《資料》

海外データアーカイブの動向 3
—IASSIST 年次大会の報告から—

五十嵐　彰
髙橋　かおり

【要旨】　社会調査データは今後の社会の発展に寄与する公共財であり，広くデータが利活用される環境を整備する必要がある．本稿では 2019 年 5 月に開催された IASSIST 年次大会で報告された内容をもとに海外のデータアーカイブで行われている先進的な取り組みを紹介する．これらを踏まえ，今後の CSI 業務ならびに RUDA 運営に対しての方針と示唆を提案する．

キーワード：データアーカイブ，データ利用，データ共有

Ⅰ　はじめに

近年日本でもオープンサイエンスの動きが加速し，自然科学のみならず人文科学・社会科学におけるデータ利活用が政策的・学術的論点となりつつある．とりわけ国立情報学研究所（NII）が中心に進めているオープンデータサイエンス構想においては，各拠点間の連携を果たしつつ，共通してデータを共有・利用しあう基盤の構想が進められている（廣松 2019）．しかし日本におけるデータ利活用の議論とその動きは諸外国と比べると圧倒的に遅れているのが現実である．

CSI 社会調査部会では，社会科学に関する情報技術・データサービスに従事する専門家の国際組織である IASSIST が開催する年次大会に部会業務として毎年参加している．本稿では，前田・朝岡（2017），朝岡・髙橋（2019）に引き続き，IASSIST にて開催されたワークショップや研究報告をもとに海外のデータアーカイブで行われている先進的な取り組みを紹介し，RUDA 運営を含めた今後の CSI における業務の指針を提案する．

今回の年次大会のテーマは“Data down under”であり，オーストラリアの別称である Land down under をもじったものである．データはあらゆる意思決定の根幹にあるものであり，そうしたデータをいかに整備し利用を促すかといった側面に焦点を当てた発表が多く見られた．参加者は世界各国の大学教育部局スタッフ，ライブラリアン，メディアセンター職員などである．Business Meeting で示されたデータからは，大学のライブラリアンが占める割合が多いことが明らかになった．

本年度の研究大会では，全体に質問紙調査を保存する旧来的なデータアーカイブの役割を越えて，多様化するデータや調査の形式への対応や，データ・ライフ・サイクルの推進，寄託されたデータの活用やリテラシー教育，さらにはデータを利用した大学改革など，データ保存を前提として，いかにデータを活用するのかということが議論の中心であった．さらに南半球での開催ということもあり，ヨーロッパ・北アメリカ中心のデータアーカイブの動

向ではない，オルタナティブな実践や試みが紹介されていた [1].

Ⅱ.保存データの多様化への対応

1. データの質の維持

2018 年に EU 一般データ保護規則（GDPR）の適用が開始されたことにともない，各地のデータアーカイブもその対応が課題となっていた．とりわけ今回の開催国のオーストラリアでは，EU の規則を参考にしながらも，その問題点を改善しつつ，独自の保護規定を作る動きがある（Russell, 2019）．すべてのデータを公開に向けて保存するのではなく，オープンデータ化は推進しつつも，歴史的背景や文脈に即した利活用の方法と段階，時期の判断が求められる．

さらに，寄託されるデータの質の担保も複数のデータアーカイブにおいて運営上の課題となっていた．例えば UK Data Service ではある程度自動的に寄託者がデータの質を確認できる「データ・ヘルス・チェック」と言われる制度が導入されつつあった（Corti, 2019）．寄託から公開までの過程を短縮する試みは，ICPSR でも行われている．これらの過程を一部自動化することは，アーカイブ側の時間や労力の省略に寄与するものの，すべての自動化は不可能であり最終的には専門家の判断が必要になるという見解は共通していた．

