《資料》

海外データアーカイブの動向 6
―IASSIST 年次大会の報告から―

池田 岳大
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 2022 年 6 月に対面，オンラインの併用で開催された IASSIST 年次大会での報告内容をもとに，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 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行っている（朝岡・髙橋 2019，五十嵐・髙橋 2020）．昨年度大会は，新型コロナウイルス感染拡大の影響により，全面オンライン開催となりオンライン参加であった（髙橋・五十嵐 2022）．今年度は対面とオンラインの併用での大会開催となったが，出張は行わず，オンラインでの参加となった．なおスウェーデンのイェーテボリで大会が開催され，報告者もオンライン参加が可能であったが，報告者に関しては現地報告を行う人が多数派であったように見受けられた．開催時期は 2022 年 6 月 7 日から 10 日までであり，Whova というプラットフォームを用いてリアルタイムで発表の配信ならびに閲覧ができ，参加者同士のチャット機能や資料の共有などが可能であった．報告内容はアーカイブとして後日閲覧が可能であった [1)]．

2022 年大会のテーマは Data by Design: Building a Sustainable Data Culture であり，データデザインとその持続可能性が強調された大会であった．各セッションでは，データポリシーとデータ管理，チームビルディングに関する取り組み，データの可視化に向けた教育モデュールの作成や，マイノリティや弱者（少数民族や青少年）なども含めた個人情報保護とデータ利用の拡大を両立させる仕組みの議論，またデジタルフットプリントデータの活用や，CESSDA の学際的な取り組みと今後の展望などについても議論がなされていた．

本稿では，データキュレーション，定量データクリーニングや文書化にかかる時間の予測因子の検証，データの活用と保護のバランスに関する報告や動向について紹介する．

Ⅱ 論文出版前のデータキュレーションや再現性に関する議論

American Journal of Political Science（以下，AJPS）は，2015 年より論文出版前に著

者にデータ，コード，資料等を提出させて，第 3 者が論文中に示された分析結果を再現できるような方法を提供することを義務づける政策を採用している（Verification Policy という [2]）．量的分析は，ノースカロライナ大学の Odum Institute for Research in Social Science（以下，Odum），質的分析はシラキュース大学の Qualitative Data Repository が検証作業を行っている．

そのうち Odum は，キュレーションと再現性の確認を行っている．キュレーションとしては，レプリケーションパッケージの完全性の確認，信頼性や著作権の問題の確認，変数ラベルや値ラベルの不完全さ，不一致さ，誤りの確認，記述的メタデータの強化，長期保存に適したファイル形式の評価を行っている．再現性の確認については，報告された結果を再現するために必要なコマンドとコメントが含まれているか，コードのレビュー，コードのコンパイルと実行，実行不可能なコードにあるエラーの特定，原稿中の表，図，その他報告された結果とアウトプットの比較等を行っている．

論文掲載の条件は，検証可能な資料をすべて提出し，その内容がすべて検証されることにある．分析の際のコマンドなども原則，公開されることとなっている．実証分析をほとんど含まない論文は上記の限りではなく，個人情報を保護する場合も検証結果は非公開となりうる（編集者の承認が必要）．

これにより，論文出版までの労力が増え，出版までの時間がかかることが想定されるが，この政策がこうしたコストに見合うものであるのかどうかの検証も同時に進めており，その有効性や問題点を洗い出す取り組みがなされている．

そのために，Odum では政策以降，パッケージの特性や欠陥，計算エラー，制限付きデータの使用，検証時間，再投稿数など，各投稿原稿に関するデータを収集しており，これらのデータを用いて，再現性チェックにかかるコストの分析が可能となる．AJPS の 409 本の論文のうち，再現性チェックを 1 回で通過するのは 14 本にすぎない．第一段階でのチェックにかかる作業時間は平均で 6 時間程度であった．

報告では，再現性のチェックに関する検証可能性の強化や作業時間の短縮化のために，再現性のチェックの際によくするエラーの検証，さらには検証にかかる時間の規定要因やどういったエラーが検証回数を少なくする要因であるかといった点に焦点が当てられた．そこで 2007 年～2019 年までの AJPS の原稿を定性的にコーディングし，エラーコードや検証にかかった時間を分析した．用いるデータは，エラータイプの分類コード，用いたコードの数や言語，コードの量，著者の情報などがある．

この研究では昨年度の報告から続き，まずは再現性チェックの際にみられたエラーのタイプを質的にコーディングしている．分析の結果，23 種類のエラータイプに分類することができ，そこから大きく以下の 7 タイプにまとめられる．

・Documentation：コードブックの変数とデータの整合性
・Coding：コードそれ自体のエラーやコードの実行エラー
・Files：ファイルの破損等でコードが実行されない，データが確認できない
・Technologies：ファイルのエンコードエラー，特殊な計算環境で実行したことによって再現不可能となる
・Data：外的データへのアクセスエラー，外的データの引用エラー，ライセンスや使用許可エラー

