データアーカイブと社会科学研究

Data Archive and Social Science

佐藤 博樹

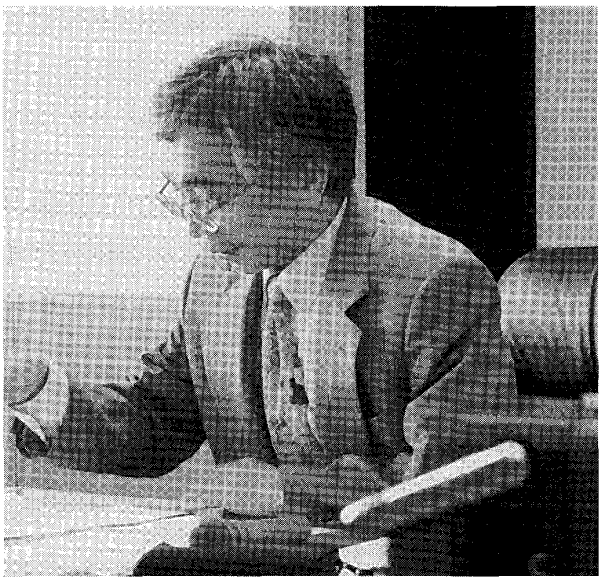
ご紹介いただきました東京大学社会科学研究所の佐藤博樹です．よろしくお願いします．

初めから，２つお詫びしなければいけません．昨日はどうしても夜まで拘束される仕事があったものですから，午前の大事なお二方の報告を聞けなかったことです．もう１つは，レジュメをつくってあったのですが，添付するのを忘れてしまいました．レジュメがないので，少し話の筋立てがわかりにくくなってしまうことがあるかと思います．

最初に，自己紹介をさせていただきたいと思います．原先生も佐藤健二さんも社会学の分野の第一線で活躍されている先生なのですが，私は一応社会学の出身だと思っているのですけれども，実際にやっている分野は，最近は人的資源管理というふうにいわれています人事管理です．社会科学研究所の教員は経済学部と法学部の大学院で教えているのです．私は経済の方の所属で，人事管理と組織行動を教えています．

ですから，今日お話することと実際の自分の研究分野とでは相当な距離があります．どちらが本業なのかといわれると困るわけでありますけれども，食べる上ではこれが本業でありまして，研究の上では人事管理が本業です．

なぜ，人事管理とか労働関係の研究をしている者がデータアーカイブを始めたのかについてはあとでお話をしたいと思います．



佐 藤 博 樹 氏

まず，データアーカイブとは何かをご説明します．基本的にはコンピュータ処理可能な磁気データを収集・保存し，その保存も単に保存しておくだけではなくて，第三者が利用可能な形で保存し，そのデータを使いたいという方に提供することが広い意味でデータアーカイブです．アーカイブというと文書資料館という意味で，歴史分野では昔からあるわけです．これに対して，データアーカイブの保存対象は，文書資料ではなく磁気データです．コンピュータ処理可能な磁気データの資料館だとご理解いただければよいと思います．

コンピュータで処理するということですから，これまでは大規模調査を集めることが多

SATO Hiroki 東京大学社会科学研究所附属日本社会研究情報センター教授

かったわけですが，最近海外では質的データのアーカイブもでき始めました．ヒアリング調査なども保存対象になっています．ヒアリングテープを起こして保存し，二次分析に提供するというものができ始めています．

では，データアーカイブというのは，社会科学研究においてどのような機能を果たすのかということであります．文書アーカイブもデータアーカイブも機能は基本的に同じです．あたりまえですが，１つは，データを保存していくということです．後世に残していくということが大事な機能です．

しかし，それだけではなくて一番大事なのは何かといいますと，社会科学研究における再現性を担保することです．つまり実証研究では，第三者が同じ手続きを踏んで同じ作業をすると，同じ結果が出てくるということが大事なのです．ですから資料の扱いについても歴史の分野であれば第三者が同じように資料分析すれば，同じ結果が出るという，その手順を示すというかたちが大事なわけです．ではデータを使った実証研究の場合における再現性を担保する条件は何かということです．ある研究者がある調査をやって，ある結果を出した，では本当にそのような手続きで同じ結果が出るのかということを担保するためには，どうしたらよいのか．１つの方法は，研究資金を豊富に与えて同じ調査をもう一度やればいいのです．しかし，それは研究資金の無駄です．そこでもう１つのやり方は，ある研究者がある結論を出した，その元になったデータに第三者がアクセスできるような環境を準備することです．第三者が同じ手続きを踏んで同じ結果が出るという状況があることが，再現性を担保することの仕組みです．それを社会的に保証する仕組みがデータアーカイブだと思います．私はこれが非常に大事な機能だろうと思います．

データが共有される仕組みができてくることによって，研究者が同じデータセットに基づいて，いろいろな仮説を立て，いろいろな角度から議論することができる．今までは，違う結果が出ると，それはよって立つデータセットが違うのだということになった．データが共有されると，同じデータセットの上に立ってある仮説を検証し直したり，異なった仮説で現実を説明するための，共通の議論の土俵ができてくる．この点が，非常に大事なことなのではないかと思います．

もう１つは，特に若手の研究者にとっての重要性です．大規模データセットを使った実証的研究をやろうと考えている大学院生とか助手クラスの若手の研究者には，自分で研究資金を取って大規模な調査をやって，データを収集し，そこから自分の仮説を検証することはなかなか難しいです．若手研究者にとっては，そういうデータセットにアクセスする機会が社会的に提供されていることが大事なのではないかと思います．もちろん，データセットは既存のものですので，自分の仮説にあてはまる変数がすべて収められているというようには中々いかないと思います．自分の仮説を検証するに足りるだけの，ある程度の変数が含まれているようなデータセットにアクセスし，それによって自分の仮説を検証し，論文を書くわけです．

つまり，若手研究者が蓄積されたデータセットに基づいた実証的研究をする機会を広げることができるのです．私は労働関係の仕事をしているので，労働経済の研究者とのつき合いが多いわけです．日本の場合はいわゆる国がつくっている指定統計等に，なかなか若手研究者はアクセスするということができないわけです．たとえば，最近，原先生は，社会的な格差，所得格差等について議論されています．働いている人たちの賃金格差を分析する賃金構造基本統計調査のデータは非常に重要です．では，大学院生がそのデータセットを使って，企業に雇用されている人たちの賃金格差が，最近の人事制度の改定，いわゆる

成果主義賃金の導入という議論の中で，広がっているのかどうかを分析するとすれば，その個票データにアクセスできるのが一番いいわけです．けれども，日本の場合は統計法の制約があって（もちろん目的外申請の手続きをふめばいいわけですが），若い研究者では手続きを踏んでもなかなか使えない状況があります．そうすると，日本についてはデータが手に入らないので，そういうデータが手に入る海外，たとえばアメリカ等のデータを使って研究するということが起きるのです．

たとえば，IT 技術を使えるか使えないかが所得格差を生む要因であるかどうかというときには，本来，日本のデータを使って研究してもらえばいいのですが，アメリカのデータを使ってそういう仮説的な議論をする．日本ではそういう議論はできない．そういうことが起きてしまうということが実際にあるわけです．

こうしたことが他の分野でも起きかねない．そういうことを少しでもなくしていくためにも，データアーカイブをつくっていくことが非常に大切だろうと思います．

もう１つは，教育上の利用です．たとえば，データの分析ということであれば社会調査とか，社会統計の授業ということになると思います．様々な統計分析の手法を教える際に，データアーカイブが充実されれば，実際の調査で集められた生のデータを使って様々な統計手法を教えることが可能になるわけです．現在でも，学生が適当につくったデータや先生がつくったデータを使って統計分析の手法を教えることなどが行われています．そうではなくて，実際におこなわれた社会調査やそれらのデータを使って，いろいろな統計手法を勉強することができるのです．

また，データアーカイブのデータを分析することによって，統計手法の勉強だけではなくて，それを分析した結果自体にも意味があるわけです．実際の社会現象を反映したデータですから，分析だけでなく，結果の解釈についても教育上の利用が可能なのです．

海外のデータアーカイブを見ますと，スタディーパックというものがあります．教育用のデータセットの提供が行われているのです．授業で使いやすい，社会調査なら統計学の授業とか，そういうときに使えるようなデータセットを提供しています．あまり件数が多くても困るわけで，ある程度件数をしぼった，授業で使えるデータセットです．

