始動したJDCat
データインフラの活用が切り開く知の世界

第２回から第10回までJDCatサロンのインタビュアーを務めていただきました、国立情報学研究所の南山泰之先生と、日本学術振興会にて人文学・社会科学データインフラストラクチャー構築推進センターのセンター長を務めていらっしゃる廣松毅先生に、これまでの座談会の振り返りや日本のデータインフラを取り巻く状況について対談いただきました。

広がるデータの活用事例

――人文学・社会科学分野のデータにアクセスする新たなインフラとして、2021年11月にJDCat[1)] の運用がはじまりました。比喩的に言えば、JDCatは研究室の書庫やコンピュータに眠っているデータに息を吹き込み、研究室や研究分野の壁を超えて活用する仕組みです。そこでお二人にまず伺いたいのですが、分野の違いを超えたデータの活用事例で印象に残っているものはございますか。

廣松：私の専門である計量経済学は、経済活動に関するデータを用い、経済学の理論にもとづくモデルを統計的に分析する学問分野です。したがって政府が公表する公的統計には馴染みがありました。もちろん経済統計が主でしたが、それ以外でもしばしば利用したのが、地域ごとの人口、就業状況などを把握するために行われる国勢調査のデータです。国勢調査は元々、人口問題や福祉、災害対策などの行政施策に利用されていますが、面白いと思ったのは、都市計画や民間の出店計画など年々用途が広がっていることです。

――廣松先生は、国勢調査の他、各種の公的統計について議論する政府関係機関の委員を長年務め、特に統計データの利用を一般に広げるための制度作りに尽力されてきました。

廣松（ひろまつ） 毅（たけし）

― 所　属 ―

日本学術振興会
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センター
センター長

― 略　歴 ―

1972年、東京大学大学院経済学研究科修士課程修了。ハーバード大学イェンチン研究所 客員研究員（1982～84年）、東京大学教授（1989～2009年）等を経て、2009年より東京大学名誉教授、情報セキュリティ大学院大学教授に就任し、2018年より現職。

南山（みなみやま） 泰之（やすゆき）

― 所　属 ―

情報・システム研究機構
国立情報学研究所
オープンサイエンス基盤研究センター
特任助教

― 略　歴 ―

2005年より国立極地研究所情報図書室に勤務。2007～08年、第49次日本南極地域観測隊に参加。東京大学駒場図書館（2011～14年）、国立極地研究所情報図書室（2014～18年）、東京財団政策研究所（2018～19年）を経て現職。2022年、総合研究大学院大学複合科学研究科情報学専攻博士課程修了。

1）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。https://jdcat.jsps.go.jp/

廣松：はい。研究レベルで扱うのは公的統計や経済統計に限られていましたが、大学で統計学の講義を担当しはじめてからはアンケートなどによる社会調査や、民間企業で利用される統計データも多く目にするようになって新鮮味を覚えました。特に興味を引かれたのは製造業で生産する製品をコントロールするのに使われていた品質管理に関わるデータです。品質管理は元々、アメリカの統計学者エドワーズ・デミング博士が日本に導入した手法ですが、日本で独特の発展を遂げました。統計データに基づく分析を核にしつつ、生産現場を小グループに分け、各グループからの提案を重視して改善を図る方式が、1980年代の高度経済成長を支えたと考えられています。トヨタのカンバン方式も有名です。

——南山先生は、図書館員としてデータを扱う実務に携わった経験をお持ちですね。

南山：国立極地研究所の情報図書室に勤務していた頃、地球科学分野における研究データのキュレーション[2)]に取り組みました。オーロラ、地磁気、大気などの研究データを扱いましたが、各々の性質は違っても、メタデータの次元で扱うと情報をうまく繋げることがあり、そこに面白さを感じていました。

日本で社会調査データの共有が遅れたワケ

——理系の研究分野では、規模の大きい観測装置や実験装置で得られたデータを研究者間で広く共有して、多角的に分析する仕組みが出来上がっています。それに比べると、人文学・社会科学系の研究分野ではデータを共有して活かす仕組みの整備が遅れているように見えます。なぜでしょうか。

