JDCat サロン

データインフラの最前線

JGSS が JDCat に連動して データアーカイブを構築する意義

佐々木尚之（ささき・たかゆき）

大阪商業大学公共学部　准教授

2000 年アリゾナ大学人間発達学専攻卒業。2008 年テキサス大学オースティン校人間発達学博士課程修了。大阪商業大学 JGSS 研究センターポストドクトラル研究員、主任研究員、総合経営学部助教、講師を経て、2018 年より現職。

大阪商業大学 JGSS 研究センターにて JGSS や EASS に関する業務に携わられている佐々木尚之さんに、JGSS 研究センター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。

私自身の研究テーマは、ライフコースの観点から生涯発達を読み解くことです。特に、家族形成のプロセスについて大規模調査データを解析し、子どもを生み育てやすい社会を構築する方策について検証してきました。

――現在はどのような仕事を担当されていますか？

大阪商業大学 JGSS 研究センター[1]には博士号取得後の 2008 年 4 月から参画し、JGSS プロジェクトの調査設計、研究会・シンポジウムのなどの企画、海外共同研究機関との交渉、データクリーニング、公開データの整備などプロジェクトに係わるすべての業務に携わってきました。現在では、JGSS 研究センターの運営委員として、Japanese General Social Survey（JGSS）[2]および East Asian Social Survey（EASS）[3]に関する業務を中心に行っています。

――ではまず、JGSS 研究センターが扱うデータについてお聞かせください。

本センターでは、代表サンプルを用いた質の高い反復横断大規模社会調査データを扱っています。データには、学歴、収入、家族構成、政治信条、行動パターン、宗教などプライバシーに関わるデリケートな情報が含まれるため、個人が特定されないよう徹底したデータ管理を行っています。また、公開用のデータには ID 番号をランダムに付与する、居住する都道府県情報を削除する、といった匿名化処理[4]を実施し、データから個人を特定できない状態にして提供しています。

---

[1] 大阪商業大学 JGSS 研究センター
https://jgss.daishodai.ac.jp/index.html

[2] JGSS（Japanese General Social Surveys）は、大阪商業大学 JGSS 研究センターが実施している、日本人の意識や行動を総合的に調べる大規模な社会調査。調査項目は、就業や生計の実態、世帯構成、余暇活動、健康状態、犯罪被害の実態、政治意識、家族規範、死生観など多岐にわたっており、2022 年 4 月現在、計 20 回の調査が実施されている。
https://jgss.daishodai.ac.jp/introduction/int_jgss_project.html

[3] EASS（East Asian Social Survey）は、欧米の研究者が中心になりがちな国際比較調査において、東アジア社会に特有な問題や関心に基づいた共通の設問（モジュール）を設定し、国際比較分析を行おうとする取組み。EASS は、独自の国際比較調査項目を新たに作り出すのではなく、それぞれの国・地域ですでに継続的に実施されている社会調査の中に、共通の設問群（モジュール）を組み入れることで国際比較を行う点に特徴がある。
https://jgss.daishodai.ac.jp/introduction/int_eass_project.html

[4] 匿名化処理とは、特定の個人を識別することができないように個人情報を加工し、当該個人情報を復元できないようにする処理を指す。

――個人情報が含まれるデータは、公開に当たって特に気を遣うポイントになりそうですね。

JGSS 研究センターでは、個人情報の流出リスクを基準に公開データ、限定公開データ、オンサイト利用データの 3 つを用意しています。公開データは、上述したように匿名化処理が加えられたデータで、特に利用制限はありません。限定公開データには、居住する都道府県情報が含まれており、利用には研究計画書の提出と本センターによる承認が必要となります。都道府県情報だけではデータから個人を特定することもほぼ不可能ですが、リスクを最小限にするための措置です。オンサイト利用データには、サンプルの抽出地点、つまり住所情報が含まれています。限定公開データと同様に事前申請でのデータ利用となりますが、承認された利用者のみが利用できること、本センターへの来所が必要なこと、の 2 点に違いがあります。センター外へのデータ持ち出しは禁止されており、セキュリティを強化したオンサイトルームで、本センターが貸与する PC を利用して分析を行ってもらいます。

