社会調査データ・アーカイブ SRDQ の公開

松　川　太　一
樋　口　耕　一
川　端　　亮

目　次

社会調査データ・アーカイブSRDQの公開

松川　太一
樋口　耕一
川端　亮

1. 社会調査データとデータ・アーカイブ

1.1 データ・アーカイブとは

データ・アーカイブとは調査データの収集、編集・加工、保管、普及のための専門機関である。調査実査者による調査データの寄託、あるいはデータ・アーカイブ自身による調査の実施によって、データ・アーカイブは調査データを収集する。データ・アーカイブは、収集されたデータを2次分析が可能な形式で公開しなければならない。そこで2次分析を可能にするため、収集されたデータはなんらかの統一的な形式へと編集・加工される。データ・アーカイブは、そのような2次分析が可能な貴重な調査データを保管し、貴重な調査データが有効活用されるよう普及活動をおこなう。

1.2 海外のデータ・アーカイブと国際組織

データ・アーカイブは社会科学の分野において調査データの2次分析を推進するうえで重要な役割をはたしており、ながい歴史をもった専門機関でもある（Scheuch 1990）。海外では、まず第2次大戦後すぐにアメリカ合衆国で、Elmo Roperが作成したIBMカード形式の調査研究データがWilliams Collegeに寄贈され、世論調査のデータ・アーカイブが整備された。このデータ・アーカイブが現在のRoper Centerの前身である。その後1960年に西ドイツでZentralarchiv（ZA）がケルン大学の研究機関として設立された。そして1960年代なかばから、北米やヨーロッパ各国でデータ・アーカイブが設立されるようになった。

このように海外ではデータ・アーカイブはながい歴史をもっており、おおくのアーカイブが国際組織に参加しながら、ゆるやかに連携している。そのような国際組織として代表的なものが、IFDO（International Federation of Data Organizations）とCESSDA（Committee of European Social Science Data Archives）である。

IFDOは1977年に設立された国際組織である[1]。世界各地にちらばったデータ・アーカイブが国際協力することによって、社会科学における調査研究が充実するという発想にもとづいて設立された。IFDOに加盟している組織は北米やヨーロッパのデータ・アー

カイブが中心であるが、日本からも後述のSSJデータ・アーカイブが2001年に加盟を承認された。このようにIFDOは、全世界的なひろがりをもった国際組織である。またIFDOは、設立以来、UNESCOの国際社会科学協議会（International Social Science Council）の準会員となっている。

いっぽうのCESSDAであるが、こちらは1976年に設立された[2]。CESSDAはヨーロッパのデータ・アーカイブ間の非公式な協力関係として位置づけられている。ヨーロッパのデータ・アーカイブのほとんどがIFDOとCESSDAの両方に加盟している。ヨーロッパのアーカイブにとってIFDOはいわば政策決定機関にあたる。それに対してCESSDAは、CESSDAに加盟するヨーロッパのデータ・アーカイブが協力してオンライン用のデータ処理技術（たとえば後述するNESSTAR）を開発するなど、国際組織としてIFDOより密接なアーカイブ間の協力関係を形成している。

1.3 日本のデータ・アーカイブ

海外ではデータ・アーカイブはながい歴史をもっているが、日本の歴史は比較的みじかい。日本の場合、1990年代にはいって調査データを公開して2次分析に役立てようという動きが社会科学の諸分野で本格化してきた。そのような社会科学における2次分析の推進を背景にして、いくつかのデータ・アーカイブが日本でも誕生している。ここでは、札幌学院大学社会情報学部のSORD（Social and Opinion Research Database)[3]、そして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のSSJデータ・アーカイブ（Social Science Japan Data Archive)[4] というふたつのデータ・アーカイブをとりあげる。

1.3.1 SORD

札幌学院大学社会情報学部のSORDは「社会・意識調査データベース作成プロジェクト（Social and Opinion Research Database Project)」によって構築された社会調査に関するデータベースである。このプロジェクトは、おもに日本社会学会会員によって実施された社会調査に関する情報の収集・蓄積を1990年からおこなっており、「日本最初の社会調査データベース」をうたっている。また毎年、社会・意識調査にかかわる研究者を招待して札幌学院大学でワークショップを継続して開催している。

