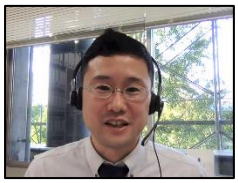
第 1 回 SPARC Japan セミナー2020
「研究データ公開:フルオープンと制限公開の境界線」

# 社会科学分野におけるデータ提供の実態 ―データアーカイブ機関と利用者の最前線―

仲 修平
（東京大学社会科学研究所）

**講演要旨**

本報告の目的は、データアーカイブ機関と利用者の関係に着目して社会科学分野におけるデータ提供の実態を共有することである。具体的には、(1) どのようなフローでデータを利用者へ提供しているのか、(2) データの提供に際していかなる検討を実施しているのかという点である。両者について、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が構築している SSJ データアーカイブ（Social Science Japan Data Archive）を事例としてみていきたい。そのことを通して、学術目的で利用するデータの公開をより一層進める一方で、利用者へ提供する際に残されている課題を考察する。



**仲 修平**

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助教。日本学術振興会特別研究員を経て2018年4月より現職。専門分野は、社会階層論、計量社会学。主な研究関心は、雇われない働き方の実態を量的／質的調査に基づいて検討すること。主著は、『岐路に立つ自営業―専門職の拡大と行方』（勁草書房、2018年）。データアーカイブ研究センターでは、調査基盤研究分野（公開データの準備や提供等に関わる業務）で活動している。

## 本報告の目的

本報告の目的は、社会調査・データアーカイブ機関の役割、個票データを提供するフローの実態、利用者の申請に対する検討の内実という 3 点を通じて、社会科学分野におけるデータの公開と制限の境界線を共有することです。

先に結論を申し上げておきますと、原則として利用者へ個票データをより広く公開していこうというスタンスで運用しています。ただし、一定の条件を満たさない場合に限って制限することがあります。重い制限をかけるのはあくまで例外的な運用であることをご理解いただきたいと思います。

## 所属機関の役割

私が所属している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は、四つの研究分野から成り立っています（図 1）。調査基盤研究分野は、データの保存・収集・公開のプロセスに携わっています。その中でも私は利用者にデータを提供するプロセスに関わっています。社会調査のデータを作っていく社会調査研究分野や、データの利用を促進するためにセミナーや研究会を開いたりする計量社会研究分野もあります。また、国際調査研究分野は、世界のデータアーカイブ機関と連携していくために、学会に行ったり、実際に人と会ってネットワークを構築したりすることを始めており、特に東アジアのデータアー

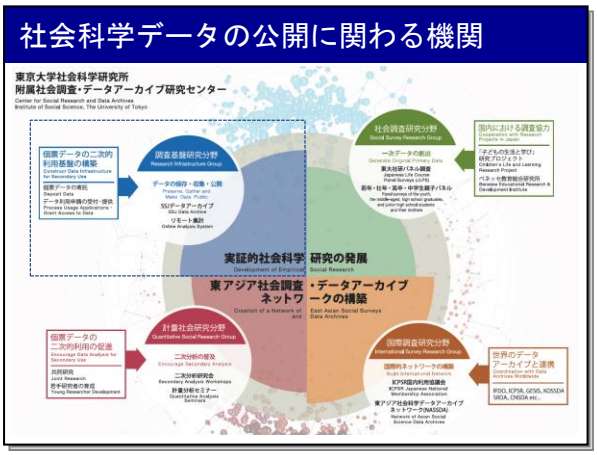
カイブの連携が進みつつあります。これが全体の状況です。

## 社会科学分野で扱う個票データ

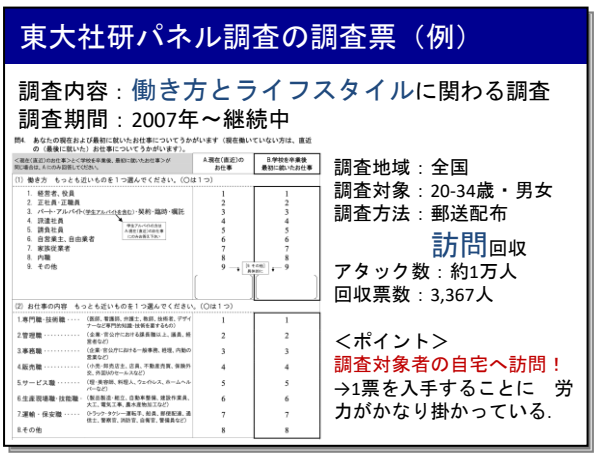
まず、社会科学分野で扱う個票データとは何なのかということを簡単に説明します。図 2 は東大社研パネル調査の調査票の一部です。パネル調査とは、同じ個人を何年も追跡していくタイプの調査です。例えば、現在の仕事について伺いますという形で、働き方や仕事内容について質問項目を立てています。それに対して、回答者は数字を選んでいきます。例えば正職員で専門職であれば、「(1) 働き方」では 2 番に、「(2) お仕事の内容」では 1 番に丸を付けます。そういったタイプのデータを作っていきます。