データアーカイブがあることは，単にデータを保存するということだけではなくて，実証研究としての重要な再現性を担保するとか，研究者同士が共通の基盤で議論するための研究基盤をつくっていくとか，あるいは若手研究者に新しい研究機会を提供したり実証研究をやろうという院生等を育てるためのいろいろなツールを提供する．そういった機能があるだろうと思います．

データアーカイブの歴史を見ますと，先進国ではだいたい 1960 年代，70 年代にデータアーカイブをつくるという運動がありました．ほとんどの主要先進国では既にデータアーカイブが設立されています．

ヨーロッパの主要先進国にはデータアーカイブがあります．アメリカにもたくさんあります．その中でもローパーセンターというのが一番古いものです．それ以外で大きいのは，ミシガンに ICPSR という，大学が集まってつくった大きなデータアーカイブがあります．他の国では，オーストラリアにありますし，南アフリカは５～６年前にできたと思います．日本では，これまでデータアーカイブというのはできてこなかった．最近，こちらの SORD でありますとか，私どもの SSJ データアーカイブが 90 年代の後半になってやっとできました．もちろん，日本でそれまでデータアーカイブがなかったのかというと，そういうことはないのでして，１つは政治学の分野では，神戸大学にいらした三宅一郎先生の

お仕事や今東大にいらっしゃる蒲島先生がレヴァイアサン・データバンクという選挙関係のデータアーカイブをつくられています．

ただし，これは，いわゆる投票行動・政治意識，ご自分たちが選挙関係，つまり政治意識についてやられた調査を研究者に提供するためにデータバンクをつくるもので，政治分野だけを対象としています．もう少し社会科学全体についてのデータアーカイブというと，これまではできてこなかった．それに取り組んだのが，SORD と我々のところということです．

なぜ日本で長い間データアーカイブができてこなかったのかということは，重要だろうと思うのです．社会科学の分野でいえば，海外での研究動向を一早く把握し，同じような研究テーマで日本でやる人が出てくるわけでありますし，分析手法も海外ではやり始めた分析手法を日本でも使うという状況があります．海外で 60 年代，70 年代にデータアーカイブをつくる運動があったにもかかわらず，なぜ日本でできなかったのかということです．

海外から日本にデータアーカイブをつくれという働きかけは，いくつかあったようです．私は良く知らないのですけれども，たとえば ZA があります．ドイツの ZA をつくるときに尽力された先生方などが日本の研究者に働きかけて，富永健一先生や統計数理研究所にいらした林知己夫先生などにアクセスされたというお話を伺っています．林先生などは，データアーカイブをつくろうという議論を長くやられていたようであります．それはどこの場で行われているかといいますと，１つは日本世論調査協会という世論調査を実施する会社と，あとは世論調査を委託しているマスコミ関係が集まった組織です．新聞社，テレビ局が主に入っています．もちろん研究者も入っているのですけれども．そこは統計数理研究所との関係が深いということで，日本でどのようにデータアーカイブ（そこではデータライブラリーといわれていましたが）をつくるかという議論がありました．

そこが一番長く早い時期から日本でデータアーカイブをつくるべきだという議論をしていました．しかし，30 年ぐらい議論をしていたかと思うのですが，実際にはできなかった．なぜなのかは非常に難しいのですけれども，外から見ている限りでは，日本世論調査協会で議論したのが限界になったのかなという気もします．調査会社は調査を受託しているのであって，調査会社がいくらデータを出すといってもデータは出てこないのです．多くの場合，データを持っているのは新聞社です．ですから，新聞社がデータを出すといわないと，データアーカイブはつくりにくいのです．では新聞社がデータを出すのかというと，これは非常にセンシティブでありまして，新聞社などマスコミは，世論調査なり，選挙データなりについては一番ガードが堅いのです．ですから，そこの場で議論をしたのは，結構難しかったのかというような気がします．

では研究者の側はどうだったのか．研究者はつくろうとしたのか，しなかったのか．これは原先生の方がお分かりかもしれません．日本でも，SSM をはじめとして大規模な調査があります．そうした調査のデータを，研究者に公開するということを言われる先生はもちろんいました．それは否定しません．しかし，組織としてできなかった．その理由は何なのだろうかというと，１つは，日本における大規模調査の実施のされ方にあるのかと思うのです．日本の場合，SSM は最近では違ってきましたけれども，やはり自分たちで調査を実施するわけです．科研費を取り，調査会社を一部使うにしても，基本的には自分たちで，学生，大学院生たちと一緒に調査をします．そのあとのデータのクリーニングからデータファイルの作成まで自分たちでやって，コードブックまでつくる．

まさにデータというのは，自分たちが時間

とエネルギーを相当投入してつくったというものであるということが大きいのではないかと思います．さらに，大きな調査では組織も大きくなりますので，やはり参加した研究者全員が合意してデータを公開するということがすごく難しいという状況が起きやすいのではないかと思っています．

海外についていいますと、基本的には大規模調査をする場合は調査会社を使うのです。研究に使えるようなちゃんとした調査会社があるわけです。アメリカでいえば、たとえばゼネラルソーシャルサーベイ（GSS）をおこなっているNORC（シカゴ大学）は、大学の中にある機関なんですね。そうした機関はミシガン大学（ISR）にもあって、大学がそういう調査機関をもっていて、一定の質を担保した調査が実施できるわけですけれども、日本はいわゆる民間の調査会社で、それについては質の問題もいろいろある。

それともう１つは，日本の場合，科学研究費で大規模調査をやりましても，公開する義務はないわけです．海外で研究費をとると，一般的に大規模調査であれば公開するということは義務です．海外の場合ですと，研究資金を取り調査会社を使って調査を委託する．基本的にはコードブックもついて調査結果が出てくるわけです．そして，その調査は，何年後かにはデータを公開しなければいけないという仕組みがあります．日本では上述したような条件がみんな欠けていたのではないか．ですから，なかなか調査を公開しない．

人的資源をかなり動員しないと調査ができないわけですけれども，逆にデータを公開してしまうと次の調査をやるときに人的資源が集められないということがあるかと思います．つまり，調査メンバーに入ることによってデータを使えるということにすることで，多くの人を集めて調査を実施してきたという面もあるのかと思います．

それともう１つは，一部にはご自分でデータを公開された方もいます．統計数理研究所についても一時期，国民性の調査を出したことがあるのです．ただ，いろいろな問題が起きた．それはどういうことかといいますと，自分たちは良かれと思ってデータを公開してみたら，１つは非常に問い合わせが多くて，それに忙殺されるわけです．それだけではなくて，データセットについて，いろいろとクレームをつけてくる．そんなことを言われるくらいであれば公開しない方がよかったと．個人的に公開された人は，コリゴリだというような，そういうお話を伺いました．

なぜこのような問題が生じたのかというと，データを公開するためのデータアーカイブがないですから，自分で全部公開する．そうするとそれについての事後対応も全部研究者個人でやる．それは非常に負荷がかかるわけです．そうすると，もう公開しないということになるわけです．

ですから私は，公開が進むためにはいくつか条件があると思います．１つは，調査実施からコードブック作成まである程度信頼のおける調査機関に寄託できる仕組みというのが必要です．それともう１つは，人を集めて調査をおこなう組織を研究者が全部管理をして，研究以外の時間を使って調査をしなければいけないという状況を変えない限り，研究者自身がデータを公開してもいいという気持ちになりにくいということです．

もう１つは，データアーカイブにデータを寄託すれば，データの問い合わせの対応をやってくれるような仕組みがないと，つまり公開をしたことによって研究者の負担が増えて研究時間がなくなるような状況があったらとても無理ですから，そういうデータアーカイブができないとやはりデータを公開するというのは無理です．もう１つは，公的な資金の問題です．やはりデータ公開を義務づけるような研究費の配分を考える必要があるのではないかと思います．日本の場合は，科学研究費

の審査の基準にそういうものを入れることがあります．これについては，文部科学省になる前の文部省に聞きに行ったことがあるのですけれども，それは文部省としてはやらない，それは学術会議あるいは学会で決めるべきだと言われました．文部省としてそういう基準を入れることはできないと言われました．

私は学会としてそういうふうにすると決めればいいのではないかと思います．科研費を申請するときに何年後には公開しますというふうに書いた人を，同じようなテーマであれば，そういう人にお金を配分することをやればいいのではないかと思います．そうしないと，なかなかデータの公開は進まないことになります．

では，私どものデータアーカイブはどうなっているのかという点をもう少しお話します．なぜ社研がデータアーカイブをつくろうとしたかをお話した方がいいかもしれません．だいたい5年ぐらい前に，社研に，付属施設として日本社会研究情報センターができました．私はそこの教授です．そのセンターのコンセプト自体は私が採用される前に決まっていたものです．

