

特集

社会調査の 再現可能性と データ標準

特集担当

中野康人

関西学院大学社会学部 教授

科学的・公的な性格を備える社会調査では、近年その成果の再現可能性が重視されるようになってきている。その背景には、データをオープン化・標準化することによって研究成果の公正で有効な共有の促進をめざす世界的な科学界の潮流がある。ところで、データが共有されて再現可能性を備えるには、その内容を記述する共通の形式が重要になる。なぜなら、共通の形式で構造や要素が記述されたデータであれば、他者がそれを読み解いて利用することが容易になるからである。そこで注目されるのが、メタデータを含むデータの形式を整理して取りまとめた、データ標準である。医薬品情報や地理情報など、いくつかの学問領域ではデータ標準とその共有が確立しているが、社会調査の分野においてはどうなっているだろうか。

今回の特集では、オープンサイエンスや再現可能性に関する議論の現状を概観したうえで、社会調査の分野における、データ標準を整備しながらデータの共有を進めている事例を紹介する。読者諸氏がこうした内容にふれることにより、社会調査の共有知が発展することを期待するものである。



特集論文

1

社会と繋がるオープンサイエンス オンラインプラットフォーム上で完結する 社会調査に向けて

船守美穂

国立情報学研究所 情報社会相関研究系／オープンサイエンス基盤研究センター 准教授

オープンサイエンスの勢いがなかなか収まらない。2013年のG8科学大臣会合でオープンサイエンスへの言及が初めてなされてから（G8, 2013）、欧州中心にオープンサイエンス政策が展開され、日本においても内閣府を中心に政策が推進された（内閣府, 2015）。とはいえ、「オープンサイエンス」という用語が極めて曖昧であったため、一時期の流行として過ぎ去るだろうと思っていたが、なかなかたち消えない。

オープンサイエンスはもともと、デジタル技術の可能性を追求する過程で生まれた概念である。一義的には、研究成果や研究活動をインターネット上でオープンにすることを指す。しかし、第2章に説明するように、さまざまな社会的環境条件の下で推進されたことから、オープンサイエンスは社会からの要望に基づく、これら社会的課題を緩和する解決方法として発展し、この概念の持続性に繋がったようである。

本小論では、オープンサイエンスの概念がさまざまな社会的環境条件の下でどのように発展しているかを紹介する。また、後半では本誌の読者への示唆となるよう、オープンサイエンスと社会との関係を論考し、想定される未来の社会調査のあり方を最後二章に描く。

1 オープンサイエンスとは

オープンサイエンスは、とても分かりにくい概念である。直訳すると「オープンな科学」あるいは「オープンな学問」であるが、学問はそもそもオープン、つまり万人に共有され発展すると考えられているから、分かりにくい。オープンサイエンスに明確な定義がないことも、オープンサイエンスのわかりにくさに拍車をかけている。OECDの報告書にも、「オープンサイエンスには明確な定義はない」とある（OECD, 2015）。

オープンサイエンス政策の端緒となったEUの政策文書“Open Innovation, Open Science, Open to the World—a vision for Europe”は、オープンサイエンスを次に挙げる様々なオープンな取り組みの総称として定義づける。

例えば、論文のオープンアクセス化（OA化）、オープンデータ、オープンコード、オープンなラボノートブック、オープン引用などが挙げられている。オープンとは付かないが、プレプリント、科学ブログ、協働の目録作成、市民科学、さらにはデータ集中科学や、評価システムも挙げられている（European Commission, 2016a）。

この説明は分からないではないが、「オープンなものは何でもオープンサイエンス」とされてしまうと、どこまでがオープンサイエンスの範囲か

がわからず、途方に暮れる。

オープンサイエンスには学術的な定義があるわけではなく、政策立案者が政策のキャッチとしてこの用語を用いている。このため、政策によってオープンサイエンスの説明の仕方に濃淡があるのは致し方ない。

しかし一方で、各政策文書をよくよく吟味すると、政策横断的に共通項もある。表1に示すのは、筆者が複数の政策文書を吟味し、抜き出した共通項である。この共通項を通して、オープンサイエンスの理解に迫りたい。

まずポイントとなるのは、オープンサイエンスが「デジタル技術によって可能となる」ことである。つまり、オープンサイエンスは「学問がオープン」であることの一般通念を指しているのではなく、学術コンテンツなどがデジタル技術によってインターネット上でオープンに流通することを指す。

次にポイントとなるのは、オープンサイエンスが「学術研究および学術を整理するための新しい方法」であるということである。「オープンサイエンス」というネーミングの発端となったEUの「サイエンス2.0」に関する意見募集では、“‘Science 2.0’ describes the on-going evolution in the modus operandi of doing research and organising science.”としている（European Commission, 2014）。つまり、デジタル化の進展により、学術研究の方法が徐々に変わっていく様子をこの概念は表している。日本では近年、DX（デジタルトランスフォーメーション）が流行りであるが、オープンサイエンスはまさに「研究のDX」とも言える。

ちなみにEUでは、このサイエンス2.0の意見

募集において、「サイエンス2.0ではなく、オープンサイエンスというネーミングの方が良い」という指摘があり、2016年以降に大展開する政策の名称として「オープンサイエンス」を選んだ。なお、欧州では「サイエンス2.0」以前は「e-サイエンス」という名称で政策を推進しており、デジタル時代における研究のあり方を描き出すことに長い年月を費やしている。

最後のポイントは、オープンサイエンスが「アカデミアにおける価値基準を変える」ということである。これについては、各政策文書の執筆者も具体的なイメージはないままにこのように記述したと思われる。一方で、EUにおけるオープンサイエンス政策開始から5年が経ち、オープンサイエンスの議論の中心が研究評価や研究のパラダイム転換にシフトしていることを考えると、これもあながち誤っていなかったと言える。

このように、オープンサイエンスはデジタル技術によって可能となる新たな研究スタイルの模索であり、オープンであることにやや重点をおきながら、デジタル化の進展に伴ってその有り様も変化させていく概念と理解できる。

2 オープンサイエンスの駆動力

オープンサイエンスを「変わりゆく研究スタイル」として理解する場合、その有り様をかたどる外的な駆動力を認識することが重要である。オープンサイエンスがデジタル技術によって成り立つと言っても、技術的な可能性が開けるのみでは人間社会はなかなか変わらないことは、歴史の示すところである。人の心や、社会の慣習は、一番変わりにくい。これらに変化が生まれるのは、社会環境が変わり、人々の物の見方がそれに伴って変わってきたときである。

以下に、オープンサイエンスの議論で頻繁に論点として挙げられる社会的環境条件や外的な力を、ポジティブとネガティブな駆動力に分けて紹介する。なお、これらの駆動力については拙著（船

表1 オープンサイエンス関連の政策文書に見られる「オープンサイエンス」の説明の共通項

- | |
|---------------------------|
| 1. 学術研究および学術を整理するための新しい方法 |
| 2. デジタル技術により可能となる |
| 3. アカデミアにおける価値基準を変える |

（出典）筆者がOECD, 2015; European Commission, 2014, 2016a, 2016bの政策文書から抽出



守,2017)により詳しく論じているため、関心のある方はそちらを参照されたい。

A. ポジティブな駆動力

①データの有用性への着目(データ集中科学)

「データは新しい石油である」とも言われるように、データは学術の世界においても重要な資源となりつつある。

これまでのデータの利用には、学問分野によって濃淡があった。人文社会科学系の分野は、理工系の分野に比して扱うデータ量が少なかった。また、理工系の分野においても、宇宙・地球科学や一部の情報科学の分野では膨大なデータを扱っていたのに対して、生物系や実験系の分野においては、扱うデータ量は比較的少なかった。

しかし近年は、どの分野においても、扱うデータ量が多くなっている。人文学ではデジタルヒューマニティーズという学問領域が形成され、社会調査においても大規模な調査でないと国際誌に掲載されにくくなってきた。いまだに分野ごとに扱うデータ量の濃淡はあるものの、あらゆる分野において、膨大なデータを使いこなすスキルが、研究能力を左右するようになっていく。このためトニー・ヘイ(2009)は、豊富なデータに基づいて実践される科学研究を「データ集中科学」と呼び、これを「実験科学」「理論科学」「計算機科学」に続く「第四のパラダイム」とした(Hey, 2009)。

2013年のG8の共同声明やEUも、政策の重点を研究データの管理・共有においている(G8, 2013; European Commission, 2021)。オープンサイエンス政策は全体としては、1) 学術論文のOA化、2) 研究データの共有、3) オープン査読等、その他もろもろのオープン化の三本柱で進められているが、1)は過去20年以上、学術雑誌購読価格の高騰を背景に世界的に進められてきているため(本章⑥参照)、オープンサイエンス政策としての新しさは2)にあるのである。

研究データやその取り扱いのしやすさが研究

活動の要となったため、豊富な研究データを使いこなすためのツールやプラットフォームなどの研究環境の整備を通じて、国の研究力の強化が図られている。社会科学分野においても、社会科学データアーカイブやプラットフォームなどが各国にて整備されつつある。

なお、研究データとは、研究活動中のデータ生成・収集・加工・解析過程に関わるあらゆるデータのことを指し、社会における一般的なデータとは区別される。

②学際領域科学、イノベーションへの期待

あらゆる分野でデータが豊富に使われるようになったと言っても、研究者が研究に利用するデータが、研究者自身が生成・収集した研究データに限定されていると、データの恩恵は限定的である。他の研究者が生成した研究データを異なる観点から解析したり、異なる研究領域のデータをつなぎ合わせた学際的な分析をしたりしないと、新たな発見には繋がらない。また、研究枠組みについても、アカデミアの関心の範囲内に留めるのではなく、産業界や一般社会の視点を組み込んだほうが、新たな研究領域への展開やイノベーションにつながる可能性が高まる。

複数の分野のデータをつなぎ合わせたり、これまでと異なる視点からデータ分析がなされたりするためには、研究データがその生成者より広い範囲で共有され、再利用されなくてはならない。このためオープンサイエンスでは、研究データの共有に重きをおいている。

なお、研究データには個人情報等の機微なデータが含まれる場合があり、また、データ生成者がデータ生成の労に報いられる必要もあるため、研究データを何が何でも公開することは想定されていない。EUでは「可能な限りオープンに、必要な範囲でクローズドに(as open as possible, as closed as necessary)」という標語を用いている。「オープンデータ」ではなく、「研究データの共有(share)」という表現が用いられるのも、限定し

た人や特定のデータのみを制限公開することが一般的であるためである。

③社会とアカデミアの協働（市民科学）

オープンサイエンスにはしばしば「市民科学」の視点も含まれる。マイケル・ニールセンは『オープンサイエンス革命』という著書において、市民科学の視点を強く打ち出した（ニールセン、2013）。

市民科学とは、例えば希少動物の発見など、人手を必要とする研究において、一般市民の協力を得る研究の手法である。研究枠組みはアカデミアが設定するものとして説明されることが多い。しかし、⑤で説明するように、研究成果に対する社会からの圧力が全般に高まっているため、社会的課題の解決や、社会主導の研究なども今後、市民科学の一部として拡大していくものと思われる。

市民科学はオープンサイエンスの文脈において初期の頃から登場するものの、政策においては積極的に進められた形跡がない。しかし、2021年11月にとりまとめられる予定のユネスコの「オープンサイエンス勧告」のドラフトでは、「社会とのオープンエンゲージメント」という視点が強く打ち出されており、「社会との協働」がオープンサイエンスの今後の流れとなることが示唆されている（UNESCO, 2021）。

B. ネガティブな駆動力

④研究の透明性の確保、再現性の危機

オープンサイエンスは、研究の透明性を高める観点からも言及される。研究不正が世界的に多発する現状において、論文の根拠データを確実に保存し、問題が発生した場合に証拠として提示できるようにしておく必要性が生じている。国際的な学術雑誌においては「サプリメント（supplements）」として、論文に関連するデータやその他の情報を論文に付随させる仕組みが導入されている。

研究再現性の危機も指摘されている。『ネイチャー』誌が1,500名の研究者を対象に2016年に実施したアンケート調査では、「研究再現性の危機がある」と回答した研究者は9割に上った（Nature, 2016）。研究再現性の危機とは、自身あるいは他者の行った研究を再現できない状態のことを指す。分野によっても異なるが、同アンケート調査において再現性が最も低かった化学分野においては、「他者の研究が再現できなかった」とする研究者が9割近くおり、「自身の研究が再現できなかった」とする研究者も6割以上いた。心理学分野の学術論文100本を対象とした別の研究再現性の検証では、研究再現性が確認された論文は39本に留まり、6割強の論文は再現性が低かった（Nature, 2015）。

今日の学問は、他者の研究成果が正しいという前提のもとに、その知見の上に新たな知見を築くという「積み上げ型」のスタイルを採る。このため、個々の研究内容が再現できないということは、学術の健全な発展にとって致命的な状態である。研究不正は一部の研究者のみに関わるが、「研究再現性の危機」は研究者なら誰もが関わる。このような状態を是正するためにも、研究データを研究活動期間中から適切に管理し、研究終了後も確実に保存するという、「研究データ管理」が必要とされている。

過去に取得された社会調査データを提供する社会科学データアーカイブでも、利用者が不完全なデータを元に解析を行い、誤った社会科学上の知見が拡大再生産されないよう、提供するデータの確認を念には念を入れて行っている。

⑤研究成果の説明責任、重複投資の軽減

論文や研究データなどの研究成果を共有・公開したり、研究データを確実に管理したりするという意味でのオープンサイエンスの最も強い駆動力となっているのは、じつは社会からの圧力である。OECDの報告書や内閣府の報告書においても、「公的資金を得た研究成果」につい



ての社会への共有・公開の必要性が謳われている(内閣府,2015; OECD,2015)。

研究助成機関が論文等の研究成果の共有・公開を研究者に求める「OAポリシー」は、米国において社会からの圧力により成立した。

重病患者らが、自身の病状を調べるために学術論文にアクセスしようとしたところ、学術雑誌の購読費が高額すぎてアクセスできなかった。しかし、科学研究の多くは公的資金、つまり納税者の負担の上に成り立っている。科学研究が推進されるよう税金を負担しているのに、その研究成果を閲覧するには費用を再度負担しなくてはならないのか？ というのが、重病患者らの主張であった。この主張は米国議会において受け入れられ、研究助成機関におけるOAポリシーの導入につながり、世界に広がった。

社会調査も多くは公的資金を用いて行われている。大学の研究者が行う社会調査の原資となるのは、国から得る研究費であることがほとんどである。また、国勢調査など国の実施する調査は全て国家予算が原資である。そのうえ、社会調査の調査対象は国民であり、つまり、これらの調査のスポンサーである納税者である。

医療分野と同様、国民は社会調査のテーマに対してより発言権をもって良いし、その調査結果の共有やフィードバックを要求して良いのである。

⑥商業出版社対策

①～⑤のオープンサイエンスの駆動力は、アカデミア内の各研究グループに閉じている研究成果や研究活動を、アカデミア内も含む、社会全般に対して共有・公開するものであった。これに対して⑥は、商業出版社に囲い込まれている「学術論文を中心とした研究成果」をアカデミアの手に取り戻す動きである。

学術論文は研究者が執筆しているが、論文の査読を経て出版される際には「著作権」が出版社に譲渡されるため、研究者は学術雑誌を読み

たい場合は「購読料」を支払わないと、自身の論文にさえアクセスできない。ある意味、アカデミアは自身の生産した論文を買い戻していると言える(船守,2018)。

購読料を支払うという仕組みは、印刷体として学術雑誌が流通する時代においては、出版社が行う組版や印刷、郵送などに関わる出版コストを回収するための合理的な方法であった。しかし、デジタル時代となり、論文レイアウトも含め著者が自身で調整でき、インターネットを通じてほぼコストなしで論文を流通させることができるようになると、この仕組みは不合理である。加えて、商業出版社の要求する購読料が毎年4-5%値上がりすることもあり、世界のアカデミアは論文の「OA運動」を過去20年来、繰り返し続けている。

OA運動は、論文をインターネット上で公開(OA化)することにより、商業出版社による論文囲い込みの問題を解消しようとする動きである。商業出版社は当然のことながらこれに対して猛反発しており、20年以上にわたる激しい攻防の結果として、毎年出版される論文のうち何らかの形でOAである論文は約5割と言われている。

なお、OA運動は学術論文を中心に繰り返されているが、商業出版社は次のビジネスとして「研究データ管理」や、先行研究からデータ解析、研究成果の出版までを含む「研究プラットフォーム」の提供を視野に入れている。

気がつけば、商業出版社のサービスなしには研究活動ができず、さらには、こうしたサービスが毎年値上がりするという状況も想定されるため、商業出版社を排した研究環境の整備もアカデミアの一部関係者において検討されている(船守,2019)。

社会科学データアーカイブは、現状では国や大学などが運営するプラットフォームとして構築されていることが多い。しかし、商業出版社がいつ食指を伸ばしてもおかしくない状態であ

る。商業出版社が社会科学データアーカイブを運営するようになれば、国民は、自分たちのことを示すデータであるにもかかわらず、それを閲覧・利用するのに利用料を払わなければいけなくなる可能性が高い。そうならないように、脇を締めておかなければならない。

3 オープンサイエンスと社会

ここまで、オープンサイエンスが、デジタル技術が可能とする「変わりゆく研究スタイル」として理解されること、ポジティブ・ネガティブを含むいくつかの駆動力により徐々にその輪郭が生まれていることを紹介した。とくに研究データの価値に注目が集まるなかで、研究データの利用の範囲を拡げ、新たな研究領域やイノベーションを生むことに期待がかけられている。

これに加え、研究透明性や再現性の確保、社会に対する説明責任、重複投資軽減の必要性などが、学術を社会に対して開く圧力となっている。このような社会背景においては、研究データ等の学術コンテンツが研究者や企業、政府、社会のさまざまな人々の間で広く共有され、オンラインプラットフォーム上で協働作業が展開されることが国力に繋がるため、各国ともその環境整備への投資を拡大している。

学術の世界は長いあいだ、「象牙の塔」と表現されるように、社会から一定の距離をおいて発展してきた。社会が知識基盤社会に移行し、大卒者が社会に溢れるようになると、社会から学術界への要求が強くなり、社会連携講座や産学連携活動などが大学において行われるようになった。また、コンプライアンスや説明責任を果たすことの重要性が鮮明となり、多くの報告書や説明書類が作成されるようになった。

それでも、学術が育まれる場が、大学の研究室やゼミ室、論文が掲載される学術雑誌など、物理的な場や媒体にあった時代には、社会からの要求がいくら強くなっても、社会が学術に介

