Special Edition

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 ―[注1)]

伊藤 伸介 | Ito Shinsuke

㈱日本学術振興会 人文学・社会科学データインフラストラクチャー構築推進センタープログラムオフィサー（研究員）／中央大学経済学部教授

■九州大学大学院経済学府博士後期課程単位修得退学。博士（経済学）。2014 年中央大学経済学部准教授。17 年 4 月より同教授。18 年 11 月より、㈱日本学術振興会人文学・社会科学データインフラストラクチャー構築推進センタープログラムオフィサー（研究員）。

前田 幸男 | Maeda Yukio

㈱日本学術振興会 人文学・社会科学データインフラストラクチャー構築推進センタープログラムオフィサー（研究員）／東京大学大学院情報学環教授

1. はじめに

独立行政法人日本学術振興会（以下、「学振」と略称）は、2018 年 10 月から「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」を展開してきた。本事業の課題の一つは、人文学・社会科学の分野横断的なデータカタログの整備にある。こうしたカタログの整備を行うにあたって、海外のデータアーカイブ施設において進展してきたデータの保存と共有の現状とメタデータ（データに付随する属性情報を表すデータ）の作成状況を把握することは有益と考える。

そこで、本稿は、海外におけるデータの保存と共有およびメタデータの整備状況を紹介した上で、学振で現在進めているデータカタログの現状について議論していきたい。

2. データの保存と共有に関する現状と課題

1960 年代以降、社会学や政治学の分野においては、主として、社会調査データに関する二次分析（secondary analysis）を可能にするために調査データの共有が行われてきた。例えば、アメリカにおける調査データの共有については、1962 年にミシガン大学に設立された ICPSR（= Inter-university Consortium for Political and Social Research）に遡ることができる。また、イギリスでは、1967 年にエセックス大学に UKDA（= U.K. Data Archive）が設置され、調査データの保存と共有が進展した。ドイツにおいては、現在 の GESIS（ = Leibniz-Institute for the Social Sciences）の前身にあたる Zentralarchiv が 1960

注 1）本稿は、2019 年度統計関連学会連合大会（2019 年 9 月 9 日、於：滋賀大学）における筆者の共同報告に基づいている。

年に活動を開始している。

一例を挙げると、欧米諸国のデータアーカイブ施設の中で、イギリス最大の人文学・社会科学に関するデジタルデータ（digital data）のアーカイブ施設であるUKDAは、社会調査の個票データや公的統計のサーベイデータ等の様々なデータの収集・保存を行っており、データ提供サービス機関であるUKDS（＝U.K. Data Service）を通じて、7,000以上のデータセットの提供が行われている。イギリスでは、1970年代初期に実証的な社会科学研究が進展したものの、データの共有に関する文化的環境の未成熟さ、データの寄託に関する基準の厳しさ等もあり、サーベイデータの収集は困難な状況にあった。さらに、1980年代になると、社会科学研究に対する政府予算が削減されたが、そのことが、イギリスにおける調査データの二次分析の促進とデータの保存と共有のさらなる受容をもたらした。その後、データの保存・共有の機能を拡充させることによって、UKDAは現在ヨーロッパにおける主要なアーカイブ施設となっている（UK Data Archive（2007），伊藤（2011），伊藤（2016））。

ヨーロッパには、UKDAの他にも、ドイツのGESIS、オランダのDANS（＝Data Archiving and Networked Services）、フィンランドのFSD（＝Finnish Social Science Data Archive）、ノルウェーのNSD（＝Norwegian Centre for Research Data）等、様々なデータアーカイブ施設が存在し、一定の方針に基づいて社会調査データを中心に、収集・保存・提供を行っている。これらのアーカイブ施設は、ヨーロッパにおけるデータアーカイブ施設のネットワークであるCESSDA（＝Consortium of European Social Science Data Archives）に加盟している[注2)]。

ところで、社会調査データは、どうして保存・共有されるべきなのであろうか。佐藤・石田・池田（2000）や前田（2019a）によれば、社会調査データを保存･共有する理由として、①社会調査のデータには資料的な価値があること、②大規模な標本調査の実施に当たり多くの費用がかかること、③学術的な社会調査における回答者の時間的な負担が大きいことが指摘されている。こうしたことから、諸外国において社会調査データをデータアーカイブ施設によって保存・共有する体制が進められてきた。

