# JDCat サロン

## データインフラの最前線

# 社会調査データアーカイブの本流をめざして

胡中孟徳（こなか・たけのり）

東京大学　社会科学研究所
附属社会調査・データアーカイブ研究センター　特任研究員

東京大学社会科学研究所附属社会調査・データアーカイブ研究センター学術支援専門職員（2017 年～2019 年）を経て、2019 年より現職。

東京大学社会科学研究所にてメタデータの作成に携わられている胡中孟徳さんに、SSJ データアーカイブの取り組みについてお聞かせいただきます。

**――ご自身の研究についてお聞かせください。**

子どもの時間の使い方が、子どもの現在と将来両方のライフチャンスに影響を及ぼすという観点から、小学生から高校生くらいの子どもの生活時間のデータを用いて、時間の使い方が、親の社会経済的地位によってどのくらい異なるかを分析しています。

**――人社データインフラ事業では、どのような仕事を担当されていますか？**

東京大学社会科学研究所附属社会調査・データアーカイブ研究センター（CSRDA）が運営する SSJ データアーカイブ（SSJDA）のメタデータを、社会調査メタデータの国際規格である DDI メタデータスキーマ[1]へマッピングするための調査を行っています。JDCat メタデータスキーマ[2]は DDI メタデータスキーマのバージョン 2.5 をベースに設計されているため、まずは DDI 2.5 への対応を目指しています。

**――ではまず、SSJDA が扱うデータについてお聞かせください。**

SSJDA では、東京大学社会科学研究所で保存していたデータをはじめ、民間の調査研究機関や大学等の研究者が実施した社会調査の個票データを収集し、キュレーションを行った上で、二次的な学術的研究のために研究者・大学院生に提供しています。日本では他の分野のアーカイブを含めても歴史が長く、社会科学系では最大規模のアーカイブでもあるため、国際的な利用が見込まれるデータも多く扱っています。

**――具体的には、どのようなデータが対象になるのでしょうか。**

---

[1] DDI メタデータスキーマとは、社会科学、行動科学及び経済学のデータを記述するために設計されたメタデータの国際規格。調査研究、データ収集方法、質問と回答、変数、データセットへのリンク、及び研究とデータセットとの関係性に関するメタデータが含まれ、これらのメタデータを Linked Data として共有するための RDF 表現も実装されている。2022 年 2 月現在、DDI メタデータスキーマの最新バージョンは 3.1。
https://ddialliance.org/

[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）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

社会学系だと、社会階層と社会移動全国調査（SSM調査）シリーズ[3]、全国家族調査（NFRJ）[4]、東大社研パネル調査[5]などが挙げられます。また、政治学系ではJapanese Election Study[6]などが該当すると考えられます。これらのデータは、一時点のナショナルサーベイとしての価値もありますが、ある程度シリーズとして寄託されていることによって、時系列的な変化を検討できる調査でもあるため、多様な観点からの国際比較に耐えうるデータと言えます。その他、日本の縮図になっているような理想的なデータや、すでに海外のジャーナルに掲載された実績があるデータ、あるいはアジアンバロメーター調査[7]のように、最初から国際比較の調査として実施されているデータは国際的な利用が見込めると考えています。

**――既に国際的な利用実績が多くあるのですね。利用促進に向けてどのような取り組みをされているのでしょうか。**

日本を含むデータで国際比較を行う場合、データ、メタデータともに英語化されていると使い勝手が良くなります。特に海外研究者の場合、ある程度日本語が分かる方でも問い合わせ対応は英語で行われることも多いため、英語化することでこちらの想定を超えたニーズに応えられる可能性があると考えています。過去には人文学系の海外研究者による利用希望もあり、英語化によって思いもよらないニーズを掘り起こすことにも期待しています。また、平行してCoreTrustSeal認証[8]の取得準備を進めています。

**―― CoreTrustSeal 認証は、国内では極めて先進的な取り組みですね。**

国際的に信頼できるデータアーカイブであるという認証を得ることで、海外からの利用促進に繋げることが狙いです。副次的な効果として、審査のための書類準備を進めるうえでアーカイブの体制改善に向けた気づきも多くありました。たとえば、中長期的な事業の継続可能性の保証として、アーカイブが運営できなくなったときのサルベージのプランなどを提示することが求められている点などです。また、本審査のプロセスではより詳細なアドバイスが期待できるため、認証取得に向けたモチベーションの１つとなっています。

**――海外展開を強く意識された取り組みを進められている、と感じました。**

SSJDAでは、海外の社会科学分野データアーカイブと連携したいという考えはある程度前から検討していました。本事業を通じて、データ検索・利用システ

---

[3] 社会階層と社会移動全国調査（SSM調査）は、1955年の第1回調査以来10年ごとに実施されている大規模な社会調査の一つ。社会階層、社会移動、不平等、職業、教育、社会意識などの解明を目的としており、2022年2月現在、計7回の調査が実施されている。

