JDCat サロン

データインフラの最前線

RUDA における多様なデータアーカイブ活動とその展望

髙橋　かおり（たかはし・かおり）

立教大学　社会情報教育研究センター　助教

立教大学社会情報教育研究センターにて社会調査データアーカイブ「RUDA」の運営・管理に携わられている髙橋かおりさんに、データアーカイブ活動についてお聞かせいただきます。

早稲田大学大学院文学研究科博士後期課程単位取得満期退学ののち、早稲田大学文学学術院社会学コース助手（2014 年～2017 年）、東京大学大学院情報学環特任助教（2017 年～2018 年）を経て、2018 年より現職。

――ご自身の研究についてお聞かせください。

インタビューや参与観察などの質的調査に基づいて、芸術活動を行う人たちの活動状況やキャリアプランに関する調査を行っています。さらに、量的データやアーカイブ化されたデータを補完的に用いて、「芸術家」とされる方の実態把握も行っています。これまでプロフェッショナルの芸術家を対象とした調査はいくつか行われてきましたが、「芸術家」をどのように定義するのか難しいという問題があります。通常、専門家と呼ばれる人たちはその専門的な業務を職業として自認するはずですが、芸術活動に関わる人たちの中には独自の基準があり、調査の切り口によっては「芸術家」を名乗らないことがあります。そのため、調査結果を相互に補い合って読みながら、調査対象としたい「芸術家」の範囲を明らかにしています。

――立教大学のデータアーカイブ事業では、どのような仕事を担当されていますか？

立教大学では、立教大学社会情報教育研究センター（CSI）が運営する社会調査データアーカイブ「RUDA」[1]の運営・管理業務を行っています。日常業務としては利用者の申請受付と寄託者とのやり取り、ならびに寄託データの公開に向けたクリーニングを行っています。また、国内外のデータアーカイブの紹介やデータ分析の実例などをトピックとして学内向けのセミナーを企画しています。

――ではまず、RUDA が扱うデータについてお聞かせください。

RUDA では、社会学者が 2000 年代以降行ってきた地域調査データを中心に扱っています。調査データは立教大学関係者からの寄託が多いですが、それ以外の研究者でも寄託を受け付けており、大学の実習や演習、科研費をもとにした共同研究による、都道府県や市区町村を単位とした郵送調査データを多く保有しています。

――どのような層が利用者になるのでしょうか。

利用者は、社会科学分野の学生や研究者がほとんどです。大学の授業での実習・演習といった教育利用だけではなく、卒業論文や修士論文執筆の際に二次分析を行ったり、類似の社会調査を実施する前にプレ分析を行ったりする利用もあるようです。また、最近は学

[1] RUDA（Rikkyo University Data Archive: ルーダ）とは、貴重な公共財産である社会調査データを収集・整理・保管し、学術的な二次分析といった研究目的での利用、および授業での教育利用のために、広く公開していくことを目的とするデータアーカイブ。https://ruda.rikkyo.ac.jp/dspace/

内でデータサイエンス関連の取り組みが始まったため、将来的にこういった分野での研究利用も始まるかもしれません。

――続いて、RUDA のデータ整備方針についてお聞かせください。

RUDA では、寄託者の意向やデータの特性に合わせ、最小限度のデータクリーニング[2]を中心に実施しています。データクリーニングでは、寄託されたデータと報告書や書籍といった既刊物との間に齟齬がないか確認するほか、調査票またはコードブックに存在しないコード（オフコード）のチェック、回答者が限定されている質問や複数回答を許容する質問の整合性チェック、変数型のフォーマット最適化といった処理を行っています。

――データクリーニングの際、データに含まれる個人情報はどのように扱われていますか。

RUDA ではデータを全て公開するため、個人情報が含まれる可能性がある自由記述や細かい地点情報などは原則として削除しています。一方で、RUDA は地域情報を扱うデータアーカイブなので、一定以上の地点情報を消してしまうと意味のないデータになってしまう場合もあり、その都度寄託者と協議しながら取扱いを決定しています。自由記述の存在自体が重要となるデータの場合には、自由記述欄のあり／なしを記述する変数を追加して対応したようなケースもあります。