SORDを運用している社会・意識調査データベース作成プロジェクトは「社会調査の概要情報」を収集し、その調査概要情報をデータベース化して公開している。このような収集・データベース化の目的は、社会調査に関する情報の学術的な活用にある。データベース化された調査概要情報は報告書というかたちで出版されているだけでなく、情報提供者から許可されたものに関してはWebサイトでも公開されている。そのため現在では、インターネットを通じて調査概要情報を検索・閲覧することが可能である[5]。

検索可能な調査概要情報には、調査の名称、調査の領域、調査の概要、調査開始年、

母集団地域、調査方法といった項目がある。それぞれの項目についてキーワードを入力することで、各自の興味・関心にそった調査情報を調査データベースから検索することができる。もしデータベースの検索をおこなって2次分析してみたい調査情報がみつかった場合は、SORDの事務局を通じてその調査をおこなった調査主体と連絡をとることができる。このようにSORDは、分析用データの提供・公開を基本的におこなっておらず、2次分析希望者への分析用データの提供は実際に調査をおこなった調査主体にゆだねられている。なおSORDのWebサイトによれば、現在1,000以上の調査概要情報がこのデータベースに登録されているとのことであるが、1998年以降のものはかなり少なく、2001年以降のものは登録されていないようである。

1.3.2 SSJデータ・アーカイブ

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のSSJデータ・アーカイブは日本における社会科学の実証研究を支援することを目的として設立された。1996年5月、東京大学社会科学研究所の付属施設としてまず日本社会研究情報センターが設立された。SSJデータ・アーカイブのWebサイトによれば、このセンターは「研究実証データと研究関連情報を蒐集・作成して提供すること、国際的な研究ネットワークを形成するための情報基盤を構築すること」を目的としている。そしてSSJデータ・アーカイブは、1998年4月から学術目的での2次分析のために教員や大学院生等に調査データの提供をおこなっている。またSSJデータ・アーカイブは、さきに取りあげたIFDOへの加盟を2001年2月に承認された。

さきにふれたSORDと同様に、SSJデータ・アーカイブでもアーカイブに収録された調査の検索をインターネット上でおこなうことができる。おもな検索項目としては、調査名、データ寄託者、調査対象、主な調査事項、調査の概要、さらに調査につかわれた質問文等が検索できる。これらの検索機能をつかって、2次分析に関心をもつ教員や大学院生は自分の研究上の興味・関心を満たす調査データがSSJデータ・アーカイブに収録されているかどうかを調べることができる。2次分析をしたいデータがSSJデータ・アーカイブに収録されている場合、分析者はアーカイブにデータ利用の申請をおこなう。そして利用申請後、アーカイブと調査データ寄託者の承認を得てから分析者へデータが提供されることになっている。なおSSJデータ・アーカイブでは、2004年9月現在で479の調査データが公開されている。

2. これまでのデータ・アーカイブの意義と問題点

2.1 これまでのデータ・アーカイブの意義

国内外におけるデータ・アーカイブの経緯についてふれてきたが、このようなデータ・アーカイブは社会科学の領域において研究と教育の面で意義をもっている（佐藤ほか

2000)。まず研究上の意義とは、研究をすすめる際にデータ・アーカイブに保管された過去の調査データを有効活用できる点である。たとえば既存のデータを有効活用することで多額の調査経費が不要になる。また過去の調査データの2次分析が普及すれば、おなじような目的の調査をくりかえし実施する必要もなくなり、調査対象者が似たような質問の調査に回答する負担も軽減される。

教育における意義は、質のよいデータで社会調査法の学習が可能になる点である。おおくの学生にとって、さまざまな分野で実施された質のよい調査データを分析しながら社会調査法を学習することは、データ・アーカイブが提供するデータによってはじめて可能になる。社会調査法の授業では「有意水準が5%で帰無仮説が棄却」というような推測統計学にもとづいた説明をすることがおおいが、このような説明は無作為抽出をおこなった標本調査データを分析したときにはじめて意味をもつ。学生がみずからの手で無作為抽出にもとづいた質のよい標本調査を実施することはむずかしいから、データ・アーカイブが質の良い調査データを提供することは教育においても意義がある。

2.2 これまでのデータ・アーカイブの問題点

これまでのデータ・アーカイブは、現時点でも研究・教育にとって重要な意義をもっている。しかしながら同時にいくつかの問題点もある。ここでは問題点を3点だけ指摘しておく。