・Modeling：GIS などの使用方法エラー，分析モデルの設定エラー
・Results：原稿やアペンディクスのエラー，出力結果と原稿との齟齬

本報告で示された結果は上記まででであったが，このコーディングを用いて今後は Time-to-Event モデルの検討，つまり具体的に上記のようなイベントが生じたか否かとイベントが生じた際の所要時間やその後，検証に合格したか否かとその時間の分析などを実行する予定であるという．

Ⅲ　データクリーニング・文書化の所要時間の予測因子

本稿では，データクリーニング作業とデータドキュメンテーションの作業の所要時間の予測因子に関する検証結果に関する報告について概観する．先行研究（Treloar & Klump 2019）は，データ管理タスクはプライベート，グループ，持続（persistent），パブリックの 4 領域に分けるモデルを提示しており，うち最初の 2 つは研究グループ内，後の 2 つは研究グループ外に永続的かつ公共性を保持した形でデータを共有する方法を検討することである．研究グループ外へのデータの移行に焦点を当てると，他の研究者がデータを再利用するためには，データをクリーニングや文書化に対するコストを測定することの重要性が指摘できる．

先行研究を検討する中で，学習効果（追跡調査における繰り返し質問が登場することによる慣れ，あるソフトやツールを使用することによる慣れなど），変数の数やケース数（ただし，回答情報が増えるとその組み合わせの分，回答者の再識別を行う必要があるため，偽名化やクリーニングなどの所要時間は変数数やケース数だけ単調増加するわけではない），機密情報のチェックなどが所要時間を増加させる主たる要因であると考えられる．

2016 年 12 月から 2017 年 9 月にかけて GESIS の中の Data Archive for the Social Sciences がデータキュレーションにおけるコスト要因に関する検討プロジェクトを開始し，そこで得られた 3 つのパネル調査データ（データ 1 は 2010, 2012, 2014, 2016，データ 2 は 2007, 2009, 2011, 2013，データ 3 は 2014, 2016 のもの）から，キュレーションにかかる所要時間について検討を行った．

主要な分析結果としては次の 2 点である．第 1 に，データサイズとデータに含まれる個人情報の数が時間を増加させる要因となっている．第 2 に，変数の数が増えると，クリーニングの文書化にかかる時間が増加している．また，自由記述回答がある場合，機密情報の確認作業があるために，より時間を増加させている．一方で，wave が連続した調査データの場合には，クリーニング作業が効率化し，学習効果が高まることが分かった．

Ⅳ　データの活用と保護のバランス

EU における個人情報保護の高まり（GDPR の発効・適用）とオープンデータ化の高まりの中で，公開できる情報とデータを保護することの両方のバランスが各機関には求められている．今日，自らのデータや論文をウェブ上に一般公開することは技術的に容易になったが，データアーカイブに寄託することの利点は，検索や活用の可能性の拡大とデータの

保護・管理を担保することの 2 点にある．さらにこの公開・活用と保護のバランスは国や地域による差がある．各種統計や社会調査におけるイギリス国内の地域差と，それらのデータの統合可能性については，Emma Gordon による基調講演（Administrative Data Research UK: The journey so far）でも論点となっていた．データの形式や管理の仕方が異なることは検索や活用における利便性を下げるが，データの合成を行うことで個人情報保護が損なわれる可能性があることは，日本国内のミクロデータ活用においても共通する問題である．

検索や活用の可能性については①検索基盤の充実，②寄託データの広報・宣伝を，データの保護・管理については③RDM（Research Data Management）教育の必要性と④寄託プロセスの改善についてそれぞれ傾向を紹介する．

1. 検索基盤の充実

各アーカイブからの報告では検索の簡易化に関する事例が紹介されていた．検索の簡易化の工夫は，キーワードやテーマの見やすさ（ビジュアライズ）と横断検索の導入に大別できる．例えば Roper Center の取り組みでは，変数レベルでの質問のカテゴライズをプロジェクトチームで行い，その結果をもとに各質問項目をビジュアライズして質問同士の関係を可視化できるようにしていた（Joyce & Weldon 2022）．変数レベルでのメタデータ化においては単に検索可能なテキストデータやメタデータにするだけではなく，その検索の利便性を担保するためにデータアーカイブ側で編集する取り組みを行っていた [3)]．

2. 寄託データの広報・宣伝

調査実施者に対して寄託への動機づけをすることはどのデータアーカイブにおいても課題である．ここでは，各データアーカイブを対象に 2019 年秋に行われた調査をもとに，データアーカイブに求められる 8 つの対応についての仮説が検証された報告を見ていこう（Hayslett & Jansen 2022）．寄託者にとってはスタッフや追加でのキュレーションのサービスの充実が促進要因となると同時に，ソーシャルメディア活用への要望が寄託への動機付けになっていることが明らかになった．いずれにしても，自らが寄託したデータがほかの研究者へ利活用されたり，宣伝されたりすることは，データアーカイブだからこそ果たせる役割である．そのため，調査者の予想以上に宣伝や広報に関しての研究者からの期待は高い．寄託する研究者や研究グループが個別にはできない作業やサービスを提供できることが，アーカイブへの寄託を促す作用になっていよう．