入る機会はそれほど多くなかった。

しかしデジタル時代となり、インターネットを介して人々が時間的にも空間的にもユビキタスに繋がるようになると、アカデミアがアカデミアの中だけに閉じて学術を展開する必然性は薄れ、社会から聞こえてくる要求にもより直接に反応できるようになった。ある意味、インターネットが、これまで潜在的にあった社会からの欲求を学術に繋ぐパスとなったと言える。

前章③で紹介した「市民科学」では、当該研究テーマに関わるオンラインプラットフォームが、市民と研究者の協働の場となっている。また、新型コロナウイルス感染症の世界的な拡大は、人々の活動をインターネット上に押しだし、インターネット上の協働活動に加速をかけている。

社会調査は、市民の協力のもとに社会の状況を把握するという観点からは、究極の市民科学と言える。オープンサイエンスの動きを捉え、社会調査を「市民とのコラボ」という図式の下に発展させることに期待がかかる。

ただし、社会調査の場合、調査を受ける市民の側に「学術に協力している」「研究者との協力のもと、社会の状況を可視化している」という意識が希薄なため、社会調査を市民科学と言えるようになるまでには、あともう一步のように見える。この意識が育つと、「○○という社会課題を可視化するための社会調査をともし企画しよう!」といった前向きな働きかけが、社会からアカデミアに対して寄せられるようになることが想像される。

現状においては、社会調査士や社会調査会社とアカデミアとのコラボを中心に、社会調査が成り立っていると言える。次章では、オンラインプラットフォームを介した「オープンサイエンス時代の社会調査」が、どのように発展していく可能性があるかを描いてみたい。



4

未来の社会調査への第一歩 オンラインプラットフォーム上で完結する 社会調査

現在、各国はオンラインプラットフォームの構築にしのぎを削っている。効果的なオンラインプラットフォームを構築することにより、情報の共有や研究活動、人々間のコラボレーションが進み、他国に抜きん出ることができるためである。オンラインプラットフォームは、材料科学やバイオ系の分野、地球・宇宙科学、人文系の分野など、多様な分野で構築されつつある。

オンラインプラットフォームの機能は分野に応じて多様であるが、概ね、①データや情報を保存・整理・共有・発信する機能、②それらデータを解析する機能、③掲示板やSNSなどを通じて人々がコミュニケーションをとりながら協働する機能などからなる。

社会科学分野においても、過去に行われた社会調査データを保存・提供する社会科学データアーカイブや、それらのコンテンツを横断的に検索可能とするデータカタログなどが形成されつつある。

国内においても、日本学術振興会の実施する「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」の一環として「人文学・社会科学総合データカタログ (JDCat)」が構築され、2021年7月に公開されたばかりである。このデータカタログは、四つの社会科学系データアーカイブと、一つの人文学系のデータアーカイブが提供するデータセットを横断検索可能とする（独立行政法人日本学術振興会、2021）。

JDCatは一方、社会調査の最終的なアウトプットを保存・提供するだけのため、上に挙げた①の機能のみを満たすに留まる。現在、ここで提供されるデータセットを解析するための環境が別途用意されつつあり、②の解析機能も近い将来、満たされる。他方、オープンサイエンスの醍醐味である③の人々が協働できる機能は依然として開発の予定すらない。

しかし今後、国内で行われる社会調査の多くが共通のオンラインプラットフォーム上で行われ、このプラットフォームが調査を受ける社会の構成員と、社会調査を企画運営する研究者や社会調査士、社会調査会社とのコラボの場となることができれば素晴らしいことではないか。

具体的には、社会調査専用のオンラインプラットフォーム上でインターネット調査が行われ、そこで得られたデータが同プラットフォーム上に保存され、利用に供せられると良い。データは、プラットフォーム上の解析ツールで解析され、さらに、解析結果や報告書、論文なども同じプラットフォーム上に公開される。解析に用いられたデータもプラットフォーム上にあるため、同じデータを利用して、解析を追証したり、新たな視点で解析したりすることで、一つのデータセットからの研究成果を重層的に発展させていくことが可能となる。

また、社会調査の前段階としては、社会調査で利用された設問や回答が同プラットフォーム上に蓄積されることになるため、過去に利用した質問群と選択肢が「クエスチョンバンク」としてデータベース化され、新たな社会調査を企画する際に参照可能となる。これらの質問群には、過去に行われた社会調査の回答結果や回答者の属性なども同時に蓄積されているため、同じ質問に対する異なる社会調査の回答結果の比較など、メタ研究も可能となる。

社会調査では依然として、紙面による訪問や郵送調査の方が信頼性が高いとされるが、現代の若い世代が中堅以上の年齢になる頃には、高齢者も含め、人口の大半がデジタル端末を使いこなしており、インターネット調査が違和感のないものとして普及しているであろう。

また、ここでは社会調査として、アンケート調査を想定した説明をしたが、インタビュー調査などの質的調査においても、このようなオンラインプラットフォームは有用となる。例えば、音声や画像データを集積し、これをテキストマイニ

ングするなど、AIを通じた解析ができる。AIは、学習データが豊富であるほど精度が上がるため、単一のプラットフォームにデータが集積されることは極めて有用である。

5 オープンサイエンス時代の 社会調査体制の構築に向けて

こうした将来の社会調査の実現に向けては、解決すべき課題もある。

社会調査のデータがオンラインプラットフォーム上で集積・公開される場合、個人情報の取り扱いについて注意を払う必要がある。これについては、データの機微の度合いに応じて、匿名化処理や制限公開等の適切な手続きを行う必要がある。また、オンラインプラットフォームは、情報セキュリティ上のパッチを随時更新していく必要がある。

個人情報や情報セキュリティへの対応はオンラインプラットフォームを運営する者にとっても荷が重い。しかし、個々の調査グループがこれらの対応をしていくことは不可能な場合が多いし、可能であったとしても、個々の調査グループがこれらに対して重複した努力を行うのは無駄であることは、誰しもが認めるところであろう。

それよりは、社会調査においては共通のプラットフォームを用い、そのような手続きはオンラインプラットフォームの事務局に任せの方が良い。そうすれば、個々の調査グループは研究内容を深めることに注力できる。

では、このようなオンラインプラットフォームを誰が担うのか。上述のように、オンラインプラットフォームの運用は相当に労力のかかる仕事である。片手間でできるものではない。専門性も要求される。情報インフラを運営するITの技術的知見とともに、社会調査のニーズも十分に理解できる必要がある。一人の人間が双方の知見を持たないにしても、どちらかの専門性は有し、もう一方の言っている内容を理解し、社会科学側のニーズと技術のバランスを調整できるだけ

の能力は必要である。

また、オンラインプラットフォームの運営だけでなく、各調査結果のデータクリーニングやデータキュレーション、メタデータや説明文章の付加、第三者がデータを手に取った時にわかりやすいように調整する担当者なども必要である。これらは専門職の仕事として発展させていかないと、アーカイブやプラットフォームの質も高まらないし、持続性も危ぶまれることになる。EUでは現在、こうした専門職のキャリアを確立すべく検討を進めている（European Commission, 2017；European University Association, 2021）。

日本でも、「人文学・社会科学総合データカタログ（JDCat）」が日本学術振興会の事業として構築されつつある。ただし、この事業では初期構築は行いが、その後の運営については保証されていない。また、事業継続があったとしても、日本学術振興会には社会調査コミュニティの細かいニーズをくみ取ることは難しい。

現状においては、委員会形式などで人文学や社会科学、情報技術の知見を取り入れて構築を行っているが、社会調査コミュニティにとって真の意味で使い勝手が良く、コミュニティの将来の発展につながるインフラを構築するには、社会調査コミュニティの主体的なコミットメントが必要であることは理解されるであろう。プラットフォーム運営に専門職が必要なことから、「社会と調査」の読者である社会調査士の方々などにも期待が集まる。

海外で推進されている「オープンサイエンス」は、とても急進的にみえる。しかし、その流れの背景を吟味すると、それがデジタル技術のみならず、社会との関係性も深まった現代の学術活動を効果的に進める上での必然であることが理解される。デジタル時代における「オープンサイエンス」という新たな研究スタイルの潮流を「対岸の火事」として眺めるのではなく、コミュニティの新たな未来を創るツールとして捉え、正面から向き合うことができるならば、日本の学術も健やかに



伸びる可能性が開かれる。

ぜひ、社会調査コミュニティとしても、自分た

ちの活動のしやすさやコミュニティの将来の発

展を考え、主体的に検討して欲しい。

文献

- 独立行政法人日本学術振興会, 2021, 「人文学・社会科学総合データカタログ(JDCat)」, (2021年8月8日取得, <https://jdcata.jsps.go.jp/>).
- European Commission, 2014, “Consultation on ‘Science 2.0’: Science in Transition”.
- European Commission, 2016a, “Open Innovation, Open Science, Open to the World—a vision for Europe”.
- European Commission, 2016b, “Amsterdam Call for Action on Open Science”.
- European Commission, 2021, “Open Science Monitor”, (2021年6月11日取得, https://ec.europa.eu/info/research-and-innovation/strategy/strategy-2020-2024/our-digital-future/open-science/open-science-monitor_en).
- European Commission Working Group on Education and Skills under Open Science, Directorate-General for Research and Innovation, European Commission, 2017, “Providing researchers with the skills and competencies they need to practise Open Science”.
- European University Association, 2021, “Reimagining Academic Career Assessment: Stories of innovation and change”.
- 船守美穂, 2017, 「オープンサイエンス推進に関わる学術機関の役割と課題」『情報知識学会誌』27(4): 309-322.
- 船守美穂, 2018, 「学術誌をアカデミアの手に取り戻す——オープンアクセスの最新動向と岐路に立つ日本」, NII Today, 82, (2021年6月11日取得, <https://www.nii.ac.jp/today/82/4.html>).
- 船守美穂, 2019, 「デジタル時代の学術情報流通変革期における電子ジャーナル問題の所在と対応」『大学マネジメント』, 173: 6-19.
- G8, 2013, “G8 Science Ministers Statement London UK, 12 June 2013”, Department for Business, Innovation & Skills, Prime Minister’s Office, 10 Downing Street and The Rt Hon David Willetts.
- Hey, Tony, 2009, “The Fourth Paradigm: Data-intensive Scientific Discovery” Microsoft Research.
- 内閣府, 2015, 『国際的動向を踏まえたオープンサイエンスに関する検討会』報告書.
- マイケル・ニールセン(高橋 洋訳), 2013, 『オープンサイエンス革命』紀伊國屋書店.
- Nature, 2015, “Over half of psychology studies fail reproducibility test”, (2021年6月11日取得, <https://www.nature.com/articles/nature.2015.18248>).
- Nature, 2016, “1,500 scientists lift the lid on reproducibility,” (2021年6月11日取得, <https://www.nature.com/articles/533452a>).
- OECD, 2015, “Making Open Science a Reality”.
- UNESCO, 2021, “Recommendation on Open Science (Draft)”, (2021年6月11日取得, <https://en.unesco.org/science-sustainable-future/open-science/recommendation>).

2

再現性ポリシーの 意義・現状・課題

三輪 哲

東京大学社会科学研究所 教授

下瀬川 陽

作新学院大学人間文化学部 講師

打越文弥

プリンストン大学大学院社会学部 博士候補
人口学研究所

工藤沙季

タンベレ大学大学院社会科学研究所 博士課程

1 はじめに

データを用いた経験科学的な研究に対して、再現性を問う声は近年とりわけ大きくなっている。日本においても、データの捏造など研究不正の事件が社会の耳目を集めたことは、多くの人びとの記憶に今なお刻まれているであろう。他にも、心理学系の論文を対象に分析結果の再現を試みたところ、再現できなかったものが多数を占めたことを報告する論文が有名誌に掲載されたことも広く知られている (Open Science Collaboration, 2015)。再現性は、科学の信憑性を確たるものにするための条件であるといえ、非常に重要であることは論を待たない。

ところで、そもそも「再現」とは何を意味するのか。ここで Freese and Peterson (2017) にしたがって、まとめておこう。彼らは、用いるデータが同一か否か、用いる分析手法が同一か否か、という二つの軸をもとにして、以下のように整理する。

第一に、同一のデータを同一の手法で検証し、まったく同じ結果が出るかどうかをみるもので、これを「検証可能性 (verifiability)」という。第二に、同一のデータに異なる手法を適用して類似した結果が出るかどうかをみるのが「頑健性 (robustness)」。第三に、異なるデータを同一手

法で分析するのが「反復可能性 (repeatability)」。そして第四に、異なるデータを異なる手法で分析しても類似した結果になるかどうかをみるのが「一般化 (generalization)」だとする。

これらのように、再現にもさまざまな類型がありうるが、本論で扱うのは主に第一に挙げた検証可能性、いわば狭義の再現性の問題である。

さて、社会調査の再現性を確認するには、必要なことが少なくとも二つある。それは、調査データが公開されることと、分析のコード (プログラム) が公開されることである。前者については、進みつつあるデータアーカイブの発展が後押しするものとなっているが、それでも公開されていないデータは実際にはあまたある (独立行政法人日本学術振興会 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事業運営委員会作業部会 [データ保有状況調査], 2019)。後者については、社会科学分野では分析結果を再現するために必要な手続きが必ずしも共有されているわけではなく、論文の記述だけから結果を再現することは困難である (板倉・尾崎, 2002)。研究成果が公表される段階において、分析結果を再現するための手続き自体、具体的には分析に用いたコードがデータと併せて共有されることが望ましいとされる (Key, 2016)。

再現性を担保するための指針のことを、「再現性ポリシー (replication policy)」と呼ぶ。そ



のポリシーのもとで、データやコードの共有のルールが明文化され、学術誌の投稿規定などに反映される。再現性ポリシーは、1980年代から議論が重ねられており、世界の学術雑誌や出版社において実際に定められるようになってきている。だが、現実問題として、それがどこまで浸透しているのかを理解している者は決して多くはあるまい。そこで本論では、再現性ポリシーに関して、その意義・現状・課題を概観すること、そして、研究者が自身の研究の再現性を高める方策には何がありうるかを紹介することを目的とする。

2 再現性ポリシーの意義

まず、再現性ポリシーはどのように必要とされるのか、その意義をまとめておこう。最も大きくいえば、再現性ポリシーは反証可能性 (Popper, 1959=1971) を担保するという意味で、科学の発展に寄与するということができるだろう。さらに細かい論点として、再現性ポリシーは研究者コミュニティに対してはもちろんのこと、論文を投稿する研究者自身に対しても、以下のようなさまざまな意義があることが提起されてきた。

第一に、ミスを減少させ、科学研究の質を守るものとなる。再現性ポリシーが遵守され、分析に用いたデータやコードが公開されれば、その分析が再現できるかどうかのチェックが可能になる。そして仮に誤りがあれば、それを発見し、研究成果そのものの撤回や再検討もできるようになる。著者はそうしたプレッシャーにさらされるために、分析をより注意深く行い、自身で事前に修正をするようになる (Anderson et al., 2008)。

第二に、「仮説適合性バイアス」を防ぐことができる。研究者は通常、たくさんのモデルの分析結果を出力するものの、実際に1本の学術論文に掲載するのは、最終的に選択されたモデルの結果に限られる。そうした状況では、自身の

提示した仮説に最も適合的な分析結果を選択する誘因が生じる。これを仮説適合性バイアスという。もし仮説が正しいのであれば、最終的に選択されたモデル以外でも、結果が頑健に確認されるべきである。こうしたバイアスを防ぐうえで、再現性ポリシーは重要な役割を果たすとされる (Firebaugh, 2007)。

第三に、再現性ポリシーの普及が将来の研究者、教育者にとって利益をもたらす (King, 2007)。データやコードの公開により、最新の優れた研究を再現することからはじめて、模倣し、スキルアップし、ひいては自身の研究に活かすことができる。学部や大学院における統計分析の授業では、学生の関心に近い研究の分析結果を再現するトレーニングを行うこともできる。つまり、研究者の育成や、学界全体のレベルアップの礎となるわけである。

そして第四に、被引用数を増加させるという、著者に対するメリットがある。Gleditsch et al. (2003) によれば、データを提供することを義務付ける雑誌に掲載された論文は、そうでない論文に比べて約2倍引用される傾向にある。さらにデータだけではなく、プログラムも提供する、すなわち再現性ポリシーを定めている雑誌に掲載された論文は、そうではない論文に比べて約3倍引用されることが報告されている。

3 再現性ポリシーの展開

管見の限り、社会科学の分野において、再現性に関する議論が表面化したのは、1980年代以降のことである。たとえば経済学においては、1986年に *Journal of Money, Credit and Banking* 誌に掲載された論文の再現性検証がなされた (Dewald et al., 1986)。これを受けて、*American Economic Review* (以下、AER) 誌で編集委員会の方針が発表され、再現性に関する条件が著者に求められるようになった (Ashenfelt et al., 1986)。ただし、そのときにAERが定めた「他

の研究者の要請に応じて著者がデータを提供する」形式は、再現性ポリシーの基準に照らして考えれば弱いものであったといえる。

その後、McCullough and Vinod (2003) によって、1999年のAER掲載論文が再現性ポリシーを守っているかどうかを検討された。その結果、半分の論文がこのポリシーに反していたことが明らかとなり、2004年からは再現用のデータとプログラムの共有が義務付けられるようになった (Bernanke, 2004)。AERの決定に続き、*Econometrica* や *The Review of Economic Studies* などのその他の経済学主要誌でも、データとプログラムの双方を共有することを義務付けるようになった。

政治学における再現性ポリシーの展開は、1990年代のアメリカで始まった。King (1995) によれば、1994年にアメリカ政治学会 (American Political Science Association) の政治学方法論 (Political Methodology) 部会において、政治学分野の雑誌編集者に対して、論文掲載の際に再現用データの提供を要求することを決議したとされる。

その後、アメリカ政治学会の機関誌 *American Journal of Political Science* (以下、AJPS) を皮切りに、再現性ポリシーをもつ学術雑誌は増加していった。1995年時点ではまだAJPSと *Political Analysis* のみで (King 1995)、2013年時点でも政治学の主要120誌のうちの18誌しか再

現性ポリシーをもっていなかったが (Gherghina and Katsanidou, 2013)、2015年に46誌が共同声明の形で再現性ポリシーを持つことを宣言した。一部のジャーナルではデータとコードを共有するにとどまらず、出版前に再現性をチェックする取り組みも既に始まっている (Stockemer et al., 2018)。