――厳密なアクセス管理がなされているのですね。データを公開した後にも気を付けていらっしゃる点があるのでしょうか。

法改正や官公庁の運用変更によって公開可能な範囲が変わることがあり、常に注意を払っています。例えば、JGSS では設問ごとの単純集計や変数コードをコードブックとして公表していますが、以前は調査対象の自治体情報なども掲載していました。一方で、住民基本台帳法の一部改正に伴う総務省令の一部改正（2006 年）により、自治体の多くが、調査対象者を抽出した地点の範囲を公表し始めました。その情報だけで特定されることはありませんが、いくつかの情報を組合せて特定しようとする可能性があるので、JGSS 研究センターでは過去のデータに遡って調査対象の自治体情報を削除し、総務省に注意喚起しました。その他にも、寄託後に必要に応じて修正を加えることはあるため、継続的なメンテナンスは欠かせません。

――続いて、JGSS 研究センターのデータアーカイブ機能についてお聞かせください。

JGSS 研究センターと他の拠点機関との大きな違いは、本事業開始までにデータアーカイブ機能を有していなかった点にあると考えています。これまで、JGSS データの利用希望者は、我々がデータを寄託している国内外の 4 つのデータアーカイブ（SSJDA[5] 、ICPSR[6] 、GESIS[7] 、EASSDA[8]）のいずれかに利用申請し、データを取得していただいていました。しかしながら、我々の成果報告として重要である、データ利用者に関する情報を外部データアーカイブから入手することが困難になってきたことや、データを寄託してから公開までの時間が 1 年以上かかるようになってきたことなどから、自前のデータアーカイブ（JGSSDDS）を構築することを決断しました。時を同じくして、国立情報学研究所（NII）で開発中の WEKO3[9]では、データの公開機能が実装される予定であったことから、NII と共同で制限機能を付加した新たなデータ提供システムの開発に着手しました。さまざまなトラブルに見舞われてはいるものの、3 月下旬から一部が稼働を開始しました。

――自前のデータアーカイブ構築は、非常に大きな決断だったのではないでしょうか。

大規模な社会調査を継続的にやっていく際、一番大変なのが調査資金の確保になります。JGSS 研究センターでは複数の助成金を獲得しながら調査を実施しているため、毎年数種類の申請書と報告書を書く必要があります。そして、資金提供者へ報告する際には、

---

[5] SSJ データアーカイブ（Social Science Japan Data Archive）：https://ssjda.iss.u-tokyo.ac.jp/Direct/

[6] Inter-university Consortium for Political and Social Research (ICPSR)：https://www.icpsr.umich.edu/web/pages/

[7] GESIS：https://www.gesis.org/home

[8] East Asian Social Survey Data Archive：https://www.eassda.org/

[9] WEKO3：https://rcos.nii.ac.jp/service/weko3/

JGSS が作成したデータが海外の研究者にどのくらい使われているのか、大学院生の二次利用は進んでいるのか、といった利用者別の情報が重要です。このような情報は自前でデータアーカイブを持っていればすぐに集計が可能ですが、外部から取得する場合は手続きが煩雑になります。さらに、最近では海外データアーカイブにおける利用者情報の扱いが厳格化し、利用者情報の取得が難しくなってきたことも背景にあります。

――助成金の獲得に必要な情報が取得できない、というのは強い後押しですね。その他、構築によるメリットはどのようなものが考えられるのでしょうか。

自前でデータアーカイブを持つことによって、データ公開までの時間が大幅に短縮されることが期待できます。JGSS のデータを海外のデータアーカイブへ寄託する場合、長い時には公開まで 1 年以上かかってしまうケースがありました。データ利用者から公開時期に関する問い合わせも受けていましたが、自前のデータアーカイブであればキュレーション終了後にすぐ公開が可能なため、これらの声に応えられるようになります。また、データが手元にあることで、有用な分析アプリケーションとの連携を独自に進めることが可能になります。JGSS 研究センターでは R Shiny[10] を用いたアプリケーションの開発を進めており、大学の統計演習授業などでの活用を見込んでいます。WEKO3 によって、データを利用するために必要な許諾申請もオンライン上で完結できるため、教員にとってより使いやすい環境を提供できると考えています。