廣松：その背景の一つは、戦前戦中の反省です。特に社会調査については、政府による世論操作への懸念から、公的統計は国民の意識に関わるべきではないという考えになり、戦後、政府は社会調査から撤退しました。その代わりに旧文部省直轄の研究所として生まれた統計数理研究所が、1953年に「日本人の国民性調査」を実施（5年ごとの継続実施）したのを嚆矢として、社会学者の方々がさまざまな社会調査を開始しました。その他に日本社会学会を中心とする「社会階層と社会移動全国調査（SSM調査）」などの大規模な調査もありますが、大部分は個別の研究者や小規模な研究グループによる調査分析に留まり、多数の研究者によるデータ共有の例はほとんどありませんでした。

南山：新たに社会調査を設計する場合には、まず既存の調査を参照する必要があります。ところが、分野横断的に既存の調査をカバーしようとしても、当時はデータを共有するための便利なインフラもなかったため、参照するために膨大な手間がかかることは想像に難くありません。結果として、自分の元々の研究分野の範疇の中で過去の調査を参照し、調査項目を設定せざるを得な

2）データを整理・統合・編集して活用しやすいように加工すること。

かったのではないでしょうか。

――どちらかというと各種散らばった社会調査を統合するよりは細分化する方向に進んできたわけですね。

南山：はい。さらに言えば、もともと多くの調査研究は新しい事象を見出したい、理解したいという動機からスタートするので、調査データを共有するよりも一から調査を設計するほうが、当初の関心に沿ったデータを得られやすいといった事情もあるかと思います。一方で、同じ研究分野における国内外のデータ共有はむしろ活発に進んでいたようです。日本と海外の研究者が共通の関心を持つことはよくあることで、現在でも社会調査データの国際比較はよく行われています。

廣松：個人による地道なデータ収集と分析の威力を示した事例として歴史的に有名なのは、19世紀半ばイギリスのブロード・ストリートで流行したコレラの発生源を突き止めた医師ジョン・スノウです。当時は病原体すら知られていませんでしたが、スノウはコレラ患者の居住地を地図に丹念にプロットすることで、コレラがブロード・ストリートの公共の水ポンプから広がったことを特定し、感染症の流行を食い止めるのに貢献しました。その意味で、スノウは近代の公衆衛生学の創始者とされています。また、彼は近代の麻酔学の創始者としても知られています。

――疫学や公衆衛生の先駆けですね。

廣松：スノウの例は、限られた知識、限られたリソースしかない時代の一人の医師による画期的な成果ですが、データ共有の仕組みがあれば、もっと容易に病気の原因を突き止めたり、対策を考えたりできたでしょう。

オープンデータの潮流

――Covid-19の場合は、中国武漢で最初の原因不明の肺炎の発生が報告されてから数日で新型のコロナウイルスであることが特定され、それから１年足らずでワクチンまでできた。遺伝子を高速で解析する技術などの進歩もさることながら、世界中で得られたコロナウイルスに関するデータをアップロードし、公開するプラットフォームの存在も、ウイルスへの理解から対策まで迅速に進められた要因に挙げられます。

南山：命を助けるという差し迫った課題があることも、医療・保健分野でのデータ共有が先行した理由でしょうね。一方で、多くの分野では学会ごとに大きく異なる研究対象を扱っているので、研究者が蓄積するデータを学会間で標準化して連携しようという動きはなかなか生まれませんでした。しかし情報通信技術の進展に伴い、各学会で進められてきた取り組みを広げる形で、他の学会とのデータ共有が少しずつはじまってきた印象です。さらに異なる分野のデータを組み合わせたり、部分的に用いたりすることで、新たな課題を解決する事例がいくつか出てきました。これらの事例をきっかけに、近年、データ共有への期待が急速に高まってきたと感じています。