4 アメリカの現状

それでは、さまざまな議論を経て、2021年現在のアメリカにおける再現性ポリシーの実施状況はどうか、確認してみたい。表1に、社会科学諸分野の主要学術雑誌の現状を整理して示した。これらはいずれも、アメリカのみならず世界のトップジャーナルとみてよい。

対象とした10誌のうち、再現性ポリシーを明記して運用開始しているものは6誌であった。社会学分野ではまだポリシーとして定めるには至っていないようである。データの公開を必須とするものは5誌、推奨とするものは3誌となっており、データの公開について言及するのが多数派といえる。しかし、コードの公開についてはそこまでではなく、経済学と政治学の5誌にとどまる。なお、再現性のチェックまで要求する雑誌は二つであった。

表1 米国の主要ジャーナルにおける再現性ポリシーの状況

雑誌名	分野	データの公開	コードの公開	出版前の再現性 チェック	ポリシーを導入 した年
American Journal of Political Science	政治学	必須	必須	有	1994年
Journal of Politics	政治学	必須	必須	有	2015年
American Political Science Review	政治学	必須	必須	----	2010年
American Economic Review	経済学	必須	必須	----	2004年
Econometrica	経済学	必須	必須	----	1991年
Demography	人口学	推奨	----	----	1994年
Sociology of Education	教育社会学	推奨	----	----	NA
American Sociological Review	社会学	推奨	----	----	NA
American Journal of Sociology	社会学	----	----	----	NA
Social Forces	社会学	----	----	----	NA



5 北欧の現状

これまでにみてきたアメリカの事例は、世界の学術界のフロンランナーに位置づけられるだけのことはあって、議論、実践ともに先進的である。ただし、海外の諸国がそれにキャッチアップしているかといえば、必ずしもそうではないかもしれない。そこで、英語圏ではない海外の事例を確認するため、フィンランドを中心に北欧の現状をみてみよう（表2）。

北欧における経済学、政治学、社会学の主要な学術雑誌では、再現性ポリシーやデータ公開に関して明文化されているかどうかを調べてみた結果が、表2である。最も厳格に再現性ポリシーを定めているのは、*The Scandinavian Journal of Economics* 誌であった。データの種類によっては例外も認められようだが、原則としてはデータおよびコードの公開が必須とされている。ただし、再現性のチェックまでは求められていない。それに準ずるのが社会学分野の *Acta Sociologica* 誌で、データとコードの公開を推奨することが明記されている。英語で刊行される学術雑誌は、たとえばSage社のSAGE Research Data Policy (option1) のように出版社がポリシーを定めていることもあって、再現性ポリシーが

定められることが多いようである。

その一方で、フィンランド国内の学術雑誌については、再現性ポリシーがまだ未整備であることもわかる。ただし、これらのなかにも、*Kansantaloudellinen aikakauskirja* 誌のように、再現性を検証する論文の投稿を促すものもあるようだ。

6 日本の現状

続いて、日本国内の社会科学系分野について、再現性の担保がどの程度ルール化されているのかを検討するべく、主要な学術誌の投稿規定および執筆要項を調べた¹⁾。

その結果、明確に再現性ポリシーについて表明しているのは、海外の出版社がかかわる三つの英文誌のみであった。それらのうち、日本経済学会の機関誌である *The Japanese Economic Review*、および行動計量学会の機関誌である *Behaviormetrika* はSpringer社から出版されており、共通のResearch Data Policy (type1) に則っている。すなわち、「著者は、そうすることが可能である・適切である場合には、研究結果を裏付けるデータを公開リポジトリに寄託することが奨励される」というものである。また、日本社会学会の英文誌である *International Journal*

表2 北欧の主要ジャーナルにおける再現性ポリシーの状況

雑誌名	分野	データの公開	コードの公開	出版前の再現性チェック	言語
<i>The Scandinavian Journal of Economics</i>	経済学	必須	必須	----	英語
<i>Acta Sociologica</i>	社会学	推奨	推奨	----	英語
<i>Scandinavian Political Studies</i>	政治学	推奨	----	----	英語
<i>Nordic Journal of Working Life Studies</i>	労働研究	----	----	----	英語
<i>Scandinavian Journal of Work, Environment & Health</i>	学際	----	----	----	英語
<i>Kansantaloudellinen aikakauskirja</i>	経済学	----	----	----	フィンランド語、英語
<i>Sosiologia</i>	社会学	----	----	----	フィンランド語、スウェーデン語、英語
<i>Politiikka</i>	政治学	----	----	----	フィンランド語、スウェーデン語

of Japanese Sociology」では、「論文中の分析に使用したデータやスクリプト、その他のアーティファクトについて、それらがデータリポジトリを通して公開され利用可能な場合、著者は論文内で、そのマテリアルの場所への参照を含める必要がある」としている。ただし、3誌とも、再現性の担保に必要なマテリアル（データやコードなど）の公開を義務付けているわけではなく、あくまでも著者の裁量にゆだねており、推奨に相当するといえる。

確認した学術誌のうち和文誌には、投稿規定や執筆要綱のなかで明確に再現性の担保を求めているものはない。しかしながら、この点をもって、広義または狭義の再現性が日本の社会科学では軽視されている、とはいうことはできない。ルールとして定めるには至っていないが、再現性についての議論自体は、各分野においてみることができる。たとえば、社会学分野では板倉・尾崎 (2002) や打越・三輪 (2018) が、心理学分野では池田・平石 (2016) が、それぞれの分野における再現性問題をめぐる議論の整理を行っている。

7 再現性ポリシーの実施にまつわる課題

既に述べたように、再現性ポリシーには理念的にはひろく認められるに値する意義があり、それに則ったデータやコードの公開も進められている。しかしながら、現実には実施する段階になると、少なからず課題が残ることもまた事実であろう。

第一の論点は、どの段階で再現性のチェックを行うか、である。大きく分けると、(1) 論文執筆の前の段階、(2) 執筆後だが出版前（審査時）の段階、(3) 出版後の段階、の三つである。

心理学分野では、異なるデータを用いても同じ結果を再現できるかという点が重視され、それを阻害する要因として、統計的妥当性を損なうような研究デザインの選択や、論文として

提出する際の出版バイアス（仮説適合性バイアスも含む）の存在が挙げられている。ゆえに、再現性担保のための方策として、論文執筆の前段階での「事前審査つき事前登録制度（pre-registration system）」が挙げられている（池田・平石, 2016）。ただし、これは狭義の再現性（検証可能性）に関しては対象外であり、あくまで新たな一次データ収集に伴う広義の再現性（反復可能性、一般化など）とかかわるものとみてよい。

出版前の段階においてチェックをするのであれば、提出あるいは公開されたデータおよびコードに基づいて再現性を確かめることになる。出版後の段階でチェックをする場合でも、やはり公開されたデータおよびコードを使用した追試により、再現性が確かめられる。調査データの二次利用が広まってきた社会科学では、一般に、これらの段階でのチェックのほうが現実的である。

第二の論点は、誰が再現性のチェックを行うか、である。出版社の立場からすれば、計量分析の論文が出版後に批判される可能性を案じ、査読者ならびに編集側が再現性チェックを行う必要があると考える。だが、査読者が再現性チェックを行うことになれば、場合によっては査読を拒否する研究者も出てくるかもしれない、あるいは査読前に編集委員会の判断で論文をリジェクトする可能性が高まるかもしれないと、懸念する意見もある（Abbott, 2007）。

他の選択肢としては、アップロードされた再現用ファイルをデータアーカイブのスタッフが検証するという方法もある。AJPSでは検証プロセスにて、ノースカロライナ大学チャペルヒル校のオーダム社会科学研究所のアーカイブ・スタッフが、分析結果が再現できるかどうかを検証し、この検証が成功しなければ、論文は掲載されない。ただし、その再現性チェックに平均46.32日を要することが報告されている（Jacoby et al., 2016）。

もちろん、出版後に任意の研究者が再現性を



検証することもありうる。実際、既にみたようにいくつかの雑誌は、再現性を検証する論文の投稿を推奨している。だが、それでは事後的に問題の発見がなされるだけで、出版された論文の保証にはなりえない。出版前に再現性チェックを行わない経済学においては、データとプログラムが提供されたとしても、分析結果を完全に再現できたものは、経済学主要誌に掲載された論文でさえ、わずか14%程度という指摘もある(Gertler et al., 2018)。

第三には、どこまで厳しく、再現性の手続きを行うか、である。Transparency and Openness Promotion (TOP) Guidelines (<https://osf.io/9f6gx/wiki/Guidelines/>)では、データやコードだけでなく、分析デザイン、引用、分析プランの事前登録などさまざまな面についての透明性と公開性の基準をレベル分けしている。Liu and Salganik (2019)は、それを、レベル1：リクエストに応じて著者が公開、レベル2：出版前のアクセス、レベル3：出版前の検証、と整理する。つまり、最も厳格に再現性チェックを行うのであれば、きちんと検証をして再現性が確かめられてから出版されるべきである(レベル3相当)。もう少し緩やかにするならば、出版されるときにはデータやコードを公開しておき、チェックの機会を開いておく(レベル2相当)、さらに緩めるなら、求められたときにはデータやコードを開示する(レベル1相当)、というようにレベル設定を調整することができる。

そして第四に、再現性ポリシーは、現実の科学の営みの進化を踏まえ、つねにアップデートが必須であること、である。Liu and Salganik (2019)は、現代の計算社会科学の発展に伴って、異なる計算環境では分析結果の再現がより困難になることを予想する。データのサイズや分析手法の複雑化に対応するように、再現性ポリシーをアップデートしていくことも間違いなく必要であろう。

これらの論点は、じつのところ重なり合っ

おり、独立に考えることは難しいだろう。ただ結局のところ、再現性ポリシー実施の課題は、重い負荷をどのように扱うか、という問題に尽きる。2021年現在の日本の現状に照らして検討してみると、データアーカイブのスタッフが再現性チェックを担うのは事実上不可能である。では、学術雑誌の編集委員会や査読者が担えるかといえ、それにも現実味はあまり感じられない。そもそもデータ公開自体が十分ではないことも踏まえると、先述の“TOP Guidelines”のいうレベル1、すなわちリクエストがあれば公開できるようにすることを推奨する、というところからめざすのが現実的と思われる。

かつて板倉・尾崎(2002)は、再現性担保のためにとりうる方策を次のように述べた。すなわち、論文の著者・報告者には「変数に関する情報および分析に使用したソフト・プラットフォームに関する情報を示すこと」を、データベース(アーカイブ)には「詳細な項目の説明と欠損値の適切な指定およびその符号と内容、度数の公開」を²⁾、そして学協会には「投稿規定にデータやシンタックスなどのプログラム公開について明記し、その提出を義務付けること、および提出されたデータ・シンタックスをアーカイブすること、著者に求める内容を学会として示すこと、再現性に関する研究を評価するようなシステムの導入など、追試を促進するために配慮すること」を求めたいとした。20年後の現在、その提言よりもずいぶん前の位置からとなってしまうが、世界的潮流をも見据えつつ、再現性ポリシーの導入と実施に向けて、現実的な歩みを始めるときがきたと受けとめている。

8 再現性を高めるための研究者の方策

再現性ポリシーが実効力をもつようになれば、当然ながら、個々の研究者もそれへと対応するよう変化が要求される。ここでは、研究者レベルでどのような取り組みを行えば、再現性

のある計量的研究をしやすくなるのか、Long (2016) に基づいて考察したい。米国のデータアーカイブICPSRの行っているサマープログラムにおいて、Longは2016年からStrategies for Reproducible Researchと題する短期コースを担当しており、以降の記述はその講義内容を参考に、筆者が抜粋・要約したものである³⁾。

検証可能性(狭義の再現性)というならば科学の最低目標である。同じデータに対し同じ手法で分析して、同じ結果が出るかどうかをみるだけだから、再現できて当然のように思えるかもしれない。しかし、実際はそうではないことはすでに述べた。

なぜ、再現できないのか。それにはさまざまな「落とし穴」があるからである。コードについていえば、「どのコードが最終版かわからなくなった」、「最終版としたものに誤りがあった」、など。データに関しては、「どのデータが利用したものかわからなくなった」、「上書きしていたら元に戻れなくなった」、「データのバージョンが変わっていた」、など。ソフトウェアについても、異なるソフトウェアを用いたり、同一ソフトウェアでもバージョンが異なれば、仕様が変化するなどして出てくる結果に違いを生じさせることさえあるのだ。

対処する方策のうち、特に強調されるのは、フォルダ(ディレクトリ)のつくり方である。まず、フォルダの階層構造を適切につくる。すなわち、当該の研究プロジェクトのフォルダの下に階層に、現在進行中のファイル用のもの、関連文献用のもの、テンプレファイルを入れておくもの、フォルダ内のファイルは絶対に変更しないいわば「貯蔵庫」として用いるもの、などに分ける。さらにそれぞれのフォルダ内で次なる階層のフォルダへと分ける。たとえば、現在進行中フォルダの下には、データ用のもの、プログラム用のもの、図表用のもの、原稿用のもの……などというように整理するのである。

これらのなかで再現性にとってとりわけ重要

なのは「貯蔵庫」フォルダである。このフォルダの下に階層には、初稿時点での確定ファイル用のフォルダ、査読後の修正稿提出時のファイル用のフォルダ、というように各々の時点でのバージョンの原稿、データ、コードなどを入れておくわけである。くわえて、各々の確定がいつ(年月日)であるかの情報をフォルダ名につけておくことも強く推奨する。

そのほかの方策としては、次のものが挙げられる。ファイル名を体系的につけること(例: Miwa2021-06-30mobilityunidiff.do)、保存とバックアップ(自宅PC、職場PC、クラウドの3箇所ほどを推奨)を忘れずにこまめに行うこと、作業の意図をメモに残すこと、コードを小分けにすること(例:階層分類作成のdoファイル、データ配列変換のdoファイル、対数線形モデル用のdoファイル、それら個別のdoファイルを順々に走らせるためのマスターdoファイル)、などである。

研究結果の再現性を確かなものとするためには、以上のような方策は大いに役立つことだろう。ただし、これらを実践するには、相当の準備と心づもりがないとできないのも事実であろう。筆者としては、まずは「貯蔵庫」フォルダだけでも用意するとよいのではないかと思う。それにより、いつどこに出した結果を、どのセット、どのバージョンならば再現できるのかが遡れるようになるからである。この点が、再現性を高めるための第一歩だといえる。

9 おわりに

本論では、再現性ポリシーの意義・現状・課題について、包括的に整理をした。再現性ポリシーは、研究者コミュニティにとっても著者にとっても意義のあるものであるが、現状では米国の政治学、経済学では広まっているものの、日本国内に関してはまだごくわずかにとどまる。再現性ポリシーを実施するにはかなりの負担が



生じるので、解決すべき課題は少なくない。そしてまた、社会科学をとりまく状況は刻々と変わりゆくので、ポリシーをアップデートしていくことが絶えず要求されるのである。

課題が残るとはいえ、再現性ポリシーの重要性は、今後高まることはあってもその逆は考え難い。そして、社会科学者たちも、国際化から目を背けないのであれば、再現性ポリシーへの対応は避けては通れなくなるだろう。既に日本においても、一部の英文雑誌では再現性ポリシーが導入されているのである。この先、日本の学

協会においても再現性ポリシーの検討が必要になる日がやがてくるように思われるが、その際に、本論が議論の素材となるならば幸いである。

付記

本稿は打越・三輪 (2018) をもとに大幅に改稿したものである。本稿の議論は、2020年度課題公募型二次分析研究会「高等教育機関進学者における周縁的集団の出身背景とライフコースにかんする二次分析」および科学研究費 (19H01556) の研究成果の一部である。

注

- 1) 対象とした学術誌は、政治学系では『年報政治学』と『国際政治』、心理学系では『心理学研究』と『教育心理学研究』、教育学系では『教育学研究』と *Educational Studies in Japan*、経済・経営学系では *The Japanese Economic Review* と『日本経営学会誌』、社会学系では『社会学評論』、*International Journal of Japanese Sociology* および『家族社会学研究』、『教育社会学研究』、『理論と方法』、そのほかに『人口学研究』、『組織科学』、『行動計量学』、*Behaviormetrika*、『日本労働研究雑誌』である。
- 2) このような「データに付随する属性情報を表すデータ」をメタデータと呼ぶ。日本学術振興会

では、メタデータの国際標準である DDI (Data Documentation Initiative) を採用したデータカタログの整備を進めている (伊藤・前田, 2019)。データカタログの整備が進めば、狭義の意味での再現が行いやすくなるにとどまらず、調査の企画や実施について詳細情報が得やすくなり、また複数の調査データ情報の比較が容易になることにより、広義の再現性の確保にも寄与する可能性がある。

- 3) こうしたコースが2016年になって開設されたこと自体、米国はじめ世界の社会科学研究者にとって、再現性ポリシーとそれへの対応が重要な話題となったことを含意している。

文献

- Abbott, A., 2007, "Notes on Replication", *Sociological Methods and Research*, 36: 210-219.
- Anderson, R. G., W. H. Greene, B. D. McCullough and H. D. Vinod, 2008, "The Role of Data & Program Code Archives in the Future of Economic Research", *Journal of Economic Methodology*, 15(1): 99-119.
- Ashenfelt, O., 1986, "Editorial Statement", *American Economic Review*, 76(4), v.
- Bernanke, B. S., 2004, "Editorial Statement", *American Economic Review*, 94(1): 404.
- Dewald, W. G., J. G. Thursby and R. G. Anderson, 1986, "Replication in Empirical Economics: The Journal of Money, Credit and Banking Project", *American Economic Review*, 76(4): 587-603.
- 独立行政法人日本学術振興会人文・社会科学データインフラストラクチャー構築推進事業運営委員会作業部会(データ保有状況調査), 2019, 『「人文・社会科学系学術的調査データの現状に関する調査」報告書』(2021年6月29日取得, https://www.jsps.go.jp/j-di/data/datainfra_houkokusyo.pdf)
- Firebaugh, G., 2007, "Replication Data Sets and Favored-Hypothesis Bias (Comment on

- Jeremy Freese [2007] and Gary King [2007]), *Sociological Methods and Research*, 36: 200-209.
- Freese, J. and D. Peterson, 2017, "Replication in Social Science", *Annual Review of Sociology*, 43: 147-165.
- Gertler, P., S. Galiani and M. Romero, 2018, "How to Make Replication the Norm", *Nature*, 554(7693): 417-419.
- Gherghina, S. and A. Katsanidou, 2013, "Data Availability in Political Science Journals", *European Political Science*, 12(3): 333-349.
- Gleditsch, N. P., C. Metelits and H. Strand, 2003, "Posting Your Data: Will You Be Scooped or Will You Be Famous?", *International Studies Perspectives*, 4: 89-97.
- 池田功毅・平石界, 2016, 「心理学における再現可能性危機——問題の構造と解決策」『心理学評論』59(1): 3-14.
- 板倉宏昭・尾崎万枝, 2002, 「二次分析における再現可能性」『理論と方法』17(1): 41-51.
- 伊藤伸介・前田幸男, 2019, 「分野横断的なデータカタログの整備に向けて——現状と課題」統計情報研究開発センター『Estrela』308: 8-14.
- Jacoby, W. G., R. N. Lupton, M. T. Armaly, and A. Enders, 2016, Report to the Editorial Board and the Midwest Political Science Association Executive Council. (<https://ajpsblogging.files.wordpress.com/2016/05/ajps-editors-report-on-2015.pdf>).
- Key, E., 2016, "How Are We Doing? Data Access and Replication in Political Science", *PS: Political Science & Politics*, 49(2): 268-272.
- King, G., 1995, "Replication, Replication", *PS: Political Science and Politics*, 28(3): 444-452.
- King, G., 2007, "An Introduction to the Dataverse Network as an Infrastructure for Data Sharing", *Sociological Methods and Research*, 36(2): 173-199.
- Liu, D. M. and M. J. Salganik, 2019, Successes and Struggles with Computational Reproducibility: Lessons from the Fragile Families Challenge, *Socius*. January 2019.
- Long, J. S., 2016, *Lecture and Lab Notes for Strategies for Reproducible Research*(ICPSR サマープログラム講義資料) .
- McCullough, B. D. and H.D. Vinod, 2003, "Verifying the Solution from a Nonlinear Solver: A case study", *American Economic Review*, 93(3): 873-892.
- Open Science Collaboration, 2015, "Estimating the Reproducibility of Psychological Science", *Science*, 349(6251), aac4716.
- Popper, K., 1959, *The Logic of Scientific Discovery*, New York: Basic Books (大内義一・森博訳, 1971, 『科学的発見の論理(上・下)』恒星社厚生閣.)
- Stockemer, D., S. Koehler and T. Lentz, 2018, "Data Access, Transparency, and Replication: New Insights from the Political Behavior Literature", *PS: Political Science & Politics*, 51(4): 799-803.
- 打越文弥・三輪哲, 2018, 「社会科学分野における再現性ポリシーの概要と今後の課題——経済学・政治学・社会学を中心としたレビュー」SSJ Data Archive Research Paper Series No.66, <https://csrda.iss.u-tokyo.ac.jp/csrda/rps/f2018.html> (2021年6月21日最終アクセス).