まず、知りたい分析結果が公開されていないという問題点がある。これは度数分布表やクロス集計表による調査データの公開を想像すれば容易に理解できる。たとえば政府の統計局などが実施したセンサス・データは、そのおおくがクロス集計表の掲載された報告書のかたちで公開されている。このような場合、報告書には男女別の集計結果は掲載されていたとしても、世帯収入別の集計結果は掲載されていないことがおおい。この場合は、独立変数としての世帯収入がさまざまな従属変数に対してどのような効果をもっているのかという研究上の関心を分析者がもっていても、報告書はそのような分析者の関心にはこたえてくれない。

第2の問題点は、知りたいことがすぐ分析できないという点である。第1点目に指摘したように、度数分布表やクロス集計表のかたちでの調査データ公開はすぐにみることができるが、知りたい分析結果が得られない点が難点であった。そこで、そのような集計表の元になっているデータファイルをそのままCD-R等のメディアをつかって分析者に提供するデータ・アーカイブが登場する。このようにデータ・アーカイブが調査データのデータ・セットの提供をおこなうことで、分析者は各自の関心にもとづいて調査データを自由に分析できるようになった。

しかし、分析者へのデータファイルの提供には時間がかかるという問題点がある。たとえばSSJデータ・アーカイブの場合、さきほども紹介したように、分析者による利用申請のあとSSJデータ・アーカイブと調査データ寄託者の承認を得てからデータが

提供される。そのためデータの利用申請をしてからデータを入手するのに2ヶ月程度を要する場合がある。このようなデータ提供方法は、利用者がデータを入手するのに時間がかかるという問題点があるにもかかわらず、海外のおおくのデータ・アーカイブでもこの方法が採用されてきた。

第3点目の問題として、分析データを提供するSSJデータ・アーカイブは日本語のみでデータを提供している点があげられる。つまり日本語の能力のある研究者でないと日本の社会調査データを利用できないのである。日本における社会調査の質は向上し、世界的に見ても高いレベルにあるとおもわれる。したがって、もし海外の研究者がそのような日本で実施された高いレベルの調査データを利用することができれば、日本についてよりいっそう知識を得ることができるのではないだろうか。このようにデータを海外に向けて発信するという観点からみれば、日本語だけでデータを提供している日本のデータ・アーカイブには問題点があるといえるだろう。

3. 社会調査データのオンライン分析

3.1 オンライン分析の登場

これまでのデータ提供方法にかわって、調査データをWeb上で直接分析できるシステムが登場した。このような調査データのオンライン分析システムは、これまでのデータ・アーカイブがもっていた問題点を解消している。調査データを直接分析できるので、分析者が知りたい分析結果を得ることができるのは当然である。そして調査データをWeb上で分析できるから、知りたいことをすぐに分析することもできる。ここでは、社会調査データのオンライン分析を可能にした代表的なシステムとしてSDA（Survey Documentation and Analysis）とNESSTAR（Networked Social Science Tools and Resources）を紹介する。

3.1.1 SDA

SDAは調査データの文書化とオンライン分析のための一連のプログラムである。この一連のプログラムは、カリフォルニア大学バークレー校のComputer-assisted Survey Methods Program（CSM）が開発したシステムである[6]。現在、このSDAはアメリカ合衆国の複数のデータ・アーカイブで試験的に導入されている。SDAをつかったアーカイブで公開されている代表的なデータにはGeneral Social Survey（GSS）データがある。このようなオンライン分析システムの登場が社会科学における調査研究において画期的であったことは、このSDAがAmerican Association for Public Opinion Research (AAPOR)の「Innovators Award」やAmerican Political Science Association (APSA)の「Best Instructional Software Award」を2000年に受賞している点からもわかる。

SDAは、おおきくわけて3個のプログラムから構成されている。まず「Codebooks」

プログラムによって、公開する調査データのコードブックを印刷や閲覧用に作成する。コードブックとは変数の定義などが文書化されたものである。第2のAnalysisプログラムによって、データの分析をオンラインでおこなうことができる。この「Analysis」プログラムはWebブラウザで動作するように設計されたパッケージであり、クロス集計表、平均の比較、相関分析、回帰分析といった多変量解析がWebブラウザだけで実行可能である。そして第3の「Subsetting」プログラムをつかうと、データ・アーカイブ上の調査データをダウンロードすることができる。なお調査データのダウンロードの際には、SASやSPSSやStataといった統計パッケージで分析できるように変換されたデータのサブセットがダウンロード可能になっている。