わが国でも、東京大学のSSJデータアーカイブ（＝Social Science Japan Data Archive）が創設されており、社会調査やアンケート調査の個票データの提供が1998年4月に開始された。しかしながら、わが国においては、研究者が保有する社会調査データがデータアーカイブ施設に寄託・収集され、データを保存・共有する体制が十分に展開されてきたとは言いがたい。こうしたデータの保存と共有を阻む要因としては、以下の5点を指摘することができる（前田（2019a））。

⑴ 研究者文化の問題

研究者はデータへの愛着（所有意識）が強いことから、外部の研究者と調査・研究データを共有したがらない傾向にある。したがって、自分だけがデータを正しく解釈できるという思い込みがあると思われる。

注2）CESSDAの加盟国は、以下の20か国である。
オーストリア、ベルギー、クロアチア、チェコ、デンマーク、フィンランド、フランス、ドイツ、ギリシャ、ハンガリー、オランダ、北マケドニア、ノルウェー、ポルトガル、セルビア、スロバキア、スロベニア、スウェーデン、スイス、イギリス。

⑵　データの所有権・著作権の考え方

データの所有権･著作権といった調査･研究データの保有に関する権利関係についての理解が確定していないということが考えられる。そのため、どのような手続きでデータを共有できるのかについての認識が十分でない可能性がある。さらに、物故者がいる場合、データの保存・共有に関する手続きはより煩雑になる。

⑶　データを保存・共有する誘因の欠如

第三者がデータを利用するためには､調査設計、調査実施日等、調査研究の過程について詳細な文書を残すことが必要である。しかしながら､調査･研究データを理解する上で必要なメタデータが適切に残されていないことが考えられる。これらのメタデータの保存が自身の研究に対するメリットをもたらさないと研究者が考えれば、研究者が調査・研究データを保存・共有する誘因は発生しないだろう。

⑷　データの保管に関する技術的な問題

調査・研究データを保管している磁気媒体に劣化や破損があった場合、それは、データの保存が技術的に困難になる。さらにデータが古いソフトウェアで保管されていると、現在使用可能なソフトウェアで､古いソフトウェアの形式に従ったデータを読むことができないという問題点もある。

⑸　社会調査データの匿名化をめぐる問題

社会調査データにおける匿名化の基準は必ずしも明確ではない[注3)]。さらに、情報技術の発達や利用できるデータの増加に伴い、要求される匿名化の方法も変化すると思われる[注4)]。したがって、社会調査データの匿名化の基準をどのように考えるかは、社会調査に含まれる個人情報の取り扱いも含め、検討する必要があるだろう[注5)]。

ところで、保存されたデータを効果的に共有するためには何が必要であろうか。利用者が必要なデータを発見しやすく、そして、入手したデータを利用しやすくするためには、メタデータを充実させる必要がある。メタデータが整備されると、広範な学術研究のデータを一括して検索することが可能になる。それによって、研究に必要なデータの探索および比較研究を行うこともできる。それは、データを共有するための誘因になると言える。なお、メタデータとしては、例えば、以下のような項目が該当する。

・研究代表者
・調査実施機関
・母集団とサンプリングの方法
・調査実施時期
・調査の詳細（回収率）
・調査員へのインストラクション
・調査票 / 回答票
・自由回答の分類（コード）
・理想的にはすべての変数について度数分布表

注 3）わが国の公的統計のミクロデータの場合、2018 年 6 月に公布された改正統計法の全面施行後に、「匿名データの作成・提供に関するガイドライン」（総務省政策統括官（統計基準担当）決定）が改正された（2019 年 6 月 27 日改正）。なお､ガイドラインには、統計作成部局が統計調査の匿名データを作成するにあたっての参考となる匿名化の基準について記載がなされている。
注 4）公的統計データにおける匿名化の方法の現状については、例えば伊藤（2018）を参照。
注 5）社会調査の調査個票データが個人情報に該当しうると判断される場合、本人の同意なく第三者提供を行おうとすれば、個人情報保護法の条文に明記されている匿名加工情報として提供することが第三者提供の要件となり、それゆえに、同法施行規則第 19 条第 1 号～第 5 号の条文を満たすことが求められる可能性がある。これについては､個人情報保護法における調査データの位置づけ、さらには施行規則に基づく匿名加工の手続きの可能性など、検討すべき課題は少なくないと思われる。

3. 海外におけるデータカタログの整備の現状

海外のデータアーカイブ施設では、保存されている社会調査データや研究データ、さらには公的統計ミクロデータをウェブ上において一括で検索することが可能なデータカタログを整備している。データカタログのサイトでキーワードを入力することによって、利用者が比較研究を行う上で必要なデータの一覧表を表示することができる。データアーカイブ施設でこうしたデータカタログを作成するためには、メタデータのエレメント（項目）を設定する必要がある。