廣松：誰もが自由にデータを入手し、分析できる「オープンデータ」（もちろん、機微なデータは例外ですが）、研究プロセスの透明化を担保した上で、複数の専門家、組織に加え、市民の力も結集して研究活動を行う「オープンサイエンス」へ向けた取り組みも世界中で活発化しています。この潮流に乗り、諸外国のように日本も国レベルでデータを蓄積、管理して、利用を促進するインフラを作るべきであるという声が高まってきました。それが本事業[3)]がスタートした理由の一つです。

3）学振が平成30（2018）年度から実施している「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」のこと。https://www.jsps.go.jp/j-di/index.html

データ活用の鍵を握るメタデータスキーマ

――この事業から生まれたのがJDCatですが、人文学・社会科学系のデータインフラを構築する上で、苦労されたのはどんな点ですか？

廣松：大変だったのは、メタデータスキーマの策定とデータの内容を表す用語を決める作業でした。異なる分野のデータを横断的に検索するには、各データに対して、それがどんなカテゴリーに属するかを表す用語を与える必要があります。これを専門的には「統制語彙」と呼ぶのですが、メタデータスキーマの策定と統制語彙の選定が最初の大きな課題でした。

――JDCatは、ユーザーが自由に入力できるキーワード検索欄の他に、対象地域、対象時期、トピックなどあらかじめ運用者側が用意した切り口で検索するファセット検索欄がありますね。たとえばトピックには「人口移動」「消費と消費者行動」「義務教育と就学前教育」「薬物乱用、アルコール、喫煙」「労働と雇用政策」「エネルギーと天然資源」「政治的イデオロギー」「エリートとリーダーシップ」「マイノリティー」「社会福祉政策」「社会変動」「運輸と旅行」などさまざまなジャンルのキーワードがざっと100個ほど並んでいます。

廣松：トピックに入れる内容は主に、欧州の社会科学アーカイブコンソーシアム「CESSDA（the Consortium of European Social Science Data Archives）」によって分類された用語[4)]を翻訳し、日本の事情に合わないものを捨象するなどして選定しました。本事業のセンターの研究員が諸外国の情報を収集した上で、CESSDAを採用したのですが、JDCat版を作るときに苦労したのは、同じトピックに対して研究分野ごとに異なる意味づけがなされている場合のすり合わせです。

南山：用語の選択は難問ですね。分野によってはデータを特徴づける用語の付与自体が研究の対象になる場合もあり、時代によってその解釈が変わることも珍しくありません。情報システムで扱う際には、結局は選択の問題としてエイヤッと決める他ないのですが、適当に決めると利用者にとって思うような検索結果が得られず、データ共有の事業自体が先に進みません。

廣松：どの分野でも言われることですが、こういう決め事は決めた途端に陳腐化がはじまります。したがって一度決めた用語もたえず更新しなければなりませんが、その作業には人手も手間もかかるので、慎重に用語を選択したら数年間はそのまま使うのがいいでしょう。

――メタデータスキーマの設定は終わったんですか。

廣松：社会科学系のメタデータスキーマの策定は終わり、現在は各拠点にそのスキーマに従い、それぞれのデータにメタデータを与える作業をしてもらっています。一方、事業開始して２年目から取り組んだ人文学系のメタデータスキーマの策定はまだ終わっていません。先に作った社会科学系のメタデータスキーマを人文学系にも引き継いだのですが、人文学系の研究者のみなさんから、「自分たちのデータと合わない」という意見が出てきました。この問題は今なお未解決です。

――社会科学系と人文学系のデータではかなり性質が違うんですね。

南山：研究で利用されるメタデータは、一義的にはデータを理解し、説

4) CESSDA Controlled Vocabulary for CESSDA Topic Classification https://vocabularies.cessda.eu/vocabulary/TopicClassification