これらの取り組みから，データの種類が複雑になるにつれて，専門職者の果たす役割も増す一方で，寄託数の増加に伴い，利用者や寄託者側にも最低限のリテラシーや判断基準を持つことが求められていることがわかる．日本ではそもそも「データの寄託」という意識が共有されていない状況であるが，オープンデータ化を目指す前にデータを収集し保存すること，そしてそれを利活用するのはなぜなのか，という研究に関わる根本的な意識改革を進めることが不可欠であろう．

2. 多様な形式のデータ保存に向けて

データ利用者からの報告においては，既存のデータセットに収まらない形のデータ分析の例が示されている．ここには，旧来から存在する量的／質的という単純なデータの種類の問題だけではなく，技術の進歩によってこれまで取得できなかった，あるいは想像もしていなかった種類のデータの入手が可能になったことが要因にある．

例えば，画像データはその 1 つであろう．文化財保護の文脈において，写真や図面の保存に向けた取り組みの紹介があった（Kirby, 2019）．建物そのものだけではなく，建物に関するデータや保存は，万が一災害や事故などによって現物が破損や消失したとしても，修復や復元を可能にする助けとなる．ただし実際にデータを記録・保存するのであれば，そもそも何のために保存するのか，単なる搾取ではなく，コロニアリズムへの反省を踏まえた取り組みになっているのか，という問いかけもなされる．情報保存するという目的を定めるだけではなく，保存データに関しての著作権・肖像権などの法的・倫理的な利権関係も明らかにしておかねばならない．

あるいは，オンラインゲームのプレイデータといったビッグデータの収集も可能になっている．犯罪組織分析の代替モデルとして，犯罪をテーマにしたオンラインゲームデータの

分析をした事例においては，分析の可能性と同時に，IP 情報や利用者 ID をめぐる利用者のプライバシーの保護などが今後の展開に向けた課題としてあげられた（Fleet, 2019）．

さらに，社会科学データアーカイブの老舗である ICPSR においても，ソーシャルメディアの情報をどのように扱うのかということが議論されていた（Hemphill, 2019）．ソーシャルメディアの情報は可能性に満ちているとはいえ様々リスクがあり，質の良いデータを保存しつつ提供してくプラットフォームの構築には各所との調整が必要であるという報告がなされた．

新たなデータの保存・活用の事例からは，以下の点が課題として提出されている．第一に，データの保存の目的である．二次利用を目的とした保存であったとしても，ソーシャルメディアやゲームのように同時代の人に向けて活用促進を図るのか，あるいは文化財情報のように将来の人のために向けて保険として残すのかによっても保存のための制度設計は変わってくる．第二に，保存の際に生じる倫理的・法的手続きである．個人情報の保護だけにとどまらず，データの所有権の明確化も必要な手続きである．そして第三に，どのような形式で提供するかである．画像やネットワーク情報など，従来のデータセットよりも複雑なデータ形式になるのであれば，その利活用の際の説明やチュートリアルなどの充実も図らなければならない．

このように，取得できるデータが多様化する中で，これまでのデータアーカイブとしては形式が異なる資料やデータについても，その保存方法を検討していく柔軟な姿勢が求められるともいえよう．

Ⅲ　データ利活用の実際――研究以外での応用可能性

2019 年の年次大会においては，大学でのデータの利活用に関する教育的セミナー報告が複数見られた．例えばイーストミシガン大学の「Datathon」は，プロジェクトベースの市民向け学習プログラムであり，行政データの再分析や解釈を行いながら地域の水問題への理解を深める取り組みであった（Brodsky & Kelly, 2019）．参加者は高校生や高校教員などが多く，受講生がデータへの興味を持つ一助となっていた．プログラム運営においては実際にデータに触れて分析する時間の少なさや運営時における財政問題など指摘されてはいたが，これは大学の地域連携の新たな形の 1 つといえよう．

別の例では，学生が自学自習の過程で適切なデータにたどり着くための指針や援助のあり方が紹介された（Sheley & Arave, 2019）．インディアナ大学では CRAAP という指標を制作し，適切に流通しているか（Currency），妥当性をもっているか（Relevance），適切な機関や団体が発行しているか（Authority），正しいか（Accuracy），そして知りたいことの目的にあっているのか（Purpose）の 5 点を確認できるワークシートを学生に提供するようになった．

