《資料》

海外データアーカイブの動向 5
―IASSIST 年次大会の報告から―

髙橋 かおり
五十嵐 彰

【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 2021 年 5 月にオンラインで開催された IASSIST 年次大会で報告された内容をもとに海外で行われている統計教育や実践の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

社会調査部会は例年 the International Association for Social Science Information Service and Technology（IASSIST）が開催する年次大会に参加し，社会科学に関する情報技術・データサービスに関する情報収集をすることを業務としていた（朝岡・髙橋 2019，五十嵐・髙橋 2020）．昨年度は試験的に American Statistical Association が主催する Joint Statistical Meetings（JSM）に参加し，とりわけアメリカにおける調査技術や統計手法や，アーカイブのデータ活用に関する動向調査を行った（五十嵐・髙橋 2021）．今年度は例年通り IASSIST への出張参加を予定していたが，新型コロナウイルス感染拡大の影響でオンライン開催となり，出張は行わずに各自オンライン参加を行った．開催時期は 2021 年 5 月 17 日～20 日であり，セッションはリアルタイム配信され，事後に録画と資料が公開された．オンライン開催においては Whova というプラットフォームが用意され，リアルタイムで発表の配信が閲覧できただけではなく，セッション後の質問や参加者同士との交流ができるスペース，参加登録者へのメッセージが送信できる機能もあった．

これまでは同時間帯に複数のセッションが行われていたが，IASSIST 2021 においては同時間帯には 1 つのセッションのみであったかわりに開催日数が 1 日延びた．2020 年の IASSIST 年次大会が中止になったこともあり，報告の多くはコロナ禍以前からの取り組みに焦点を当てたものであった．セッションテーマについては，アーカイブの運営や教育，複数機関での連携，検証可能性など，例年と同様の傾向であった．ただし，一部の報告においてはコロナ禍におけるアーカイブの変化や，デジタル化への取り組みも紹介されていた．次回大会以降コロナ禍における変化を踏まえた報告がより増えるものと推測できる．

本稿では，データマネジメント活用の拡大，検証可能性，公開性と匿名性の問題のトピックを取りあげ，報告や動向を紹介する．

Ⅱ　データマネジメント活用の拡大

IASSIST の年次大会では例年，社会科学のデータにとどまらず，人文科学におけるデータ活用も事例として紹介されている．以下では，デジタルヒューマニティーズなど，人文学においても統計的手法や分析が広がる中で，社会科学にとどまらない統計活用の活用例の報告を取りあげる．

Luo（2021）ではボストンカレッジにおける取り組みとして，歴史学者やデータライブラリアンなど複数の専門家が共同して行った歴史学専攻の学生に対するデータ活用カリキュラムが紹介された．このプロジェクトでは 18 世紀の交易や人の移動に関する統計資料をデジタルデータ化することにより，紙媒体でしか残っていない歴史資料をコーディングし（収集），数値化したうえで（操作と分析），図示をするための材料を作成した．その結果一つの資料を丹念に読み込む従来の解釈学的分析では見えてこなかった通時的な変化が明らかになり，地域や集団ごとの類型化にもつながった．さらに複数の資料群において統一されたコードを付与することで，資料を管理するプラットフォームを作成し，歴史学における研究や探究の素材を共有・継承できる形式にしていた．

このプロジェクトの背景には IASSIST で標榜されるデータライフサイクルの考え方があり（例えば Corti et al.，2014: 17-23），人文学においてこの考え方が共有されることでデータアーカイブが扱う資料の範囲や幅も広がることが予想される．この際，データライブラリアンが人文学者とコードブック作成の際に議論や相談をすることでフォーマットを定めることが，のちのデータ保管・管理において重要な要素となろう．

Ⅲ　検証可能性

今回の IASSIST では，検証可能性（Verifiability）について 4 つのセッションが立てられていた．検証可能性は 2 年前の IASSIST でも取り上げられ，当時も紀要にまとめたが，今年の IASSIST ではより進んだ議論が展開されていた（Hettne et al., 2021; Thompson & Christian, 2021; Sawchuk & Khair, 2021．また本稿では取り上げないが，Kellam & Kowalski, 2021）．そのためデータアーカイブと検証可能性について，改めてここでまとめておく．