特集論文

3

人文学・社会科学総合データ
カタログ(JDCat)について

前田幸男

東京大学大学院情報学環 教授／日本学術振興会 研究員

1 はじめに

学術研究においてはつねに資料や文献を参照し、論文執筆時には適切な引用を通じて読者が原典を参照できるようにするのが常識である。それらの資料を保存する機関としての図書館や資料館も整備されている。学術誌の多くは詳細に参考文献の記載方法を定めている¹⁾。これには出版物としての体裁を統一するという要請だけでなく、確実に典拠に到達するために必要な情報を提供するという意味もある。そこに掲載されている情報を頼りに、私たちは検索を行い、引用されている文献を自ら手にすることになる。

しかしながら、史資料、書籍、論文等と異なり、社会科学分野の標本調査データや心理学の実験データに関しては、そもそも、それらのデータが保存されているのか、保存されているとしたら、どの機関に所蔵され、どのような手順を経ると利用可能なのかについて、体系的に検索することは難しいのが実情である。この背景には、データを引用あるいは参照する場合の形式的規則や、そのデータに関する情報を構造的に整理して提供する方法について考察する機会が乏しかったことがあるように思われる。

日本では、様々な機関や関係者の努力もあり、過去四半世紀の間に社会調査データの保存と共

有について相当の進歩が見られた。しかしながら、データを記述する形式的ルールや、データをより見つけやすくするための体系的な検索の仕組み作りについては、手つかずであったというのが実情ではないだろうか。

本稿では、そのような欠落を埋めるために日本学術振興会が構築したデータ検索システムの内容について概観する。次節で政策的背景について触れた後、第3節以降で、具体的な内容について説明したい。

2 社会科学における
オープンサイエンス

データを広く共有し、利活用するという課題は社会科学に限られた問題ではない。広く学術研究全体に影響を与える政策動向に目を転ずると、研究データは、そのデータを取得した研究グループでの利用に限定されるのではなく、広く共有され、利活用されるべきであるというオープンデータの考え方が近年強調されるようになってきている。例えば、OECDは、公的な資金により作成された研究データについては広くアクセスを保証し、公開すべきであるという方針を示したガイドラインを2007年に公表している(OECD, 2007; 2020)。オープンデータあるいはさらに上位概念のオープンサイエンスに関する政策動向については、本特集の船守論文に詳細

を委ねるが、2013年には主要国首脳会議で「オープンデータ憲章」が採択され、2016年のG7における科学技術大臣会合「つくばコミュニケ」では、オープンデータを推進する方針についての共同声明が出されている（NBDC研究チーム、2018）。実際、欧米諸国の研究資金配分機関は、研究費申請書においてデータ管理計画（DMP：Data Management Plan）の策定を義務付け、データの共有を要請することで、オープンデータの実践を奨励している（Lyon、2016）。

OECDは2015年には「オープンサイエンスを現実にする」というレポートを公表しているが、そのレポートにおいては、オープン・リサーチ・データについてとくに一章を割り、様々な分野のデータについて議論している（OECD、2015：52-70）。社会調査データ（Survey Data）については、天文学等のデータと同様に、ある特定の時期について収集されるものであり、全く同じデータを、別の機会に改めて集めることができない以上、保存すべきものとして位置づけている。実験データも、再度同じ実験を行うことが実際的ではない（あるいは倫理的ではない）場合を考えるならば、保存することが重要だとされている。これらの議論は、人を対象にした研究におけるプライバシーや匿名性の保護を前提に、データを共有することを求めており、プライバシーの心配がない数値データや自然科学データと同じ意味で社会調査データをオープンにし、公開することを求めているわけではない点は理解しておく必要がある²⁾。

オープンデータの具体的な意味だが、OECDの文書では、「技術的あるいは法的な制約なく、誰もが使えるデータ」と簡潔に説明している（OECD、2015：55）。本稿では、技術的あるいは法的な制約を取り除くための方策を詳述する余裕はないが、データが「誰にでも使える」ためには、データ収集の背景や具体的内容、利用条件など、データを理解し、入手し、そして分析するために必要な情報も「誰にでも使える」状態でな

ければならないことは指摘しておく。一般的に、データの作成者や収集・作成期間、測定のために用いられた質問票など、従来はコードブックという冊子に収められていたデータに関する情報をメタデータと呼ぶが、オープンデータを実現する上では、データがソフトウェアなどで利用しやすいフォーマットで整理されているだけではなく、データを理解する上で必須のメタデータもコンピュータープログラム処理できるように構造化され、様々な機械的な文脈に対応できるように設計されなければならない。すなわち、構造化されたメタデータにより、データの幅広い共有が可能になるのである。

3 JDCatの全体像 複数の機関のデータをハーベストする仕組み

本節以降では、日本学術振興会が人文・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一環として開発した人文・社会科学総合データカタログ（以下、JDCat）について紹介したい³⁾。行論上、メタデータそのものについて解説する前に、メタデータを活用するカタログの仕組みについて概観する。

図1に示したのは、JDCatの構成である⁴⁾。JDCatはデータそのものを所蔵しておらず、ここで検索できるのはあくまでメタデータである。ただし、図の下段に位置している5つの拠点機関が提供するデータのメタデータを機関および分野を横断して一括検索することが可能になっている⁵⁾。メタデータ自体は各拠点機関が作成しており、拠点機関のローカルなカタログにおいても検索・閲覧は可能である。しかしながら、JDCatはそれらのメタデータを機械により定期的に自動収集（ハーベスト）することで、カタログの内容を最新の状態に保っている。また、機械的な連携を通じてCiNii Researchの側から、論文や図書などの研究成果やその成果を生み出した研究者・研究プロジェクトとデータとを結びつけることもできる。

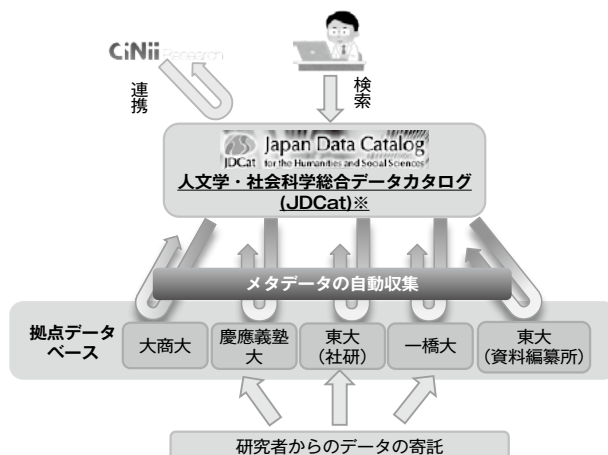


図1 人文・社会科学総合データカタログ (JDCat) の構成
※ NIIが開発したリポジトリソフトウェアWEKO3を利用して構築

より具体的なJDCatのカタログとしての特徴だが、メタデータの規格として、諸外国の主要社会科学データアーカイブの多くが採用しているData Documentation Initiative (以下、DDI)のDDI-Codebookという規格を採用している^{6,7)}。メタデータは基本的に日本語と英語の2言語で作成している。日本のデータを探すときに、個別に各機関のカタログを検索するのではなく、JDCatを使って英語による横断検索ができることは、従来から比べると格段の進歩である。

また、JDCatが拠点のデータベースからメタデータをハーベストするように、外部のカタログがJDCatのメタデータをハーベストすることも設計上可能にしている。JDCatがハーベストするメタデータはクリエイティブ・コモンズのCC0⁸⁾を条件としているので、JDCatのメタデータをハーベストしたい機関は、日本学術振興会や参加機関からの許諾を得ることなく自由にメタデータをハーベストできる。メタデータの自由な流通を可能にすることで、日本のデータに関する情報がより多様な経路で参照され、さらには、データそのものが国内外でより積極的に利用されるようになるのではないかと期待している。

なお、JDCatのように複数のデータ提供機関

を連携させるカタログ自体はとくに珍しいものではない。考え方としては日本中の大学図書館の目録をOPACで一括検索していることと同じである。社会科学のデータアーカイブを横断するデータカタログとしてはヨーロッパの20以上のデータアーカイブを包含するコンソーシアムであるCESSDA (Consortium of European Social Science Data Archives)のカタログが著名である⁹⁾。CESSDA CatalogueではCESSDA加盟アーカイブのデータを横断検索できるが、CESSDA Catalogue自体にはやはりデータは所蔵されていない。各国のアーカイブがDDIの規格に基づいて作成したメタデータをCESSDA Catalogueが機械的に自動収集することで、横断検索が可能になっている。利用者が検索結果を閲覧したのち、自身が利用したいデータを取得する段階では、CESSDA Catalogueの検索結果上に表示されるリンクをクリックして、データを所蔵する各アーカイブのカタログに遷移する必要がある。規模もユーザーインターフェイスも異なるが、JDCatの基本的な設計は、CESSDA Catalogueと各国のアーカイブとの関係を参考にしている。

なお、複数の機関が連携してデータカタログ

を作る例としては、複数のカタログから上位のカタログにハーベストするのではなく、複数機関が共同して1つのデータカタログを運営する場合もある。例えば、スウェーデンのSwedish National Data Serviceは9つの大学のコンソーシアムとして運営されているが、データとメタデータの準備自体はコンソーシアムに加盟している各大学が担当する分野の専門的な観点から行い、検索カタログの設計やデータ提供の仕組み作りを全体の統括・調整を担うイエテボリ大学のチームが行っている¹⁰⁾。社会科学だけに閉じるのではなく、分野を超えて地理学や人文学等を含む分野横断的なデータ提供システムを構築する場合、特定分野を担う複数のデータ機関を連携させることは、今後のデータインフラストラクチャーを整備する方向性として有効な選択肢であろう(Recker et al., 2015)。複数のカタログからハーベストするか、1つのカタログを複数機関で運営するかの違いはあるが、将来的には各分野のデータを専門的に扱う機関を連携させる仕組みを採用する国が増えてくるのではないと思われる。

4 DDI-Codebook に基づく JDCat のメタデータ・スキーマ

次に、より具体的なメタデータの設計について紹介する。データを記述する上では、例えば作成者やデータ収集の方法等が重要であることは直感的に明らかであろうが、データについて具体的に「何を記述するべきか」を定めたルールを一般的にメタデータ・スキーマと呼ぶ。JDCatのメタデータ・スキーマを作る上では、海外の主要社会科学データアーカイブ、具体的には、ICPSR (米)、UKDS (英)、GESIS (独)、DANS (蘭)、そしてADA (豪)のカタログを比較検討し、各カタログから共通項を抜き出すところから出発している¹¹⁾。個人的な観察であるが、DDI自体は1990年代から欧米の社会科学アーカイブが協力するプロジェクトとして立ち上がっており(Vardigan, Heus, and Thomas, 2008)、その経

緯を考えると、DDIから各国のカタログが作成されたと言うよりは、とくにDDI-Codebookについては各国のカタログの共通項が集約された、あるいは、共通化の努力が行われたという側面もあるように思われる。

検討の結果を経て、最終的には31項目からなるメタデータ・スキーマを設定した。その内容を表1に示す。

簡単に解説すると、メタデータの項目(エレメント)はカタログの掲載において、必須、強く推奨、推奨、任意の四段階に分類されている。例えば、必須とされている5項目の「1. タイトル」は調査やデータの名称、「3. 作成者」は、調査等の代表者、「5. URI」(URL)等はWeb上のデータを見つけて再利用する上では文字通り必須の情報である。順番は前後するが「4. 配布者」は実際にデータを配布する機関、そして「2. 整理番号」は「配布者」がデータに付与する整理番号となる。

JDCatのメタデータ・スキーマで特徴的なところを確認したい。第一に、DDI-Codebookが社会調査データのカタログに由来するため、標本調査の個票データを記述する項目が多い点である。任意の項目ではあるが、例えば「17. 母集団」「18. サンプリング方法」「19. 調査方法」「25. 回収率」等は、標本調査との関連から導入されている記述項目である。

第二の特徴であるが、多様なデータを扱う関係で、データに関わる主体についての項目が多い点である。「3. 作成者」「4. 配布者」は先述の通りだが、この他にも「30. 編集者」と「31. 所蔵者・寄託者」が存在する。「30. 編集者」は、調査主体、調査代表者とデータの編集者が異なる場合に利用する。とくに、アグリゲートデータが共有される際は、原データから様々な加工を経ている場合があるので必要になる項目だと考えた。また、「31. 所蔵者・寄託者」は、データの電子化と提供を他機関に委ねた、あるいは、許可した人や機関に関する情報を記載するために利用される。例えば画像のデジタルデータなどは、画



表1 人文学・社会科学総合データカタログのメタデータ・スキーマ

No.	JDCatエレメント名	記入レベル (社会科学)	記入レベル (人文学)	統制語彙の 有無	意味	JPCOARスキーマ名称
1	タイトル	○	○		調査名／タイトル	タイトル
2	整理番号	○	○		調査番号／資料番号／資料ID	
3	作成者	○	○		調査主体／調査代表者／作成者／著者	作成者(作成者姓名)
4	配布者	○	○		配布者／提供者 ※提供機関名	寄与者(寄与者姓名)
5	URI	○	○		ランディングページURI	識別子
6	トピック	○	○	○	トピック	主題
7	概要	○	○		調査概要／資料概要	内容記述
8	対象時期	○	○		対象時点／対象期間	時間的範囲
9	対象地域	○	○	○	調査地域／位置情報／地理的範囲	位置情報
10	データタイプ	○	○	○	データタイプ	内容記述
11	アクセス権	○	○	○	アクセス制限(利用条件)	アクセス権
12	権利情報	○	○		データのライセンス	権利情報
13	DOI	○	○		DOI	識別子
14	データの言語	○	○	○	データの言語	言語
15	バージョン情報	○			データのバージョン	バージョン情報
16	観察単位	○		○	観察単位	内容記述
17	母集団	△			母集団	内容記述
18	サンプリング方法	△		○	標本抽出方法	内容記述
19	調査方法	△		○	調査方法	内容記述
20	研究助成機関				研究助成機関	助成機関名
21	研究費番号				研究費番号	研究課題番号
22	その他のタイトル				調査別名／その他のタイトル	その他のタイトル
23	シリーズ				調査シリーズ／資料シリーズ	
24	調査日				調査日／収集日	内容記述
25	回収率				回収率	内容記述
26	引用上の注意				引用上の注意	
27	データファイルURI				データファイルURI	本文URL
28	関連情報				関連データ	関連情報
29	関連文献				関連文献(データに基づく文献)	関連情報
30	編集者				編集者	出版社
31	所蔵者・寄託者				所蔵者・寄託者	寄与者

○:必須 ○:強く推奨 △:推奨 空欄:任意

像の撮影対象(古文書や古地図)の(歴史上の)著者と、デジタルデータの提供主体、古文書そのものを所蔵する組織(例えば寺社)とが異なるのが通常であろう。社会調査データにおいても、データの作成者、配布者、編集者以外がデータを寄託する場合に利用することが考えられる。

第三の特徴は、相互運用性への配慮である。データカタログの規格は、ユニークであるよりも、様々な規格との連携が取れていることが望ましい。JDCatのメタデータ・スキーマは、日本国内のオープンアクセスリポジトリ推進協会(JPCOAR: Japan Consortium for Open Access Repository)¹²⁾のメタデータ・スキーマとも関連付けがなされており、国内外の他のカタログとの相互運用性を担保するように努力している。

第四の特徴は、いくつかの項目について、統制語彙を指定している点である。この点については、次節で説明する。

5 統制語彙の作成

前節まで議論してきたメタデータ・スキーマが、データについて「何を記述するか」のルールであるとすれば、統制語彙(Controlled Vocabulary)は、メタデータを「どのように記述するか」のルールである。例えば、「イギリス」と「英国」は、日本語を解する人には同じ意味だが、機械には全く別の記号と解釈される。統制語彙を設定するのは、一つの事象は一つの言葉で記述するというルールを徹底することで、誤解の余

