JDCat サロン

データインフラの最前線

公的統計データの利活用

田中雅行（たなか・まさゆき）

一橋大学　経済研究所　准教授

1993 年、図書館情報大学図書館情報学部卒業。総務省（旧総務庁）事務官（1993～2019 年）を経て、2019 年より現職。

一橋大学経済研究所にて公的統計データの整備に携わっている田中雅行さんに、公的統計データの利活用についてお聞かせいただきます。

――ご自身の経歴についてお聞かせください。

これまで、統計センター[1]、総務省統計局[2]等で公的統計調査の実施、集計、二次利用などの統計実務を担当してきました。2019 年より一橋大学経済研究所（以下、経済研究所）社会科学統計情報研究センター[3]に在籍し、学術機関の立場から公的統計ミクロデータの利用促進に関する業務に取り組んでいます。

――人社データインフラ事業へ参画されたきっかけは何ですか。

経済研究所においては、『長期経済統計（LTES）』等の歴史統計を始めとする様々なデータベースを構築・提供してきましたが、個々のデータベースが個別の環境やシステムで整備・運営されており、データを一元的に保存・共有する基盤がありませんでした。今回、データインフラ事業を通じて、新たに経済研究所データリポジトリ[4]を構築し、データアーカイブの整備を進めています。

――経済研究所が保有するデータの概要についてお聞かせください。

経済研究所は政府統計機関との繋がりが強く、長年、政府統計ミクロデータの利用促進や整備に関わってきており、公的統計分野のデータ利用に長けていることから拠点機関に選ばれた経緯があります。公的統計は幅広い分野で活用されることを前提に作成されており、経済学をはじめ、経営学・政治学・社会学を含めた社会科学全般、教育や社会福祉を含めた人文学全般、さらには統計学や情報学等の自然科学分野も含めた幅広い層で活用されることが理想です。また、公的統計データを扱う以前から収集してきた各種データも保有しており、経済学分野を中心に利用されています。

――現在、公的統計データはどのような分野で活用されているのでしょうか。

例えば、国勢調査データは経済学や人口学ばかりでなく、最近では GIS（地理情報システム）と連動させ、都市工学、都市計画の分野でも良く使われています。また、5 年に一度の社会生活基本調査データは、余暇活動の現状を把握するために始まった調査ですが、最近ではより一般的な生活行動や生活習慣などを把握するデータとして、社会学分野などで使われているようです。最新の調査が今年行われましたが、ここ 1 年のコロナ禍でこのデータは大きく傾向が変わるは

---

1 独立行政法人統計センター. https://www.nstac.go.jp/

2 総務省統計局. https://www.stat.go.jp/

3 一橋大学経済研究所 社会科学統計情報研究センター. https://rcisss.ier.hit-u.ac.jp/Japanese/

4 一橋大学経済研究所データリポジトリ. https://d-repo.ier.hit-u.ac.jp/

ずで、調査結果が大いに注目されます。その他、サンプルデータとして統計学の授業で用いられたり、匿名化の研究で用いられたりする例もあります。

――既に様々な分野での活用事例があるのですね。どのように二次利用促進の取り組みを行ってこられたのでしょうか。

データは、整備しただけではなかなか利用実績が伸びないため、利用者側である研究機関のネットワークを通じて利用を呼び掛けたり、経済研究所のデータを用いた共同研究に自ら参加する、などの取り組みを行ってきました。また、海外における利用促進にも力を入れてきています。公的統計ミクロデータは統計法のしばりがあり、調査票情報（生データ）は運用上国外へ持ち出せないことになっています。一方で、海外ジャーナルへの論文投稿を行う際にデータの提出を求められるケースもあり、利用者からの要望が高まっています。

――海外研究者からはどのような要望が寄せられているのでしょうか。

海外との共同研究では、海外研究者が日本のデータを利用しやすい環境を整備することも重要な課題です。経済研究所ではオックスフォード大学等の研究者と共同研究を行っており、利用促進の一環として、ミクロデータの符号表等メタデータを翻訳、整備しています。また、政府はデータのオンサイト利用を推進しており、今後は、データのリモートアクセスを見据えた動きにも対応する必要があります。また、利用者がデータを直接さわることなく、計算結果のみを返すプログラム送付型の『秘密計算システム』などの利用を見据えた取り組みも行っています。

