手順を完結させることによって、提供側、利用側双方の時間短縮が期待できます。

――想定される利用者は、どのような層が考えられるのでしょうか。

特に、教育面で活用されることを想定しています。社会科学データを扱う授業では、データ分析の実習を行うために学生自身がデータをダウンロードする必要があります。そのため、授業の冒頭では、データの申請と統計ソフトへのアップロードを説明しますが、この内容だけで 2 回分の授業を使ってしまうことがあります。また、申請したデータを紛失する、といったセキュリティ上の懸念や、人数分の統計ソフトを準備するための手間や予算といった課題も常に付きまといます。JDCat 分析ツール機能の開発、および JDCat との連携で、こういった現場の課題解決に繋がることを期待しています。

――利用者の認証、承認手続きはどのように行われるのでしょうか。

利用者の認証、承認手続きについては、WEKO3 で開発中の制限公開機能と組み合わせて行うことを考えています。特にプライバシーに関わるデータを扱う場合、データ提供機関としては利用目的の確認等が必要になりますので、何らかの承認プロセスを設定する必要があるでしょう。

一方で、JDCat／JDCat 分析ツール連携はデータの申請からデータ分析環境の提供までを全てオンライン上で扱える手軽さが特徴ですので、追加の承認プロセスを要求する制限公開機能とは少し相性が悪い面もあります。英国のデータアーカイブ UK Data Archive[4]

[4] https://www.data-archive.ac.uk/

などでは、承認プロセスを利用者側のセルフチェックに委ねる、といった簡略化も試行されており、日本でも柔軟な申請方法について議論を進められると良いと考えています。

――JDCat の運用面についてお聞かせください。JDCat の運用はどのように行われているのでしょうか？

JDCat は、提携するリポジトリのメタデータを自動収集し、自動収集したメタデータに対して言語別に絞り込み検索を用いて横断検索を行うことができることを目標に設計しています。この設計は、ヨーロッパの社会科学データアーカイブポータルサイト CESSDA Data Catalogue [5]を参考にして構築しており、将来は人文学や社会科学専門の研究者でも運用できるようにする構想があります。

――開発元の国立情報学研究所（NII）が運用するのではないのですね。

現在は試験段階ということで、メタデータがルール通りに収集されているかどうかを確認しながら NII が運用しています。一方で、拠点機関内での運用を見ると、JDCat メタデータスキーマ をそのまま利用しているところは少なく、ある程度カスタマイズして利用されているようです。また、具体的な要望としてメタデータスキーマや統制語彙の拡充も提案されており、メタデータの検索処理にはもうひと工夫必要と思われます。個人的には、コミュニティによる運用体制の構築を NII が支援する形が望ましいのではないかと考えています。

――コミュニティによる運用体制を念頭に置く場合、JDCat への登録作業はどのように行われることになるのでしょうか？

JDCat の開発当初は、人文学・社会科学分野の研究者が JDCat へ直接データを登録する機能（セルフアーカイブ機能）を実装する構想がありましたが、個人研究者のデータは拠点機関が収集し、キュレーションして公開するという方針になったためこの機能は不要になりました。一方で、拠点機関で提供できる研究データの公開量は限られており、今後の研究データの公開希望に耐えられないことが予想されます。現に、データキュレータを多く置いている ICPSR[6]などでも、公開待ちのデータが多数存在しているようです。上述した ICPSR では、このような状況に対応するために OpenICPSR[7]という、セルフアーカイブ機能を有するリポジトリを立ち上げています。JDCat においても、拠点機関でデータ公開作業を集約する方針は維持しつつ、自分で責任を持ってデータを公開するという選択肢が必要になるかもしれません。

――関連して、JDCat を取り巻くコミュニティ形成についてお聞かせください。

JDCat を取り巻くステークホルダーとしては、データ配布機関、データ提供者、データ利用者が考えられます。現在は拠点機関の方々がデータ提供者とデータ配布機関を兼ねている状態ですが、今後、研究グループや学会がデータ提供者となるケースが増えていくものと思われます。その際、データ提供者はどのデータ配布機関にデータを預ければいいのかが課題になると思います。つい最近も、オープンアクセスに則って自分たちのデータを公開できるリポジトリを探しているが、見つからないといった悩みや、退職された研究者の貴重なデータが学会に残されており、保存先がないといった声をお聞きしています。

――拠点機関外の研究グループや学会に所属する方々は、どのように JDCat でデータを公開できるようになるのでしょうか？

人社データインフラ事業では、拠点機関の方々にメタデータを提供いただくにあたり、分野別メタデータのハーベスト機能を開発しています。これまで NII が扱ってきた JPCOAR スキーマ だけではなく、新たに JDCat メタデータスキーマを扱う際にはいくつかの課題がありましたが、拠点機関の方々のご協力によって解決することができました。この場を借りて御礼申し

---

[5] https://datacatalogue.cessda.eu/

[6] https://www.icpsr.umich.edu/

[7] https://www.openicpsr.org/openicpsr/

上げます。

分野別メタデータのハーベスト機能は、現在のところ拠点機関が保有するデータアーカイブのみを対象としていますが、将来的にはJAIRO Cloud を利用している機関からもJDCatにメタデータを提供していただくことが可能になると思います。

――拠点機関だけではなく、各大学・研究機関の機関リポジトリがデータ配布機関になるということですね。

はい。今後、研究代表者が所属する大学図書館にデータを預け、キュレーションを経てデータを公開するというケースが増えていくと思われます。一方で、大学図書館の方からも、自分たちで所属する研究者のデータをきちんと整備して、公開できるのか不安だという声をお聞きしています。こういった方々も含めた形で、JDCat を中心としたコミュニティを形成していく必要があると思われます。

―― JDCat のコミュニティには、どのようなことを期待されているのでしょうか。

まず、データ提供者が研究データを預け、整備し、公開するまでの標準的なフローを構築、共有する必要があります。現在NIIでは、（インタビュアーの）南山さんがデータキュレーションのタスクを管理するシステム開発に着手しています。このシステムでは、各機関で実践が難しい専門的なデータキュレーションを、外部の専門家に依頼する機能を備える予定です。拠点機関の皆さまには、キュレーションを行う専門家としてコミュニティに参加していただき、研究データの整備に協力していただけると良いなと思います。

また、人文学・社会科学のデータ提供の現場では、個々の利用者のリクエストへの対応に追われ、利用者全体の要望を汲み取ることに苦心されているように感じます。このような課題に対しても、学会単位でコミュニティに働きかけていただくことで、データ寄託から公開までの手続きの円滑化や、JDCat の検索ルーチンやメタデータ入力ルールの更新といった対策に繋がるのではないかと期待しております。さらに、こういった取り組みを通じて、個別の機関では対応が難しいデータに対して、どこの機関が持つのか、キュレーションはどこが行うのか、承認はどうするか、コミュニティとしてユースケースを作っていければと考えています。

（座談会開催：令和４年５月１２日／聞き手：南山泰之）