地を減らし、かつ、検索の効率を上げるためである。

統制語彙についてもいくつかの国際標準が存在するので、JDCatで利用するためにいたずらに新しい語彙を作るのではなく、できるだけ既存の統制語彙を利用するようにした。その観点から、DDIに対応する適切な統制語彙がある場合は採用している。具体的には、「観察単位」、「サンプリング方法」、「調査方法」等はDDIの統制語彙¹³⁾の翻訳である。ただし、日本におけるこれまでの議論や慣行を踏まえ、ライセンスなどはJPCOARの統制語彙に準拠している。データタイプについても、日本では学術社会調査と公的統計とを分けて議論した経緯を前提に、独自の統制語彙を設定している。

研究トピックの統制語彙については、CESSDA Catalogueで採用されているCESSDA TC (CESSDA Topic Classification) を利用することにした¹⁴⁾。CESSDA TCには100前後の項目が設定されているが、それらの用語ならびに定義を訳出した¹⁵⁾。社会科学の学術用語集として国際的に利用されているものに、14言語の対訳が整備されているEuropean Language Social Science Thesaurus (ELSST)¹⁶⁾も存在するが、3,300前後の用語が存在し、学術利用に耐えうる翻訳をすることは時間と資源の制約上厳しいこともあり、CESSDA TCを採用した。CESSDA TCでカバーすることが難しい人文学分野のトピックについては日本十進分類法を援用する形で対応している¹⁷⁾。また、主に公的統計からの集計結果を念頭に『日本統計年鑑』¹⁸⁾の目次情報を援用する形でトピックとしている。カタログを公開する初期段階に検索可能なデータセットは限られているが、長期的には幅広い分野のデータの関連トピックの情報を記述することができるのではないかと考えている。

6 結びにかえて

以上、日本学術振興会が準備してきたデータカタログ (JDCat) の仕組みとメタデータ・スキーマについて解説をしてきた。JDCatは、データおよびメタデータを提供する機関との協力を通じて、国内外でデータがより広範に共有されるようになることを目的として構築されている。最後に、結びにかえて、今後の課題について触れておきたい。

デジタルデータあるいはメタデータを扱うインフラストラクチャーは、情報技術の変化や革新につねに対応する必要があるため、定期的な更新が不可欠である。また、各拠点機関は今後もデータの受け入れを続け、メタデータを追加し、カタログの内容を充実させてゆく予定である。ただし、カタログを閉じたものにする積極的な理由はなく、他にも技術的な条件と組織的体制が整った機関が存在するのであれば、長期的には参加できる機関数を増やして、より多くかつ多様なデータの保存、共有、利活用を促すことが望ましい。ただし、ここで問題になるのは、この長期的運用を誰が (あるいはどの組織が) 担うのかということである。

筆者が理解する限り、日本には、海外のデータアーカイブやレポジトリで通常働いているようなデータの保存、共有、利活用に特化した専門家の数が極めて少ない (あるいは皆無である)。今日においても社会科学や人文学の研究者が、本業の研究教育の片手間に、あるいはやむを得ず本業との関係を一時的には逆転させて、データの保存、共有についての活動をしているのが実情だと思われる。しかしながら、デジタルな研究資源の保存と共有については、人文学はおろか自然科学を含めて研究分野ごとの敷居が下がり、環境が劇的に変化している。国際的なデータ・レポジトリの認証基準は、以前は社会科学系のData Seal of Approval と自然科学系



のWorld Data Systemの二つが存在していたが、2018年には、Core Trust Sealという一つの基準に統合された¹⁹⁾。領域固有の知識は今もなお重要だが、それを超えて、デジタルな研究資源をいかに保存、共有、利活用するかという観点からの専門家育成が急務になっている。専門の人材

の育成と組織基盤の強化、そのうえで、研究者に対して政策的にデータの共有を促し、より多くの研究者が多様なデータにアクセスできる環境を整えることが、日本の人文学・社会科学研究の水準を高めるために、必要である。

注

- 1) 例えば、日本社会学会の『社会学評論スタイルガイド』(<https://jss-sociology.org/bulletin/guide/>)、あるいは日本心理学会の『執筆・投稿の手びき』(<https://psych.or.jp/manual/>) 等がある（共に2021年7月5日確認）。
- 2) データの「公開」というと、利用方法並びに利用主体の面で全くの制約がないデータの提供形態と誤解される可能性があるので、ここでは、社会調査データの利活用の慣行に基づき、日本語では、データの「共有」という表現を使う。
- 3) 事業の概要は、日本学術振興会のウェブサイトをご覧ください。<https://www.jsps.go.jp/j-di/index.html> (2021年7月5日確認)。
- 4) 事業自体の全体像および関連資料は、文部科学省のウェブサイトにも掲載されている。<https://www.mext.go.jp/kaigisiryō/content/000120425.pdf> (2021年7月5日確認)。
- 5) 5つの機関は、大阪商業大学JGSS研究センター、慶應義塾大学経済学部附属経済研究所パネルデータ設計・解析センター、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、一橋大学経済研究所（以上、社会科学分野）、東京大学史料編纂所（人文学分野）である。
- 6) <https://ddialliance.org/Specification/DDI-Codebook/2.5/> (2021年7月5日確認)。
- 7) DDIにはもう一つ高次のDDI-Lifecycleという規格もある。DDI-Codebookがデータ取得後の記述に力点があるとする、DDI-Lifecycleはその名の通り、データ収集の計画段階から最終的なアーカイブへの寄託までの全過程を包含し、より柔軟な記述ができるようになっている。ただし、その内容が複雑であるため、データカタログでの運用において適応できていない機関も多いようである（Förster et al., 2020 : 5）。
- 8) Creative Commonsが提案する利用許諾条件の1つ。著作物に関する権利を放棄することで情報の自由な流通を保証する。詳しくは、以下のウェブサイトを参照されたい。<https://creativecommons.jp/licenses/> (2021年7月19日確認)。
- 9) <https://datacatalogue.CESSDA.eu/> (2021年7月5日確認)。
- 10) <https://snd.gu.se/en/catalogue/search> (2021年7月5日確認)。
- 11) 人文学分野と対比すると社会科学分野の方が、メタデータ・スキーマに関しては共通理解が成立しているように思われたので、社会科学ベースでメタデータの設計をし、その上で、人文学分野に拡張した。
- 12) <https://schema.irdb.nii.ac.jp/ja> (2021年7月5日確認)。
- 13) <https://ddialliance.org/controlled-vocabularies> (2021年7月5日確認)。
- 14) <https://vocabularies.CESSDA.eu/vocabulary/TopicClassification> (2021年7月5日確認)。
- 15) 最終的にはCESSDAのMetadata Officeに翻訳を登録する予定である（Förster et al., 2020 : 10）。
- 16) <https://elsst.CESSDA.eu/> (2021年7月5日確認)。
- 17) 日本図書館協会分類委員会（改訂）, 2014a, 『日本十進分類法』新訂10版, 1本表・補助表編, 日本図書館協会。
- 18) <https://www.stat.go.jp/data/nenkan/index1.html> (2021年7月5日確認)。
- 19) <https://www.coretrustseal.org/about/history/> (2021年7月5日確認)。

- Förster, André, Kerrin Borschewski, Sharon Bolton, and Taina Jääskeläinen, 2020, "The matter of meta in research data management: Introducing the CESSDA Metadata Office Project," *IASSIST Quarterly*, 44 (3), (<https://doi.org/10.29173/iq970>) .
- Lyon, L.iz, 2016, "Transparency: the emerging third dimension of Open Science and Open Data," *LIBER Quarterly*, 25 (4): 153-171, (<http://doi.org/10.18352/lq.10113>) .
- NBDC 研究チーム, 2018, 「データ共有の基準としてのFAIR原則」, バイオサイエンスデータベースセンター (NBDC) , (DOI:10.18908/a.2018041901) .
- OECD, 2007, *OECD Principles and Guidelines for Access to Research Data from Public Funding*, Paris : OECD Publishing, (<https://doi.org/10.1787/9789264034020-en-fr>) .
- , 2015, *Making Open Science a Reality*, Paris : OECD Publishing, (<https://doi.org/10.1787/5jrs2f963zs1-en>) .
- , (日本学術振興会訳), 2020, 「公的資金による研究データへのアクセスに関する原則およびガイドライン」, (2021年7月5日取得, <https://www.jsp.go.jp/j-di/data/jigyo/OECDGuidelines.pdf>) .
- Recker, Astrid, Stefan Müller, Jessica Trixa, and Natascha Schumann, 2015, "Paving the Way For Data-Centric, Open Science: An Example From the Social Sciences," *Journal of Librarianship and Scholarly Communication*, 3 (2): p.eP1227, (<https://doi.org/10.7710/2162-3309.1227>) .
- Vardigan, Mary, Pascal Heus, and Wendy Thomas, 2008, "Data Documentation Initiative: Toward a Standard for the Social Sciences," *The International Journal of Digital Curation*, 3 (1): 107-113, (<https://doi.org/10.2218/ijdc.v3i1.45>) .



特集論文

4

Data Documentation Initiative (DDI) の意義と利用

JDCat への応用と今後の課題

朝岡 誠

国立情報学研究所 特任助教

林 正治

国立情報学研究所 特任助教

1 はじめに

近年、情報通信技術の高まりにより、データ駆動型研究や研究データの共有・再利用に関心が高まっている。オープンサイエンスともよばれるこの取り組みは、各学術分野における情報インフラのスタンダードを一新し、従来の教育研究環境を大きく変えつつある。

日本の社会科学分野では、社会調査データや統計データを公共財とみなす慣習があり、研究プロジェクトで集められた社会調査データはSSJデータ・アーカイブをはじめとするデータ・アーカイブで公開されている。公的統計データについても「政府統計の総合窓口 (e-Stat)」を通じて広く利用されており、研究データの共有・再利用は広がりつつあるが、まだ研究データは共有、再利用されやすい環境にあるとはいえない。

例えば、研究データの再利用を希望する研究者は、データを探し出すために複数のデータセンターにアクセスする必要がある。そのため、再利用しているのはデータセンターについて熟知した一部の研究者に限られており、この状況は潜在的なデータの共有・再利用を妨げているともいえる。この問題を解決するには、メタデータ

基準の標準化を行い、データセンター間でメタデータの相互運用ができる体制を整えなくてはならない。

ヨーロッパ社会科学データ・アーカイブ協議会 (CESSDA) は、ヨーロッパ中のデータ・アーカイブが所蔵するデータの横断的検索システム Data Catalog を構築するにあたり、Data Documentation Initiative (DDI) を採用した。DDI は社会科学、行動科学、経済科学、健康科学の分野における調査や実験などの観察によって得られたデータを記述するための国際標準メタデータ規格であり、その入力ルールは厳密に定義されている。ヨーロッパの各データ・アーカイブは、DDI で定義されたルールにしたがってメタデータを整備し、整備したメタデータをシステムに提供することで、アーカイブ間の横断検索を実現した。

このCESSDAの取り組みを参考に、日本学術振興会 (学振) と国立情報学研究所 (NII) は人文学・社会科学総合データカタログ (JDCat) を構築し、分野横断的検索システムとして2021年7月に公開した。JDCat は、大阪商業大学JGSS研究センター、慶應義塾大学経済学部附属経済研究所パネルデータ設計・解析センター (PDRC)、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)、一橋大学

経済研究所 (IER), 東京大学史料編纂所が整備したメタデータを使って運用しているが, このメタデータはDDIに準じて設計されたものである。

本特集の前田論文では, 日本版DDIともいえるJDCatメタデータスキーマについて紹介しているが, 本論文では, このJDCatメタデータスキーマが具体的にどのように記述されているのかを説明する。そして, JDCatと連携するデータセンターが実際にどのような方法でメタデータを整備しているのかについて触れつつ, 横断的検索システムとしてのJDCatについて紹介する。

2 DDIメタデータの機械可読性と相互運用性

学振とNIIは, JDCat運用のためにDDI Codebook (以下, DDI-C) とよばれるバージョンをもとにJDCatメタデータスキーマを作成した。DDI-Cは, DDIの4つのバージョンのうち2つ目のバージョンにあたり, 海外の社会科学データセンターのシステム運用で利用されている (朝岡・林, 2019)。DDI-Cをはじめ, メタデータの多くはXMLスキーマとよばれる形式で定義されているが, これはFAIR原則にしたがい, 記述した内容の機械可読性と相互運用性をもたせるためである。

この仕組みを調査タイトルの記述を例に説明する。

<titl xml:lang="ja">情報通信技術活用に関する調査2020</titl>

タグ内の「titl」は要素とよび, 検索エンジンにtitlが調査タイトルを表す要素だと定義することで, システムはタグで囲まれた「情報通信技術活用に関する調査2020」を調査タイトルだと理解することができる。また<titl>タグ内に書かれた「xml:lang="ja"」は属性とよび, 要素に付随する情報を表現している。「xml:lang」はXMLスキーマにおいて, 記述言語と定義されて

おり, jaは日本語と定義されているので, JDCatメタデータスキーマでは, 「xml:lang="ja"」という記述は「情報通信技術に関する調査2020」というタイトルが日本語であることを示している。JDCatは日本の研究者だけでなく, 海外の研究者にも研究データ検索に利用されることを想定して設計しているので, JDCatに協力するデータセンターにはメタデータの日英併記を依頼している。また, JDCatは「xml:lang」を利用することで, JDCat上で表示するメタデータの表示言語を制御している。

図1の記述はJDCatメタデータスキーマで, 調査タイトル, 調査番号, データ作成者の情報を日英併記した例である。図1を見ると, 「citation」という要素の中に「titlStmt」という要素があり, さらにその中に「titl (調査タイトル)」「altTitle (調査別名)」「IDNo (整理番号)」という要素が階層的に組み入れられていることがわかり, 「rspStmt」という要素の中に「AuthEnty (作成者)」が複数組み込まれている事がわかる。DDIでは調査や実験に関連する項目が階層的に定義されており, DDIに準じてメタデータを記述する際にはこの記述様式を遵守する必要がある。メタデータ作成者が新たに項目を追加するようなことはできない¹⁾。

また, 「citation」に記述言語属性「xml:lang」を利用し, 日本語と英語を分けて記述しているが, これはFinnish Social Science Data Archive (FSD) がCESSDAの横断的検索システムにDDI-C規格のメタデータを提供する際に採用した記述方式である。学振とNIIは, 同じデータを指し示すメタデータが言語によって複数存在し, 同整理番号のメタデータが複数存在することを避けるために, この記述方法を採用した。

NIIはJDCatメタデータスキーマを策定するにあたり, NIIで運用する科研費データベースやNII学術情報ナビゲータ文献検索システム (CiNii) との連携を考慮して, 研究者の科研費番号などの識別子を属性値として入力できるようにした。



```

<citation xml:lang="ja">
  <titlStmt>
    <titl xml:lang="ja">情報通信技術活用に関する調査 2020</titl>
    <altTitl xml:lang="ja">ICT2020</altTitl>
    <IDNo xml:lang="ja" agency="DOI">12345/0000001234</IDNo>
    <IDNo xml:lang="ja" agency="RCOS" xml:lang="ja">RCOS0012</IDNo>
  </titlStmt>
  <rspStmt>
    <AuthEnty ID="70583839" xml:lang="ja" affiliation="National Institute of Informatics">朝岡, 誠</AuthEnty>
    <AuthEnty ID="90552084" xml:lang="ja" affiliation="National Institute of Informatics">林, 正治</AuthEnty>
  </rspStmt>
  ...
</citation>
<citation xml:lang="en">
  <titlStmt>
    <titl xml:lang="en">Survey on the use of information and communication technology</titl>
    <altTitl xml:lang="en">ICT2020</altTitl>
    <IDNo xml:lang="en" agency="DOI">12345/0000001234</IDNo>
    <IDNo xml:lang="en" agency="RCOS" xml:lang="en">RCOS0012</IDNo>
  </titlStmt>
  <rspStmt>
    <AuthEnty ID="70583839" xml:lang="en" affiliation="National Institute of Informatics">Asaoka, Makoto</AuthEnty>
    <AuthEnty ID="90552084" xml:lang="en" affiliation="National Institute of Informatics">Hayashi, Masaharu</AuthEnty>
  </rspStmt>
  ...