ここで強調しておきたいのは、全ての対象者に対して調査票を郵送配布し、訪問によって回収しているデータであるということです。東大社研パネル調査のサンプルサイズは、おおよそ 5000 人です。調査対象者の自宅に調査員が行って回収しますので、1 票のデータを得ることにかなりの労力がかかっています。ですので、データを公開するときにも、こうしていろいろな人の苦労が積み重なって集まったデータを保存していますので、データがとてもかわいく見えてきます。そういったデータを扱っています。

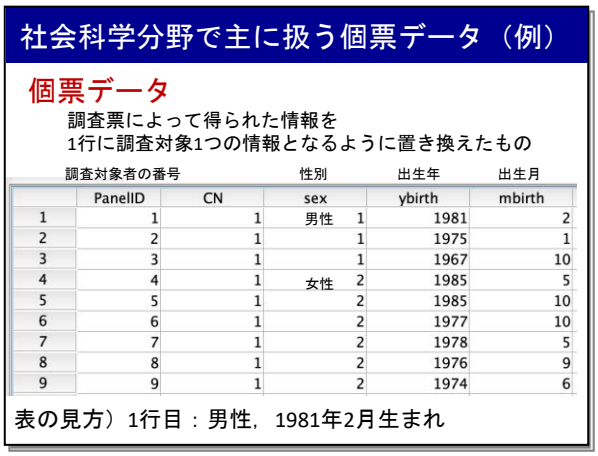
実際に個票データを配布するときには、図 3 のような電子ファイルで提供しています。表の見方としては、例えば 1 行目であれば、性別に関しては 1（男性）、出生年に関しては 1981（1981 年生まれ）だということが分かります。社会科学分野で主に扱うデータの形は、基本的にはこのように 1 行に一つの単位（個人や組織）が入っていて、列方向にそれぞれの質問項目が入っているというものです。こういった数値情報の集まりを社会調査データとして公開しています。




（図 1）




（図 2）




（図 3）

## Social Science Japan Data Archive （SSJDA）の仕組みと利用状況

では、実際にどういうデータプロセスによって公開していくのかという模式図が図 4 です。まず各主体が調査を実施して、それを SSJDA に寄託して、公開できるようにデータの処理をして利用者に使っていただくというプロセスで進めています。

図 5 は、実際に集められたデータと利用している方のおおよその数です。SSJDA は 1998 年からスタート

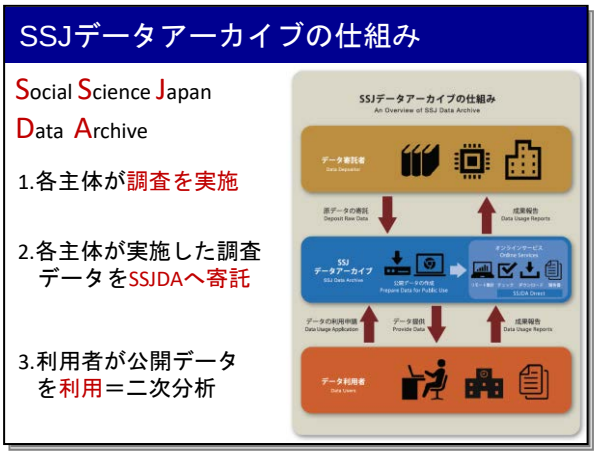
して、現在では公開データセット数や利用者が年々非常に増えていることがご覧いただけると思います。直近 3 年では約 70 件ずつ新規のデータセットを公開しています。研究者はもちろん、大学院生や、教育目的で学部生にも利用していただいています。また、日本に関心のある海外の研究者の利用も最近になって増えてきています。

図 6 は SSJDA を利用した研究業績です。主に学術著書や論文、学位論文です。2 次分析による研究成果は増加傾向にあり、若手研究者による学術書の出版、卒論、修論、博論等でご活用いただいています。図書館職員の方からも問い合わせがあります。学部生の教育目的で使っていただく機会が最近になって増えてきているという特徴があります。調査や研究に困った学生がいれば、SSJDA を紹介すると何か研究ができそうだと思っていただけると幸いです。