表 1　ICPSR のデータカタログに含まれるメタデータのエレメント

1. 調査のタイトル（Title, Alternate Title）
2. 調査番号（Study Number）
3. 主要な調査者（Principal Investigator）
4. 資金提供機関（Funding）
5. 調査データに関する文献資料の引用（Bibliographic Citation）
6. 調査データのシリーズに関する情報（Series Information）
7. 調査の概要（Summary）
8. キーワード（Subject Terms）
9. 調査対象地域（Geographic Coverage）
10. データが提供される期間（Time Period）
11. 調査時点（Date of Collection）
12. 調査単位（Unit of Observation）
13. 調査対象（Universe）
14. データの類型（Data Type）
15. 標本抽出方法（Sampling）
16. 乗率（Weights）
17. データの収集方法（Mode of Collection）
18. 回答率（Response Rates）
19. 加工の程度（Extent of Processing）
20. データアクセスの制限状況（Restrictions）
21. データのヴァージョンの履歴（Version History）

出典）http://www.ddialliance.org/training/getting-started/data-catalog

表 1 は、ICPSR のデータカタログに含まれるメタデータのエレメントを示したものである。表 1 を見ると、調査のタイトル（Title, Alternate Title）、調査番号（Study Number）、主要な調査者（Principal Investigator）、調査の概要（Summary）、キーワード（Subject terms）、調査時点（Date of Collection）、調査単位（Unit of Observation）、調査対象（Universe）といった項目が列挙されている。こうした ICPSR で提供されているメタデータのエレメントは、データアーカイブ施設で提供すべきメタデータを検討する上での素材の一つとなりうるものである。

同様に、表 2 は、CESSDA、UKDS、GESIS、DANS において提供しているデータカタログの項目の一覧表を示したものである。これらの施設におけるデータカタログを比較すると、タイトル、調査番号ないしは ID、作成者（調査者）、キーワード（検索項目）といった共通のエレメントが確認できる。それに対して、例えば GESIS のデータカタログには、「時間軸上の比較可能性（Comparability over time）」や「国家間の比較可能性（Comparability between Countries）」等、他のデータアーカイブ施設のカタログには含まれない項目も存在する。

近年、調査データを保存するデータアーカイブ施設を横断して検索できるカタログを各国が整備している。例えば、フィンランドの Estin は人文学・自然科学を含む 35 分野の研究データを一括検索することができる。また、スウェーデンの SND のデータカタログは、自らが所蔵する研究データだけでなく、他機関が所蔵する研究データの検索も可能にしている。さらに、ドイツの GESIS は、58 機関のメタデータを一括して検索するデータカタログのベータ版を試行的に運営し

表２　主要なデータアーカイブ施設におけるデータカタログの項目のリスト

CESSDA	UKDS	GESIS	DANS
Study title	Title:	Title	Persistent identifier
Creator	Study number (SN):	Abstract	Title
Study Persistent Identifier	Persistent identifier:	Data Version	Creator
Abstract	Series:	Keywords	Contributor
Country	Principal investigator(s):	Processing	Date created (ISO 8601)
Time dimension	Depositor:	Geographical Coverage	Description
Analysis unit	Sponsor(s):	Organizaiton	Audience
Sampling procedure	Grant number	Universe	Subject
Data collection method	First edition release	Sampling	Temporal coverage
Data collection period	Latest edition release	Data Collection	Spatial coverage
Language of data files	Topics	Anonymization	Identifier
Publisher	Abstract	Legal Basis	Relation
Year of publication	Dates of fieldwork:	Weighting	Type (DCMI resource type)
Terms of data access	Country:	Data Access	Format
Study number	Spatial units:	Access Conditions	Language (ISO 639)
Topics	Observation units:	Access Form	Language
Keywords	Observation unit location:	Access Contact	Source
	Population:	Data Service	Remarks
	Number of units:	Comparability	Access rights
	Method of data collection:	Comparability over time	Date submitted
	Time dimensions:	Comparability between Countries	
	Sampling procedures:	Note	
	Kind of data:	References	
	Weighting:		