```

図1 JDCatメタデータの記述例

JDCat上でデータ作成者名とその研究者の科研費番号、そのデータに関連する文献とその文献のリンクアドレスを併記することにより、データセンターが管理するデータとその作成者、科研費情報、研究成果を正確に関連付けることができ、他の検索サイトに研究データに関する情報を正確に流通させることができる。

また、データセンターのシステム変更に伴うURLの変更によるリンク切れによりデータの所在がわからなくなることがある。このような事態を回避するために、学振とNIIはJDCatメタデータスキーマにDOI (Digital Object Identifier)を導入した。海外のジャーナルでは、論文執筆時に参考文献として、利用した研究データ情報をDOIとともに記載することが慣習となっており、日本のジャーナルでも近い将来、参考文献にDOIの記載が慣習となることが予想される。

現在、JDCatと連携するデータセンターはこ

のJDCatスキーマに準じたメタデータを整備しているが²⁾、その整備は容易ではない。先の前田論文で紹介されたように、JDCatメタデータスキーマは31の項目を日英2ヶ国語で用意している。さらに一部の項目については、相互運用性を向上させるために利用できる語彙を制限している。例えば、トピックはCESSDAが利用している語彙を使って記述しており、調査時期は2021-10-01というようにyyyy-mm-ddの形式での記述に統一している。

この統制語彙は新たにメタデータを整備するデータセンターにとっては記述の指針となるが、古くからメタデータを整備してきた機関にとっては、過去のメタデータを整備し直す必要がある。例えば、社会調査データを扱うCSRDAは独自のトピックリストをもとにメタデータを整備していたが、JDCatへのメタデータ提供に際してCESSDAが利用している語彙に変更して

いる。また、公的統計を扱うIERは、これまで調査時期を「昭和58年度」というように和暦で記述してきたが、JDCatスキーマに適應するために、1983-04-01、1984-03-31と形式を変え、和暦は概要に記述する形でメタデータを記述している。統制語彙を使って統制を行うことは、メタデータ相互運用性を向上させるために重要なことであるが、研究分野を正確に表現するには、統制語彙のメンテナンスを継続していく必要がある。また時代によって研究者にとって重要なメタデータは変化するので、分野のコミュニティを中心にこれらのメタデータスキーマの改訂を行うなどの枠組みが必要となる。

3 DDI メタデータの記述方法と提供方法

本章では、JDCatと連携するデータセンターがどのようにメタデータを整備し、JDCatにメタデータを提供しているのかを説明する。DDI形式のメタデータには、商用のソフトウェアであるColectica、Nesstar Publisher、CSRDAが開発したEasy DDI Organizerがあり、汎用のXMLエディタソフトを使って整備することも可能である(前田・朝岡,2018)。しかし、これらのソフトウェアはFSDが採用した多言語記述に対応していない。また、JDCatメタデータスキーマで設定した統制語彙に合わせてメタデータを整備するには、統制語彙通りに入力するためにバリデータ(文法チェック)を使って入力した方がよい。

JDCatと連携するデータセンターは、整備したメタデータをOAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting) というメタデータ交換通信プロトコルを通じて提供することになっており、各データセンターは、自分たちのデータ提供システム(リポジトリシステム)にOAI-PMH機能を実装する必要がある。現在、JDCatと連携しているデータセンターのうち、CSRDAは自分たちのリポジトリシステムにDDIメタデータ編集機能とOAI-PMH機能を実装し、

史料編纂所はResourceSyncというプロトコルを使ってメタデータを提供しているが³⁾、他のデータセンターはリポジトリソフトウェアWEKO3を使ってDDIメタデータ編集とOAI-PMHによるメタデータ提供を行っている⁴⁾。

リポジトリソフトウェアWEKO3は、オープンアクセスリポジトリ推進協会(JPCOAR)とNIIが共同で提供するリポジトリクラウドサービスJAIRCloudのベースシステムであり、研究データの公開のために利用されることが期待されている。

WEKO3は様々な学術分野のメタデータに対応するため、登録するメタデータのテンプレートを自由に編集できるように設計されており、メタデータをブラウザ上で編集することができる。NIIは、WEKO3上でJDCatメタデータスキーマ用のテンプレートを開発したので、連携するデータセンターはこのテンプレートを利用することでJDCatに提供するメタデータを整備することができる。

図2はWEKO3のコンテンツ登録画面である。この画面では、論文や研究データなどのコンテンツと一緒にメタデータを登録することができる。このテンプレートには、学振とNIIで設定した統制語彙もあらかじめ設定されているので、統制語彙が設定されている項目は、ブラウザ上から該当するキーワードを選択することができる。メタデータは要素ごとに複数登録が可能であり、言語などの属性とセットで登録することができる。

また、永続的なリンクアドレスであるDOIを取得するには、世界に10機関存在するDOI登録機関が提供するシステムを利用して登録作業を行う必要があるが、WEKO3はいくつかのDOI提供機関と提携しており、登録したコンテンツにDOIを付与することができる。なお、登録したコンテンツにアクセス制限をかけることもでき、JGSS研究センターはこのアクセス制限機能を利用したデータ提供サービスを2021年から開始する(朝岡ら,2020)。

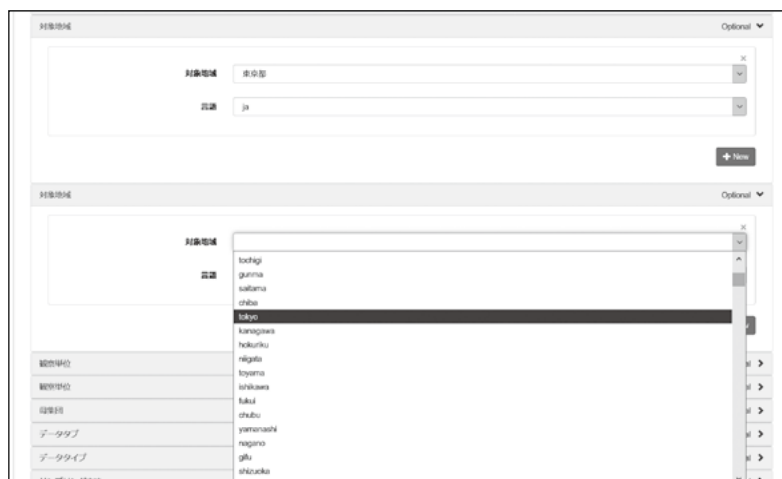


図2 WEKO3のコンテンツ登録画面

現在, JDCatは先述した5つのデータセンターのメタデータを対象にしており, 自分の研究データのメタデータをJDCatに提供したい研究者は, 5つのデータセンターのいずれかに研究データを預ける必要がある。将来的には他のデータセンターからのメタデータ提供に対応することを検討している(前田, 2019)。その際には, JAIRO Cloudで公開された人文学・社会科学データのメタデータ蒐集も検討されることが考えられる。

4 人文学・社会科学 総合データカタログJDCat

本章では, JDCat (人文学・社会科学総合データカタログ) について説明する。JDCatは5つのデータセンターの研究データの横断検索システムであり, リポジトリソフトウェアWEKO3を元に構築されている。本システムには, 機械可読なDDIメタデータを活用した機能が実装されている。

図3はJDCatのトップページ画面である。CESSDA Data Catalogと同様に, 画面右上の言語切替ボックスから日本語と英語を切り替えることができる。検索方法は, 一般的なテキスト

検索だけでなく, 詳細検索とファセット検索を用意している。一般的なテキスト検索で検索を行う場合, 調査時期を数値で検索しても, サンプルサイズなどの数値を含めて検索してしまうが, JDCatではDDIを使って要素を統制して記述しているので, 項目ごとの検索が可能である。

さらに, トピックなどいくつかの要素はファセット検索に対応している。ファセット検索とは, 検索システムがあらかじめ用意した検索条件を使った絞り込みを目的としたナビゲーションシステムのことである。トップページ画面左下のファセット検索ウィンドウを開き, キーワードを入力すると, ヒット件数の多い順に検索条件が表示され, この検索条件をクリックすることで, 検索した内容を絞り込むことができる。このファセット検索を利用することで, 専門的なキーワードを知らない学生や人文学・社会科学分野外の研究者が手軽に研究データを検索することができる。

トップ画面右下は検索結果であり, 検索されたデータのタイトルをクリックすると, そのデータの詳細画面を確認することができる(図4)。図4はCSRDAが提供している「第3回全国家



図3 JDCatトップページ



図4 調査データの詳細画面

族調査」の詳細情報であるが、図中に楕円の線で囲んだリンクアドレスをクリックすることでCSRDAのランディングページにアクセスできる。また、図右下のExport欄の四角の線で囲んだボタンをクリックすると、JDCatメタデータを確認することができる。

5 むすび

最後に、DDIの最新動向について述べ、むすびとする。ヨーロッパでは、ECによってEuropean Open Science Cloud (EOSC) という学術分野を



超えたデータ共有のためのインフラの構築が進められており、学術分野間連携を念頭においたメタデータとして、Data Documentation Initiative-Cross Domain Integration (DDI-CDI) の利用が検討されている。本稿で紹介したJDCatは質問文などの変数情報は取り扱っていないが、海外では、データセンターを超えたオンライン分析システムの構築のために、変数情報（質問文、回答選択肢、回答結果の簡単な記述統計）をDDI-CDIの規格に基づいて整備し、オンライン分析システム上で、複数のデータを統合して分析できるように設計を行っている（Gregoryら、2021）。今後社会科学ではDDIを標準規格とした研究インフラが設計され、DDIが研究者の研究活動

に大きな影響を持つことになると考えられる⁵⁾。

本稿では、国際メタデータ標準規格DDIを利用して構築したインフラとして、JDCatを紹介した。現在、JDCatの検索対象は5つのデータセンターで整備・公開している研究データに限られているが、将来は他のデータセンターや大学図書館が運用する機関リポジトリで管理・公開している研究データも検索対象にする可能性がある。また、公的な研究助成を受けた研究において得られた研究データを機関リポジトリなどで公開することが推奨される可能性もある。近い将来、調査や実験を行う際には、DDIの項目や統制語彙を使って調査情報を記録することが慣習となると考えられる。

注

- 1) DDIで定義されていない要素を記述するための要素として<note>がある。JDCatスキーマを策定するにあたり、DDI-Cにはデータの言語を記述する項目がなかったので、この<note>を利用した。
- 2) 人文学分野のメタデータはJPCOARスキーマに準じてメタデータを整備している。JDCatメタデータスキーマはJPCOARスキーマとの互換性を持つように設計しているため、JDCatでの運用は可能である（池内、2020）。
- 3) 史料編纂所はResourceSyncとよばれるウェブコンテンツ同期プロトコルを利用してJDCatにメタデータを提供している。OAI-PMHがメタデータの交換に限定されているのに対し、ResourceSyncはメタデータだけでなく、データファイルの交換が可能であるため、OAI-PMHの後継として注目されている（林、2015）。
- 4) リポジトリソフトウェアWEKO3はオープンソースソフトウェアであるので無償で利用することができる。PDRCは慶応義塾大学メディアセンターが用意した環境にWEKO3をインストールして運用しているが、IERとJGSS研究センターはJAIR Cloudサービスを利用して運用している。JAIR Cloudは全国600以上の教育研究機関が利用しているサービスであり、主に大学図書館が利用しており、大学図書館から機関に所属する研究者の研究データが公開されることが予想される。詳しくはJPCOARのホームページ（<https://jpcoar.repo.nii.ac.jp/page/42>）を参照。
- 5) 欧米のデータセンターは、社会科学リポジトリシステムDataverseや汎用リポジトリシステムDSpaceを利用して研究データの公開、メタデータの交換を行っている。両者ともオープンソースソフトウェアなので、無料で利用することができる。EUはEOSCのインフラとの連携のため、DDIに対応したDataverseによる研究データ公開を推奨している。

- 朝岡誠, 林正治, 2019, 「人文学・社会科学分野におけるデータインフラストラクチャー構築の推進」 ESTRELA 308 : 15-20.
- 朝岡誠, 林正治, 藤原一毅, 岩井紀子, 船守美穂, 山地一禎, 2020, 「汎用的データリポジトリにおける制限公開機能の検討と実装」 情報知識学会誌30(2) : 168-175, https://doi.org/10.2964/jsik_2020_014.
- Gregory, A., S. Hodson, J. Wackerow, 2021, "The Role of DDI-CDI in EOSC: Possible Uses and Applications", <https://doi.org/10.5281/zenodo.4707263>.
- 林豊, 2015, 「ResourceSync: OAI-PMHの後継規格」 カレントアウェアネス323 : 17-21, <https://doi.org/10.11501/9107589>
- 池内有為, 2020, 「人文学分野におけるメタデータの整備状況」2020年度統計関連学会連合大会報告資料, https://www.jsps.go.jp/j-di/data/jigyo/3_ikeuchi.pdf.
- 前田幸男, 2019, 「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取組について」 第4回 SPARC Japan セミナー2018「人文社会系分野におけるオープンサイエンス——その課題解決に向けて」報告資料, https://www.nii.ac.jp/sparc/event/2018/pdf/20190129_2.pdf.
- 前田豊, 朝岡誠, 2018, 「RUDAのDDI対応に向けた取り組み」『社会と統計』4 : 25-35. <https://doi/10.14992/00015575>



特集論文

5

データアーカイブにおける DDIの意義と課題 RUDAの実践を事例として

高橋かおり

立教大学社会情報教育研究センター 助教

五十嵐 彰

立教大学社会情報教育研究センター 助教

1 はじめに

著者の二人が所属する立教大学・社会情報教育研究センター（略称CSI）は、RUDA（Rikkyo University Data Archive）と呼ばれる量的社会調査データのデータアーカイブを運用しており、先ごろ新たにDDIメタデータ事業を始めた。本稿では、RUDAの事例をもとに、データアーカイブにおけるDDIの導入・運用の実際について紹介する。

以下では、まずRUDAの現状や事業を簡単に紹介したのち、次に、2017年度より構想され、2018年度から公開を開始したDDIメタデータ事業の実際と課題を紹介する。なお、DDI事業については朝岡誠と前田豊（ともに元立教大学助教）によって既に詳細な論稿（前田・朝岡、2018）が報告されているため、そちらも参照されたい。本稿では既発表の論稿を踏まえつつ、2020年度末時点の状況をもとに、データアーカイブにおけるDDI導入の意義と課題を論じる。

2 RUDAについて

RUDAは、管理主体をCSIとする量的社会調査データのデータアーカイブであり、2010年に学内公開、2011年に一般公開を開始した。CSIの

設立の経緯についてはすでにいくつかの論稿があるが（松本、2012；菊池ほか、2020）、データアーカイブ事業の他に、社会調査法や統計学などの統計教育授業の展開や、社会調査士や統計検定などの資格の取得支援、あるいは公的統計の活用事業などを実施している。

RUDAは2020年度末時点で64のデータセットを公開しており、利用登録者数は809人である。メタデータと調査票は利用者登録や申請をせずとも閲覧できるが、データセットを利用する場合は、利用者登録と申請が必要である。データセットの利用は毎年度20件弱である。利用者は立教大学内にとどまらず、他大学における授業での教育利用や卒論や修論での研究利用などにも及んでいる。また、英語で提供されているデータセットも一部あるため、海外からの問い合わせも年に数件程度ある。

3 RUDAにおけるDDI事業の実際

(1) 導入の経緯

ここでは、朝岡・前田（2015）ならびに前田・朝岡（2018）を踏まえ、RUDAにおけるDDI導入の経緯をまとめる。

RUDAでは、Dspaceと呼ばれるリポジトリに対応したオープンソースのソフトウェアをカスタマイズしてアーカイブシステムを構築している。

DspaceはDublin Core (DC) をメタデータ基準としているが、RUDAの構築の際にはDCの基本要素に独自の拡張を加えている。具体的には、DCに備わっている15の基本要素（タイトルや作成者など）そのものを利用するのではなく、その基本要素の下位分類である装飾子 (qualifier) を利用して独自のメタデータ基準を定めることで、社会調査データ情報の管理に特化したメタデータを整備し、検索環境を整えている（朝岡・前田、2015：28）。

しかし、すでに前田・朝岡（2018：28）で問題提起がされているように、RUDA独自のDCの仕様はデータアーカイブ「内」での整合的な情報管理や検索環境の整備においては問題がないものの、アーカイブ「間」での情報資源の相互運用という点から見れば弊害もある。すなわち、RUDAでは独自のカスタマイズを加えたメタデータ基準を採用しているため、他のデータアーカイブとの横断検索をはかる際にメタデータ情報の意味が損なわれることがある。

そのため、将来的に構築が望まれる他のデータアーカイブとの横断検索を見込み、RUDA運営においては国内アーカイブ間での共通メタデータ基準の統一が以前から課題として認識されていた。横断検索の先行例には、各国のアーカイブのデータ横断検索を可能にするCESSDAの“Data Catalog”がある¹⁾（前田・朝岡、2018：25）。利用者が自身の関心に沿って横断的にデータセットを検索できることは、データアーカイブの利便性の向上につながりうる。

(2) 運用状況

このような横断検索を実現するためのパイロット事業として行われてきたのが、RUDAにおけるDDIメタデータの提供である。具体的には、前田・朝岡（2018：30）で示したマッピング案に基づき、寄託された社会調査のメタデータをDDIへと変換する事業を2017年度から行ってきた。この作業はデータの寄託時に寄託者か

ら提供されたメタデータを参照の上、リサーチアシスタントとして大学院生6名の協力を得て実施した²⁾。

2020年度末時点では、RUDAにて公開している64データのうち、63データでDDIのメタデータを提供している。なお、DDIには複数のバージョンが存在するが、RUDAではDDI-Codebookを採用しており、提供するメタデータはDDI-CodebookにおけるDocument DescriptionとStudy Descriptionの2つの情報に限定している。また、RUDAが寄託時に記入を依頼したメタデータ記入シートにおいて把握できていない情報については空欄のままである。この理由と課題については後述する。

図1はRUDAのウェブサイト上で閲覧可能なDCメタデータ、図2はダウンロードによって閲覧可能なDDIメタデータである。これらのメタデータはいずれも利用者登録せずに閲覧可能である。

また、表1では図1と図2で示した2種類のメタデータの対応関係について、前田・朝岡（2018：30）で図示したマッピングをもとに、データセットの社会調査に関する情報の記述であるstdyDscr (Study Description) の情報を整理した。

表1の第1列 (DDI-C) がDDIメタデータであり、第2列がDC基本要素、第3列がDC装飾子である。例としてRUDA0001「東京版総合社会調査（高齢・少子社会における都市居住と家族・親族関係に関する調査）」を取り上げた。これまでは第2・3列のDC基本要素ならびにDC装飾子に依拠してメタデータを整備していたが、DDIに基づいたメタデータを別途公開し、より網羅的で互換的なメタデータの提供を行っている。

(3) 公開作業に向けての課題と問題提起

このようにRUDAではDDIメタデータの公開作業を行い、現在ではほとんどのデータセットにおいてDDIメタデータの提供を行っているが、実際の事業運営を通じていくつかの課題が出て



アイテムの簡略レコードを表示する

TO 東京版総合社会調査「高齢・少子社会における都市居住と家族・親族関係に関する調査」 [Japanese only]

調査番号	0001	ja_JP
調査名	東京版総合社会調査「高齢・少子社会における都市居住と家族・親族関係に関する調査」	ja_JP
調査略称	TGSS2000	ja_JP
調査主体 / 調査 代表者	東京立大学都市研究所松本康	ja_JP
調査資金	東京立大学都市研究所共同研究 I	ja_JP
寄託者	松本康	ja_JP
調査概要	高齢少子化・高度低成長・グローバル化などの趨勢のもとにおかれている大都市東京の都市社会構造と社会意識の変化を経年的に明らかにする。	ja_JP
キーワード	社会的ネットワーク	ja_JP
	都市社会構造	ja_JP
	社会意識	ja_JP
研究分野	urban sociology 都市社会学	ja_JP
	sociology 社会学	ja_JP
母集団	20歳以上70歳未満の東京都民（島嶼部は除く）	ja_JP
調査時期	2000年9月	ja_JP