同様の例として，カールトン・カレッジでは「Data Squad」というデータ利活用に特化した学生スタッフを置き，おもに学部生からの質問に答えられるような制度を整えていた（Lackie, 2019）．

この 3 つの実践からは，データを利用できるようにするだけではなく，どのように利用す

ることが可能か，そしてその利用方法の適切さをいかに確認するのかということが，データ利活用教育において欠かせないということが明らかになる．CSI の各種セミナーやコンサルティングもこのような役割を果たしているといえるが，今後の展開や活動の広報をする際の参考になる事例であった．

さらに，アメリカのロチェスター大学では，データチーフという役職に抜擢された学外の専門家が，学内のデータ活用の実態とその散逸の状況を把握し，必要なデータを適切に提供する取り組みを行ったことで，大学改革につながったという報告があった（Cannon 2019）．この事例を立教大学に応用できるのであれば，データや資料を保有する複数の部局が互いの状況を知り，活用しあうことが，学術面のみならず大学のマネジメントの変化にもつながることが考えられる．

各機関のデータ利活用の取り組みからは，データ利用におけるコミュニケーションの受容さと，単に「データを分析する」という作業面だけではなく目的や活用例，そしてその影響も考えたマネジメントプランを考える思考を利用者側も分析者側も持つことの重要性が明らかになった．

Ⅳ　大学間ネットワークの構築

複数の報告で，大学間の協同ネットワークやその構築を志向していることが目立った．大学間ネットワーク構築の目的として，データアーカイブそのものの共有というよりは，アーカイブに関する様々な業務を共有して負担を減少させることを目的としているようであった．

例えば Data curation network（DCN)というデータキュレーション（データチェック・クリーニングなどの一連の過程）を行うネットワークを 9 つの機関で構築しているという報告があった（Blake & Herndon, 2019）．発表時の参加機関としてジョン・ホプキンス大学，デューク大学，ミネソタ大学，セントルイス・ワシントン大学，コーネル大学，ペンシルバニア州立大学の 8 つの大学，そしてデータリポジトリの DRYAD が挙げられていた．ネットワーク参加者の内訳として，20 人以上のデータキュレイター，7 人の代表，そしてアドミニストレータとプログラムディレクターとが一人ずつということであった．活動内容はキュレーションのステップの標準化，キュレーションツールやトレーニングデータの共有，ワークショップなどを通したキュレイターの育成，ベストプラクティスの共有などであり，事例などを何本かの論文にまとめることもしている．

2 つ目の事例として，ICPSR のリサーチャーパスポートが挙げられる（Hemphill, 2019）．データアーカイブには，個人を特定できる情報などを含むデータに対して制限を加えており，こうしたデータが 1,482 件蓄積されている．制限データにアクセスするには，研究者個人がオンラインアプリケーションを行い，ICPSR の職員がそれを閲覧（必要であれば修正の要求）し，アクセスを承認するという流れがある．このアプリケーションと承認のプロセスを簡略化するために，リサーチャーパスポートを提案している．リサーチャーパスポートはデータ利用者それぞれに与えられるアカウントで，データ利用者の情報（博士号の有無，データ利用に関する受講の有無，利用経験など）を搭載している．リサーチャーパスポートはデータ利用者がどの程度信頼できるかを測るバロメーターとして機能し，データ利用の

許可の判断をリサーチャーパスポートが肩代わりすることができる．現在は ICPSR のみの適応となっているが，今後はデータアーカイブ間でパスポートを共有することも構想しているとのことだった．

以上のように，アーカイブ間の共同では，データの共有ではなくそれにまつわる業務についての共有を行っていることが目立っていた．またこうした構想に共通しているのは，研究者ではなく，データキュレイターなどの技術者による先導だろう．技術者の専門的な知識と経験をもとにした主導により，こうしたネットワーキングや新たな構想が可能になるのではと考えられる．