――自前のデータアーカイブを持つことで、海外データアーカイブとはどのようなすみ分けになっていくのでしょうか。

JGSSDDS が稼働すれば、いち早く海外の研究者へもデータを提供できる優位性が期待できます。一方で、データアーカイブにはそれぞれ特徴があり、データ利用者はメンバーシップの制限、データ利用料の有無、サポートの充実度、成果報告申請の容易さなど複数の要因に基づいてデータアーカイブを選択しています。JGSS は、戦略的にアメリカとヨーロッパでもっとも利用されている ICPSR と GESIS に英語データを寄託したことにより、海外利用者がもっとも多い日本の社会科学データの一つになったと確信しています。引き続き彼らに馴染みのあるデータアーカイブに寄託しつつ、共存していく形が望ましいのではないでしょうか。

――次に、JGSS における海外発信・連携機能の強化についてお聞かせください。

今回、新たに European Social Survey（ESS）[11] と MOU[12]を締結し、共通する調査項目の組み込みといった研究協力に向けた連携が始まりました。海外の機関と連携して調査を実施する際の課題は、継続して調査費用を確保する難しさ、調査票のすり合わせ、データクリーニングの考え方の違いなど枚挙にいとまがありませんが、この連携を通じて、日本を含めた国際比較研究の促進だけではなく、JGSS や EASS の認知度の向上や JGSSDDS の利用者の促進につながると期待しています。

―― ESS と協定を結ぶきっかけは何だったのでしょうか。

ESS から EASS チーム全体に対しては、以前から何度も共同研究の呼びかけがありましたが、ESS の求める共通設問の多さから実現していませんでした。ESS と JGSS の協定のきっかけとなったのは、COVID-19 の感染拡大による調査モードへの影響でした。ESS で

---

[10] Shiny とは、RStudio Inc.が開発している R 言語のパッケージの一つ。Shiny を用いることで、R 言語のみで簡単な web アプリケーションを作成することが可能。

[11] European Social Survey（ESS）：https://www.europeansocialsurvey.org/

[12] MOU（Memorandum of Understanding）は組織間の合意事項を記した了解覚書のこと。

は各国の調査において、調査対象者宅での面接法を基本としていたため、COVID-19 の感染拡大が大きく影響したようです。一方で、JGSS では、感染拡大までは面接法と留置法を併用し、緊急事態宣言時には留置法のみで調査を実施していたため、調査を継続できていました。ESS が作成した COVID-19 の設問の一部を JGSS に組み込むことを JGSS が連絡したことを契機に、ESS からは、留置法の有効性や実施手順について尋ねられました。その後、調査モードの相違解消に留まらない意見交換が続けられています。各国で望ましい調査方法について、住民リストの有無、人口密度、社会インフラ、国民性など各国で調査環境が異なるため、望ましい調査方法はそれぞれだとは思いますが、将来的により良い調査方法は何なのかを議論する上で有意義な協定になると思います。

――続いて、JGSS データの整備についてより詳しくお聞かせください。

JGSS 研究センターでは、1 年ごとに生成される調査データのほか、時系列変化を分析するための累積データを整備、公開しています。対象となるデータは JGSS データ、EASS データ及び一部のパネルデータですが、中心となるデータは前者の JGSS データ、EASS データです。JGSS データと EASS データに組み込まれている変数は大きく二つに分類することができ、一つは複数回にわたり同じ設問を尋ねる継続設問、もう一つは、それぞれの調査年で設定している研究テーマに沿って追加した時事問題です。利用者は、何らかの時系列変化を分析したい場合は累積データを利用し、あるトピックについて分析したい場合は、そのトピックが含まれる調査年データを利用することになります。EASS では 10 年ごとに同じ設問を尋ねていたり、JGSS では女性天皇や税負担に関する設問など調査時点での社会情勢によって不定期に尋ねていたりする変数もあります。

――累積データは、変数の扱いが難しい印象です。

JGSS は 20 年以上にわたり継続して調査しているため、蓄積された変数は膨大になっていますが、半数程度はコアな設問として同じ内容を尋ねています。継続設問には同じ変数名を付与しているため、初心者であっても時系列に分析することが比較的容易なように整備されています。一部、より正確な測定をするために、調査を重ねるにつれて尋ね方が少しずつ変わっている設問もありますが、そのような場合は変数名を分けるとともに注意書きをつけています。

