Special Edition

オープンサイエンスの取り組みとしての人文学・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推進事業

朝岡 誠 | Asaoka Makoto

国立情報学研究所オープンサイエンス基盤研究センター特任助教

■東北大学大学院文学研究科博士課程単位取得退学。立教大学社会情報教育研究センター助教を経て、2019 年 4 月より現職。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携わる。主な業績は「ワンステップ内で伝わる評判の効果」（『理論と方法』49、2011 年）「評判によって信頼が生成されるか」（『社会学研究』86、2009 年）等。asaoka@nii.ac.jp

林 正治 | Hayashi Masaharu

国立情報学研究所オープンサイエンス基盤研究センター特任助教

■北陸先端科学技術大学院大学知識科学研究科博士後期課程修了、博士（知識科学）。一橋大学情報化統括本部情報基盤センター助教を経て、2016 年 12 月より現職。NII 研究データ基盤のひとつ WEKO3 の開発に携わる。mhaya@nii.ac.jp

1. はじめに

近年における情報通信技術（ICT）の発達は人々の生活、行政、ビジネスのあり方に大きな変化をもたらしている。これは学術研究についても例外ではなく、ICT の発達は研究のスタイルに大きな変化を生みつつある。インターネットを通じて研究成果や研究データを流通させることにより、学術分野や国境をこえて研究者の研究成果に容易にアクセス可能な環境が整いつつある。これは学術分野をこえた共同研究やデジタル人文学のような新しい研究を生み出し、行政、ビジネス、市民活動に携わる人々と研究者の連携を深め、新しい科学技術やイノベーションの発展を促進する源泉となっている。

オープンサイエンスともよばれるこの変化は2013 年の G8 科学技術大臣会合の共同声明で研究成果や研究データのオープン化が合意されたことを契機に、世界的な潮流として広がっている。日本でも「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」の報告書 [1] や第 5 期科学技術基本計画 [2] において、国をあげてオープンサイエンスを推進する体制を構築していくことが明言されている。

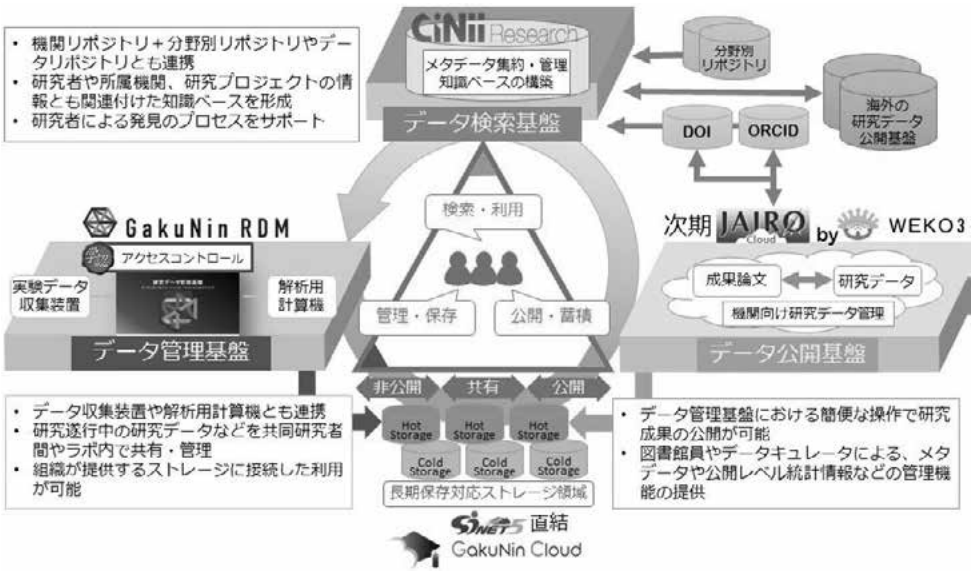
本稿で紹介する「人文学・社会科学データインフラストラクチャー構築推進事業」（以降、データインフラ事業）はこのオープンサイエンスに関わる事業であり、本稿では日本の人文学・社会科学分野におけるオープンサイエンスへの取り組みの一例として本事業の取り組みを紹介する。