そのときは，今できているデータアーカイブのイメージはなかったと思います．センターをつくるときにどのようなものを構想していたかと言いますと，1つは日本の社会学研究を英文で海外に発信することです．日本の社会科学についての研究成果を海外に発信し，海外の研究者と議論の場をつくっていくことが1つ目的です．これについてはオックスフォードから Social Science Japan Journal (SSJJ) という英文の雑誌を年2回出しています．

もう1つは，英語のフォーラムです．インターネット上のフォーラムをやっています．これは今，800人ぐらいが参加しています．政治関係が多いのですが，モデレーターを置いて，日本人，外国人，海外の方を含めて英語で日本の様々な社会問題について議論する場をつくっています．これは1つの情報発信です．

もう1つが，日本の社会科学研究を支えるような研究基盤の支援です．この中にデータアーカイブがあります．そういうものをなぜ考えたのかと言いますと，社会科学研究所には労働関係の調査の伝統がありまして，戦後様々な労働関係の調査を実施しています．社会科学研究所の地下に当時実施した調査票が保存されていたのです．これについては，東大出版会から戦後日本の労働調査という本が出ています．

当時は大規模な調査をやっても，手作業で集計するということがあって，充分に分析されていなかったデータもあるのです．そこでまず，センターをつくるときに社研として考えたのは，保存されていた調査票をもう1度コンピューター処理し，利用可能なかたちで再整理すること．その上で第三者に提供するということをやろうと考えたのです．その中で1つ，社研の石田浩さんが中心になって，1953年に行われた新規学卒，中卒労働者の調査が整理されて公開されています．

もう1つ，なぜ私が呼ばれたのかということなのですけれども，私は労働関係の分野では，ヒアリングとかのアンケート調査等で調査をするという研究をやっています．特に，企業や雇用者についての調査を，年間毎年20本ぐらいやるということをしてきました．自分で研究費を取ってやることもありますし，政府のいろいろなプロジェクトで，要望研究的にやることもあります．特に，要望研究的なものについては，報告書を次から次へと書いていかないと間に合わないので，せっかくいろいろな方の協力を得て調査をしても，充分に分析し尽くすことができない．もう少し時間があればと思いながら次の調査をやるということをずっとやってきています．そういう中で，個人的には，そういうデータを第三

者が分析できるようなかたちで保存して公開する仕組みが大事だということを，いくつか小さいものを書いていたりしていました．佐藤がそういうことを考えているのであれば，やってもらったら良いではないかというようなことがあったそうです．そういうきっかけでデータアーカイブの仕事を始めたわけです．

当初，社会科学研究所の人たちは，社研にあるデータの整理をすることを考えたようです．それも非常に大事ですが，戦後におこなわれた調査を整理するためには，単にそのまま入力するわけにはいかなくて，そのデータを使って分析しようとする人が，一票一票慎重に判断しながら再コード化して入力することが必要なのです．

今のような調査票のかたちになっていないものが，すごく多いのです．たとえば，当時の職業分類をそのまま使っていいかどうか，再分類しなければならないのか，ということがあります．そうすると相当手間暇がかかります．ですので，それだけではいつになってもデータアーカイブはできあがりません．

私は，そういうことも続けながら，同時に調査データの寄託をお願いして，それを整理するかたちでデータアーカイブをつくることが大事だと考えています．たとえば，地下に眠っている調査票についてはどうなっているかと言いますと，研究者の方でこの調査を使って研究したい人が現れれば，是非それを使って下さい，ただし条件は，第一次利用優先権はその人にありますが，数年間その人がデータを使ったら第三者の利用に提供することです．それで良ければどうぞ使って下さいというかたちにしています．今，3つぐらい整理が進んでいます．

今，どのくらいのデータセットがあるかといいますと，358データセットが提供されています．社会科学研究所のデータ，および他の様々な研究機関がおこなった調査のデータです．

最初は，私が関係しているところにお願いして回るというかたちで増やしてきました．たとえば，生命保険文化センターについては，ほとんどのデータを寄託していただいております．先程お話しました三宅一郎先生が寄託者として，データを提供しているものもあります．

全体でいえば，一つは社研が独自に集めたもの，社研が昔やったデータということです．2つ目は，我々が集めたデータです．3つ目は，三宅先生が集められたデータです．ちょっとおもしろいのは，三宅先生が集められたデータに，「国民生活に関する世論調査」が入っています．これは総理府の世論調査室がやったものです．

なぜここにあるのかというと，先程，アメリカにローパーセンターというデータアーカイブがあるという話をいたしました．ここで，1960年代後半ぐらいに，世界中のデータを集めようとしたのです．それがきっかけになって，ヨーロッパでは，アメリカにデータを持っていかれるぐらいでならば，ヨーロッパでデータアーカイブをつくるということになりました．そのとき，日本からもデータを集めるというので，コンピューターカードをローパーセンターが持って行ったのです．その当時は，世論調査室もデータを出したのです．それをローパーセンターに集めて公開するという予定だったのですが，日本語の調査票だったという理由で，実際は整理されないまま置かれていました．それを三宅先生が日本に持ち帰ってきて，もう1度コンピューターで読めるかたちにしたのです．それをローパーセンターにも戻して，かつ，三宅先生は日本でもそれを利用する権利があるというかたちになったのです．これはローパーセンターにも提供されていて，アメリカでも使えるようになっています．現在では，これを我々のデータアーカイブを通じて公開するというかたちになっています．

我々の仕事の流れとしては，１つにはデータの寄託をお願いするということがあります．日本の場合，データの寄託についていろいろな研究機関にご理解をいただいています．１つは自分のところでデータがきちんと保存されていない，だからデータを寄託して整理してもらおうという考えがあります．もう１つは，いろいろな方に使ってもらう方が，調査実施機関の名前が知られるようになるし，自分たちが分析できなかったような分析がやれるということは，自分たちが次の調査をやるときに参考になる，ということです．

ただ，最初にお話ししましたたように，データを第三者が利用できるようにデータが保存されていないのです．データがきちんとクリーニングされていなかったりすることもあります．ですから，基本的に我々がもう１度全部チェックして，データクリーニングして，SPSS ファイルにして，簡単なコードブックをつくる作業をしています．

データセットについては，Web サイトの検索画面で選択肢レベルまで検索できるようになっています．たとえば，「結婚」と検索しますと，結婚についての質問が全部検索できるようになっています．

ただし，我々は寄託のお願いをしていますが，基本的にはデータを預かって保管しているだけです．利用の手続きはどうかと言いますと，利用の申請が来たら形式的な審査をします．一部学生にも提供していますが，基本的に我々は大学院生以上の研究者に提供するという基準にしています．

申請した人がその基準を満たすかどうかを形式的に審査しそれを満たせば，寄託者のところに書類を回して，こういう先生あるいは研究者の方がデータを利用したいと言っているけれども，利用を許可して良いかを問い合わせます．OK であればデータを提供するという仕組みです．

利用者の義務は，論文を書いたらその論文を二部，我々データアーカイブに送ることです．一部は我々が寄託者に渡します．もう１つの義務は，寄託者名，どこを経由してデータを利用したか，つまりデータアーカイブを経由してデータを使ったということ，これらを論文に必ず書いて下さいということだけです．別にお金も取りません．データを使って論文が出ると，その成果は Web ページの関連論文等に入れることになります．そうすると次の利用者は，先行研究が何であるかがわかることになります．

それでは，データの利用状況はどうかということなのですが，当初はあまり芳しくなかったというのが事実であります．１つは，日本の場合，研究に誰でも自由に使えるようなデータセットがなかったので，小さい規模のお金でも取って，自分で小規模な調査を実施してきたという事情があります．データを利用できることがだんだんわかってきたことで，去年の半ば以降ぐらいから利用者が急激に増えてきています．とりわけ，昨年の暮れに『社会調査の公開データ－２次分析への招待』という本を東大出版会から出したということがあります．この本では，公開されたデータを使った分析や研究，教育の仕方を紹介しています．また，どのようなデータセットが世の中に存在するのか，海外のデータアーカイブはどこにあって，どういうデータセットをどうやって取り寄せたら良いのかを紹介しています．それもあって，利用者が増えてきているところです．

さて，データアーカイブとして，どのようなことをこれからの課題と考えているかについてお話しします．１つは，２次分析をしようとする研究者を増やすための教育です．毎年いくつかのデータセットを取り上げて，それについて分析したいという人を集めて，１年間かけて１つのデータセットについて 10 人ぐらいで論文をそれぞれ書こうというプロジェクトを始めております．