3.1.2 NESSTAR

NESSTARは、インターネットによるデータ普及の基盤技術となるために開発されているシステムである[7]。このシステムはイギリス、ノルウェー、オランダのデータ・アーカイブが中心となってヨーロッパで共同開発されている。NESSTARの特徴は、専用ソフトウェアによって複数のデータ・アーカイブに保管されている調査データに関する情報を一覧できる点にある。世界中のいたるところにアーカイブのサイトが多数存在していたとしても、NESSTARをつかえば各サイトで公開されているデータに関する検索結果の一覧をとりだすことができる。

このようなNESSTARのシステムを実現するために NESSTAR Publisher、NESSTAR Server、NESSTAR WebViewという3種類の専用ソフトウェアが開発されている。手持ちのデータや関連文書をインターネット上で公開するための書式へと変換するソフトウェアが NESSTAR Publisher である。多数のデータを対象にした検索システムを実現するためには、このように共通の書式によって標準化したデータを準備することが必要である。NESSTAR Publisherによって標準化されたデータや関連文書をインターネット上のサーバーで公開・普及するためのソフトウェアが NESSTAR Server である。そしてNESSTAR Serverによってインターネット上に公開されたデータや関連文書は NESSTAR WebView というソフトウェアをつかって検索や閲覧ができる。またNESSTAR WebViewにはオンライン分析の機能がついているので、公開されているデータについてクロス集計表の作成や回帰分析をおこなうことができる。

3.2 ふたつのデータ

ここで、データ・アーカイブが収集・保管するデータにはおおきくわけて2種類ある点にふれておきたい。そのふたつのデータは、それぞれデータとメタデータとよばれている。もっとも一般的な意味での分析のための調査データ、つまりSPSSやSASで分析するデータ・ファイルといったもの、それがデータである。これについて説明は不要だろう。

もうひとつのメタデータとは、データ・アーカイブで収集・保管される個々の調査データについてのデータである。たとえば調査対象はどのような人や組織なのか。全数調査なのか標本調査なのか。標本調査であれば、どのような標本抽出法をつかって標本を抽出したのか。ほかにもさまざまな項目をあげることができるが、これらが調査データのメタデータとよばれるものにあたる。メタデータについてはDDI（Data Documentation Initiative）という仕様書を討議するための国際的なプログラムが始動しており、いくつかのデータ・アーカイブがDDIにもとづいてメタデータを提供するための準備をすすめている[8]。

調査データのオンライン分析にとって、メタデータのとりあつかいは重要な意味をもっている。データ・アーカイブに保管されたおおくの調査データのなかから、分析目的に適したデータをすばやくみつけだすためにはメタデータの情報を参照することが必要不可欠だからである。さきほど紹介したNESSTARは、専用ソフトウェアにメタデータの検索機能がそなわっている。このようなメタデータをあつかうための機能によって、NESSTARはネット上にいわば仮想アーカイブをつくりだすことが可能になった。また一方のSDAは、現時点ではメタデータの検索機能はないが、個々の調査データに関するメタデータをSDAシステムの書式からDDIに変換するためのプログラムが開発されつつある。

3.3 SDAとNESSTARの問題点

なによりも、SDAによって、GSSデータなどがWebブラウザだけでオンライン分析可能になったのはすばらしい。ただし分析の際、コードブックを参照して変数名の直接入力が必要な点など、SDAはインターフェイスに改良の余地がある。また先述したように、SDAは現時点ではメタデータの検索ができない点も問題である。

それに対してNESSTARは、データ分析だけでなくメタデータについての情報収集もオンラインでおこなうことができる。しかし、NESSTAR専用のソフトウェアをダウンロードしなければならないという問題点がある。

4. あたらしいデータ公開システムの開発

4.1 SRDQ（Social Research Database on Questionnaires）の概要

そこで、これまでのオンライン分析システムの問題点を解消するために、あたらしいデータ公開システムを開発した。SDAはメタデータの検索ができなかった。NESSTARは専用ソフトが必要という欠点があった。それに対して、わたしたちが開発したあたらしい公開システムは、専用ソフトなしで、Webブラウザだけでデータとメタデータの分析が可能なシステムになっている。