そもそも検証可能性とは，研究において用いられたデータと同じデータを，同じ分析手法を用いて，研究で報告されている結果と同じ結果が得られることを指す．以下に Freese & Peterson（2017）が作成した四象限を示すが，このうち左上に位置するのが本稿で取り上げる検証可能性である．検証可能性は再現性という訳語でも議論されることがあり，必ずしも一定していない（打越・三輪 2018）．また，例えば心理学における「再現性の危機」として紹介される論文（Open Science Collaboration 2015）では，提唱された理論を，新しいデータを使い同じ手法で検証している．すなわち Freese & Peterson（2017）の言でいう repeatability（反復可能性）の問題として扱われている．IASSIST にて発表のあった Hettne ら（2021）では，図 1 の verification を reproducible，repeatability を replicable としている．こちらのほうが日本語と整合的かもしれない．

	同じ手法	異なる手法
同じデータ	Verifiability（検証可能性）	Robustness（ロバストネス）
新しいデータ	Repeatability（反復可能性）	Generalization（一般化可能性）

図表 1　再現可能性の四象限（Freese & Peterson (2017)をもとに筆者加筆修正）

IASSIST では，政治学の American Journal of Political Science（AJPS）誌における検証可能性ポリシーに関する発表があった（Thompson & Christian, 2021）．AJPS では，アクセプトされた論文の分析に用いられたデータとコードを第三者が再現できるよう提供するポリシーがある．発表者が所属している，University of North Carolina at Chapel Hill の Odum Institute では，独立した第三機関として，AJPS に提出されるデータとコードを確認している．

Odum Institute が担う役割は主に 2 つ，キュレーションと再現である．キュレーションは，パッケージ（ここではデータとコード）が完全であるかの確認，守秘義務や著作権にまつわる問題の特定，不完全な変数やラベルの特定，そしてパッケージのフォーマットが長期に渡る保存に耐えうるかの評価，である．次に再現だが，これは報告された結果を再現するために必要なコマンドやコメントがコードに含まれているかの確認，コードのコンパイルと実行，実行できないコードの検証，そして報告された結果と再現された結果との比較である．

再現を試みた際の結果が報告されていた．2021 年 5 月までに，Odum Institute が扱った AJPS に採択された論文は 340 以上だが，うち一回目の再現性チェックで結果が再現できた論文はわずか 11 件であった．その後再提出を求めるのだが，最終的に結果を再現するのに要したラウンド数は平均して 2.29 回であった．すなわちほとんどの論文が，提出された形では一度で再現できなかった．

ここで，報告者は 2017 年から 2019 年に提出された原稿 105 件について，検証可能性を妨げるものは何かを検証した．エラーのタイプは文章作成，コード，ファイル，テクノロジー，データ，モデリング，結果の 7 つに大別されており，例えば文章作成の問題として，変数に関する記述やファイルの説明に関する問題が指摘されている．コードの問題として，ファイルパスの問題やコードの欠落などが報告されている．テクノロジーに関する問題では，分析環境の違い（ソフトウェアやパッケージなど）やエンコーディングの違いからくるものが報告されている．さらに報告では，研究者とライブラリアンの間に生じる摩擦の原因として，研究者はデータやコードをあくまで研究の過程とみなしているのに対し，ライブラリアンはそれ自体を最終的なプロダクトとみなしているという違いが指摘されていた．

検証可能性の問題はアメリカのみならず，他国においても議論されている．IASSIST でみられたもう一つの報告はオランダのライデン大学のライブラリアンによるものだった（Hettne et al., 2021）．発表では研究者がコードの公開を躊躇する理由について紹介しており，最も多い理由が，共有のためのコードを作成するのに時間がかかりすぎる，というものであった．次に多かったのがソフトウェアやシステムに依存したコードであること，そして共有のためのコードを用意する自身の能力への懸念，そして知的財産の保持の必要であった．

こうした懸念は AJPS の検証可能性に関する諸問題の裏返しともいえるだろう．研究者はコードなどをあくまで研究の過程としてみなしているために，第三者と共有するためのコード作成や自身の能力に懸念を覚える．システム依存のコードなども Odum Institute が報告していた問題の原因となるものである．

こうした諸問題に対し，ライブラリアンにできることは多いだろう．例えばライデン大学のライブラリアンは，研究者を対象にしたレクチャーやワークショップを行っている．内容は検証可能性のためのパッケージを用意する方法や研究者の能力についてであり，パッケージ共有のための懸念を取り払うことに成功している．さらにワークショップのみならず，Odum Institute のような機関がキュレーションの過程で知的財産権についての確認を行うことにより，研究者の負担を減らすことができる．