去年は国民生活金融公庫，昔の国民金融公庫が毎年やっている新規開業実態調査を取り上げました．みなさんご存じのように，今，日本では，雇用をどうやってつくるのかということがすごく大事です．雇用機会をつくるために新しい企業をどんどん増やさなければいけないわけです．どうすれば新しい企業が増えるのか，できた企業はどうしたら成長するのかということが，雇用創出という点で研究の関心になってきています．それを分析するのに非常に良いデータセットなのです．国民金融公庫が新しく創業する人たちにお金を貸すことをやっているわけです．お金を貸した人たちに調査をしている．この調査の10年分データが寄託されましたので，それを分析することを行っています．

その研究会では，データを寄託した，つまりデータをつくった人にも来ていただいて，お話をしていただき，それについて若手研究者がいろいろな視点からテーマを出して分析しています．どういうデータセットなのかを充分に理解した上で分析するという場をつくろうとしています．

全国大学生活協同組合連合会が学生の生活実態調査をしていまして，こちらのデータセットも10年分寄託されております．このデータセットを分析する研究会を今年は立ち上げています．それともう1つは，生命保険文化センターがおこなった生命保険についての調査を分析する研究会があります．両方で30人ぐらいの研究者が集まって，今年1年かけてそれぞれのデータセットについて分析しようとしているところです．

このように，データセットの内容を十分理解した上で分析するための機会を提供しながら，データセットを使った二次分析を少しずつはじめています．1年が終わると報告書をつくります．そこには論文とそのデータセットを使う上での留意事項が書かれたものを出しますので，そういうものができれば，次の方はそれを見ながら分析するということができるのではないかと思います．

2次分析による研究の普及に加えて，もう1つは，既存データを使った分析手法について勉強をしてもらおうとしています．アメリカのICPSRは，サマースクールというかたちで，2カ月弱の期間，様々な社会分析の手法について，きちんとしたセミナーをやっています．そこまでは我々はできませんので，それについては日本から派遣するというかたちで，ある程度短期間のセミナーをやりたいと考えています．今年はシカゴ大学の山口一男先生が日本にいらして，社研で講義をしてもいいということなので，秋から5回連続で社会分析の手法についてセミナーをやります．

もう1つは，データを集めるだけではなくて，自分たちでデータをつくることを大阪商業大学と一緒に始めました．ご存じのようにアメリカには，ゼネラル・ソーシャル・サーベイ（GSS）という，アメリカ人の意識や行動についての汎用的なデータセットがあります．今，汎用的なデータセットだとお話しましたが，たとえば政治意識なんかも聞いているわけです．政治意識だけを分析するのであれば質問として深くは聞いていないわけですが，家族や労働，他の領域にまたがって分析することができます．

ですから，政治意識と他の領域についての関係を分析するということについては非常に使いやすいデータセットですし，あるいはいろいろな分野について調査していますので，授業にも使いやすいのです．家族について分析したいときには家族について分析できるし，政治についてなら政治，労働なら労働というように，教育にも使いやすい．人間というのは，政治だけで生きているわけでも労働だけで生きているわけでもありません．意識の相互関係なり行動の相互関係なんかを分析するときには，非常に便利なデータセットになっているわけです．

そういう意味では，教育とか院生等が自分の仮説を検証しようとするときに，いろいろな変数が入っていますので，とりあえずそれを使って分析してみる．ある程度できたら，お金を取って大きな調査をやるとか，そういう使い方ができるデータセットなのです．

そういうデータセットが，アメリカでは社会調査の授業で使われているわけです．そのようなことを，日本でも実施すればよいと思っています．我々のデータアーカイブにもたくさんのデータセットがあるのですが，それぞれが特定のテーマのデータセットなのです．ですので，なかなか授業などには使いにくいのです．そこで，授業でも使いやすく，若手研究者が自分の仮説を検証しようというときに，いろいろな変数が入っていて使いやすい調査をやろうということになりました．一昨年から予備調査を始めて，去年，第1回の本調査を実施しました．全国4,500人で，結局回収率は約65パーセントでした．

これは，二次分析をしようとする研究者，あるいは若手大学生を育てるため使っていただこうということ，もう1つはそういうデータを共有する研究環境を少しでも広げていって，できるだけ研究者の方に自分のデータを寄託しようという気持ちになってもらおうという意図もあるわけです．ですから，基本的には調査が終わったらすぐに公開することを原則にしています．

ただ，調査を終了してから1年で公開しようと思って始めたのですが，なかなか1年で公開するというのは時間的に言っても難しいことがわかりました．本調査を去年の11月に調査会社を使って実施し，調査会社からある程度クリーニングが終わったデータファイルがやってくるのが2月の上旬ぐらいなのです．その後に，職業とか産業のコーディングは，我々がやっています．その作業が終わるのが，だいたい3月の末ぐらいです．それからデータファイルをSPSSファイルにしたり，ロジカルチェックや何やらをし，研究者が実際に動かしてみて，問題がないかチェックします．そこで問題があれば，また原票に戻ったりということをやりますと，7月の末ぐらいにどうやらみんなで使えるデータファイルができるというわけです．これが終わると，今度はコードブックを作り，その印刷が終わると公開ということになります．公開は，だいたい来年の2月の初めということで，どうもやっぱり1年3カ月ぐらいかかるという感じです．それでも他の調査に比べると早いかなとは思っています．

現在は2回の予備調査分を公開しています．第2回の予備調査（有効回収数790人）を今年の2月に公開しまして，これについて15大学ぐらいから，授業で使うという理由で申請が来ています．利用者ベースでいうと100人から150人ぐらいになっています．ですから本調査が公開されると，相当授業で使ってもらえるのではないかと思っております．

先程，日本では調査会社を使っても，自分たちでコードブックをつくったり，データトレーニングをやったりしないといけないということを指摘しました．将来的に我々はどうしようと思っているかといいますと，1つは調査のコーディングとかコードブックつくりだけをデータアーカイブで受託できたらいいという考えを持っています．大規模調査についていえば，データのコーディングとか職業や産業の分類とか，コードブックをこちらでつくるようなかたちでデータアーカイブでやらせていただく．その代わりデータの公開を早くしてもらうというやり方もあり得るかと思っています．つまり，職業分類などを1カ所でやってそのノウハウなど蓄積していくということがすごく大事ではないかと思っております．本当はどこかで，アメリカのNORCみたいな，大学の中に調査を実施できるところがあればいいと思うのですけれども．

今回，我々は職業や産業の分類をかなり自動化しようと試みて，人によるコーディングと同時にコンピュータの自動コーディングを組み合わせておこないました．もちろん，機械を動かすことによってコーディングの精度を上げるということをやって，かなりノウハウを蓄積してきたので，他の調査にも応用できるのではないかと思うのです．本当にそういうことがやれるようになれば，データの公開を研究者がする上で，少しは役に立てるかなと思っています．

実際に，少しずつ状況が変わってきています．初めは，ほとんど研究機関しか寄託をしてくれなかったのですが，政治学関係の寄託や日本家族社会学会の全国家族調査研究会が1999年におこなった大規模調査のデータセットも寄託されました．10,500人がやって回収数が7,000人ぐらいの大きな調査です．当時はこの研究会でしか使えないということだったのですが，公開されましたので誰でも使えるようになりました．ですから，こういう状況に少しずつなっていくと思っています．そのうち，寄託できないデータセットというのはちゃんとした調査ではないのではないか，というふうになることが期待できるのではないかと思っています．

いろいろと行ったり来たりしていますが，データアーカイブの機能というお話をしました．1つはデータアーカイブがあることによって，データの作成とかコードブックづくりの標準化が，ある程度進んで行くと思うのです．データセットを第三者が使えるようにするためには，メタ情報をきちんと整理していく．それは今までは，ただ研究者のノウハウという秘伝になっていたわけです．秘伝ゆえに，第三者たちが間違った分析をするから公開できないなどと言われていたのです．しかし，基本的には科学ですので，やっぱり秘伝では困るわけでして，誰が使っても同じような結果が出るようにコードブックを整理することが必要だと思います．

では，コードブックにどういう情報を求めるのか，あるいはどのようにデータクリーニングをするのかということは，日本では整理されてこなかった．ですからデータアーカイブができることによって，ある程度一定のスタンダードができる，つくらなければいけないと思っています．JGSSでは詳しいコードブックをつくっていまして，かなり厚いものなのです．データセットに加えてデータに関するどのような情報を公開していくかをきちんと整理しなければなりません．本当は学会できちんと議論してもらえれば良いのですけれども．