V　再現性確保へのデータアーカイブの役割

社会科学の分野において，昨今再現性の問題が盛んに取り上げられている．トップジャーナルに掲載された心理学の研究のうち 39%しか再現に成功しておらず（Open Science Collaboration, 2015），大きな議論を巻き起こした．これらは異なるデータを用いて同じ結果を示す反復可能性（repeatability）を指しており，理論の形成と学問の発展にとって重要なものであるといえる（Freese & Peterson, 2017）．他方でより根本的な，研究者が用いたデータと手法で同一の結果を示す検証可能性（verifiability）に関してはより整備が進んでおり，政治学のジャーナルでは論文採択時にデータを求めるものが多くなってきている．仮に検証可能性が確保されないと，最悪のケースでは研究者による捏造などを見過ごすことになる．こうした風潮に対応するために，データアーカイブが再現可能性（検証可能性）を促すためのトレーニングなどのサービスを行っているという事例がいくつか散見された．

コーネル大学のデータセンター（CISER）では，再現性の確保のための新たな事業を立ち上げている（Heslop, 2019）．CISER では，投稿前の段階で研究者が実際に用いたコードとデータを用いて結果のダブルチェックを行い，さらにデータとコードをパッケージとしてまとめ，投稿後のデータ投稿要求に即座に答えられるように体制を整えている．利用者は 2018 年の段階では 5 件，2019 年では 20 件と順調に対象を拡大している．現在では研究者に限定されているが，今後は修士や博士の学生向けにサービスを拡大する計画のようである．

次に，イエールとコーネルから，Data CURE Training Program の紹介があった（Christian, Thompson, Peer, & Arguillas, 2019）．このプログラムは研究者に対する再現可能性に関するトレーニングを行うと同時に，情報専門職の職員に対してはデータマネジメントの方法に関する授業やワークショップを，リサーチスタッフに対してデータキュレーションに関するトレーニングを行っている．研究者に対するトレーニングで再現性に関する基本的な考え方を授ける一方で，再現性確保のために研究者以外のスタッフに対して，データの整備やコードの解読ができるように教育している．再現性の確保が大学の評判に関わるという考えからか，全学的な取り組みとして再現性確保のためのシステムが構築されつつあるといっていいだろう．ただこれはあくまでコーネルを中心とした取り組みであって，その他の大学についてどこまで当てはまるかは不明である．

IASSIST に関する事例ではないが，共著者の一人である五十嵐が過去に所属していたオランダのユトレヒト大学では，再現性に関する取り組みが 2010 年代初頭から行われていた．

2012 年にオランダ人の社会心理学者であるディーデルク・スターペル（Diederik Stapel）による大規模な不正が発覚して以降，オランダでは再現性に慎重になっているといえる．ユトレヒトでは修士論文を含む全ての論文はデータとコード，そしてその目録をフォルダにまとめて大学に提出し，毎年ランダムにそれらのファイルを選んで再現するという取り組みが行われていた．こうした再現を行うのは研究者ではなくマネジメントスタッフであるため，今回のコーネルにおける取り組みのように，彼らに対するコードやデータに関するトレーニングは必須であったのだろうと思われる．

Ⅵ　データ寄託の行動分析

最後に，アーカイブへのデータの寄託行動に関する分析を紹介する．データアーカイブにとってデータの寄託数を増やすことが 1 つの重要な懸念事項と言えるが，なぜ研究者はデータを寄託するのか（しないのか）という問いに対し，GESIS の研究者が計量分析を行っていた（Akdeniz & Perry, 2019）．Ajzen (1991)の計画的行動理論をもとにし，寄託行動への態度，主観的な規範の認知，そして知覚された行動の統制可能性の三つの観点から寄託意図への効果を分析している．発表者達は 2012 年から 2014 年に社会科学のジャーナル（10 の社会学，そしていくつかの政治学のジャーナル）から出版された論文の著者にオンラインサーベイを送付し，459 人からの回答を得た．寄託行動への態度は 3 つの変数，規範認知は 2 つの変数，そして統制可能性は 1 つの変数によって測定されており，それぞれが寄託意図に与える効果を分析している．