注）CESSDA、UKDS、GESIS と DANS におけるデータカタログをもとに作成した。

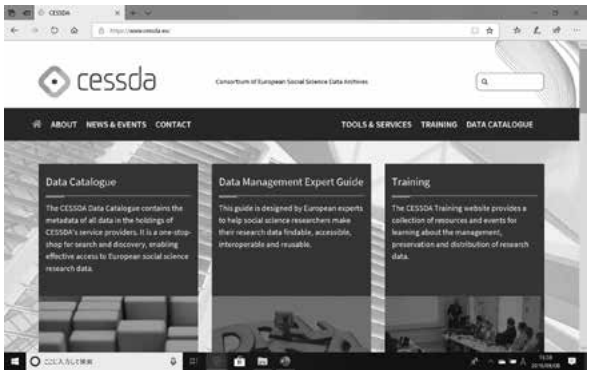
ている（前田（2019b））。

つぎに、CESSDA を例に、横断検索可能なデータカタログのイメージについて説明したい。図 1 は、CESSDA の HP におけるトップページを示している。画面左側にある「Data Catalogue」をクリックすると、検索画面が出てくる。そこで、例えば「labour force」と入力してみると、labour force が含まれる社会調査データや公的統計データの一覧表が出てくる（図 2 - 1）。これは、CESSDA に加盟しているデータアーカイブ施設において個票データが提供されている社会調査、さらにはメタデータが利用可能な社会調査データや公的統計データに関する一覧を示している。その中で、例えば、「Labour Force Sample Survey 1982, 2nd quarter」をクリックすると、Labour Force Sample Survey 1982, 2nd quarter という調査データに関するメタデータを閲覧することができる（図 2 - 2）。具体的には、表 2 の CESSDA のデータカタログのリストの中に含まれる調査のタイトル、作成者、調査の ID 等を図 2 - 2 において確認することができる。

このように CESSDA においては、アーカイブ施設間の横断的な検索が可能になっている。こうした組織間の横断検索においては、メタデータの形式が共通である（あるいは互換的である）と同時に、OAI-PMH（＝ Open Archives Initiative Protocol for Metadata Harvesting）を通じて、メタデータの自動的な収集（ハーベスト）がなされることが重要である。

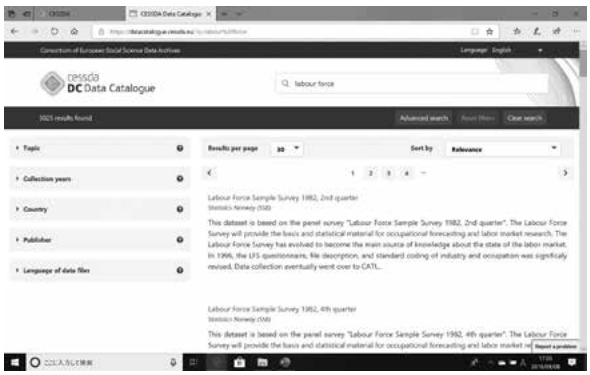
図書館のカタログとしての OPAC では、書誌情報が一定の形式に従って記述されることで、効果的な検索が可能になっている。同様に、海外のデータアーカイブ施設では、それぞれのルールにしたがってメタデータの整備を行っている。その

図 1　CESSDA のトップページ



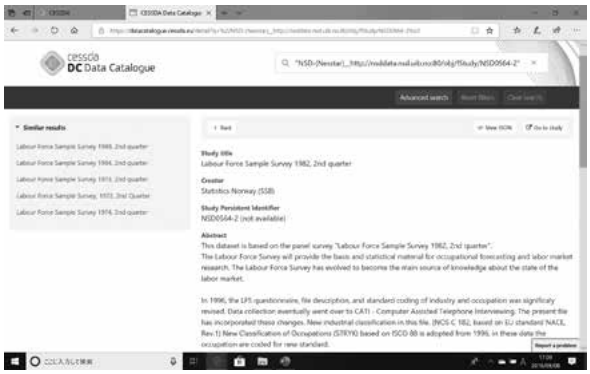
出典）https://www.cessda.eu/

図 2-1　CESSDA における検索結果 ― labour force と入力



出典）https://datacatalogue.cessda.eu/?q=labour%20force

図 2-2　CESSDA における検索結果 ― Labour Force Sample Survey 1982, 2nd quarter を選択



出典）https://datacatalogue.cessda.eu/detail?q=%22NSD-(Nesstar)__http://nsddata.nsd.uib.no:80/obj/fStudy/NSD0564-2%22

ため、調査・研究データの効率的な検索を可能にするために、データ記述に関する国際的な基準である DDI（= Data Documentation Initiative）が採用されていることが少なくない。また、UKDA や FSD 等、多くのデータアーカイブ施設で、メタデータの記述に必要な専門用語とその定義を含む DDI の統制語彙（Controlled Vocabulary）[注6] が採用されており、利用者にとって利便性の高いデータ検索を可能にするデータカタログの開発が進められてきた。このように、メタデータのエレメントを適切に記述するための専門用語の定義を定めることも、メタデータの整備においては求められる。