そういったスタンダードをつくればデータの公開もしやすくなるのではないかと思っています．データ公開することによって，調査会社についても，調査会社の質を高めるということにもなるわけです．公開できるような調査をちゃんとやってもらうということがあるわけです．ですから公開することによって，調査の質が上がっていくだろうと思います．

ただ，そうは言っても，なかなか公開できないいろいろな問題がある，例えば公開した結果，当初の分析と違う結果が出てきて困るのではないかということを言う方もいます．僕は公開しないで間違いがわからないよりも公開して間違いがわかった方がいいじゃないかと思うのです．100パーセント完全なコーディングなんてないのです．誰かが使って，問題があればデータファイルを直していけば良いのです．

あとでお話しますが，海外のデータアーカイブの仕事をしていると良くわかります．このデータは，ここが問題だから直して下さいという連絡があるのです．それは，やっぱりいろいろな人が使うからわかるのです．みんなが使うことによってデータファイルの質が上がっていく．データを公開することで調査の水準も，メタデータの水準も，データファ

イル自体も改善されていくということがあるのではないかと思います．

データアーカイブがやっているもう1つの仕事は，日本国内のデータをできるだけ使いやすいようにしていこうということです．あるいは，使えるような環境を整備する，使えるスキルを持った研究者を育ててもらえるような仕組みをつくることです．また，海外のデータを日本国内でできるだけ使いやすいようにするということもやっています．

それで，海外のデータアーカイブを，直接皆さんが，たとえばイギリスのデータアーカイブを検索し，イギリスの家族についての調査を探して取り寄せることができます．我々のデータアーカイブも日本人だけに提供しているわけではないので，海外から利用申請があれば提供します．海外にも相当出しています．ただ，今のところ，アメリカの大学に留学している日本人の大学院生が多いです．

その点に関して，僕は英語しかできない人に全部英語でサービスするということはやめた方がいいと思っています．全部英語に調査票を直してくれるなら使うなんていうのは，僕は怠慢だと思っています．ある程度日本語や日本についてわかる人でないとマイクロデータは使えないということもあって，大多数はアメリカに留学していて，向こうで日本とアメリカの比較で論文を書きたい，日本のデータセットを手に入れたいという人たちです．

アメリカのICPSRというデータアーカイブは，基本的に大学の研究用，教育用にデータを集めて提供するという組織です．アメリカのデータセットだけでなくて，ヨーロッパのデータもあります．EUがやっているユーロバロメーターという価値観の調査がありますけれども，つまりデータアーカイブ同士でデータを交換しているわけです．ここにアクセスすると，かなり主要な国のデータが手に入ります．

では，このデータを使うにはどうしたらいいのかというと，ここの会員になればいいのです．大学として会員になれば，たとえば札幌学院大学がICPSRの会員になると，札幌学院大学の学生，教員が基本的にフリーで使えるようになる，ただ，大学院を持つ大学が加盟する場合，今1万ドルかかるのです．

これまでどのような状況が起きていたかというと，それぞれの大学に，これまで海外のデータの二次分析をしようという研究者がいなかったので，ある程度の規模の大学しかICPSRに入れなかったのです．入ったとしても数人の先生が使うだけです．昔，北大が入っていたのですけれども，今は入っていません．それは先生が移動したからです．北大にいた先生が東北大学に移りました．

日本はデータアーカイブがなかったので，仕方なくそれぞれの大学がバラバラに入っていたのです．その結果，すごいお金がかかっていたのです．トータルで5万ドルぐらいアメリカに払っていたのです．けれども，ご存じのように英語のデータセットですので，5万ドルを払う分だけデータを使っているかというと，そんなことはないわけです．それぞれの大学の1人か2人の先生が，データをいくつか取り寄せて利用するということです．

2年前にそれぞれの大学に呼びかけて，まとまって入り直そうという話をしました．日本でICPSR国内利用協議会というのをつくりまして，19大学でアソシエーションをつくり，まとまってICPSRに加盟するというかたちにしました．その時のネゴシエーションで，5万ドル払う代わりにメンバーを増やしてもいいかという交渉をやりました．ですから，今19大学で払っている額は5万ドルです．

我々のSSJDAがハブ機関になりまして，それぞれの大学で，たとえば学習院大学がデータを欲しいと言ったときに，我々がアメリカのデータを取り寄せて，学習院大学の先生にデータを渡すということをします．ですから，

19大学の先生は大学がお金を払っていますので，基本的にはその中の先生は自由に授業や自分の研究に，ICPSRを利用してデータセットを使えるのです．是非，札幌学院大学，東北大学にも加盟していただきたいと思います．

東北大学規模ですと年間で40万です．単独で加盟した場合の3分の1です．よく，大学加盟というと，大学全部で同意しなければいけないと誤解する方がたくさんいますが，これはどこかが払えばいいのです．どこかが40万払えば，全学で使えるようになります．

それぞれの大学で誰が窓口になってくれるかがすごく大事だということです．今のところ，2種類あるのです．図書館が窓口になってやってくれているところと，ある先生が窓口になっているケースがあります．大学が窓口なっている場合は利用者が多くなります．ところが教員が窓口になっている場合は，その先生しか使わない．あるいはその先生の院生しか使えないという問題が起きてきます．つまり，自分が一生懸命アメリカのデータを使いなさいと宣伝すると，自分の仕事が増えてしまうからです．

ですから，やはり学部内でデータを取り扱うセクションをつくってもらうことが大事です．問題なのは，図書館がそれを嫌がることと，もう1つはデータセットの扱いが，磁気データとして，百科事典のCD-ROM等と同じように扱われることです．

基本的に，データはハブ機関である我々を通じて渡すことになっています．ICPSRでは，四半期ごとにCD-ROMを出しています．この中に，新しくICPSRで公開されたデータセットのコードブックからデータまで，全部入っています．そのCD-ROMをコピーして各大学に配っています．そのCD-ROMに入っているデータについては，直接そこから渡していいのです．が，その時にCD-ROMを渡されると困るのです．そこから，学生なり教員が使うというデータだけを取り出して渡します．かつその時に，誓約書をとっておくということが大事です．そういうことをやるのは，今までの図書の扱いと違いますので，その理解がなかなかうまくいかないことがあります．

ついでに言うと，基本的には社研でお金を集めていますが，集めたお金はアメリカに全部払ってしまっています．ICPSRのデータ提供について管理費は一切取らないかたちでやっています．

それともう1つ，データアーカイブの国際組織があるのです．先程新國さんが報告していたインターナル・フェデレーション・データ・オーガナイゼーションというのがあって，去年，そこにやっと加盟できました．これまでなぜ加盟できなかったかという理由はいろいろあるのが，やっと加盟できました．

あっちに行ったり，こっちに行ったりしてきましたが，ここで今後の課題を少しまとめてお話したいと思います．

1つ目の課題は，何度も言いましたように，研究者が，特に大規模データセットについて，社会科学分野であれば，社会科学分野の研究財産として共有するという雰囲気をどうやってつくっていくのかということです．データアーカイブの器はできたけれども，なかなか研究者のデータ寄託が進まないという状態があります．これを，どういうふうに進めていったら良いのかということになります．

状況は変わりつつあると思います．データを早めに公開しないと，お金が取りにくいという方向に変わって行くのではないか，あるいはデータを公開しないと，ちゃんとしたデータセットとして社会的に認知されないというふうに移っていくのではないかと思います．特に若い研究者は，アメリカ留学から戻ってくると，アメリカではそうだったと，当然わかっている研究者が増えています．そういうふうになって行くのではないかと思います．

もう１つの課題は，やはり二次分析を研究の手法として確立して行くということです．やはり，社会学でも，自分で調査データを集めるところから始めるというのが，まだまだ根強いように思います．既存のデータに関する二次分析でも立派な修士論文が書けるわけでありますから，そういうものを研究手法としてきちんと認めて行くということが大事なのではないかと思います．

もう１つは，二次分析の研究手法について教育することです．確かに社会学部ですと，社会調査実習というのはデータを集めるということで，データを集めて分析をするまでを教えるということがメインなのです．僕は，大学院教育では当然データの仮説を立て，調査票をつくり，データを集め，データのクリーニングからコーティングをやってということがすごく大事だと思うのですが，学部教育で学部生全てにそれを要求するのは，如何なものかなと思っています．

もちろん社会調査士みたいなものがあるわけですけれども，時間の限られた授業しかやれない大学であれば，僕は既存データの分析手法をきちんと教える方が就職してからの役に立つのではないかと思います．