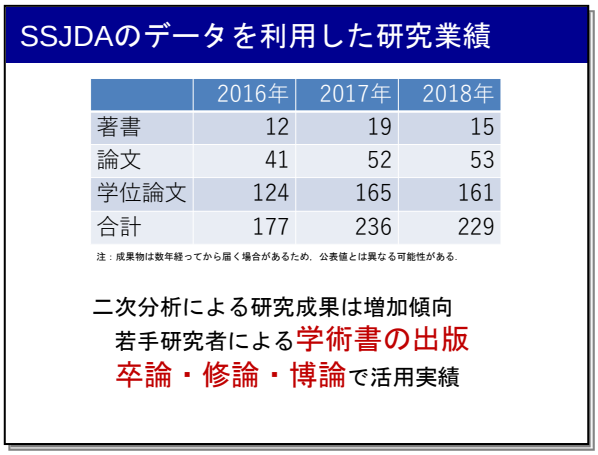
## 保有データ

SSJDA で保有しているデータは、経済学と社会学のデータが多くなっています。それ以外の分野でも、経営学や教育学、政治学など、幅広い分野を扱っているという特徴があります（図 7）。

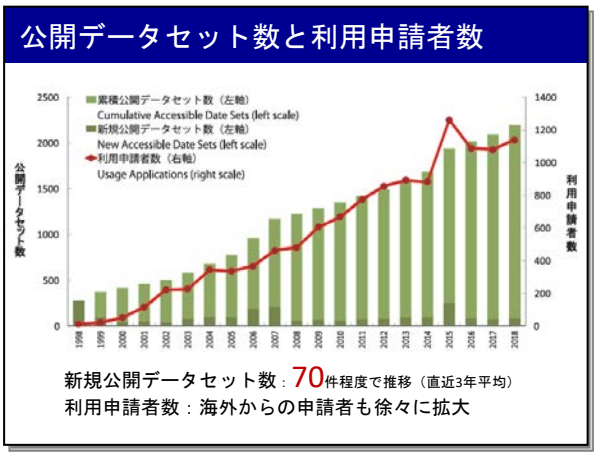
現在保有しているデータは、多くは公開中ですが、全てを公開しているわけではありません。公開に至るまでには、データのクリーニングや、ローデータに対して後からコードを付与するアフターコーディングと呼ばれるプロセスもありますので、そのような公開準備をしているもの、あるいは寄託の段階から保存目的で寄託していただいているデータもあります。保有している多くのデータは公開を前提としていますが、公開に向けていろいろなプロセスの途上にあるデータがあります。



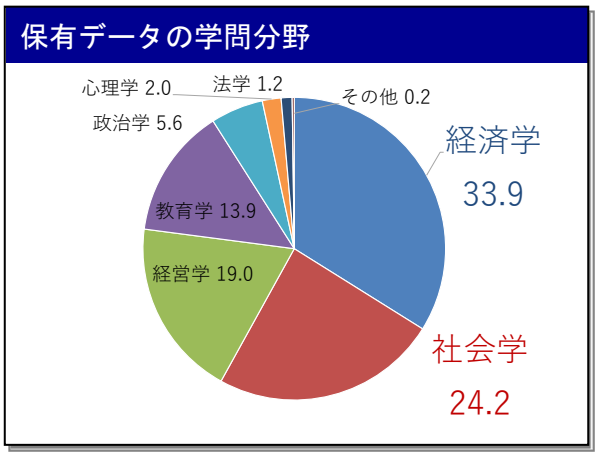

（図 4）




（図 6）




（図 5）




（図 7）

## データ提供フロー

先ほどお示ししたデータ提供の三つのプロセスの中で、特に本セミナーに関わるのが、SSJDA と利用者との間でどのようなやりとりがなされているかという点です。このプロセスにおいて、セミナーのテーマである公開と制限が行われています。データ提供フローは、図 8 のような三つのステップになっています。

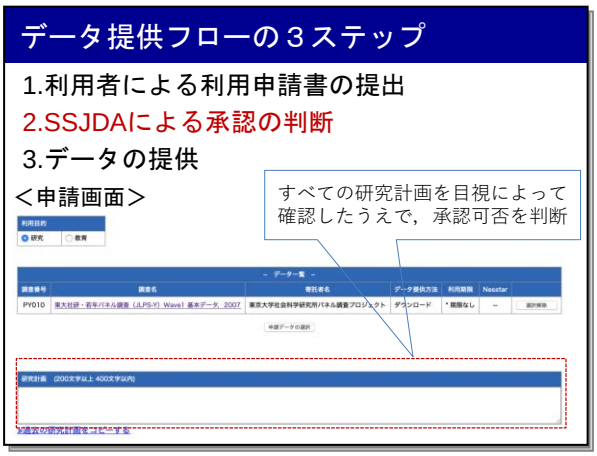
一つ目は、利用者による利用申請書の提出です。実際の申請画面では、利用の目的や、どのデータを使うかということを選択していただいた上で、研究計画という形で、なぜこのデータが必要なのか、どのように使うのかということを記述していただきます。