寄託行動への態度は 1）他の研究者が新たな発見を行う可能性への態度，2）自分の研究の評判が上がることへの態度，そして 3）データの誤解釈の 3 つで測定されている．自分の研究への評判は寄託意図へ有意な正の関連，それ以外の 2 つは有意な負の関連をもっていた．次に主観的な規範として，データの寄託への期待とそれに応えること，そしてデータ寄託をするのがより一般的であるという考え方のそれぞれが寄託意図と正の関連があった．最後に，統制可能性を反映する，適切なデータリポジトリが利用可能だという尺度が寄託意図と正の関連をもっていることが明らかとなった．これらの結果に加えて発表者達は追加分析を行い，自分の研究やデータが認知される必要性という新たな変数が完全に規範を媒介していることを示し，データ認知の必要性に訴えること，そして必要性が低い場合は，インセンティブを与えることが重要であると議論していた．

研究者がデータを共有することはある意味で規範的な行動とも言えるが，他の研究者による新たな発見や誤解釈といったリスクも同様に認知し，寄託意図を妨げている．他方で自分の研究に対する評判や認知が向上することは寄託への誘引として働いているといえる．そのため．また外的なインセンティブも重要であり，この結果に従うと，例えば SSJDA の寄託者表彰などはデータの寄託を促進するといえる．

Ⅶ. まとめと今後への指針

本稿では IASSIST において紹介されていた取り組みをいくつか紹介した．リサーチャー

パスポートのように早急に取り入れることが難しいものも多いかもしれないが，例えばキュレーションに関するネットワークなどは大学間・アーカイブ間で取り組みをしても良いかもしれない．特に現段階ではアーカイブに関する共有の構想は盛り上がっているものの，その他の業務に関する共有は比較的下火だと思われる．また再現性に関する取り組みも，研究者に対するトレーニングに加えて，アーカイブが再現性確保のために協力することも将来的には構想することも可能だろう．

画像やゲームのプレイデータなど，従来データアーカイブが想定していた量的データとは異なる形式のデータも保存が可能であるということは，質的研究に大きく依拠している日本の社会学のアーカイブを考える上で大きな収穫であったと思われる．今後は量的データに拘らない，新たなデータアーカイブの形も模索することが求められるだろう．例えば本号の髙橋による報告は，日本における質的データアーカイブの可能性についてまとめられている．こうした海外の非量的データアーカイブについての知見を参考にした，非量的データアーカイブを構築することにより，将来的に社会学のデータの保全に貢献することができるだろう．

注

1) IASSIST 年次大会での発表資料は会議記録よりオンラインで入手が可能である．
https://iassistdata.org/conferences （2019 年 12 月 26 日取得）
なお，本大会における報告は（名字, 2019）で表記している．

参考文献

Ajzen, Icek, 1991, The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, *50*(2), 179-211.

朝岡誠・髙橋かおり，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.

Freese, Jeremy, and Peterson, David, 2017, Replication in social science. *Annual Review of Sociology*, 43, 147-165.

廣松毅，2019，「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』について」『ESTRELA』308:2-7.

Open Science Collaboration, 2015, Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716.

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.

Summary

The Trend in Foreign Data Archives 3
：From the Presentations at the IASSIST Annual Conference

Akira Igarashi
Kaori Takahashi

Social research data is a public good that contributes to the development of future societies, and thus we need to facilitate an environment in which people can use data appropriately. This paper reports on cutting-edge research and projects presented at the International Association for Social Science Information Service and Technology (IASSIST) held in May 2019. Based on these presentations, we suggest future directions for the Center for Statistics and Information (CSI) and Rikkyo University Data Archive (RUDA).

Key words: data archives, data usage, data sharing