2. データインフラ事業とオープンサイエンス

現在、日本学術振興会（学振）が実施しているデータインフラ事業は、人文学・社会科学に係るデータを分野や国をこえて共有・利活用する総合的な基盤の構築を目的とした事業である。この事業の背景には、2015 年の「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」の報告書 [1] で示されているように、我が国の人文学・社会科学に係る研究データ共有が進まないことによる、人文学・社会科学研究における日本離れや、世界的な研究活動の効率化から取り残されることによる国際競争力の低下に対する危機感がある。

欧米諸国をはじめとする諸外国はオープンサイエンスを国家戦略と位置づけ、研究データの公開について、大きく分けて二つの政策を実施している。一つは研究データの原則公開義務化であり、もう一つは研究支援のための共通情報基盤の整備である。アメリカの科学技術政策局（OSTP）は 2013 年に「公開助成研究成果オープンアクセス司令」[3] を発令し、年間 1 億ドル以上の研究開発費を有する研究助成機関に対して「論文と科学データへのアクセス拡大計画」の策定を指示した。これを受けて国立衛生研究所（NIH）などの研究助成機関によるパブリックアクセスプランの策定が進んだ。我が国でも、研究データは原則公開する方向で進んでいる。2016 年の第 5 期科学技術基本計画 [2] で公的資金による研究データについては各研究分野で扱う研究データの特殊性を考慮し、国益等を意識したオープン・アンド・クローズ戦略を各分野の研究コミュニティで検討することが必要であるという方針が打ち出されている。この方針を踏まえて日本医療開発研究機構（AMED）や海洋開発研究機構（JAMSTEC）などの研究機構は、機関における研究データについてのデータポリシーを定め、データ公開についての方針を定め始めた。

図 1　研究データ基盤の概要（https://rcos.nii.ac.jp/service/ より引用）



また、オープンサイエンスの一環として、研究データを管理し、共有するための情報基盤を整備しているプラットフォームの整備も進んでいる。EU では、2015 年より、European Open Science Cloud というプロジェクトがスタートし、クラウド環境やオンラインストレージを研究者向けに提供し、研究データ共有のためのシステムを構築している。我が国でも国立情報学研究所でオープンサイエンスのためのインフラ、研究データ基盤の開発が進められている（図 1）。

国立情報学研究所（NII）が開発を進める研究データ基盤は、それぞれ独立したデータ管理基盤（GakuNin RDM）、データ公開基盤（WEKO3）、データ検索基盤（CiNii Research）の三つのシステムで構成される。研究活動の研究ライフサイクルの段階に応じてそれぞれの基盤が利用されることを想定している。

データ管理基盤は、研究プロジェクト実施中に収集されたデータを管理し、研究グループ内での研究データ管理・共有のためのシステムである。研究プロジェクト終了後、研究者が公開すると判断したデータや関連の資料は公開基盤を用いて公開する。そして、公開基盤で公開された研究データや他のレポジトリで公開されたメタデータ情報は検索基盤にて集約・管理され、利用者は CiNii Research にて横断的に研究データを検索することができ、検索したリンク先より公開基盤にアクセスできる。これらの研究データ基盤は 2020 年度の運用開始を予定している。

ここまでは、国策としてのオープンサイエンスの取り組みについて紹介したが、オープンサイエンスの推進にはトップダウン方式による決定だけではなく、研究者コミュニティからのボトムアップ方式による決定も必要であり、双方がお互いを補完しながら連携する必要がある [4]。社会科学では International Federation of Data Organizations（IFDO）という社会科学データアーカイブの国際組織や The International Association for Social Science Information Services and Technology（IASSIST）といった社会科学に関する研究・教育にかかわる情報技術専門家の国際組織が存在し、社会科学における研究データ共有やそのデータ利用の促進についての議論が行われており、後述する Data Documentation Initiative（DDI）という社会科学のためのメタデータ規格もこれらの議論の積み重ねのもと策定されている。データインフラ事業は先で紹介した弊所の研究データ基盤を使って、研究データを公開する試みであり、人文学・社会科学のコミュニティによるオープンサイエンス推進事業ととらえることができる。したがって、国際的な動向をつかみつつ、人文学・社会科学研究コミュニティの要望をもとにこの事業を推進していく必要がある。

3. データインフラ事業の活動内容

本事業の活動は学振が主体となって行うデータ利活用システムの構築事業と学振が公募を通じて拠点となった大学組織が行うデータ共有基盤の構築事業の二つに分けられる（表 1）。