[4] 全国家族調査（NFRJ）とは、日本家族社会学会全国家族調査委員会が実施している、確率標本による全国規模の家族調査。主たる目的は、研究者が利用可能な無作為抽出に基づく全国確率標本データを定期的に構築すること、そうしたデータを多くの研究者の公共利用に供することの2点。2022年2月現在、計4回の調査が実施されている。https://nfrj.org/

[5] 東大社研パネル調査は、東京大学社会科学研究所が独自に実施するパネル調査。2022年2月現在、若年パネル調査、壮年パネル調査、高卒パネル調査、中学生親子パネル調査の4つが展開されている。https://csrda.iss.u-tokyo.ac.jp/socialresearch/project/

[6] Japanese Election Study (JES)は、投票行動研究会が実施している、日本における投票行動を研究するための全国パネル調査。JESは意識調査データ及び投票行動の説明要因となる関連データ（選挙公約・地域特性としての国勢調査・議会議事録）からなり、2022年2月現在、計6回の調査が実施されている。https://jesproject.wixsite.com/jesproject

[7] アジアンバロメーター調査は、世界規模で取得が進められている社会―政治参加・社会関係資本・民主主義の関係性に関わる、アジア地区のパネル調査。代表的な国際比較研究の一つとして、社会心理学者、政治学者、政治社会学者の共同によって実施されており、2022年2月現在、日本では5回の調査が実施されている。

[8] CoreTrustSealとは、2016年に策定されたデータリポジトリの整備・運用に関する国際的な認証基準の一つ。社会科学分野の認証団体であるData Seal of Approval（DSA）と自然科学系のコミュニティである世界科学データシステム（WDS)が連携して策定され、16項目の中核的な要件及び用語集を定めている。https://www.coretrustseal.org/

ムである SSJDA Direct[9]への OAI-PMH 実装や、DOI 登録機能の開発が進んだため、これらを活用することで Gesis Data Search[10]などと連携することができると考えています。海外の社会科学分野データアーカイブでは、DDI メタデータスキーマの採用によってスムーズな連携を実現しており、今回共通のスキーマに準拠したことで可能性が広がったと感じています。

**――引き続き SSJDA のシステム面での取り組みについてお聞かせください。**

本事業では、新たにセルフ・アーカイブ機能の導入に着手しました。現時点では、ほとんどすべてのデータで、寄託を受けてから公開するまでに必要なキュレーションを、SSJDA 側で行わなければならない状態にあります。事務的な作業やメタデータ作成には相応の時間がかかるため、スピーディーな公開を希望する寄託者側の要望に応えられないケースがありました。セルフ・アーカイブ機能の導入によって、寄託者自身によるデータのアップロードとメタデータの一次作成が実施可能になるため、こうしたミスマッチを解消できるようになると考えています。

**――担当者の作業負荷軽減にも繋がりそうですね。**

セルフ・アーカイブ機能が活用されるようになれば、SSJDA 側の作業は公開前の最終確認程度まで減らすことができると考えています。もっとも、寄託者へ全面的に業務を移す訳ではなく、自主的に出来る部分を増やすことで寄託者側の要望に応えることが目的ですので、データやメタデータの品質管理は引き続き重要な課題です。初期の運用としては、キュレーション活動に理解のある寄託者に機能を試してもらい、フィードバックをもらえると良いと考えています。例えば、寄託実績によってセルフ・アーカイブの可否を決めていくような運用も考えられるかもしれません。

**――リモート集計システムの強化も取り組み事項に挙げられていました。**

SSJDA では、DDI メタデータスキーマに対応したリモート集計システム Nesstar[11]の開発を進めています。Nesstar では、SSJDA でデータを利用する前に簡易な分析を実施できること、SPSS などの有償ソフトウェアへアクセスしづらい大学生が授業やレポートで利用できることを重視しています。そのため、初学者の利用を念頭においた GUI ベースでの分析機能や、社会調査データの分析に良く用いられるリコード機能[12]と二項分布モデルの実装を中心に開発を進めています。

**――本事業で提供されるオンライン分析システムとは、どのような関係になるのでしょうか。**

学振と国立情報学研究所が開発・提供するオンライン分析システムとは、利用する際のフェーズが異なると考えています。分析を行う際には、手早く集計結果を把握したい場合とより詳細な分析を行いたい場合があり、オンライン分析システムは後者に当たると思われます。SSJDA では制限公開の対象となるデータがあるため、オンライン分析システム導入に当たっては SSJDA に登録していない研究者の取扱いなどに課題がありますが、将来的に連携していければよいと考えています。