4.1.1 SRDQ の技術的側面

あらたに開発したシステムの名称は、SRDQ（Social Research Database on Questionnaires）である。このシステムでは、調査データの分析とメタデータの検索をWeb ブラウザだけでおこなうことができる。また調査票の質問文を検索する機能もある。SRDQ は URL：http://srdq.hus.osaka-u.ac.jp において、2004 年 1 月から試験的にシステムの運用を開始し、2004 年 4 月より正式に公開している。利用状況は 4 月から 9 月までの半年間に 1 万件をこえる訪問者があり、ヒット数の合計は 80 万をこえている。

この SRDQ においては、調査データの分析とメタ・データ検索の両方を Web ブラウザだけでおこなえるようにするために、独自のシステムを作製した。なぜなら、現在のところ、調査データ分析とメタデータ検索というふたつの目的を実現できる市販のシステムが存在しないからである。SRDQ を構築する過程では、市販パッケージをもちいての試行錯誤がなされたものの、SRDQ に要求される機能を十分に実現することができず、最終的には独自システムの作製をおこなうことになった。以下に、SRDQ を構成している独自のシステムについて、その概要を述べる。

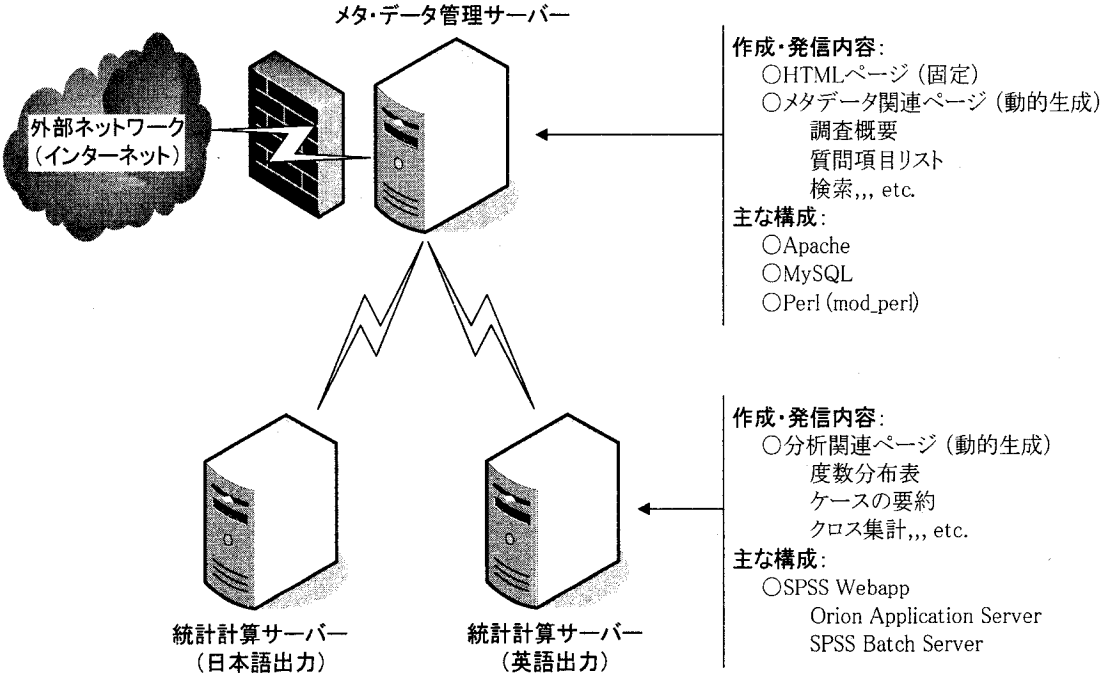

図 1　SRDQ を構成するサーバー群

この独自システムの概略は、図 1 の通りである。このシステムは、図 1 からもわかるように、調査データを分析するための部分と、メタデータを検索・表示するための部分とに、大きくわかれている。