二つ目に、書かれた研究計画を見て、SSJDA の中で承認判断をします。今のところ、全ての研究計画を目視によって判断して、場合によっては申請者に補足的な対応を問い合わせる形で進めています。その結果、承認という判断になれば、三つ目としてデータを提供していきます。

## 提供／制限の内実

提供と制限に対する基本的なスタンスとしては、原則は、利用目的に合致した場合、データを利用者へ提供します。制限は例外的で、あくまで利用条件に満たない場合を想定しています。

利用の承認については、①利用条件に合致しているか、②申請内容が妥当であるか、③利用状況が適切であるかという三つの観点から判断しています。それぞれについて、具体的にどのような点を見ているかということをお示ししたいと思います。




（図 8）

まず、①利用条件に合致しているかという点については、二つポイントがあります。一つは利用資格、もう一つは利用状況です。SSJDA は、大学または公的研究機関の研究者、あるいは教員の指導を受けた大学院生に利用資格があります。NG になるケースで多いのは、民間企業や学部学生、高校教員などの場合です。ただ、学部学生に関しては、指導教員からの申請で受け付けています。利用状況というのは、過去の申請に対してきちんと対応しているかどうかということです。利用期間は基本的に 1 年間ですが、それに対して利用報告をしているか、また、教育目的で使った場合には受講者リストをきちんと提出しているかという点から判断しています。通常、図書館でも、借りた本を先に返してから次の本を借りると思いますが、そのように、過去の申請にきちんと対応しているかということもポイントになっています。

②申請内容の妥当性に関しては、研究計画のボックス欄に書かれた内容・目的が妥当であるかどうかという点から判断しています。研究に関しては、研究計画の内容が十分に示されているか、示された内容と調査に含まれている変数が合致しているか、また、教育に関しては、卒業論文の中身や使われる授業名などの記載が必須ですので、その有無を見ています。

③利用状況の適切性については、一度の申請で大量のデータを申請する方がごくまれにいらっしゃいます。データによってはシリーズで提供しているものもあり、そういった場合にはたくさんのデータが必要になります。しかし、そうではなくて、いろいろなデータをとにかく使ってみたいということで、同じ研究計画の目的で大量のデータを申請された場合には、それほど同時に必要ないことを確認した上で、データを絞って提供するケースもあります。また、学生が教員としてアカウントを作成する、複数名で一つのアカウントを共有する、教員が自らのアカウントを学生へ貸し出すなど、アカウントの不正利用が明らかに見られた場合に

は制限をかけます。具体的に制限を加えたケースを次に紹介したいと思います。

## 制限のケース

これまでに最も重い制限を加えたケースとしては、利用者のアカウントを停止する措置をとったことがあります。そのケースは次のようなものです。まず、教員が学生にアカウントを貸し出して、その学生がデータを申請してくる、場合によってはその学生からSSJDA の担当者に問い合わせが来るという状況になっていました。この不正利用が発覚した時点で該当の教員に対して改善を促しましたが、十分な対応がなされませんでした。そこで、私たちの機関の中で検討を重ね、合意をもってアカウントを停止することになりました。基本的にはデータをどんどん使っていただきたいのですが、使い方が良くない場合には制限が加わります。ただし、その場合でも、何度も検討を重ねて、結果としてやむなく停止しているという形です。

もう一つ制限を加えた例として、学部ゼミメンバーでアカウントを共有していたケースがあります。同じゼミに所属する人たちが一つのアカウントで利用申請を繰り返したというケースです。これについては先方の先生に改善を依頼して対処されたと伺っていますが、その対応と同時に、SSJDA の中でもシステムを再検討し、2016 年 3 月からは、学部学生の申請は指導教員の先生による申請という形に限定しています。

## まとめ

提供と制限の境界線については、データ提供の原則は「常にオープン」という形になります。ただ、制限の可否は利用条件の合致を個別に確認した上で検討しています。特に重い制限は、データに対して不誠実な利用者に対してのみ行っています。

最後に、一担当者から見た提供フローの課題を示したいと思います。個々の申請全てを目視で対応しているので、人員と時間のコストが非常に大きくなっています。ですので、今後データの公開数や利用者が増えていった場合に、申請件数に対する処理にも制約が出てくるのではないかと感じています。それに対しては、申請フローを自動化あるいは半自動化にする、あるいは、現在は申請ベースで検討しているものを利用者ベースにする検討を加えるなどの形が可能ではないかと考えています。