**――続いて、JDCat についてお伺いします。まず、JDCat にはどのようなことを期待されていますでしょうか。**

CSRDA では、これまでもセンター教員が所属する学会などのコミュニティを中心にデータアーカイブや二次分析の広報を行ってきました。そうした広報による成果として、利用実績が伸びてきた面もありますが、同時に、広報していく分野の偏りが生じていたとも考えています。JDCat によって、これまでの広報ではアプローチできなかった層にも目に留めてもらえる可能性があり、こういった層の参入によって学際的な研究が活性化されることに期待しています。

また、JDCat が海外のデータアーカイブと連携する

---

[9] SSJDA Direct : https://ssjda.iss.u-tokyo.ac.jp/Direct/

[10] Gesis Data Search : https://datasearch.gesis.org/start

[11] Nesstar : https://nesstar.iss.u-tokyo.ac.jp/webview/

[12] リコード機能とは、社会調査で用いた回答選択肢の並び順を逆転させたり、適切なカテゴリに再マッピングさせたりすることで、選択肢番号を分析可能な数量変数に変換する機能。

ことができれば、各データアーカイブとの調整コストが省けて好ましいと考えており、ぜひ連携をすすめていただきたいと思っています。JDCat による海外連携が実現する場合は、SSJDA が進める直接連携と重複しないように調整しながら進めさせてもらいたいとも考えています。

**―― SSJDA では保有するメタデータを JDCat メタデータスキーマへマッピングしているとのことですが、これまでの運用からどのように変わるのでしょうか。**

実のところ、SSJDA が保有するメタデータと JDCat メタデータスキーマでは、それほど大きな違いはありませんでした。JDCat メタデータスキーマが準拠する DDI はバージョン 2.5 であり、研究の動的な側面を扱っていないのでマッピングも容易だったと考えられます。一方、データのライフサイクル全体を扱えるバージョン 3.0 を視野に入れると、かなり大きくメタデータ作成の実務が変わることが予想されます。自分の専門である社会調査教育に引き付けて考えると、データのライフサイクルを社会調査教育の中で位置づけられている教科書は少ないため、DDI3.0 に準拠したメタデータ作成を通じて研究の表現が可能になるかもしれません。本事業が終了した後になるかもしれませんが、バージョン 3.0 の適用も考えていきたいところです。

**――社会調査教育との結びつきは興味深い観点ですね。**

将来的には、研究・教育という観点からデータに付加価値をつけるような取り組みができるとよいと考えています。例えば、シリーズ調査のハーモナイゼーションや、クエスチョン・バンクの作成・運営、データの作成から公開までのライフサイクルやメタデータに関する教育を社会調査教育の中に取り入れる工夫、調査に関するおすすめ情報の提示（機械学習による類似の調査の提示や、教員による研究状況などを踏まえたコラムなど）といったアイディアを SSJDA 内で議論しています。現時点では事務作業の負担が大きいことから、なかなか着手にいたっていませんが、実現できれば社会調査に関する研究・教育としての意義は大きいと考えています。

**――そのうえで、メタデータ作成に当たっての課題はあるでしょうか。**

JDCat メタデータスキーマに準拠したメタデータ作成に当たっては、用意された統制語彙を用いることで、テキストとして従来持っていた豊かな情報を失う面があることが課題と考えています。関連して、統制語彙が欧米の社会調査環境を念頭においた語彙であるため、日本の社会調査で用いられるメタデータを十分に反映できないケースがありました。例えば、調査方法として日本では「郵送調査」といった記載が見られますが、現状の統制語彙ではこれに該当する語彙がありません。調査の質を判断するうえでこの情報を重視する利用者もいるため、SSJDA Direct では従来のフリーテキスト形式も併記できる形をとっています。この対応により、今後新規にメタデータを作成する際、若干の負担増となる可能性があると考えています。

**――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、あるいはどのような役割を果たしていきたいか、お考えをお聞かせください。**

データアーカイブはその性質上、半永久的に継続することそれ自体が役割だと考えているので、引き続き取り組みを続けることがまずは重要だと考えています。そのためには、資金・ノウハウ・人材を確保し続けることが必要と感じています。本事業のようなプロジェクトはそのための大きな手助けとなっていますが、同時に、より永続的な資金確保も必要だと考えています。また、人材の観点から言えば、海外のデータアーカイブはライブラリアンやアーキビストの背景を持った人を中心に運営されているように感じています。社会科学の研究者を中心に運営している SSJDA では、海外の動向にキャッチアップすることを主眼に置いた運営になるのに対して、海外のデータアーカイブはライブラリアン主導で、自分たちで動向を作っているというイメージを持っています。

また、社会科学の研究で用いられるデータは、質的データ、公的統計、行政データ、Web スクレイピング

によるものなどがあり、もともと多様であるし、近年より多様化しているように感じています。このような現状に対応していくためには、多様なデータを扱うアーカイブが増え、かつデータアーカイブ間での連携を実現することで、現在 SSJDA ではカバーできてないデータを保存して、利用可能性を高めていくことができるのではないかと考えています。

（座談会開催：令和４年１月２１日／聞き手：南山泰之）