明するために作られます。しかし社会科学系と人文学系では、それぞれのデータに対する理解の仕方が異なるので、そのデータを説明するメタデータの記述様式を合わせることが難しくなります。しかしJDCatのように横断検索を目的に据えたデータインフラであれば、たとえ理解の仕方を合わせることはできないとしても、共通する部分を抽出したメタデータスキーマを策定することができるんじゃないでしょうか。

――横断検索の次元で俯瞰すると、一見性質の異なるデータを繋げることができる。

南山：そうですね。一つ次元を上げないと、同じ社会科学分野であっても、隣の分野の研究者が言っていることがわからないという声は良く聞きます。

廣松：具体例で言えば、社会調査はアンケート調査により得られる量的データと、調査者による対象者に対するインタビューや、対象集団に対する観察などで得られる質的データに分けられますが、JDCatでは現状、量的データしか収録できていません。質的データの扱いはこれからの課題です。

データインフラ構築を議論のきっかけに

――質的データには、機微に触れる内容が含まれる恐れがありますね。

廣松：この事業で扱っているデータとは関係しませんが、大学で教員をしていた頃、ある修士課程の学生が、格差問題をテーマに修士論文を書きました。その中に、当事者に対するインタビューが含まれていました。論文としては出来映えのよいものでしたので、最終的に合格になったのですが、公開はされませんでした。修士論文は大学の図書館に収められ、原則として公開されますが、それに対してインタビュー対象者から公開されると思わなかったとクレームが付いたからです。プライバシーに配慮し、公開の了解をあらかじめ得ておくなどしなければならなかった事例ですが、この種の質的データについては20年、30年と時が経つうちに本人の気持ちが変わり、公開を了承するケースもあります。現時点で差し障りがあるからと廃棄したり、散逸させたりするのはもったいない気がします。

――戦中の庶民の日記は、当時の人にはそれほど価値がなかったはずですが、たとえば何をいくらで買ったといった些細な記述でも後世の人には大きな示唆を与えてくれます。

廣松：30年後、50年後にはじめて利用者が出てくるデータもよくあります。将来の情報の価値は現時点ではなかなか推し量れないので、一定のコストのかかるデータ化に二の足を踏むのも理解はできます。

南山：一方で、デジタル技術の進展により、一度データ化して整理すればストレージ費用はどんどん下がるので、まずはデータ化するまでの閾値をどう超えるかが課題です。データインフラが整備された海外の状況を見ると、データをどう共有するか、どう提供するかといった議論自体は、このような理解を前提にしているように見ています。日本では、これまで人文学・社会科学のデータインフラが少なかったので、そういう議論自体がしにくかったのかもしれません。

廣松：JDCatの運用をきっかけに、そういう議論の場ができれば我々としても嬉しいことです。日本でも、ある研究目的の下に集められたデータについて、たとえば外部に漏らしてはならないとか捏造してはならないなどのルールはありますが、いったん論文としてまとまった後のそのデータを扱うルールはこれまで曖昧でした。これらの議論が深まることで、議論を主導しているJDCat参加機関にデータの保存、管理を任せてみようという研究者が増えるといいですね。

求められる人材

――JDCatの拠点機関の方々、あるいはデータインフラに関わる研究や実務をされている方々への連続インタビュー「JDcatサロン - データインフラの最前線」は廣松先生の発案でスタートし、そのインタビュアーを南山先生が務められました。廣松先生に伺いますが、みなさんの意見をどうご覧になりましたか？

廣松：私たちJSPS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センターが決めたメタデータスキーマをもとに、各拠点機関の方々にメタデータの作成など作業をお願いしたので、それに対する不平不満がたくさん出てくるのだろうと予想していました。ところが実際には、データインフラ作りに役立ちたいという思いを語ってくれた方が多く、大変心強かったですね。ただし、全員が口を揃えて、データインフラを維持、発展させるための人材が足りないと仰っていました。この点は、大きな課題だと思います。