まず調査データの分析については、SPSS（.sav）形式の調査データをシステムに登録するだけで、そのデータをWebブラウザから分析できるというシステムを作製した。このシステムではほとんどすべてのWebページが、利用者からの要求に応じて、動的に生成される。たとえば、登録されたデータに含まれる変数の一覧が自動的に読み取られ、分析ページに利用できる変数の一覧が表示される。さらに、分析ページで利用者がおこなった選択にもとづいて、各種の集計や分析がその場でおこなわれ、結果がWebページとして表示される。このようにWebページの動的生成おこなうことによって、あらかじめ管理者が必要とかんがえる集計表を大量に作成しておくといった作業は不要になった。またデータを修正するような場合にも、単にデータを差し替えるだけでよく、集計表を修正するといった作業は必要ない。なお、このようにWebページを動的に生成することで、何らかの機能をWebブラウザだけで利用できるようなシステムは、一般にWebアプリケーションと呼ばれている。

この調査データ分析のためのシステムは、Webアプリケーション作成用の開発キット（SDK: Software Development Kit）である、SPSS社のWebAppをもちいて作製されている。SPSS WebAppでは、SPSS Batch Serverによって分析が実行され、その結果がOrion Application Serverを通してWebページとして送信される。したがって、SPSS Batch Server（SPSS Base System）でおこなえる分析はすべて、SPSS WebAppでもおこなえることになる。ただし、それぞれの分析を実行可能にするためにはカスタマイズが必要になるので、現在SRDQで提供している分析手法は後述する10種類に限られている。このカスタマイズについては、以下に示すようなプログラミング言語および記述言語をもちいておこなわれる。

- HTML
- Java Script
- XML
- JAVAおよびJava Servlet、Java Beans、JSP
- SPSS Syntax

つぎに、メタデータの検索や表示をおこなうシステムであるが、これについてもWebページを動的に生成するシステムを作製し、利用している。すなわち、メタデータについても、データベース・システムに情報を登録しておくだけで、質問項目・社会調査の一覧ページや検索ページが自動的に生成される。このシステムでは、高速性と堅牢性を追求したデータベース・システムであるMySQLを採用している。そして、Perl（mod_perl）というプログラミング（スクリプティング）言語によって記述されたサブ・システムが、MySQLに登録されたデータを取り出してWebページとしての体裁に整形する。最終的には、Webページを発信するためのシステムであるApacheを通じて、利用者に情報が送信される。

以上で述べたように、SRDQは統計分析をおこなうためのシステムと、メタデータ

の検索や表示をおこなうためのシステムという、2種類の別個のシステムによって構成されている。

図1からもわかるように、これら2種類のシステムは異なるサーバー機で稼働しているので、ふたつのシステムは物理的にも分断されていると言えよう。しかし、実際の利用画面ではこれらふたつシステムが継ぎ目なしに接合されるように両方のシステムを設計した。たとえば後掲する図5をみれば、分析のためのオプションを選択するページや分析結果が表示されるページには、メタデータを表示するためのリンクがつねに表示されている。よって、分析ページから1クリックで、調査概要や質問項目のリストといったメタデータのページへ移動できる。もちろん、調査概要や質問項目のリストが表示されるページにも移動用のリンクがあり、1クリックで分析ページに移動することができる。このように2種類のシステムは接合されているので、SRDQを利用している際に、画面上からその背後に完全に分離したふたつのシステムが存在すると想像するのはおそらく困難だろう。SRDQの通常の利用は、トップページがおいてあるメタデータ管理サーバーからはじまり、利用者の要求に応じて、それぞれのサーバーが適宜稼働することになるのだが、図1のように複数のサーバー機が稼動していることに利用者は気づかないだろう。

最後にSRDQを構成するハードウェアについて述べておこう。統計計算サーバーとしてはCPU（Xeon 3.2GHz）をふたつ搭載し、メモリについても2G bytesを備えたサーバー機を使用している。日本語の分析ページと英語の分析ページを生成・送信するために、それぞれ専用のサーバー機をもちいている直接の理由は、SPSS WebAppの仕様によるものである。もっとも、サーバー機にかかる負荷を2台に分散するという意味でも、現在の2台での構成は有効に機能していると言える。つぎにメタデータ管理サーバーについては、CPU（Pentium III 1GHz）も1個でメモリも256M Bytesと、やや性能が劣っている。だが統計計算をおこなう必要がなく、さらにMySQLとmod_perlをもちいた処理は高速でなので、現時点でこのサーバー機は十分なパフォーマンスを発揮している。