学振の役割は大きく分けて二つある。一つは研究データの共有を促進するために、拠点機関の公開しているデータを潜在的なデータ利用者が検索するためのデータカタログの整備と拠点機関が公開するデータを分析するシステムの構築であり、もう一つは人文学・社会科学分野における研究データ共有のルールを整備し、データを所有する研究者や研究機関が安心してデー

表１　データインフラ事業の活動内容（2019年６月30日現在）

	中核機関	拠点機関
主な事業	データ利活用システムの構築	データ共有基盤の構築
具体的な事業	1. データ公開、利用、権利関係等の共通ガイドラインの策定 2. 分野横断的なデータカタログを整備 3. オンライン分析システムの開発研究 4. 公開シンポジウムの開催やニュースレターの配信等を推進	a）データ・アーカイブ機能の強化（共有化） b）海外発信・連携機能の強化（国際化） c）データ間の時系列等、接続関係の整備（連結化）
参加機関	日本学術振興会 国立情報学研究所（2., 3. の事業を受託）	東京大学 一橋大学 慶應義塾大学 大阪商業大学

タを提供できる環境を構築することである。このうち、NIIは2. 横断的なデータカタログの開発と3. オンライン分析システムの開発について受託を受けており、横断的なデータカタログの開発は、データ公開基盤をベースにシステム開発が進められている。

次に拠点機関の活動内容についてみてみよう。学振は人文学・社会科学について、作成、管理、共有、提供又はそれらの支援について十分な実績のある研究機関を公募で募り、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、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、そして大阪商業大学JGSS研究センターを採択した。これらの組織は日本の社会科学において貴重なデータを作成し、そのデータを研究者に提供している機関である。現在、各拠点機関は自分たちが管理しているデータを整理し、調査概要の英語化を行っている。今後は、他の研究者や機関が所持する研究データを受け入れ、公開する体制を構築することを目指している。これらの取り組みを行うためには、拠点で管理している研究データを編集加工するだけではなく、データの内容を指し示すデータ、である「メタデータ」を意識してデータを整備する必要があるが、このメタデータの作成は学振のデータカタログの整備に大きく関わる。

4. データカタログの整備とオンライン分析システムの開発

現在、日本では本事業の拠点機関をはじめとした複数の機関がデータアーカイブを運営しており、データのダウンロード提供を行っているが、データ利用者からみれば、データを探し出すために複数のデータアーカイブにアクセスする必要がある。また分野外や海外の研究者にとってはこれらのデータアーカイブは馴染みがないため、潜在的なデータの利活用を妨げているともいえる。データカタログとは、これらの機関が管理する研究データのポータルサイトであり、日本の人文学・社会科学データを横断するデータ検索環境である。このようなポータルサイトの構築はヨーロッパの社会科学データアーカイブ協議会CESSDAが実施している。CESSDA

は Data Catalogue という検索システムを用いて加盟データアーカイブにそれぞれ所蔵されているデータ情報を共有し、利用者は用語、ないしはブラウザからトピック、キーワード、または配布アーカイブごとに調査データとその調査で用いられている変数を検索することができる。しかし、このような横断的検索システムを作成するのは容易ではなく、各データアーカイブ間でメタデータ基準の統一を行う必要がある。

データを検索するためには、そのデータのタイトル、作成者、作成年、トピック、社会調査データの場合、調査の実施方法、母集団、調査対象の選定方法などのメタデータが必要となる。例えば、「この調査は 2012 年に行われ、2018 年に公開されました」という一文からは、調査年＝2012、公開年＝2018 であることは判断できるが、検索エンジンは 2012 と 2018 という数字が何を指し示しているのかを判断することはできない。したがって検索を目的にする場合、機械が判断できるようにメタデータを作成する必要がある。たとえば、XML 形式で先程の例を記述すると、「< 調査時期 >2012</ 調査時期 >< 配布年 >2018</ 配布年 >」となり、タグによってその情報が指し示す意味を表現することができる。そして、検索エンジンはそのタグを読み取ることで、2012 と 2018 をそれぞれ調査次期、配布年と区別することができる。しかしながら、それぞれの機関が独自のタグを付けていては、機関をこえた相互検索の実現は困難である。共通のタグを用いて研究データのメタデータを定義し、機関間で共有する必要がある。このように体系化されたメタデータ定義したものが XML スキーマであり、データの相互運用を考慮した検索システムを構築する上での重要要素である。本カタログ事業では拠点機関の活動の際に挙げた研究助成機関、データを利用した研究成果、英語情報、調査地点、調査時期など、拠点機関が整備しているメタデータ情報を包含した XML スキーマを用意する必要がある。