3. RDM 教育の必要性

RDM については，既存のカリキュラムや教育の中で対応するのは難しい．ヨーク大学の教員に対して行われた RDM の教育に関する理想と現実に関する調査においては，専門教育とは別に RDM 教育を行う担当者の設置の必要性が説かれていた（Savard & Wang 2022）．教員側からすれば，教育においては専門教育の他に手が回らないので，内容（content）ではなく形式（format）を教育する専門職は別に必要だという論理である．確かに日本の大学においても検索に関するセミナーやワークショップ，論文の書き方に関する共通教育は浸透してきている．また，社会調査士カリキュラムの制度化とともに社会（科）学でも調

査の方法論に関する教育や講義は充実しつつある．しかし，調査において集めたデータをどう管理するのか，特に院生に向けた RDM やデータライフサイクルの必要性など，今後は調査データや資料の管理についてのセミナーや講座も設置する可能性はあるかもしれない．

4. 寄託プロセスの改善

GDPR の発効と適用に際し，各データアーカイブではそれまでの寄託プロセスや手順の見直しが行われている．4 か国のデータアーカイブの事例研究においては，GDPR の適用以降規約や寄託後の同意（特に個人情報保護に関する見直しや権利関係の確認）について変更する必要性が生じ，専門性を持った職員の養成や確保が追いついておらず，アーカイブ側も寄託者側も負担が増えている（Tyler 2022）．例えば，ノルウェーの Sikt（Norwegian Agency for Shared Services in Education and Research）のように，ウェブ上の指示に従えば自分の寄託したいデータにどの程度の保護が必要かわかるようにする仕組みなども，他のアーカイブで導入がなされていくだろう [4]（Kvamme 2022）．

Ⅴ　まとめと今後への指針

本稿では IASSIST2022 の報告に関していくつか紹介した．例年に引き続き，データ利用と保護，データの再現性など社会調査の発展に関する議論が多角的になされていた印象である．国，大学，研究所レベルでデータアーカイブ業務に携わる研究者，技術者の取り組みやそこでの問題点について報告があった。報告の中から見出されてきたデータアーカイブの今後の役割として，単にデータの寄託の推進と管理にとどまらず，データキュレーションやデータの再現性に関するポリシーの作成や実務に関するタスクを担うなど，多様化している印象を受けた．

さらに，集めたデータをどう管理・活用するのか，という点については，データアーカイブ側への調査でも，あるいは大学教員に対する調査でも，データアーキビストやキュレーターなど専門職の充実と教育の必要性が各種調査から説かれていたことは興味深い．日本では専門職であることすら認知されにくいデータアーカイブの業務について，IASSIST においてはその重要性を共有するところまでは議論が進んでいる．しかしいずれの調査においても「どのように専門職者を養成し，どのような場で活躍してもらうのか」という実務面での取り組みはまだ検討や紹介がされていない．今後は，長期的な人材育成を含めた調査研究の報告も展開されることを期待したい．

注

1) 基調講演やセッションの一部は後日 YouTube でも公開され，報告資料は Zenodo で公開されている．（2022 年 12 月 25 日現在）
https://www.youtube.com/playlist?list=PLD9Y_M_A24iSmUFsNkyevdZSSqF2AKjSA
https://zenodo.org/communities/iassist-2022/search?page=1&size=20&sort=conference_session

2) AJPS Verification Policy は次の URL から閲覧可能（2022 年 12 月 10 日現在）
https://onl.bz/GsXAME1

3) 参考　https://ropercenter.cornell.edu/health-poll-database-project（2022 年 12 月 10

日現在）

4) 参考 https://sikt.no/ （2022 年 12 月 10 日現在）

本文中言及の IASSIST2022 での報告

Cheryl Thompson & Thu-Mai Christian, 2022, Modeling costs of computational reproducibility and data verification in political science(UNC Odum Institute for Research in Social Science United States).

Anja Perry & Sebastian Netscher, 2022, Measuring the time spent on data curation (GESIS - Leibniz Institute for the Social Sciences Germany).

Kathleen Joyce & Weldon Weldon, 2022, Facilitating database search and analysis through the Roper Center's node network interface (The Roper Center for Public Opinion Research)

Michele Hayslett & Matt Jansen, 2022, Factors affecting deposits in data repositories (UNC at Chapel Hill Libraries; Matt Jansen, UNC at Chapel Hill Libraries)

Dany Savard & Minglu Wang, 2022, Faculty researcher perspectives on RDM and the pedagogical needs of graduate students (York University)

Allison R. B. Tyler, 2022, Can we still archive data? A comparative case study of social science data archives under the GDPR (University of Michigan School of Information)

Trond Kvamme, 2022, NSD DMP – towards a FAIR ecosystem for data management planning (NSD - Norwegian Centre for Research)

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.

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.

髙橋かおり・五十嵐彰，2022,「海外データアーカイブの動向 5――IASSIST 年次大会の報告から」『社会と統計』8: 27-33.

Treloar, A. & Klump, J., 2019, "Updating the data curation continuum," *International Journal of Digital Curation*, 14(1): 87-101.

Summary

Trends in Foreign Data Archives 6
: From the Presentations at the Annual IASSIST Conference

Takehiro Ikeda
Kaori Takahashi

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