さらに Sawchuk & Khair（2021）は分析環境を含んだ議論に踏み込んでいた．Sawchuk & Khair（2021）は検証可能性をスペクトラムとして理解するよう提唱する論文を引用し，データやコードを共有することは検証することとイコールではないと議論している．検証可能性のスペクトラムは，高い検証可能性（コードとデータ，分析環境に関する情報），中程度の検証可能性（コードとデータのみ，分析環境に関する情報はなし），低い検証可能性（アルゴリズムと結果のみを提示）を指す．こうしたスペクトラムを背景に，パッケージの共有だけでは不十分であり，第三者がパッケージを開くことができるか，そしてそれを理解し，走らせ，再利用できるかなどまでライブラリアンは考えなければならないと論じている．このためには，ファイルやファイルに関する記述を十分に整備し，第三者のスキルレベルを考慮し，幅広い分析環境に対応し，さらにソフトウェアのライセンスなどにも気を配らなければならないとしている．無論現在の研究機関でここまで求めている場所はないと思われるが，海外の機関がこうした方向性にあること，そしてそれは一カ国だけでなく複数カ国にまたがったトレンドであることは抑えておく必要があるだろう．

日本では打越・三輪（2018）が指摘するように，未だに再現可能性（検証可能性）への関心が低い．検証可能性に対する関心の低さは，データの悪用を生む可能性がある．例えば日本は論文の撤回件数が突出して多い．撤回の背景にはデータの捏造があるが，これはデータの不透明さを許してしまう構造的な理由からきている可能性も指摘できるだろう．医学などの分野に多い撤回ではあるが，同様の構造が社会科学にもないとは言い切れないだろう．

今回提示した例は政治学の例であるが，今後はアメリカの社会学にも検証可能性に関する議論は及び，将来的にはアクセプトされた論文のパッケージを提出することが求められるようになるかもしれない．その余波は日本の計量研究にも及び，日本においても検証可能性に対する関心が高まる可能性があるだろう．その際に，RUDA が担うことのできる一つの役割として，キュレーションや検証を選択肢に入れることにより，データアーカイブの独自性を示すこともできるだろう．

Ⅳ　公開性と匿名性の問題

データ共有においては情報公開と回答者保護の両立は常に課題となり，本テーマは 2018 年大会の参加報告においても取りあげた（朝岡・髙橋 2019）．今回はより実務に関わる報告

が 3 つあり，基準は統一できても判断については各調査の性質によることが共通していた．いずれの報告も基本的なルールやマニュアルはありつつ，それを最終的にどう判断するのかという担当者の基準が共有されていた．

Sullivan & Thompson（2021）ではカナダ国内の事例をもとに，不完全で整っていない調査結果やデータセットの管理について，その救済方法とそのリスクについて実例を紹介していた．データセットには直接に個人を識別しうるものだけではなく，他のデータと組み合わせることで特定がされてしまう情報がある．例えば，地域情報は個人特定につながりやすい．薬物使用に関するデータにおいては，その際，k-匿名性という基準（同じ属性の回答者が k 人未満いるという条件）を用いて，各階級に人数が 5 人未満になる場合は地域情報を消すことによって対応をしていた．あるいは水質調査における回答者においても，居住地点によってエスニックマイノリティであることが明らかになることから，公開データにおいては居住地情報を除いている．

RUDA のクリーニングにおいて，公開に際して地点データを削除する場合もあることはマニュアルで共有されている．しかし，この際の基準は現在明確ではなく，あくまで寄託者との了解によって基準が都度決められている．例えば地点情報が重要な変数になる場合は削除することは難しい．最終的に寄託者との合意があることについては共通しているが，Sullivan . & Thompson （2021）の報告のように k-匿名性の基準を使うなどの工夫を取り入れたりすることも検討してもよいかもしれない．

データ共有の考え方については研究者間での理解もさまざまであるだけではなく，学内倫理委員会（Institutional Review Board, IRB）においてもその理解の程度はさまざまである．アメリカの 17 の大学の IRB への聞き取り調査（Kirilova, Kapiszewski & Elman, 2021）からは，IRB が必ずしもデータ共有に配慮した対応や判断をしているわけではないことが明らかになった．IRB は調査の実施についての倫理性を判断するものの，その後の共有についての判断をするとは限らない．さらに，データの共有については IRB から働きかけるものではなく，研究者間の伝達や共有事項として，個人に任せられる．この点について IRB の担当者たちからは統一基準や形式の一律化を望む声も上がり，データの形式や匿名化の有無，公開保留期間に関しての基準が特に求められていた．発表者らは部局間の緊張関係を踏まえたうえで，データ共有については IRB だけではなくアーカイブや情報系の部局，法律の専門家も交えて議論すべきだと提言していた．