――メタデータはどのように作成されていますか。

RUDA では、基本的に寄託者から提供された情報をそのままウェブサイト上で公開しています。現時点では独自に付与する項目はありませんが、地域情報については寄託されたデータから抽出、タグ付けしており、当該地域や関心のある地域で過去に行われた調査の横断検索を可能にしています。また、RUDA は CiNii Research[3] による横断検索の対象になっているため、今後は研究成果と紐づいたデータの検索に必要なメタデータの整備も必要になると考えています。

――他データベースとの横断検索を行う場合、キーワードの調整が課題になりそうですね。

キーワードの調整は悩ましい問題です。語彙の選択は当該研究の専門性の現れでもあり、データアーカイブ側で最終的な選択の責任を負えないという問題があります。そのため、RUDA ではキーワード検索を提供していますが、現在のところ統制作業は行っていません。一方で、検索対象が広がった場合に効率的な検索が求められる点は認識しており、寄託者にうまく書いてもらえるような運用がないか模索しています。

――JDCat[4]については、どのように見られていますか。

RUDA のデータは、統計分析や量的データを日常的に扱う人のみならず、フィールドワーカーや地域調査を行う人たちにとっても活用の可能性があると思っています。例えば、調査対象地域や地点でどのような調査が過去に行われていたのかを予備的に知ることは、質的調査や更なる量的調査を実施するうえでの基礎資料となりえます。また、RUDA の開設当初はカレントな情報であった調査データも、時間経過とともに歴史的データに変わりつつあり、データによっては異なる観点からの二次分析が可能かもしれません。人文学・社会科学の横断検索を実現している JDCat と連携していくことで、データの新たな側面が明らかになり、異なる層の方々にも興味をもたれる可能性があると考えています。

――RUDA の運用体制についてお聞かせください。

RUDA では、大学院生の RA（リサーチアシスタン

---

[2] データクリーニングとは、分析の障害となる異常値、重複データなどを取り除き、データを分析しやすい状態にする処理のこと。データキュレーションに含まれる処理の一つ。

[3] CiNii Research とは、日本最大規模の学術情報検索サービス。機関リポジトリに登録された研究成果や論文情報のみならず、図書、研究データ、それらの成果を生み出した研究者、そして研究プロジェクトの情報などを包括して探索することができる。https://cir.nii.ac.jp/

[4] 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。

ト）と協力しながらデータクリーニングとアーカイブ管理を行っています。役割分担としては、データ寄託に関連する対外的な業務やアーカイブ管理を私が担当し、データクリーニングの実務面を RA に依頼しています。

――順番にお伺いします。対外的な業務としては、どのようなことを行うのでしょうか。

RUDA に寄託依頼があった場合、まず提出された「寄託チェックシート」に基づいてデータの確認を行います。ここでは、対象となるデータの権利関係や報告書の有無、データ形式などを確認し、受け入れの可否を判断するとともに、後の作業に必要な情報を整理しています。受け入れが可能なデータについては、さらに「メタデータ記入シート」の提出を依頼し、調査の概要やデータに含まれる主要変数といったより詳細な情報を取得しています。一連の手続きが完了した段階で、寄託契約書の取り交わしも行っています。

――アーカイブ管理とは、どのような業務なのでしょうか。

通常は寄託されたデータセットの原票を保管することですが、まれにデータセット化が難しいものの、歴史的価値のある調査票や調査資料を保管し提供する業務を行うこともあります。データセットの中には、調査票の原本を復元、再分析したり、調査資料群を再検討するような質的分析の対象となるものが存在します。こういった歴史的な研究の観点から役立つ資料については、個人情報保護や管理を前提としつつ、電子ファイル化しない形での保管や活用の方法について検討する必要があります。

――質的データを扱うデータアーカイブは国際的にも事例が少ない印象です。

RUDA においても実務的な取扱い手順はまだ確立できておらず、海外の先行するデータアーカイブである UK Data Archive[5]や Finnish Social Science Data Archive[6]の事例を調査し、取扱いを模索している段階です。また、立教大学には社会運動にかかわる歴史的資料を扱う共生社会研究センターがあるため、同センターに所属するアーキビストの方と情報交換を行い、原本の保存や整理の面で連携の可能性を探っています。