ですから，特に学部での社会調査，教育のあり方を見直しても良いのかなと思います．その時に，もし二次分析の手法を教えるとすれば，JGSS のような汎用的なデータセットがあると，教育もしやすいかなと思っています．

最後に，今日は量的なデータについての話だったわけですけれども．海外では質的データのデータアーカイブができています．いわゆる事例研究です．ヒアリングとかテープの記録という，そういうものも保存し公開していくというプロジェクトが，ヨーロッパ，特にイギリスで始まっています．質的データの保存はヒアリング記録とかテープです．そういうものも後世の研究者が使えるようにして行くということがなされています．

もう１つは，データについて言えば皆さんもご存じのように，日本にはパネルデータがほとんどありません．日本ではやっとデータアーカイブができましたが，パネルデータがない．一部でパネルデータの実施と公開が始まりましたが，学会としてもパネルデータをつくっていくことが必要です．

たとえば，我々の研究領域で言えば，最近は若年失業が問題になっています．今，いわゆるフリーターになった人たちについて，彼らの技能形成に将来どういう影響を及ぼすのかというと，やはり若年の人たちを追いかけて行くことが求められます．最近の学力低下なんかもそうですけれども，やはり時系列的に追っていくような調査が必要だろうと思います．

最後に，今日お話しませんでした官庁のデータです．官庁のデータの公開をどう考えるか．その場合に大事なのは，統計法を変えなくてはどうしようもないのです．この統計法を変えようという動きがどこからも出にくい構造があります．学者ぐらいしかそういうことを言わないのです．統計法を変えて，もう少し研究用に官庁データにアクセスしやすいような仕組みをつくっていくことが，大事なのではないかと思います．

国勢調査のデータについても，海外ではサンプリングデータが公開されているわけです．日本では，まだデータを公開することとプライバシーの保護が両立しないと思っている人が結構います．我々は，データを寄託するときに，まずその説明からしなくてはいけないのです．マイクロデータを公開するということとプライバシーを保護するということは，基本的には両立し得るのです．つまりプライバシーを保護するということは，個人が特定できないようにすれば良いわけです．我々もデータをやるときに，JGSS の調査区のデータを外すとか，いろいろやって，個人を特定

できないようにしています．

もちろん難しいものもあります．企業調査なんかは，結構難しいです．我々にも企業調査のデータがあるのですが，たとえば愛知県で，自動車関係で従業員が何人以上というと，どこの企業かすぐにわかってしまうわけです．そうすると，愛知県というコードも外すとか，規模を大きくくくるとか，そういう作業を我々のところでやっています．データを公開する前に，そうした作業をやるわけであります．

特に官庁統計の中でも，世論調査の部分は統計法にカバーされていないのです．我々のデータアーカイブに，以前の総理府の世論調査室がやっていた調査があります．政府の見解だと，世論調査は統計調査ではないのです．統計調査だけを統計法がカバーしています．ですから世論調査は，法律上は公開できるのです．そういう観点で言いますと，たとえば青少年の国際比較調査や高齢者の国際比較調査があります．あれは統計じゃないのです．ですから統計法にカバーされていません．ですから公開できるはずです．突破口は，まずは政府がやっているいわゆる世論調査の公開にあると思っています．

もう１つは，いわゆる政策評価の議論の中で，政府内部で政策評価をするだけではなくて，研究者がいろいろな立場から政策評価をするような環境をつくっていくことが非常に大事です．そうすると政策評価に係わる研究をやる場合，やはりデータにどれだけアクセスできるかということがすごく大事だと思います．その観点から政府にデータの公開を求めるということが，正攻法としてあるのではないかというふうに考えます．

レジメがなかったために，あっちに行ったりこっちに行ったりして，原先生のようにまとまった話ができなかったと思いますが，あとは質疑応答ということにさせていただきたいと思います．

司会（高橋）：ありがとうございました．研究上・教育上の観点からのデータアーカイブの意義，それからデータアーカイブの歴史，さらにSSJDAの活動の内容や，今後のデータアーカイブの課題として，データアーカイブを二次分析にどう教育上活用していくかといった観点からお話がありました．JGSSも現在展開中とのことで，これについては既に利用の申請があるということですので，大変楽しみな展開ではないかと思います．最後に，国際的なデータの相互利用という話もありました．ご意見・ご質問を伺いたいと思います．

中澤：２つほど質問をさせていただきたいのです．まず１つは，事務的なといいますか，管理上の問題で，利用できる資格の問題なのです．今のところ研究者，大学院生以上の研究者しか利用できないということになっているのですけれども，民間のリサーチャーなんかで，SSJDAのデータを使いたいという声が，たぶんお耳に届いているのではないかと思います．聞いていると，そういうリサーチャーも結構いるのです．

SSJDAの中に入っているデータセットの中には，民間の会社が委託してやっているものもあるわけで，そうすると民間の会社がやったものだから民間の研究者も利用できるようにしろというような声も出てくると思うのですが，この点はこれからどうなっていくのだろうかということが１つです．

２つ目は，今度は利用する側に対する注意の与え方の問題なのですけれども，若手の研究者がデータの再分析をする場合に，その調査の前提となっている文脈をなかなか共有できなくて，非常に文脈を踏み外した，誤ったと言っていいかどうかわかりませんが，誤った研究をする可能性というのがあるわけです．そういう可能性をできるだけ排除するためのことを，いろいろおっしゃっていたわけですけれども，お話の中でこれに関わる点がいくつか上がっていて，たとえばアメリカの研究者で英語しか理解できない人には日本のデー

タはやっぱり怖くて渡せないという話がありました.

となると，これはやっぱり日本の文脈を共有している人でないと，文脈を踏み外す危険があるという，そういう実例としてお話されたわけです．ところがもう片方で，そういうふうに文脈を共有していないと分析できないとか，秘伝としてしか研究できないということでは科学にならないという，もう1つの観点があります．この場合には文脈をできるだけ明示的に表して，誰でもどんな人でも使えるようにしなければいけないという要請があるわけです．こういうふうに文脈を理解しないで使ってもらっては困るという，今までの秘伝的な職人芸的な観点と，一方ではできるだけ文脈を明示的に表して，誰でも使えるようにしなければならないという，相反する観点が今のお話しの中にあったと思うのです．

その矛盾する点を，どういう基準でもって解決していけばいいのかと，この2点です．

佐藤博樹：鋭いですね．最初の，民間の研究者なのですけれども，たぶん海外なんかで言いますと，ICPSR は基本的にこれは大学だけという組織です．ローパーなんかは誰でも使えて，データを売っています．ローパーでは GSS のデータを売っているのです．ICPSR の方は大学が加盟していますから，研究者はタダで入手できるというかたちでデータアーカイブを管理運営しています．イギリスについても，ザ・データアーカイブと呼ばれている機関があります．あそこは，どこからお金が出ているかという研究資金の出所によってデータの利用料が違うということです．

ですから，いわゆる営利目的には提供しないということが，世界的なスタンダードになっているわけではないのです．では，我々はどうしているかといいますと，一応我々の方では国の予算でやっているといったこともありますし，まだまだ日本はデータアーカイブができたばかりなので，データを寄託するときのお願いの仕方としても，やはり研究用に使うというかたちで用途を限定した方が，データを寄託していただきやすいということが実際としてあります．ですから，当面はそのルールで行きたいというふうに考えています．あくまでも研究用で，利用者は大学か研究機関に所属している人だけということにしています．

あらためていえば，我々の機関としては，基本的には非営利目的で，大学の研究者だけに出すということでやっています．

2番目の点なのですが，海外の研究者の場合，もちろん英語しかできなくても日本人のリサーチアシスタントを使って日本語のコードブックを利用できる環境があれば，僕は使えると思うのです．たとえば我々が英語の調査票やコードブックまでつくって，英語で問い合わせに答えるということはできない．それはできないし，基本的には利用者が日本語環境で研究できるようにしておく．自分がやらなくてもいいですし，リサーチアシスタントを使っていい．そういうところにしかデータを出さないということです．文脈という言い方は，ちょっときつかったかもしれませんけれども，そういうことです．

それで，利用者が誤解した使い方をしないために必要なこととして，2つあります．1つはやはりコードブックを充実していくということです．ただ，正直に言いまして，JGSS など我々自身がデータを収集したものについては，ある程度きちんとしたコードブックを作成していますが，データを寄託してもらったものについては，コードブックとしては不十分なものです．

今，過渡期としてある程度やむを得ないであろうというふうに思っています．ですので，先程お話しましたように，いくつかのデータセットについては，データの寄託者と研究者が一緒になって研究する．そういう中で，データセットの特徴について議論をして，それを