一般に日本の学内倫理委員会は医学や生物としてのヒトに関する倫理規定を基準としており，社会生活や経歴など人を対象とした社会科学・人文科学の研究を前提にしておらず適応が難しい場合がある．しかしデジタル化によって社会科学者でも入手できるデータの質や量は増大している．社会科学の調査に適した IRB の基準を整えることも課題であるが，その際は調査の実施のみならず保存も視野に入れたガイドラインの策定が望まれる．

あるいは EU では GDPR（一般データ保護規則）の制定に伴い，個人情報の保護が厳格化していることを示した報告もあった（Valaranta, 2021）．例示されたフィンランドのデータアーカイブ（FDA）は量的調査のみならず質的調査も扱っているが，いずれにしても匿名化においての原則は変わらないという．その 5 つの要素は①対象者とサンプリングの方法，②データの内容の判断（犯罪や病歴，マイノリティなどの情報は特に守らなければならない），③データセットが作成されてからの期間（データの年齢），④ほかの資料から得られる

情報との関係，⑤使いやすさと匿名性のバランス，である．いずれにしてもデータによって匿名化処理のプランは違い，基準やガイドはあっても決定をするのは調査者やアーカイブの担当者なのである．さらにこの基準や判断はそのデータをどの程度容易に入手できるのかによっても流動的である．本報告ではその際の訓練用の資料なども掲示しつつ，原則を理解したうえで個々人が判断できるようなガイドラインも公開されている．

質的データアーカイブの可能性については髙橋（2020）でも議論したが，FDA の取り組みにおいては質的であろうと量的であろうと匿名性の基準や原則に大きな違いがないことがわかる．

V まとめと今後への指針

本稿では IASSIST の報告をいくつか紹介した．IASSIST においては例年から大きなテーマの変動はなかったが，2019 年大会より議論や手法，実践が深まっていることは確かであった．多くの報告が単一の調査や研究の結果のみならず，その先の発展的連携を提言しており，今後の CSI や RUDA の業務や研究における示唆となった．例えばデータ共有が進む中での課題となる回答者の保護や倫理的な問題を議論するためには，異なる専門分野の研究者や異なる部局の担当者との対話が可能になるプラットフォームの構築が求められる．あるいは検証やキュレーションに関する議論や支援をするためには，アーカイブごとの独自性を保ちつつ，国内外のデータアーカイブ間での情報共有が必要になろう．大学での複数の部局との連携や大学間の連携，あるいは国家規模でのプロジェクトの場合もあり，日本で実現するためには CSI や立教大学によるボトムアップの事業のみならず，より広い枠組みでの協力や連携，トップダウンの事業内での連携への参加も進めていく必要があろう．

本文中言及の IASSIST2021 での報告

Hettne, K., Proppert, R., Nab, L., Saunero, P. R., & Gawehns, D. 2021, ReprohackNL 2019: Enhancing research reproducibility at Dutch Universities.

Kellam, L., Block, B, & Kowalski, B.,2021, Crafting a Sustainable Reproducibility Service and Archive.

Kirilova D., Kapiszewski, D.& Elman, C., 2021, Optimizing Openness in Human Subjects Research: Balancing Transparency and Protection.

Luo, J., 2021, Building Data Literacy Suite in the Humanities: A Hands-on Approach.

Sullivan, C., & Thompson, Y., 2021, Mathematics, Risk, and Messy Survey Data.

Sawchuk, S., & Khair, S., 2021, Computational reproducibility: A simplified framework for data curators.

Thompson, C., & Christian, T., 2021, Computational reproducibility: Examining verification errors and frictions.

Valaranta, A., 2021, How to guide anonymisation?

参考文献

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』5:33-40.

Corti, L, V. den Veerle, E., Bishop, L., & Woollard, M., 2014, *Managing and Sharing Research Data:A Guide to Good Practice*, second Edition. London: Sage Publications. Ltd.

Freese, J., & Peterson, D. 2017 "Replication in Social Science," *Annual Review of Sociology*, 43(1), 147-165.

五十嵐彰・髙橋かおり，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.

五十嵐彰・髙橋かおり，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.

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

髙橋かおり，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.

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』 66.

Summary

Trends in Foreign Data Archives 5
: From the Presentations at the Annual IASSIST Conference

Kaori Takahashi
Akira Igarashi

Social research data is a public good that contributes to the future development of society, 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 online in May 2021. Based on these presentations, we propose future directions for the Center for Statistics and Information (CSI) and Rikkyo University Data Archive (RUDA).

Key words: data archives, data usage, data sharing