4. 学振における総合データカタログの整備状況

本稿は、海外のデータアーカイブ機関におけるデータカタログの整備状況を概括した。そこで、最後に学振で現在進めているデータカタログの現状について論じることにしたい。

学振では、現在、社会科学に関する調査データや公的統計データを対象に、四つの拠点機関が収集し、提供してい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進めている。それは、具体的には、拠点機関である東京大学社会科学研究所附属社会調

注 6）統制語彙については、例えば、つぎのような例で説明することができる。「隔年の調査」「2 年に 1 回の調査」「1 年おきの調査」は人間にとっては同じ意味で捉えられるが、機械にとっては三つの異なる言葉と認識される。ゆえに、機械が判別可能なように、例えば、「2 年に 1 回の調査」と定義することが求められる。

査・データアーカイブ研究センターが保存・共有している社会調査の個票データや大阪商業大学JGSS研究センターが提供している日本版総合的社会調査（Japanese General Social Surveys＝JGSS）、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・解析センターが実施し、提供を進めているパネルデータ、および一橋大学経済研究所が保有している公的統計の歴史的な統計データを対象に、一括して検索可能な総合データカタログを整備することによって、実証研究を行う上での利便性を高めることを指向している。そこで、学振では、以下のようなデータカタログの整備計画を進めている。

現在、各拠点機関から提供を受ける調査データや公的統計データに関するメタデータのハーベストを可能にするために、総合データカタログのシステムの開発を国立情報学研究所と共同で行っている。その前提として、社会調査メタデータの国際規格であるDDIに準じたメタデータを拠点機関から収集することを目指している。そのために、中核機関である学振は、ICPSR等のデータカタログを参考に、20程度のメタデータのエレメントを持つような形で、学振としてのメタデータの設計を進めている[注7)]。さらに、総合データカタログについては、データを保存する拠点機関と直接的に連動させることで、データの検索から取得までの流れを円滑に行うことができるようにしていきたいと考えている。

その一方で、将来的に様々な分野の研究データにおけるデータカタログへの展開を想定すると、DDI以外のメタデータスキーマにも拡張可能なようなメタデータの設計も考慮する必要があろう。また、国立情報学研究所が提供する検索サービスCiNiiとのデータ連携も図っていくことも考えられる。さらに、総合データカタログの持続的な運用を見据えると、オープンソースソフトウェアで構成することが必須であろう。拠点機関以外からのメタデータの登録も可能なデータカタログの設計についても、学振の人文学・社会科学におけるデータインフラ事業としては将来的な検討課題となるだろう。

＊参考文献

[1] 伊藤伸介（2011）「わが国におけるミクロデータの新たな展開可能性について―イギリスにおける地域分析用ミクロデータを例に―」（明海大学『経済学論集』Vol.23, No.3, 36～54頁）.
[2] 伊藤伸介（2016）「わが国における政府統計のデータシェアリングの現状と課題」（『情報管理』Vol.58, No.11, 836～843頁）.
[3] 伊藤伸介（2018）「公的統計ミクロデータの利活用における匿名化措置のあり方について」（『日本統計学会誌』第47巻第2号, 77～101頁）.
[4] 佐藤博樹・石田浩・池田謙一（2000）『社会調査の公開データ：2次分析への招待』東京大学出版会.
[5] 前田幸男（2019a）「社会科学データを共有する制度基盤」（『中央調査報』No.740, 1～5頁）.
[6] 前田幸男（2019b）「社会科学データを共有する制度基盤(2)」（『中央調査報』No.741, 1～5頁）.
[7] UK Data Archive（2007）*Across The Decades: 40 Years of Data Archiving*.
[8] CESSDAのデータカタログに関するURL https://www.icpsr.umich.edu/icpsrweb/ICPSR/【2019年9月27日アクセス】.
[9] UKDSのデータカタログに関するURL https://beta.ukdataservice.ac.uk/datacatalogue/studies/#!?Search=&Page=1&Rows=10&Sort=0&DateFrom=440&DateTo=2019【2019年9月27日アクセス】.
[10] GESISのデータカタログに関するURL https://dbk.gesis.org/dbksearch/home.asp【2019年9月27日アクセス】.
[11] DANSのデータカタログに関するURL https://dans.knaw.nl/en/researchers/search【2019年9月27日アクセス】.

注7）メタデータのエレメントの設計においては、ダブリン・コア（Dublin Core）と呼ばれるメタデータの標準規格で設定されている項目については全て対応する予定である。