図1 ruda0001のDCメタデータの一部
<https://ruda.rikkyo.ac.jp/dspace/handle/10577/57?show=full>

```

<?xml version="1.0" encoding="utf-8"?>
<?xml:stylesheet type="text/xsl" href="http://www.ddialliance.org/sites/default/files/codebook2-0.xml"?>
<codeBook xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="ddi:codebook2_0 http://www.ddialliance.org/Specification/DDI-Codebook/2.0/XMLSchema/codebook.xsd" version="2.0">
  <docDesc>
    <titStm>
      <IDNo>
        0001</IDNo>
      </IDNo>
      <tit>
        東京版総合社会調査「高齢・少子社会における都市居住と家族・親族関係に関する調査」</tit>
      </tit>
      <altTit>
        TGSS2000</altTit>
      </altTit>
      <parTit>
        </parTit>
      </parTit>
      </titStm>
      <distStm>
        <depositor>
          松本康</depositor>
        </depositor>
        <disttr>
          立教大学社会情報教育研究センター</disttr>
        </disttr>
      </distStm>
    </docDesc>
    <endDocDesc>
    <titStm>
      <IDNo>
        0001</IDNo>
      </IDNo>
      <tit>
        東京版総合社会調査「高齢・少子社会における都市居住と家族・親族関係に関する調査」</tit>
      </tit>
      <altTit>
        TGSS2000</altTit>
      </altTit>
      <parTit>
        </parTit>
      </parTit>
      </titStm>
      <prodStm>
        <fundAg>
          東京立大学都市研究所共同研究 I</fundAg>
        </fundAg>
        </prodStm>
        <rspStm>
          <AuthEntry>
            立教大学都市研究所 / 松本康</AuthEntry>
          </AuthEntry>
          </rspStm>
          <serStm>
            <serStm>
              <distStm>
                <depositor>
                  松本康</depositor>
                </depositor>
                <disttr>
                  立教大学社会情報教育研究センター</disttr>
                </disttr>
              </distStm>
            </serStm>
          </serStm>
        </distStm>
      </distStm>
    </distStm>
  </codeBook>

```

図2 xmlファイルにて作成したruda0001のDDIメタデータの一部
https://ruda.rikkyo.ac.jp/dspace/bitstream/handle/10577/57/0001_ruda.xml?sequence=27&isAllowed=y

いる。

第一に、不足情報への対応である。メタデータの変換過程において、これまでRUDAへの寄託の際に提供を依頼していたメタデータにはない項目への対応や、他のデータアーカイブとの用語の統一が必要になっている。現時点では、すでに提供されたメタデータ情報において記載がなければ空欄のままとしている。今後とくに課題となるのが、調査名称や調査に関するメタデータの多言語（英語）表記対応であろう。調査データの横断検索や国際的な検索可能性を担保するのであれば、調査の英語名は重要となる。

また、他のデータアーカイブとの検索を担保するのであれば、調査の特徴を示す「キーワード」の規模のレベルも、学問分野なのか、質問項目なのか、変数名なのかある程度基準が必要であるが、現時点ではRUDA内でもデータセットごとにばらつきがあり、統一基準が設けられていないのが現状である。

第二に、変換作業の手間がある。現在、

表1 DCとDDIの対応関係

DDI-C	DC基本要素	DC装飾子	意味	例:ruda0001のメタデータ
stdyDscr				
citation				
titlStmt	identifier	StudyNo		
IDNo			番号	0001
titl	title	display	調査名	東京版総合社会調査「高齢・少子社会における都市居住と家族・親族関係に関する調査」
altTitl	title	alternative	調査略称	TGSS2000
parTitl	-	-		
prodStmt				
fundAg	description	sponsorship	調査資金	東京都立大学都市研究所共同研究I
grantNo	-	-	(基金の番号)	
rspStmt				
AuthEnty	creator	-	調査主体/調査代表者	東京都立大学都市研究所/松本康
serStmt				
serName	relation	ispartofseries		
distStmt				
depositr	publisher	depositor	寄託者	松本康
distDate	date	issued	公開日	2010-10-01
distrbtr	publisher	distributor	配布者	立教大学社会情報教育研究センター
verStmt				
version	identifier	version		
holdings	identifier	uri	URI	http://hdl.handle.net/10577/57
dataAccs				
useStmt				
conditions	rights	AccessRights	利用条件	センターによる承認
citReq	identifier	citation	引用時の表記	「東京版総合社会調査2000」, 2000 (寄託者: 松本康 配布者: 立教大学社会情報教育研究センター)
method				
anlyInfo				
respRate	description	ResponseRate	有効回収率	32.9%
dataColl				
collMode	description	ModeOfCollection	調査方法	郵送配布・郵送回収法
dataCollector	description	investigator		
sampProc	description	SamplingProcedure	標本抽出法	層化多段系統抽出法により、各調査地から600ケースを抽出
targetSampleSize				
sampleSize	description	SampleSize	標本サイズ	3000
timeMeth	-	-	(クロスセクショナル・パネルなどの調査の対象時系列)	
notes	description	NoOfValidResponses		
stdyInfo				
abstract	description	abstract	調査概要	高齢少子化・高度情報化・グローバル化などの趨勢のもとにわたっている大都市東京の都市社会構造と社会意識の変化を経年的に明らかにする。
subject				
keyword	subject	-	キーワード	社会的ネットワーク 都市社会構造 社会意識
topcClas	description	discipline	研究分野	urban sociology 都市社会学 sociology 社会学
sumDscr				
anlyUnit	description	UnitOfObservation	観察単位	個人
collDate	coverage	temporal	調査時期	2000年9月
geogCover	coverage	spatial	調査地域	東京都港区、大田区、世田谷区、清瀬市、あきる野市
geogUnit	-	-	(最小の調査範囲水準)	
nation	-	-	(国)	日本
universe	coverage	-	母集団	20歳以上70歳未満の東京住民(島嶼部は除く)
dataKind	type	-	資料種別	
othrStdYMat				
relPubl	relation	-	関連資料(うち、データを使った研究成果)	(1) 松本康・原田謙, 2001, 「2000年東京総合社会調査の概要——調査設計、調査方法、回答率」『総合都市研究』, 76: 17-24. (2) 松本康編, 2004, 「東京で暮らす——都市社会構造と社会意識」東京都立大学出版会。
	relation	-	(うち、データを使った研究成果以外)	
relMat				
notes				

(前田・朝岡, 2018:30をもとにruda0001のメタデータを併記)



RUDAでは当該社会調査の実施概要に関するメタデータのみDDIに変換し公開している。かつては変数項目においてもDDIへの変換を試みたことがあったが、その煩雑さから現在は中断している。現在データ公開においてはDC形式とDDI形式の両方とも手動で入力・変換作業を行っており、将来的にはより効率的に作業ができる環境を整備することが望まれる。

第三に、既存のサービスや事業との連携や調整である。例えば、2021年4月から本格稼働したCinii Researchでは、ウェブサイト上のメタデータをもとにしてRUDAに寄託されているデータセットが検索可能である。これによって、当初RUDAで想定していたデータアーカイブ間の横断検索が実現されたともいえる。しかし、現在Cinii Researchで検索が可能なRUDAに関する情報は調査名や寄託者などごく一部であることに加え、SSJDAなどの他の社会調査データアーカイブとは項目が一致していない。社会調査データの検索システム構築においては、書籍・論文や科研費情報とは異なるメタデータの付与が必要である。利用者の利便性を考えれば、社会調査データアーカイブ間で提供されるメタデータは共通している方が望ましい。このような問題の解決方法として、DDIの活用を手がかりにメタデータの統一基準を作ることで、横断的な検索を可能にすることができると考えられる³⁾。

3 RUDAの実践から見える DDIメタデータ提供の課題と意義

(1) 活用・広報の課題

RUDAでは2018年度からDDIメタデータの提供を開始しているが、現在この事業はほとんど知られていない。というのも、第2節で論じたように、DDIはデータアーカイブ間の問題解決に寄与する側面が強いことに加え、そもそもDDIデータの役割や活用についての議論が日本国内では盛んではないためである。一機関の実践だけでは十二分にその意義が伝えにくい。実

際、RUDAのDDIメタデータは公開されて以降ほとんど活用されておらず、RUDAの運用現場としてもどのように有益性を訴えていくのかは今後の課題となっている。

DDIが英語圏においてすでに普及している状況を考えれば、英語でのメタデータ提供を開始することにより、海外からのアクセスやデータ利用が増えたり、海外の横断検索システムとの連携可能性を担保したりすることもできるかもしれない。しかし前述の通り、すでに寄託者から収集しているメタデータ情報において英語表記の併用を必須としていないため、現時点では海外の利用者への十分な情報提供は難しい。

一番の問題は調査名や変数名の翻訳である。アーカイブ側が基準を設定して事後的に英語情報を付けるのか、あるいは寄託者に個別に英語情報を問い合わせるのか、その判断は難しい。英語表記については第三に提起する問題とも関連しているが、いずれにしても何らかの統一基準やガイドラインを定めなければならないだろう。

(2) 運用・運営上の問題

第二に、RUDAの組織運営上の問題や課題がある。2000年代に日本各地の大学で社会学分野のデータアーカイブが構想され、運用が開始されたが、その後多くは休止や閉鎖となっている。例えば、大谷ら編(2013:56)が例示していた国内のデータアーカイブのうち、札幌学院大学のSORDや大阪大学のSRDQ、兵庫教育大学のJEDIは2021年5月時点で運用の終了や停止が確認されている。

RUDAはCSIという学内組織内の一業務として安定した運営が可能になっているが、それでも設立当初からの教職員や長く運営に関わっていた関係者の異動や退職に伴い、ノウハウの継続的な蓄積や新規事業開拓を含めた運営の見直しが課題となっている。とりわけRUDAのように小規模なデータアーカイブの運営においては、個々の運営や調査研究担当者の知識や熱意だけ

で支えるのではなく、体系的な知識や仕組みとして運営方法を共有し、引き継いでいく必要がある。

データアーカイブの運営や調査研究は、欧米においてはデータアーキビストとして専門職化が進んでいるものの、日本においては養成カリキュラムや教科書があるわけではなく、日々の実務や海外の事例⁴⁾を研究する中で身に着けていくしかない。そのため属人的にならざるを得ない面も確かにある。しかし、業界全体でデータライフサイクル⁵⁾ (Corti et al., 2014: 17-23) の活性化を目指すのであれば、DDIなどのデータ検索基盤の連携を通じて、アーカイブ間での人的交流や知見の共有がより進むことが望ましい。くわえて、個々の研究者がデータアーカイブ事業から担当が離れたとしても、このような事業の意義を社会に向けて伝えていく必要はあろう。

(3) 機関間連携の可能性

最後に、(1)・(2)の両方の課題ですでにふれているが、やはり機関間の連携、あるいはそれらの機関を総合的に取りまとめるための全国的機関と各アーカイブの連携は今後の社会調査

データ活用においては不可欠であろう。この点についてはRUDAに関する過去の論稿でもすでに強調されている(朝岡・前田, 2015; 前田・朝岡, 2018)。量的データアーカイブでの統一基準やガイドラインの整備を進めた先には、社会調査における質的データアーカイブの設立や連携可能性もあるだろう(高橋, 2020)。

今後各機関がいつその連携をしつつも、「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」により各機関の連携を集約し促進する事業が展開されることで、オープンサイエンスというデータ公開の方向のみならず、データを保存し将来に引き継ぐ事業も展開されていくだろう。このような展望を見据えつつ、データアーカイブの現場では、日々利用者や寄託者の研究に資する運営や取り組みを継続していく必要があろう。

謝辞

本稿の草稿には前田豊先生(信州大学)より有益なコメントを得ました。記して感謝を申し上げます。

注

- 1) cessa DC Date Catalogue <https://datacatalogue.cessa.eu/> (最終確認2021年6月15日)
- 2) 本事業の運営においては、多田はるみ、佐藤裕亮、渡辺浩平、李聡智、唐雯、益子亜明の各氏(いずれも当時立教大学大学院)の協力を得た。記して感謝したい。
- 3) Cinii Researchにおいては公開までに対応できたメタデータ項目に限りがあり、2021年6月現在で検索対象となっている項目は調査名、寄託者、調査主体/調査代表者、抄録(調査概要)、公開日のみである。
- 4) 海外の事例についてはIASSIST等の国際学会への参加を通じて知見を深めている。過去の参

加によって得た知見についてはCSI発行の雑誌にて報告している(朝岡・前田, 2017; 朝岡・高橋, 2019; 五十嵐・高橋, 2020, 2021)。

- 5) デジタルでの保存やデータキュレーションの実践を支えるためのコンセプトとして、1990年代から2000年代初頭にかけてとりわけ欧米で推進されてきた考え方。「発見と計画」「データ収集」「データ処理と分析」「出版と共有」「長期間にわたる管理」「データの再利用」などの活動の循環から成り立つが、Cortiらはこのデータライフサイクルのそれぞれの活動をつなぐメタデータの形式としてDDIをあげている(Corti et al., 2014: 17-19)。



文献

- 朝岡誠・前田豊, 2015, 「データアーカイブ事業の展望に関する一考察」『社会と統計』1: 49-62.
- 朝岡誠・高橋かおり, 2019, 「海外データアーカイブの動向2——IASSIST年次大会の報告から」『社会と統計』5: 33-41.
- Corti, Louise, Veerle Van den Eynden, Libby Bishop, and Matthew Woollard, 2014, *Managing and Sharing Research Data: A Guide to Good Practice*, second Edition. London: Sage Publications. Ltd.
- 五十嵐彰・高橋かおり, 2020, 「海外データアーカイブの動向3——IASSIST年次大会の報告から」『社会と統計』6: 75-82.
- , 2021, 「海外データアーカイブの動向4——JSM年次大会の報告から」『社会と統計』7: 33-40.
- 菊池進・松本康・山口和範・水上徹男, 2020, 「CSIのこれまでとこれから——開設メンバーで振り返る10年間のあゆみ」『社会と統計』6: 5-23.
- 前田豊・朝岡誠, 2017, 「海外データアーカイブの動向——IASSIST年次大会の報告から」『社会と統計』3: 27-35.
- , 2018, 「RUDAのDDI対応に向けた取り組み」『社会と統計』4: 25-35.
- 松本康, 2012, 「立教大学データ・アーカイブRUDAの始動」『社会と調査』8: 24-30.
- 大谷信介・木下栄二・後藤範章・小松洋編著, 2015, 『新・社会調査へのアプローチ——論理と方法』ミネルヴァ書房。
- 高橋かおり, 2020, 「質的データアーカイブ構想の現状と課題——数値化されていない調査データの保存と活用に向けて」『社会と統計』6: 65-74.

6

公的統計データの 研究利用の現状と展望 一橋大学における取り組みの例

田中雅行

一橋大学経済研究所 准教授

1 はじめに

一橋大学経済研究所では、かねてから社会科学統計に関する情報を収集・整備し、学術研究者に広く提供してきた。現在、研究所の附属施設である社会科学統計情報研究センター（以下、センター）では、（独）日本学術振興会（以下、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」（以下、データインフラ事業）の一拠点として、人文学・社会科学に係るデータを分野や国を超えて共有・利活用する総合的な基盤の構築を行う事業に参加している。本稿では、一橋大学におけるデータインフラ事業の取り組みについて紹介する。

データインフラ事業では、センターを含めた各拠点が保有・提供するデータを分野横断的に一括検索できるデータカタログの構築を目指し、各拠点はそれぞれの担当分野において保有・運用するデータアーカイブの構築およびメタデータの整備等を行っている。

センターの役割は、公的統計の分野において、歴史的な統計データを中心にメタデータの整備を行うことである。この詳細を述べる前に、公的統計データの研究利用の現状について振り返ってみる。

2 公的統計データの研究利用の現状

公的統計、すなわち国の行政機関や地方公共団体などが作成する統計は、社会全体で利用される情報基盤として位置づけられている。とくに、1947年に公布された統計法が全面改正されて2009年に施行された新統計法においては、公的統計は政府のためだけのものではなく、広く国民が容易に入手し、効果的に利用できるものとして提供されなければならないことが基本理念として明確に打ち出された。

公的統計の利用を大きく一次利用と二次利用とに分ける場合、一次利用は本来の目的である統計の作成のために実施された統計調査の集計結果（マクロデータ）を利用することがこれにあたる。研究者等がマクロ分析を行う場合に利用するもので、集計結果は広く一般に提供されている。

集計結果が電子データ化され、インターネット上で提供されるのが主流になった90年代後半以降、政府は統計調査の集計結果のデータ化、共有化及びオープン化を推進してきた。2000年代の中頃には、現行の大部分の調査結果がデータ化され、インターネット上で提供されるようになった。しかし、分散型の統計機構を持つ日本の統計制度ゆえ、各府省の所管において実施さ



れた統計調査の結果は、各府省が独自に運用するウェブサイト上に散在する状況が長らく続いていた。この状況は、各府省間での情報の散見やウェブサイトのデザインや構成の不統一などもあり、利用者にとっては不便を強いられる形態であったといえる。

そこで、利用者にとって便利で使いやすい統計データの提供や、各府省における統計情報システムの統一や集約を図る目的で、政府は電子政府構築の一環として統計調査業務・システムの最適化計画を推し進め、この計画に基づき「政府統計共同利用システム」が開発され、2008年から運用が開始された。このシステムは、公的統計の業務全般における複数のシステムが集約されたものであるが、統計データの利用者にとっては、システムの一機能としてのポータルサイト「政府統計の総合窓口 (e-Stat)」が最も重要である。各府省が公表する統計データ、公表予定、調査項目などの各種統計情報がこの中に一元的に集約され、利用者は必要とするこれらの情報を、調査を所管する各府省のサイトを經由することなくe-Statから直接、ワンストップサービスとして提供を受けることができるようになった。

e-Statは、調査を実施する各府省が作成した統計表を、EXCEL、CSV、PDFなどの元のファイル形式のまま収録し、ダウンロードを可能にしているほかに、国勢調査を始めとする一部の主要な統計調査については別途、データベース形式に加工して収録しており、項目の抽出、行・列の組み換えやグラフの作成等がサイト上で行え、抽出したデータをダウンロードすることもできる。その他、GIS (地理情報システム) を活用することで、各種統計データを地図上に表示し、視覚的に統計を把握できる統計GIS機能も利用でき、e-Statの運用開始以降、利用者は公的統計データを容易に入手できるようになったばかりでなく、目的に沿って自身で加工や分析をすることも可能となっている。

一方、二次利用とは集計前の個票である調査

票情報 (マイクロデータ) を利用するものである。世帯単位や事業所単位などの、集計する前の個票形式で提供されているマイクロデータを用いることで、調査実施機関が行う公式集計では提供されていない統計表を利用者が自ら作成したり、経済理論に基づいた実証分析を行うなど、自由な分析を行うことが可能である。

新統計法による制度下で新たな二次利用制度の運用が始まって10年以上が経過した。調査票情報を二次利用できる枠組み自体は、旧統計法制度下の時代からあったものである。ただし、「目的外利用」と呼ばれ、総務大臣の承認が必要とされた。中央官庁や地方自治体などの公的機関による政策目的で統計を作成する場合には、現在と変わらない相当数の利用が認められたが、こと研究者の利用に関しては、省庁の依頼による共同研究に参加する等、いくつかの限られた経路でのみ利用が可能であり、研究者側からの申請による利用、いわゆる学術利用はわずかな実績に留まっていた。