――データアーカイブのシステム面についてお聞かせください。複数データベースを一元的に統合する方針とのことですが、難しかった点はありますか。

経済研究所が保有するデータは、古くは 1970 年代のものからデジタル化、データベース化を進めてきました。現在までに長期経済統計（LTES）[5]、JIP データベース[6]などをデジタル化し経済研究所ウェブサイトで公開していますが、システム的には当時のものを個別に引き継いでおり、データを一元的に保存・共有するといったデータアーカイブの枠組みにうまく乗れていないと認識しています。また、日本統計年鑑のデジタル化の課題としては、膨大な統計表のメタデータ（標題、表頭・表側項目等）を１表１表手作業で作成するという作業量の問題があります。

――デジタル化の大きな課題ですね。

元々の表データが時代によって異なる形式で作成されているため、OCR[7]での読み取り精度が上がらず、膨大な修正ポイントが発生しています。今のところ、目次情報を参照しながら、表題、表頭（ひょうとう）、表側（ひょうそく）といった重要な部分を人手でメタデータとして結びつけています。作業の自動化を模索していますが、人的なチェックを完全になくすところまでは至っていません。

――整備したデータは、JDCat によってどのような活用が期待できそうですか。

統計調査分野においては、公的統計と社会調査の両者を含めた横断的な検索が可能になります。従来、公的統計と社会調査は提供側、ユーザ側の両方とも別々のコミュニティが形成されていました。一方、データの観点からは同種の調査によって集められたデータですので、JDCat[8]で検索プラットフォームを統合するこ

---

[5] 長期経済統計（LTES）データベース. https://rcisss.ier.hit-u.ac.jp/Japanese/database/long.html

[6] 日本産業生産性データベース：Japan Industrial Productivity Database. https://d-infra.ier.hit-u.ac.jp/Japanese/ltes/b000.html

[7] OCR（光学文字認識）とは、印刷された文字や手書きの文字などをカメラやスキャナといった光学的な手段でデータとして取り込み、それを機械的に文字認識させることによってテキストデータを作成する技術のこと。

[8] 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。本事業は、学振が平成 30 年度から実施する事業で、人文学・社会科学研究に係るデータを分野や国

とにより、データ検索の幅が広がり、かつより効率的になることが期待できます。

――経済研究所データアーカイブの構築には、他にどのような方が関わっているのでしょうか。

経済研究所リポジトリの運営や日本統計年鑑のメタデータ作成は、主に経済研究所の資料室へ配属された図書館系の人間が行っています。OCR 入力などの単純作業は外注していますが、出来上がった目次データを JDCat メタデータに対応させる、といった作業は図書館員が担当しています。

――図書館員の方々との協働可能性は、これまでのインタビューでも何回か言及がありました。

一橋大学もそうですが、機関リポジトリの運用は図書館の方が主に担ってきており、業界の知識や経験と相性がいい業務だと思っています。図書館の方も全員が機関リポジトリに関わる知識を持っている訳ではないかもしれませんが、もともと十分なデータ整理の経験がある人材は限られており、適性はあると言えるのではないでしょうか。また、図書館員は横断検索の視点を提供でき、学生や大学院生に必要な知識を提供するレファレンス業務の土台を持っています。今後、研究者を中心に、図書館員、さらには非常勤職員や大学院生を含めた幅広い層での体制構築や人材育成を進めるにあたり、協働が必要と思っています。

――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、お考えをお聞かせください。

データアーカイブの役割は共有に尽きると考えています。本データインフラ事業の中でも謳われていますが、データを研究者が専有するのではなく、共有することで学問の裾野が広がるため、共有を促進するための仕組みづくりが肝要です。データの利用促進には、整備された使いやすいデータを大量に備えていく取り組みのほか、利用者も参加しながら使いやすいデータアーカイブを構築していけると良いのではないでしょうか。

――利用者コミュニティがデータアーカイブと一体となって発展していくイメージですね。

データの共有化、オープン化の取り組みは、どの研究者や学術機関も抱え、取り組んでいる課題ですが、個々の研究者や組織が独自に共有のための環境を構築するのはなかなか難しいと考えています。人社データインフラ事業によって拠点機関のデータ整備が大きく進みましたので、今後は拠点機関以外の組織のデータとの連携が進むと、コミュニティ拡大に繋がるのではないでしょうか。例えば、自分のところでメタデータ作成が難しい機関が出てきた場合、拠点機関に相談してメタデータ作成を支援するような取り組みなどが考えられそうです。データアーカイブがより良いものになるよう、持続可能な形を探っていきたいと考えています。

（座談会開催：令和３年１１月４日／聞き手：南山泰之）

を超えて共有・利活用する総合的な基盤の構築により、研究者がデータを共有しあい、国内外の共同研究等の促進を目指している。