南山：たしかに、みなさん異口同音に人材が大事だと仰っていました。さらに、じゃあ、どうすればいいかと訊ねると、それぞれ異なる解決策を提案されているのが印象的でした。博士課程のプログラムとしてデータインフラの構築を担う人材を育成すべきだと言う方もいれば、研究者は研究に専念してデジタル技術に長けた人材による支援が必要だと言う方もいました。より実務的に、新たな人材を増やすのは容易ではないから、とにかく今いる人員の組み替えで運用体制を作るべきだと話す方もおり、期待される人材像に少しずつ違いがありそうです。

廣松：データインフラの構築に携わる人材には、デジタル技術、図書館情報学に加えて、個別分野に関する知識もある程度持ち合わせていることが望ましいのですが、そういう人材を確保するのが難しいのは確かです。

継続は力なり

南山：また、事業の今後に関する期待も多くありました。JDCatの登場で、人文学・社会科学のデータを分野横断的に検索することができるようになったわけですが、現場の研究者からは、このデータの可視化や分析をもっと容易にできるようにしてほしいという要望も出ていました。JDCatの運用がはじまったから、その先の研究を模索する動きが出てきた、と解釈しています。また、それぞれの拠点でデータ化の実作業に関わる課題が具体的に浮かび上がったのもJDCatの意義だと思います。それぞれの課題を解決していくことが、そのままJDCatの利便性の向上に繋がっていくものと考えられます。

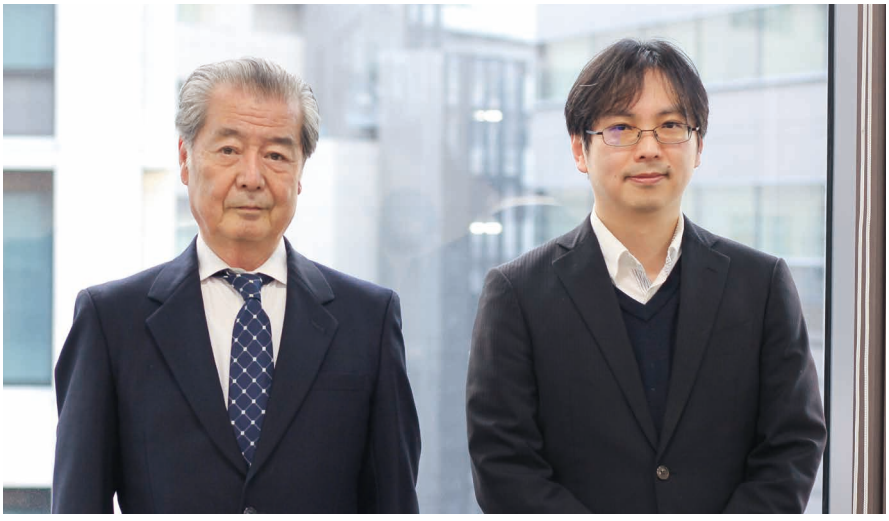
廣松：JDCatサロンのインタビューで東京大学社会学研究所の担当者の方が「データアーカイブはその性質上、半永久的に継続することそれ自体が役割である」と仰っていますが、まさにその通りだと思います。継続することが重要です。

南山：データアーカイブの継続については、特に国際共同研究の場面でより深刻な課題になると感じました。日本での研究に用いられたデータは、通常は日本のデータアーカイブに置くしかありません。そして、日本にデータインフラがなければ、こういったデータは海外から探すことができず、日本には比較対象とすべきデータがないと判断されかねません。

廣松：この事業の根底に、ご指摘のいわゆるJapan passing、Japan missingに対する危機意識がありました。

南山：これまでも、データインフラが存在しなければ生まれなかった研究成果が世界中で出ています。継続の仕方は各機関によって様々あり得ると思いますが、継続しなければ研究分野の壁を超えるような発展は望めないと考えます。

——JDCatが、思いもよらなかった異分野の繋がりの発見、社会的な課題の解決、あるいは人類の知の領域の拡大をもたらすことを期待します。どうもありがとうございました。



（聞き手、文：緑　慎也、撮影：緑　慎也、振興会事務局）