JDCat サロン

データインフラの最前線

パネルデータの利活用に向けて

石井加代子（いしい・かよこ）

慶應義塾大学　経済学部　特任准教授

慶應義塾大学パネルデータ設計・解析センターの設立時から在籍され、パネルデータ整備に携わられている石井加代子さんのデータアーカイブ構想をお聞かせいただきます。

2001 年、慶應義塾大学商学部卒業。2004 年、英国 LSE 修士号(社会政策)取得。2018 年、慶應義塾大学商学博士取得。医療経済研究機構(2007～2012 年)、慶應義塾大学商学研究科特任講師(2008～2017 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。

慶應義塾大学パネルデータ設計・解析センターには2008 年の設立時から在籍しており、家計パネル調査を用いて、所得の動態について研究をしてきました。昨年度、新型コロナウイルスの世界的流行を受けてコロナ特別調査を実施し、このデータを使ってどういった人に雇用面や所得面でダメージが大きかったかについての分析なども行っています。

――人社データインフラ事業では、どのような仕事をご担当されていますか？

企画・総括を担当しています。慶應義塾大学パネルデータ設計・解析センターの設立から時間が経ち、保有するパネルデータ[1]の規模も増してきました。センター外部からの利用も増え、本事業を通じて異なる立場の方にとっても使いやすいデータを提供するようインフラ整備を進めています。

――ではまず、センターの取り組みについてお伺いします。パネルデータの利活用について、センターではどのような取り組みを行っていますか？

基本的には、同一個人に毎年同じ質問を繰り返し行うことでパネルデータを構築していますが、細部において、質問の方法が変わったり、新しい質問を追加したりといったことがあります。利用者はデータ解析を行う際に、どういった変数が存在するのか確認する作業や、複数年のデータをパネルデータセットに整備する作業が必要となりますが、これがなかなか大変な作業になります。こうしたデータハンドリングにかかわる利用者の負担を軽減するため、最近の取り組みとして独自に変数カタログとオンライン分析ツールの開発、提供を行っています。

変数カタログは、パネル調査票に含まれる変数をデータベース化し、オンラインで検索可能にしたものです。1 回の調査の調査票は 60 ページ以上あり、分析に利用したい変数が存在するか、またどの年のデータに存在するかを確認するのは大変な作業であるため、変数を簡便に検索するためのツールとしての活用を期待しています。

オンライン分析ツールでは、オンライン上で即座に選択したデータを集計することができます。一般的に、データを解析するためには、データのダウンロード、データセットの構築、統計ソフトへの読み込みなどいくつかのプロセスを経る必要がありますが、各年の情報量が膨大で、そのうえ複数年にも渡る、縦にも横にも大きいデータを確認する過程では作業量が必然的に膨らんできます。このツールによって、縦にも横にも大きいパネルデータの中身を簡単に掴むことが可能になりますので、負担の軽減に繋がると考えています。また、統計処理に慣れていない利用者にとっては、解

[1] パネルデータとは、同一の個人などを継続的に観察し記録したデータのことを指す。クロスセクション・データや時系列データと比較して情報量が圧倒的に多く、変化の把握や因果関係の特定などに役立つ。

析のハードル自体が下がる効果が期待できると考えています。

現時点では、当センターの主要なデータである日本家計パネル調査（JHPS/KHPS[2]）のみに対応していますが、将来的にセンターが保有するパネルデータに対応していく予定です。また、利用登録は誰でも可能[3]なので、社会科学分野の研究者だけではなく、行政担当者等の活用にも期待しています。

――システムを開発、運用するうえでの課題はなんでしょうか。

データ利用者のデータハンドリングの手間を軽減するために、どんなサービスが実用的なのか、利用者目線で検討することを心掛けています。変数カタログの整備に当たっては、年ごとに変数の ID が異なっていたため、システムに落とし込むためのデータ整備にかなり時間がかかりました。オンライン分析ツールを開発する際には、適切な外部ツールの選択や、変数ごとに適した集計の見せ方などの工夫に時間がかかりました。

また、変数カタログとオンライン分析ツールに加えて、サンプルバイアスを補正するウエイトの作成・公開もしました。パネルデータでは、年々調査に参加されなくなる対象者が現われるため、データにバイアスが生じる可能性があります。それを補正するためにウエイトが必要になるのですが、諸外国のパネルデータを参考に、当センターでもウエイトを作成して、データ利用者に公開しています。データの質を高め、利用しやすく信頼性の高いデータを目指しています。

――その他、データ整備を行ううえでの課題はありますでしょうか。

諸外国の機関と比較すると、日本では研究者が研究と並行してデータを整備している状況が目立つように感じます。豪メルボルン大学が実施する家計パネル調査（HILDA）などでは、データアーキビストにあたる方が長期間にわたりデータ整備を担当しており、データアーカイブの構築に大きく貢献しています。

研究者が自身の研究と並行してデータアーカイブを構築するのは、利用者目線で携われるという良い面もありますが、時間的な制約が大きく、難しい面もあります。また、パネルデータの場合、調査が長期間にわたるため、データ整備面での情報の蓄積という観点からも、データアーキビストのような技術的人材やデータを熟知した支援職員を長期的に雇用し、分業を進めることが望ましいと思います。

――最後に、人社データインフラ事業を通じた今後の課題やご期待をお聞かせください。

JDCat[4]との連携では、パネルデータのメタデータ整備が新たな課題です。前回の記事では図書館員との連携にも言及があり、人材確保の観点でも可能性を感じています。できればメタデータ整備だけではなく、例えば変数カタログについても、データ検索の観点から何らかのデータ整備に関わる連携が出来ると嬉しいところです。その他、データの英語化の徹底や、提供データの整備・拡充、データへの DOI の付与などを進めたいと考えており、特に、データへの DOI 付与は重要な課題です。センターでは利用者へデータの利用報告をお願いしていますが、マニュアルでの情報収集には限界があるため、DOI 付与による利用状況の機械的な把握の可能性に期待しています。

こういった課題に取り組むことで、公共財としてデータの価値を高めることに繋がるものと考えております。是非継続的な支援をお願いしたいと思います。

（座談会開催：令和３年６月４日／聞き手：南山泰之）

---

[2] JHPS/KHPS は全国約 4,000 世帯を対象に 2004 年から開始された家計パネル調査。サンプル脱落を補うため、2007，2009、2012、2019 年に新規サンプルが追加されている。2004 年に開始された KHPS と 2009 年に開始された JHPS を 2015 年に統合。就業行動や貧困動態、実物資産の世帯間移転の実態など、多岐にわたる分析トピックを網羅している。

[3] 変数カタログはアカウント登録後、オンライン分析ツールはデータ利用承認後に閲覧可能。
https://www.pdrc.keio.ac.jp/pdrc/ja/

[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。