ドキュメント化していく，というようなことを始めています．たとえば，去年やった新規開業実態調査については，データセットの特性とか，そういうものについてはかなり長いものを書いてもらっていますので，そういうものはコードブックの役割を果たすのではないかと思っています．

もう１つは，既存データを使った分析の能力というものを高めなければいけないということです．今までそういう機会がなかったから，そういう能力を持っている人が少なかったというのがあるわけで，これも過渡期だろうと思います．データを使えるようになれば，そういう使い方についてノウハウを持った人が増えるだろうと思っています．データを出したって使える人がいないのだから出せないと言う方がいます．それはデータを出さないから使える人がいないわけで，僕は逆じゃないかというふうに思っています．もちろん混乱が多少起きることはあり得ると思います．間違った論文を書いた研究者が，批判されることもありうるわけです．そのプロセスで徐々に，論文として書かれていても，駄目なものは淘汰されていくわけですから，多少そういうことがあるのはしょうがないのではないかと思っています．

中澤：１つ目の件については，ただ，政治的に難しい問題をはらんでいるデータセットだと，問題がちょっと大きくなるかもしれないです．午前中にちょっと話題にしたのですけれども，都市社会学会が最近，磯村先生が昔やった売春婦の調査データというものを都市社会学会の会員向けに公開しているのです．これはかなり注意書きが厳しく付いていて，非常に興味本位の調査項目に見えると．だけれども興味本位に見える調査項目の裏には，磯村先生のいろいろな意図があるのだと．たとえば，その売春婦調査の中には，最初の性体験はいつかとか，そんな調査項目もあるのです．そういうのを見ていると，いかにも興味本位だという感じがするのだけれども，都市社会学会の人の説明だと，そうではないのだと．むしろそういう項目を調査することを通じて，結果としてその売春婦が別に特殊な境遇の人たちではなくて，恵まれないある偶然の下で，そういう境遇になってしまっただけだというような，磯村先生の意図があるのだと．そういうことを結果として言いたいという意図があるのだから，そこを読みとらずに興味本位で分析されては困るというような注意がきが付いています．その注意を読んだ上でないと，このデータを使わないでくれというような話があるのです．

そういうふうに政治的に社会的に，いろいろな微妙な結論を引き起こしかねないデータセットについては，コードブック以上の仕掛けが必要ではないかという気もするのです．一方で，都市社会学会のやることはやり過ぎかもしれないです．科学にとっては，そういう見方に最初から限定してしまって良いものなのだろうかということがあり得ると思うので．そこら辺はどうなっていくのでしょうかね．

佐藤博樹：ある程度，注意が必要なものはあると思うのです．それと利用の制限をするときに，どこまでバランスを取るのか，個々に判断せざるを得ないという気もするのです．一般論としては，そうした制限をするのは，ちょっと難しいかという感じがします．

ちなみに何で都市社会学会の会員だけに公開しているのか．もっとオープンにしてやって下されば……．

中澤：やっぱりそれなんじゃないですか．誤った使い方をされたら困るという．

佐藤博樹：いくつかのケースがあります．例えばイギリスの例ですと，学部生の場合はまず寄託者に論文を送って，それを見た後でなければ公開できないと．大学院生は良いのだけれども，学部生の場合は書いたものを先に送って下さいと．そういうようなやり方もあ

ります．

大学院生以上であれば，スタンダードなデータの使い方がある程度できる．学部生の場合，公表する際には，まず事前に許可をとりなさいというふうにコントロールする．そういうやり方もあると思います．

古村：私の専攻は社会教育なのですけれども，社会教育という分野は民間の活動家というか，住民運動をやっている方が大変に多くて，しかもそういう方々が研究を活動やったり，調査を活動やったりするということも非常に多いのです．そういう場合に今の話ですと，大学の構成員でないと全くアクセスできないということのようなのですが，おかしくありませんか．全部の調査項目について民間に開放する必要はないと思うのですけれども，少なくとも，一般市民に公開しても構わないと思われるようなものについては，アクセスできるようにするとか，市立図書館の方でもある程度は使えるようにするとか，そういったことをやっていくことによって，むしろ大学の中でもそういう活動が活発化するのではないかと思います．その辺は，まだこれからの課題だと思うのですけれども，その辺はどういうふうにお考えでしょうか．

佐藤博樹：１つは，つまりデータは誰でも使えるというわけではないということなのですね．一定のデータの使い方について，そういうノウハウを持っている人でないと使えないのです．ですから，もちろんやり方としては，そういうスキルがあるかどうかということで，データの提供範囲を変える，広げるというやり方があると思うのです．

つまり，データの個票を公開するわけですので，データ分析についてのある程度一定の手順というのがあるわけです．それを知っていないと，やはり間違った結果が出ちゃうので，そういう意味で研究者であれば一応，もちろんその中でも駄目な人もいるわけですが，そういう基準を満たしているということにしているということです．

もちろん，修士を出ているということを条件にするというやり方もあると思うのです．

集計結果はホームページに載っているので，誰でも見られるようになっています．これはあくまでも個票ですので，それを使って分析するということです．ですから，たとえばそういう住民団体が直接おこなうのではなく，大学の先生に委託して分析してくれというやり方があるのです．それを別に駄目だと言っているわけではありません．

田中：１つ伺いたいことがあるのですが．先程お話しの中で日本ではデータアーカイブというのが，どのように行われたかということがありました．私はちょうど 1972 年から 1980 年まで学術会議におりまして，そこでデータ活動に関する討論の取りまとめ役にあたっておりました．そのときにデータ活動だけではなくて，データを含む情報活動のシステムをどういう形にするかということを，随分いろいろと議論いたしました．その議論の結果，勧告として出したわけなのですが，結局それは学術審議会の方でいろいろと変わりまして，結果として出たのが今の学術情報センターなのだと思います．

その中で，１つ問題になりましたのは，どうもデータ活動を研究者自身が直接行うのは必ずしも適当ではないのではないか．それよりも図書館の司書の仕事に対応するような制度，学術情報士という制度をつくってはどうか，そしてそれはそれぞれの研究者に対する研究補助的な仕事をする立場にある．こういう人の制度をつくってはどうかということも勧告の中に入っていたと思います．そこで，データアーカイブという言葉はありませんでしたけれども，データ資料館の設置について，いろいろな発言があって議論が行われました．

１番ネックになったのは，やはりデータの共有性に関することです．その当時の社会学者の理解がなかなか得られないので，当分日

本ではデータの共有性に基づいた資料館のようなものは非常に難しいのではないかという，そういう意見が一般的な意見のように，聞きました．共有の方の議論が充分に伸びなかったと思っております．

ただ，学術会議ですから，文献だけではなしに学術資料関係全体を含めていろいろと議論をかわしたと思います．自然科学系のデータ活動自身も，必ずしも充分ではありませんでした，外でできたデータをもらって使うなんていう例は関心があまりなかったと思います．

その中で，私自身は，原子核の核反応データ，荷電粒子核反応データベースというものの作成を始めました．それが現在まで続いて行われていまして，得られたデータは全部国際原子力機関に渡しまして，国際原子力機関から，特に発展途上国でそれが使われているという，そういうことがあります．そのことについては，かなり長い間，ここにいる千葉君も関係しているので，あるいは千葉君の方からその話があるかと思います．

ただ，そういうことについて2，3伺いたいことがあるのですが．先程データの再現性という話が出ました．再現性という意味はいろいろあるかと思いますけれども，自然科学の方で再現性といいますと，対象が同一であってもデータを測定する人が違う場合には違ってまいります．実際，私が直接専門としている原子核の部分で，対象は同じで測定方法もそう違うわけではないのですが，それでも各研究機関によってデータが随分違います．世界的にそういうデータをまとめて，データマップをつくっているデータ・センタがあるのですが，ここでは，研究機関をAランク，Bランク，Cランクに分け，中で実際にデータ活動をやっている人は，Aランクのデータは平均にする，Cランクは参考にする，Bランクの機関のデータは補正に使うという扱い方を内々しているのだということを聞いたことがあります．

私が伺いたいことは，対象は同じでも得られたデータは随分違って，その意味での再現性は充分ではないのです．社会現象というのは，原子核よりは相当複雑多岐ですので，もし同一の対象に対して多くの研究者が調査をするならば，そのデータはかなり，ある幅で分布しているのではないかと思います．