そこで、欧米等諸外国における公的統計データの二次利用の状況や提供方法を参考に検討を行い、新統計法下では、統計は社会全体で利用される情報基盤であるという理念に基づき、公的統計が一層活用されるよう二次利用が明確に規定された。それを受けて、秘密の保護を図った上で、新たな統計作成や統計的手法を利用した学術研究等のために「調査票情報等」を利用できる制度が創設された。

現在、新統計法の制度下における二次利用の形態には以下の3とおりがある。

- (1) 調査票情報
- (2) 委託による統計の作成等 (オーダーメイド集計)
- (3) 匿名データ

(1) は、旧統計法下にあった目的外利用の枠組みを継承するもので、調査票情報すなわち集計前の個票データをそのままの形で利用する形態である。ただし、これまでどおり公的機関

が利用する場合のほか、公益性を有すると認められる統計的研究を研究者等が行う場合の利用が明確に認められるようになった。利用形態は、CD、DVD等の磁気媒体に複写されたデータの貸与を受けるものと、データの持ち出しができない仕組みや作業内容の監視システムなど、高度な情報セキュリティが確保された専用施設（オンサイト施設）に研究者が出向いて分析を行うオンサイト利用の2とおりがある。

(2) は、利用者からの委託を受けて、調査実施者が利用者の分析目的に応じた統計表を新たに作成する形態で、いわゆるオーダーメイド集計と呼ばれるものである。

(3) は、調査客体が特定されないよう、調査票情報に加工処理（匿名化措置）を施した匿名データの貸与を受ける形態である。匿名化措置を施すことで、個票開示リスクを低減しつつ、調査票情報よりも利用要件を緩和したもので、大学においては、教員だけでなく大学院生の利用も想定したものである。

(2)、(3) はともに、新統計法で新たに加わった二次利用の形態であり、データのセキュリティレベルに応じた複数の利用形態を用意することで、研究者を中心とする学術利用の促進を狙ったものである。ただし、学術利用を増やすには、利用者である大学を中心とする学術研究機関側との連携を強化し、行政機関側との両輪で二次利用の促進を行うことが重要である。たとえば、二次利用の提供サービスの委託を政府から受けて統計データアーカイブを構築・運営する（独）統計センターは、一橋大学を始めとする複数の学術研究機関と連携協力協定を結び、公的統計の二次的利用に関する研究・開発、普及・啓発を共同で推進している。

一橋大学は、この連携協力協定に基づく（独）統計センターのサテライト機関として、公的統計データの二次利用促進に取り組んでいる。取り組みの内容は、学内者のみならず学外者も対象とするもので、二次利用申請に関する利用相談

や手続きに対応するほか、サテライト機関に認められている匿名データの提供事務（データの複製、配布）を行っている。この匿名データとオーダーメイド集計の利用はどちらも有料であり、利用するデータや作成を依頼する統計表の分量に応じた利用料金（事務手数料等）を利用者が負担しなければならないが、このことが、原則的に利用が無料である調査票情報に比べて両者の利用が伸び悩んでいる一因でもある。そこで一橋大学では匿名データとオーダーメイド集計の利用にかかる利用料金の全部または一部を利用者に補助する取り組みも行っている。十分な研究費を得られない若手研究者や大学院生にとっては、研究費のかからない利用手段となっており、利用実績を伸ばしている。

このほか、利用環境の整備・提供も行っており、調査票情報の磁気媒体利用と匿名データの利用は、利用に際しての入退出管理の実施や、施錠可能なデータの保管場所、盗難防止措置などを明示した利用場所を事前に申請しなければならない。このような利用環境を自前で用意できない利用者に対しては、利用基準を満たしたマイクロデータ利用スペース「共用ラボラトリー」を学内に設置、提供している。また、小平国際キャンパス内に情報セキュリティを確保したオンサイト施設を設置し、調査票情報のオンサイト利用も可能となっている。共用ラボラトリー、オンサイト施設はともに、事前の申請があれば学外者も利用することができる。

一例として一橋大学の取り組みを紹介したが、他のサテライト機関においても同様もしくは独自の取り組みを行っている。このように、公的統計データの二次利用に関しては、官学一体となって促進に取り組んでおり、大学を始めとする学術研究機関は重要な役割を果たしている。



3

データインフラ事業における
一橋大学の取り組み

一橋大学では、公的統計分野におけるこれまでの実績や強みを生かし、冒頭に述べたデータインフラ事業において、「長期間をカバーする政府統計データや調査票情報等のメタデータのデータベース化・英語化」を中心に作業を進めている。具体的には、『日本統計年鑑』と国勢調査を始めとする主要な統計調査を対象に、メタデータの作成、データベースの構築を行い、経済研究所が開設するデータリポジトリに登録・公開し、同時に学振とNIIが構築する総合データカタログJDCatからの検索を可能とする作業を進めている。

『日本統計年鑑』は、総務省統計局が刊行する日本の国土、人口、経済、社会、文化などの広範な分野にわたる基本的な統計データを網羅的かつ体系的に収録した総合統計書であり、官庁統計のみならず民間調査機関などが実施している統計調査、さらには業務統計および加工統計からも基本的なデータを選択し、収録している(図1)。その名のとおり、創刊以来冊子を主形態として刊行されているが、平成12年以降は、統計表は付録のCD-ROMとしてEXCEL化されており、平成25年以降は統計局のWEBサイトから専用ブラウザでHTML5版の全文閲覧が可能となっている。

一橋大学では、データ化される以前のものについて、『日本統計年鑑』の前身である『大日本帝国統計年鑑』に遡って明治15年から平成22年のものまで、収録されている統計表のメタデータを作成し、データベース構築を行っている。

メタデータは最終的に、総合データカタログに収集されることを考慮し、社会調査メタデータの国際規格であるDDIに準じたJDCatメタデータスキーマに基づき、トピック等の統制語彙による、表題、対象とする年次、出典(資料・調査・機関等)などのキーワードを付与するとともに、主要な統計調査に基づき連続して収録されてい

る統計表にはシリーズ名を付与している。これにより、統計データの時系列での探索の支援と、データインフラ事業における他の拠点で作成するメタデータとの横断的検索を可能としている。

また、国を超えてデータを共有・利活用するというデータインフラ事業の理念に沿い、海外利用も想定し、メタデータは日本語版だけではなく、英語版の作成も行っている。『日本統計年鑑』の統計表も、近年のものは日本語・英語併記となっているが、戦前のものは日本語標記のみであるため、英語化の作業も同時に行っている(図2)。

さらに、メタデータに加え統計表も著作権保護期間が満了しているもの、利用許諾を得たものについては、OCR(光学的文字認識)機器で読み取り、PDF形式で画像ファイル化し、メタデータと結びつけて公開を行うとともに、表頭、表側、表注の対訳表を結ぶことでフルテキスト検索に対応可能としている(図3)。

これら紙ベースの統計表を電子データ化する方法としては、OCR機器で読み取った後、機械

2 1 地

1. 土 地

1. 期 間 及 面 積 (昭和15年) 統計局編

本誌に掲載の統計表は、昭和15年の国勢調査の結果に基づき、昭和15年1月1日現在の状況を基として集計した。国勢調査の結果は、国勢調査報告書(昭和15年)に掲載されている。本誌に掲載の統計表は、国勢調査の結果に基づき集計した。

期 間	国 勢 調 査 年 次	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)
期 間	国 勢 調 査 年 次	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)
期 間	国 勢 調 査 年 次	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)
期 間	国 勢 調 査 年 次	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)
期 間	国 勢 調 査 年 次	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)
期 間	国 勢 調 査 年 次	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)
期 間	国 勢 調 査 年 次	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)
期 間	国 勢 調 査 年 次	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)
期 間	国 勢 調 査 年 次	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)
期 間	国 勢 調 査 年 次	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)
期 間	国 勢 調 査 年 次	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)
期 間	国 勢 調 査 年 次	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)
期 間	国 勢 調 査 年 次	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)
期 間	国 勢 調 査 年 次	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)
期 間	国 勢 調 査 年 次	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)
期 間	国 勢 調 査 年 次	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)
期 間	国 勢 調 査 年 次	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)
期 間	国 勢 調 査 年 次	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)
期 間	国 勢 調 査 年 次	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)
期 間	国 勢 調 査 年 次	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)
期 間	国 勢 調 査 年 次	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)
期 間	国 勢 調 査 年 次	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)
期 間	国 勢 調 査 年 次	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)
期 間	国 勢 調 査 年 次	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)
期 間	国 勢 調 査 年 次	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)
期 間	国 勢 調 査 年 次	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)
期 間	国 勢 調 査 年 次	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)
期 間	国 勢 調 査 年 次	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)
期 間	国 勢 調 査 年 次	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)
期 間	国 勢 調 査 年 次	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)
期 間	国 勢 調 査 年 次	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)
期 間	国 勢 調 査 年 次	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)
期 間	国 勢 調 査 年 次	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)
期 間	国 勢 調 査 年 次	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)
期 間	国 勢 調 査 年 次	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)
期 間	国 勢 調 査 年 次	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)
期 間	国 勢 調 査 年 次	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)
期 間	国 勢 調 査 年 次	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)
期 間	国 勢 調 査 年 次	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)
期 間	国 勢 調 査 年 次	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)
期 間	国 勢 調 査 年 次	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)
期 間	国 勢 調 査 年 次	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)
期 間	国 勢 調 査 年 次	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)
期 間	国 勢 調 査 年 次	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)
期 間	国 勢 調 査 年 次	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)
期 間	国 勢 調 査 年 次	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)
期 間	国 勢 調 査 年 次	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)
期 間	国 勢 調 査 年 次	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)
期 間	国 勢 調 査 年 次	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)
期 間	国 勢 調 査 年 次	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)
期 間	国 勢 調 査 年 次	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)
期 間	国 勢 調 査 年 次	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)
期 間	国 勢 調 査 年 次	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)
期 間	国 勢 調 査 年 次	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)
期 間	国 勢 調 査 年 次	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)
期 間	国 勢 調 査 年 次	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)
期 間	国 勢 調 査 年 次	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)
期 間	国 勢 調 査 年 次	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)
期 間	国 勢 調 査 年 次	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)
期 間	国 勢 調 査 年 次	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)
期 間	国 勢 調 査 年 次	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)
期 間	国 勢 調 査 年 次	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)
期 間	国 勢 調 査 年 次	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)
期 間	国 勢 調 査 年 次	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)
期 間	国 勢 調 査 年 次	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)
期 間	国 勢 調 査 年 次	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)
期 間	国 勢 調 査 年 次	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)
期 間	国 勢 調 査 年 次	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)
期 間	国 勢 調 査 年 次	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)
期 間	国 勢 調 査 年 次	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)
期 間	国 勢 調 査 年 次	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)
期 間	国 勢 調 査 年 次	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)
期 間	国 勢 調 査 年 次	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)
期 間	国 勢 調 査 年 次	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)
期 間	国 勢 調 査 年 次	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)
期 間	国 勢 調 査 年 次	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)
期 間	国 勢 調 査 年 次	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)
期 間	国 勢 調 査 年 次	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)
期 間	国 勢 調 査 年 次	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)
期 間	国 勢 調 査 年 次	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)
期 間	国 勢 調 査 年 次	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)
期 間	国 勢 調 査 年 次	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)
期 間	国 勢 調 査 年 次	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)
期 間	国 勢 調 査 年 次	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)
期 間	国 勢 調 査 年 次	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)
期 間	国 勢 調 査 年 次	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)
期 間	国 勢 調 査 年 次	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)
期 間	国 勢 調 査 年 次	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)
期 間	国 勢 調 査 年 次	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)
期 間	国 勢 調 査 年 次	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)	面 積 (平方メートル)
期 間	国 勢 調 査 年 次	面			



可読なテキストデータ化を行うのが理想ではあるが、調査や年次、表によってばらつきのある結果表様式から、表題、集計項目、結果数値などの必要な情報の吸い出しを全自動で行うのは難しく、読み取り後の修正に相当数の手作業を伴うのが現状である。全巻全ての統計表について、膨大な手作業を費やしてこれらの作業に取り組むのは現実的ではないことから、報告書の統計表をページごとに画像データとして読み取り、PDFファイル形式で提供することとした。

前述のように、公的統計データの一次利用に関しては、e-Statの整備により全ての政府統計が一つのポータルサイトに集約され、統計表を始めとする統計関係情報のワンストップサービスとオープン化が実現しているが、いくつかの課題がまだ残されている。

e-Statに収録された統計表は、ほとんどの調査で統計表の電子データ化が進んだおおよそ2000年以降のものに限られている。一部の調査については、主要な指標が長期時系列データとして整備されており、国勢調査を始めとする総務省統計局の主要な調査については、過去に遡ってデータベース化する作業を順次行っているが、他省庁のものを含むそれ以前に公表された未収録の膨大な調査結果については、依然として報告書をめくらなければならないものも多い。実際、e-Statでは昭和50年以前の国勢調査については、PDFの画像ファイルが提供されており、検索が行えないなどの不便な点もあるが、わざわざ図書館に足を運ばなくても報告書にウェブ上でアクセスできるという点では利用者にとっては非常に有益である。

しかし、e-Stat未収録の全ての統計表をPDF化するのも大変な作業であり、現実には実施される見通しもないことから、全分野を網羅した総合統計書である『日本統計年鑑』をPDF化して提供することは、そうした課題への一つの対応になりうると考える。

e-Statに収録された国勢調査を始めとする主

要な公的統計調査についても、明治、大正期に遡り、メタデータの作成、データベースの構築を行っている。メタデータは、『日本統計年鑑』同様、調査票様式(画像イメージ)のほか、調査主体、母集団、抽出方法、観察単位などのJDCatメタデータスキーマに基づいた項目や、産業分類、職業分類といった集計用の分類、調査を実施する根拠となる法律、政令などの関連法令の情報を収集している。

こちらも、『日本統計年鑑』同様、英語化の作業を同時に行っている。政府統計マイクロデータの二次利用は、運用上の制約から大部分は国内利用が前提になっているが、学術利用促進の観点からは、外国人研究者の利用促進についても考えていかなければならない。一橋大学では、外国人研究者とのマイクロデータを用いた共同研究を多数実施しているが、その際にネックになるのは、利用するマイクロデータのメタデータであるデータレイアウトや符号表の英語版が用意されていないことである。e-Statや各府省のウェブサイトで各調査の英語版の情報を取得しようとしてもまだまだ十分とはいえない状況にあることから、メタデータは日本語・英語版両方の作成を行っている。

このようにして作成したメタデータは作業が完了したものから順次、「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/?page=1&size=20&sort=controlnumber>)に登録し、公開している。このデータリポジトリは、オープンアクセスリポジトリ推進協会とNIIが共同運営するクラウドサービスである新JAIRO-Cloudの環境下で構築したものであり、2021年6月30日から試験公開を開始、順次コンテンツを追加しつつ、改善を図っている。同時に、学振においてもデータインフラ事業の本体である総合データカタログJDCatの公開を開始し、当学データリポジトリとの連携を開始している。

また、一橋大学におけるデータインフラ事業では、専用のウェブサイト (<https://d-infra.ier.hit-u>)

ac.jp/Japanese/)において、取り組み内容やコンテンツの公開を行っている。

4 今後の展望

ここまで、一橋大学における公的統計データの研究利用にむけた学術研究機関としての関わりと、政府との連携による二次利用促進への取り組み、そして、データインフラ事業を通じての現在の取り組みについて簡単に紹介した。

政府統計データの提供は、一次利用においてはe-Statの整備によりワンストップサービスが一応実現している。操作性向上や機能拡張のための改修も随時行われており、オープン化という点では、API機能の追加やLODによるデータ提供も現在では行われているが、今後は更なるデータの集約・共有が望まれる。

二次利用においては、オンサイト利用を中心に、利用環境等のセキュリティレベルに応じたデータ提供の在り方を検討することが「公的統計の整備

に関する基本的な計画」にも示されており、オンサイト施設のさらなる展開が進められている。民間や学術レベルでは、プログラム送付型の集計やリモートアクセス型の利用などの新たな利用形態の研究も進んでおり、今後は公的統計への導入の検討も必要であると思われる。匿名データに関しても、利用可能な調査や年次が少ないことがネックであるが、新たな調査や年次の追加の作業が望まれる。

一橋大学では、(独)統計センターとの連携協力協定に基づき、匿名データの提供、オンサイト利用を中心に、今後も学術研究機関の立場から公的統計データの二次利用促進に尽力していきたい。

データインフラ事業についても、『日本統計年鑑』や公的統計のメタデータのデータベース化は、膨大な作業を今後も地道に続ける必要があるが、サテライト機関として政府の取り組みを補完するとともに、学振が掲げる分野や国を超えたデータの共有・利活用基盤の構築に向けて引き続き取り組んでいきたい。

文献

朝岡誠・林正治, 2019, 「オープンサイエンスの取り組みとしての人文・社会科学分野におけるデータインフラストラクチャー構築推進事業」『ESTRELA』(308) : 15-20.

伊藤伸介・前田幸男, 2019, 「分野横断的なデータカタログの整備に向けて——現状と課題」『ESTRELA』(308) : 8-14.

小林良行, 2012, 「公的統計マイクロデータ提供の現状と展望——一橋大学での取り組みをもとに」『日本統計学会誌』(シリーズJ?) 41(2) : 401-420.

小泉邦彦, 2012, 「政府統計共同利用システム」『統計』63(3) : 2-8.

横田直木, 2012, 「公的統計のマイクロデータ利用の現状と課題」『統計』63(3) : 9-16.

松田芳郎, 2008, 「日本におけるマイクロ政府統計活用の新しい夜明け」『統計』59(12) : 2-9.

中村英昭, 2014, 「欧米諸国におけるリモート方式による公的統計データの二次的利用——利用者の倫理に依存しない提供手段」『調査と統計』12 : 38-45.

——, 2017, 「公的統計データの二次的利用の動向：オンサイト利用を中心に」『理論と方法』32(2) : 310-320.

高部勲, 2020, 「公的統計マイクロデータの利活用状況と課題——提供者及び利用者の観点から」『統計』71(8) : 4-9.

美添泰人, 2009, 「統計の有効活用に関する展望と課題——国民の共有財産としての統計」『ESTRELA』(181) : 9-17.