現在、学振は拠点機関との協議のもと、CESSDA の Data Catalog や社会科学データアーカイブで国際的に利用されている DDI という XML スキーマをもとにしたメタデータ規格を設計している。DDI は社会科学、とりわけ社会調査データの記述に特化しており、現在本事業で導入を検討している DDI Codebook というヴァージョンでは 252 の要素が階層的に整備されている。学振はこれら 252 の要素を使って、17 の必須項目と七つの推奨項目を策定し、日本語と英語を併記できるようなスキーマを設計している。また、検索の利便性を向上させるために、DDI で設けている入力ルールをもとにトピックや調査方法について語彙を統一することも計画している。

今後、各拠点機関には DDI に準じたメタデータ規格でメタデータを整備していただき、これらのメタデータを各拠点のデータアーカイブから機械的に収集し、データカタログを構築する予定である。そして、将来的には拠点機関のメタデータだけではなく、他の研究機関のメタデータも収集し、このデータカタログに直接、研究データとメタデータを登録できるようにし、日本の人文学・社会科学のデータを検索できるようにすることを目指している。

次にオンライン分析システムについて紹介しよう。オンライン上で統計分析を行うシステムは海外の社会科学アーカイブでも提供しており、一部の拠点機関も主に教育利用を念頭にお

いてオンライン分析システムを提供している。このような分析環境は昨今、無料で利用できる統計ツールが容易に入手することができるようになり、各自でデータをダウンロードし、データ分析できるようになってきている。しかし、これらの分析結果やそのプログラムについては現状では各自で保存し、共有されるということはない。オープンサイエンスの議論の一つにこれらの分析結果やプログラムを共有できる状態にし、研究成果が再現できるようにしようという議論がある。研究者コミュニティにとって、分析結果が再現できることは自分たちの研究の透明性を保証し、研究不正を防止する手段となる。そのため海外では、論文の執筆者に研究データの提出を求め、さらに分析プログラムの提出を求めるジャーナルが増えてきている [5]。人文学・社会科学の分野ではまだこのような事例は少数であるが、今後のオープンサイエンスの流れによっては、研究データとプログラムを提出することがデファクト・スタンダードになる可能性もある。

本データインフラ事業のオンライン分析システムは、各拠点機関が整備したデータの中での特に重要なものを対象に、オンラインで統計分析を行い、これらの分析プログラムや結果を分析ノートとして利用者が設定したグループで共有することを目的としたシステムである。また、拠点機関によってはデータの保護のためにアクセス制限を設け、利用申請制度や利用期間がある。オンライン分析システムはこうした拠点機関の状況にも対応できるように設計している。

なお、これら二つのシステムは持続的な運用も見据え、オープンソフトウェアで構成している。

5. さいごに

本稿では、オープンサイエンスへの対応という観点からデータインフラ事業について紹介した。現在は社会科学データの取り扱いについて十分な実績がある拠点機関と協議を行っているが、その中でも個人のプライバシーに抵触しかねないデータについての問題が議論となっている。第 5 期科学基本計画でもオープンサイエンスの推進において「データのアクセスやデータの利用には、個人のプライバシー保護、財産的価値のある成果物の保護の観点から制限事項を設ける」と明記されており、オンライン分析システム上で拠点機関から提供されるデータに対してどのようなアクセス制限を設けるのかが課題となる。また将来的には、データカタログに直接研究データとメタデータを登録できる設計を行う場合のアクセス制限も課題である。

＊参考文献

[1] 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 http://www8.cao.go.jp/cstp/sonota/openscience/150330_openscience_1.pdf（2019 年 9 月 18 日）.

[2] 内閣府（2016）第 5 期科学技術基本計画 http://www8.cao.go.jp/cstp/kihonkeikaku/5honbun.pdf（2019 年 9 月 18 日）.

[3] OSTP（2013）Memorandum for the Heads of Executive Departments and Agencies, Subjects: Increasing Access to the Results of Federally Funded Scientific Research https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf（2019 年 9 月 18 日）.

[4] 大向一輝（2018）「オープンサイエンスと研究データ共有」（『心理学評論』no.64 ⑴, 13-21）.

[5] 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」（SSJDA- リサーチペーパーシリーズ , 2018, no.66, p.16）.