――変数の同定作業もデータ整備の対象に含まれているのですね。

はい。JGSS/EASS では 1 年ごとの調査データでも同じ基準で整備しているため、変数をそのまま接続しても累積データを作成することができます。一方で、変数名は異なるものの、場合によっては時系列の比較が可能なデータもあります。調査票やコードブックも全て公開しているため、二次分析をする場合は、注意書きだけではなく必ず調査票やコードブックをご確認いただければと思います。

――続いて、JDCat についてお伺いします。まず、JGSS データは社会科学分野以外でどのような使われ方があり得るでしょうか。

JGSS データは広範なトピックを扱っているため、これまでも医療疫学、公衆衛生学、環境学、心理学、経済学、地理学、政治学など分野横断的に利用されています。例えば、JGSS データには調査対象者が花粉症かどうか、という事項を尋ねる設問がありますが、このデータと日本の植林情報を組み合わせて、どういう地域の人が花粉症になりやすいか、といった分析が行われた事例があります。データアーカイブ機能を強化していくことによって、これまで以上に人文学・社会科学分野以外からの利用が期待できると考えています。

――JGSS データのメタデータを作成するに当たっては、どのような課題があったでしょうか。

JGSS は調査名のとおり、日本人の意識や行動を総合的に調べるための調査なので、調査項目も多岐にわたります。メタデータの作成時には、できるだけ詳細な

トピックで検索できるように、基本的に”CESSDA Topic Classification”[13] を利用して作成することを求められました。しかしながら、JDCat で扱える統制語彙が 95 件とそれほど多くなかったので、JGSS で尋ねている調査項目を十分に反映できませんでした。例えば、JGSS データには精神的健康など心理に関する設問や価値観などを尋ねた項目が多く含まれていますが、このような設問に対応するトピックが少ないと感じました。扱える統制語彙を増やしていくとともに、統制語彙でカバーしきれない項目については、自由記述も選択肢にあって良いかもしれません。

――JDCat には、どのようなことを期待されていますでしょうか。

利用者が、JDCat から人文学と社会科学のデータを横断検索することになり、本センターが公開するデータをより多くの人に活用していただくことを期待しています。一方、利用者の観点から考えれば、JDCat ではメタデータの検索だけではなく変数情報の検索もできるようになるのが望ましいと思います。例えば、ICPSR ではキーワードで検索した場合、検索結果そのものだけではなく、質問文、選択肢、回答分布、被引用情報といった関連情報が一緒に表示されます。こういった変数情報によって、例えば同じ「幸福度」といった概念を尋ねた設問であっても、選択肢によって捉え方の違いが見えることもあり、利用者が必要なデータをより正確に見つけることが出来るのではないかと思います。

――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。

2008 年の帰国当時、研究者が自由に利用できる大規模社会調査データは非常に限られており、欧米との差に衝撃を覚えたことを鮮明に記憶しています。アカデミア発展のためには、データ分析のやり方が正しいのかどうか第三者の目でチェックすることが必要です。近年では、投稿論文の根拠データ公開が義務化されつつありますが、日本においても、研究データを共有し利活用する流れがこの 10 年程度で急速に進んできたと肌で感じてきました。このタイミングで人社データインフラ事業に関われたことを光栄に思います。

そのうえで、WEKO3 に組み込む形でデータアーカイブをもつという方法は、他の大学や研究機関のモデルになると思います。特に、今回実現した制限公開の仕組みは、調査データの匿名性を担保した状態でのデータ提供が可能になるため、データの利活用促進に大きく寄与すると思います。JGSSDDS の広報に励み、同様の仕組みを希望する大学や研究機関に普及する役割を果たしたいと思います。

また、昨年秋に追加することを決めたオンライン分析アプリケーションは、統計パッケージにアクセスできない学部生や大学院生に、プログラムに詳しくなくても分析することが出来る道を開くものです。JGSS では今後、利用可能な分析の種類を徐々に増やす予定です。ただ、変数の置換だけではなく、変数の加工をする機能を付加するためには、アプリケーションを搭載する仕組み自体を変更する必要があります。人社データインフラ事業が継続するようであれば、ぜひ、その点にも取り組んで、オンライン分析アプリケーションで利用できる統計の種類をさらに拡張したいと思います。

（座談会開催：令和４年３月２５日／聞き手：南山泰之）

---

[13] CESSDA Controlled Vocabulary for CESSDA Topic Classification：https://vocabularies.cessda.eu/vocabulary/TopicClassification