しかしながら，幸いにして1度か2度しか行われないものですから，そのデータが本当の姿，本当の姿が何かということも問題でしょうけれども，どのくらいずれているかということには関心を抱く必要がなくそれが使われているのではないでしょうか．それは自然科学でいうデータの再現性という点からみれば，1つの問題を示しているのではないかという気がしないではないのです．

これは少し強調して言い過ぎた点がないわけではないと思います．その点について，ちょっと伺いたいと思います．

佐藤博樹：ちょっと私が，充分にお答えできるかどうかわかりませんが，最初再現性と言いましたのは，検証可能性というふうに置き換えても構いません．ある研究者が出した結論を，そのデータセットを使って同じ手続きをして同じ結果が出るということですね．

田中：そういうことです．いわば利用に関する再現性ということですね．

佐藤博樹：ですから，そのデータが公開されていない限り，佐藤健二さんがやった研究結果は本当に正しいのかわからないわけです．今までそれが出されていなかったわけです．やはりそれは問題だろうと．やっぱり検証の可能性がちゃんと担保されているということが必要だろうと思っています．

確かに，再現性というと，そのときは確かに先生のおっしゃられたように，データの質の問題とか，測定尺度によって，どういう質問をするかによって，結果が違ってくるわけです．それは別の議論であると思うのです．

それは検証可能性の問題と，データの質なり測定尺度を変えなければいけないという，そちらの議論はもちろんあると思います．

田中：確かにデータの質の問題があると思います．私がそのころから始めました原子核のデータでは，学術雑誌に掲載されたデータだけを使うことにしたのです．そうすると，それは一応リフリーを通っていますので，ある意味では１つのスクリーンを通ったデータだけを使っている．それ以外に，各それぞれの実験所で蓄積されているデータがたくさんあるのです．それを使わないでというやり方を取ったわけなのです．

佐藤博樹：データの質についていえば，やはり，１つはデータの利用度というのが１つの尺度だと思います．

もう１つは，先程ちょっと言い忘れましたけれども，再現性の問題とかかわります．二次分析がおこなわれるなかで，当初の調査実施者の研究意図とは違った仮説の検証に使えるデータセットというのが結構多いということがわかりました．研究申請で全然違う研究分野の人が，なぜこういうデータセットを使うのかというのを見ると，なるほどこういう分析もできるのだなと思うことがあります．

ですから，当初の調査実施者が想定していなかった仮説についても，そのデータセットを使って分析できることがありまして，それは，ある面では検証の可能性を担保する，それに加えてデータセットの有効利用ですね．新しい仮説を検証する上で，既存のデータセットが使えるというのがたくさんあるということを経験的に感じました．

原：再現性の問題なのですが，細かく考えてみると４つぐらい低下させる原因があると思います．別の言い方をすると，１回の調査結果を絶対視するということは，どだい無理だと思うのです．我々がよく知っているように，たとえば，標本調査の場合，標本抽出を同じ母集団でやり直せば，その結果は当然微妙に揺れてくるという問題があります．

それから，データの質というふうな言い方をされましたけれども，たぶん，質問文の問題というのがあると思うのです．質問文のつくり方を変えると，本当かどうかわからないのだけれども，政治問題などでは10パーセントぐらいは回答を簡単に動かせるなんていうことがいわれています．質問文が本当の意味で同一であるかどうかによっても，同じことを調べてもやっぱり結果は違ってくるわけです．

他にも私は２つぐらいあると考えています．１つは適切な質問文で正しい質問の仕方をすれば，正しい結果が出てくるというふうに考えがちなのですけれども，それはやはり幻想だと私は思うのです．というのは，私はかつて同じ母集団から２つの標本集団をつくって，一方で面接調査を，もう一方で電話調査を，同じ質問文でやったことがあります．ところが，同一の集団とはとても思えないような結果が出てくるのです．それは実験研究ですから，なるべく差の出そうな質問文，たとえば，今覚えているのは，「あなたはテレビで男女のベッドシーンを見るのが好きか」とか，そういうような質問文ばかりを集めてやったら，やっぱり２つの方法で比率が全然違ってくるのです．しかも，その研究は面接調査の費用を電話調査で浮かせられないかという研究だったのですが，どっちの方が本音で，どっちが嘘かというのは，質問文ごとには多少の想像がつくのですけれども，全体としてどっちの調査の方がよいとは言えないのです．

ですから，私の授業では，調査の結果というのは調査方法にある程度依存するものだと割り切らなくては駄目だというような言い方をしています．そういう意味での再現性というのは非常に乏しい面があります．

もう１つ，授業みたいで恐縮なのですが，学生に言っているのは，回答の揺らぎというのが結構あるということなのです．つまり同

一の人間が嘘をついているわけではないのだけれども，同じことについて聞かれても，回答が必ずしも一致しないということがあるわけです．

これは，たぶん調査で聞かれるような多くの事柄というのは，普段自分で意識することはなくて，調査になるとその質問文によって，どこか意識下から，意識の世界に引っ張り出してくるのだと思うのです．それは，いつも同一なかたちで引っ張り出されるとは限らないものなのです．

長々と言いましたけれども，再現性というか，１回の調査結果というのはそれぐらいのものだというふうに，私自身は割り切る必要があると考えています．

田中：今，おっしゃったことに賛成で，１番最後におっしゃった，取り出してくるというところ，もう少しその瞬間のコミュニケーション環境において，その人が作成するのだと，その瞬間に作り出されるのだと，あらかじめ取り出してくるというよりは，その瞬間のコミュニケーション環境に応じて作り出されるのだと言った方がいいぐらいだと思うのですが．

原：まあ，意識なんていうのは，我々の概念の構成物ですから，何となく取り出すという言い方をしたのですけれども，そういうことだと思います．

中澤：原先生がおっしゃった，最初の標本誤差のお話しでしょうか．それを，ひょっとしたら２つに分割できるかもしれないと思うのですけれども，調査を拒否する人と回答する人のバイアスというのがあります．調査に答えてくれる層はある特定の属性を持っていて，調査を拒否する層は特定の属性を持っているということがあって，回答してくれる人のバイアスということも，５つ目にあるかもしれないという気がしています．

たとえば世論調査だと，選挙のときにあなたは必ず投票に行きますかと聞くと，必ず投票に行くという答えが，実際の投票率に比べて10パーセントぐらい高く出るのです．嘘つき問題とか言われるらしいですけれども，実際には世論調査に回答してくれる人は比較的投票に行きますと答えやすいので，そのバイアスを差し引いて考えると，必ず投票に行きますというふうに答える人の割合と，実際に投票率としてあがってきた数字というのはそんなに変わらないのではないかという話を聞いたことがあるのです．政治分析をしたことがないので，詳しくはわかりません．

それは一例としてお話したのですけれども，そういうふうに，回答する人がそもそも，回答しやすい人が持っているバイアスというのも５つ目の水準として考えられないかと今思ったのですが．これは別に今結論を出すべき話ではないのですけれども．

原：事実はあるかもしれないですが，それは話のレベルがちょっと違うような気もするのです．

原：もう１つ良いですか．

さっき，東北大学のことで責められたみたいなのですけれども，弁明をしておかなくては．

私もご存じのようにICPSRの利用希望者で，お金を負担しても良いと思っている人間なのです．学内にもかなりの人数がいて，みんなの了承を受けているわけです．ところが，事務が抵抗しているのです．それはなぜかというと，とにかく誰が出すのかわからないのだけれども，国庫のお金をどこかに払うわけですよね．払う対象として，どうも事務の判断では，そのような「好い加減な」組織では駄目だという，そういう言い方をするのです．

だから私は他の大学ではちゃんとやっていると言うのですが，なんか事務官の裁量の範囲らしくて，なかなか動いてくれないのです．

佐藤博樹：東大でも払っているし，神戸も国立大学ですよね．国立でも３つぐらいやっているんですよ．なぜできないのか，おかしい

ですよね．

原：一方で他の大学でやっているから良いと判断する場合もあるし，もう一方で，いや他の大学はそうかもしれないけれども，我が大学はこうですというような．それが，私が一番頭にくるのは，事務官のある種の裁量のグレーゾーンの判断なのです．だから事務官が変わった瞬間を私は狙っているのですが．申し訳ありません．

田中：東北大学の事務官はそれほど信頼がおけないのですかね，おっしゃった通りに．

佐藤博樹：事務で，まあそんなことだろうとは思っていますけれども．

原：それから，窓口に予定している図書館でもやっていいと言っているのです．実際に研究費を拠出する側も出してもいいと言っているのです．

田中：相当強いかもしれませんね．六大学の事務は案外おおらかだと思っていたのですけれども．平均すると．

司会：佐藤博樹先生，ありがとうございました．