――データクリーニングは RA と協力しながら実施しているとのことですが、運用上の課題はあるのでしょうか。

RA は雇用期間が人によって異なり、1 つのデータに対して複数人が関わることもあるため、トラブル防止の観点から業務の引継ぎが重要な課題になります。そのため、「誰が作業を行っても同じような手順で正しい結果が得られる」ようにすることを目的としたクリーニングマニュアルを作成し、作業の標準化を行っています。

――業務別に様々な方と協力しながら実施されているのですね。他にも連携先があるのでしょうか。

学内の他機関との連携で言えば、利用者への情報提供面で図書館と協力しています。具体的には、図書館が発行する広報誌にて紹介をしてもらうほか、図書館ウェブサイト上に RUDA へのリンクを掲載してもらい、導線を整備しています。

――その他、国内のデータアーカイブ構築や運用について、お考えをお聞かせください。

社会学の分野では 2000 年代に多くのデータアーカイブが設立されたものの、大半は組織変更などの事情によって閉鎖されています。RUDA は学内で安定したセクションで運営されておりますが、いずれにせよ組織における仕組みとして継続的に運用できる体制づくりが重要です。そのためにはデータを寄託することの意義を組織内で共有していく必要がありますが、なかなか難しいところです。RUDA への寄託についても、その価値や意義を知っている研究者の方々からは自主的にご連絡を頂きますが、お声がけして初めて寄託の意識を持ってもらえる、というのが現状です。

――データの寄託が進まない理由は何でしょうか。

いくつか理由はあると思うのですが、1 つには寄託

---

[5] UK Data Archive. https://www.data-archive.ac.uk/

[6] Finnish Social Science Data Archive (FSD). https://www.fsd.tuni.fi/en/

依頼のタイミングが難しいこと、もう１つは寄託にかかる手間が煩雑であることだと考えています。

第一の点については、調査を実施した後、ある程度の時間が経過しないと「二次利用」を主張しづらい一方で、あまりに時間が経ってしまうと、データについて忘れたり、古いからといって寄託をためらう人が生じたりしてしまいます。どの時期に寄託依頼を行うのが良いか、は悩ましいですが、例えば研究助成事業に対しては５年や 10 年など一定期間が経ったらデータアーカイブに寄託を検討するようアナウンスや連絡をする、あるいはそもそもの申請時点でデータアーカイブへの寄託を前提とする、といった共有事項があると、貴重なデータが死蔵されずに済むと思います。

第二の点については、データアーカイブ側でインターフェイスや手順を簡略化する、ということももちろん大事なのですが、「面倒なことをしてでも寄託する意義がある」ことを研究者に共有していくことが大切だと思います。現時点では、そもそもの価値観を共有できる研究者の善意に頼るほかないため、寄託にかかる手間が目立ってしまう状況であると見ています。この点については、先ほどの話にもありましたデータの蓄積による歴史的な観点からの分析や、類似研究に役立つといった活用事例を強調していくことで、寄託の意義をより幅広く共有していければと考えています。

――最後に、データアーカイブ活動へのご意見をお聞かせください。

データアーカイブは社会調査や統計教育の一部として扱われていますが、実際の管理においては今あるデータから何が言えるか、どのように整備するかといった観点が重要になります。私自身、データアーカイブの運営の実際について着任前にあまり知識がなかったため、前任者からの引継ぎや過去の作業記録を確認することを通じて学んだことに加え、社会科学系の国際的なコミュニティである IASSIST[7]へ参加し、データアーカイブの国際的なトレンドを知るようになりました。昨今ではデータ分析やデータサイエンスが流行にある一方、その元となるデータをどう管理し、収集するのかについての議論がおろそかになりがちです。研究データは貴重な研究資料となるので、ぜひ共有の財産として皆で守っていければと考えています。

（座談会開催：令和４年７月５日／聞き手：南山泰之）

---

[7] IASSIST とは、データサービスや提供等に関わるネットワーク形成、社会科学のインフラ形成、専門的実務に関わる情報交換などを目的とする社会科学系コミュニティの一つ。https://iassistdata.org/