4.1.2 収録データ

現在、SRDQに登録されているデータには、統計分析できるデータとしては、2001年に実施された「情報化に関する全国調査（Japan Survey on Information Society：JIS）」と1955年にはじまり1995年まで10年ごとに実施されている「社会階層と社会移動（Social Stratification and Mobility：SSM）全国調査」（ただし威信票を除く）がある。いずれも全国を調査対象とし、無作為抽出にもとづいた標本の大きさが1,000以上の質の高い調査データである。なおこれらのデータは、海外の研究者等も利用できるように、調査票、質問文、データのラベル等の情報提供を日本語だけでなく英語でもおこなっており、日本の社会調査データを海外に発信するという意義も大きい。

質問文、調査概要などのメタデータの登録は 113 件で、おもに科学研究費補助金の助成によって実施された 1990 年代後半以降の社会学者の調査を中心に収録している。検索は、113 件全部の検索と領域別の検索ができる。この領域とは日本社会学会が会員の専門領域の分類にもちいているものであり、登録件数のおおい領域は、「地域社会（コミュニティ）」の 34 件、「家族」の 32 件、「社会心理・社会意識」の 30 件、「教育」の 26 件、「階級・階層・社会移動」の 25 件などである。

4.1.3 メタデータ分析画面

図 2 は、SRDQ のトップページである。左下の枠内「分析可能な社会調査データ」が、ブラウザで分析できるデータの一覧である。それぞれの調査に関して、「調査概要」と「分析」のページがある。中央に「質問文・選択肢の検索」という枠があり、ここから収録した調査票の質問文と選択肢が検索できる。検索方法は、「質問文と選択肢の両方を検索」、「質問文だけを検索」、「選択肢だけを検索」という 3 種類の方法があり、検索は、「AND 検索」と「OR 検索」ができる。その下に「社会調査の検索」という枠があり、ここで登録されている全 113 件の調査について調査の名称や代表者名による検索ができる。

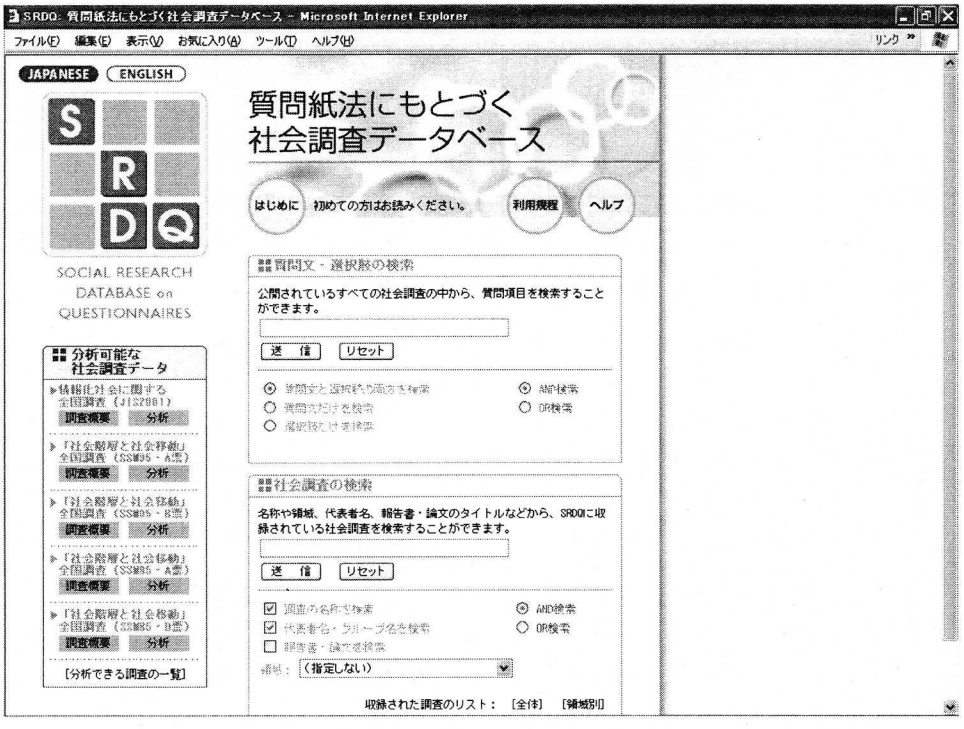

図 2　トップページ

図3は、「調査概要」のページである。

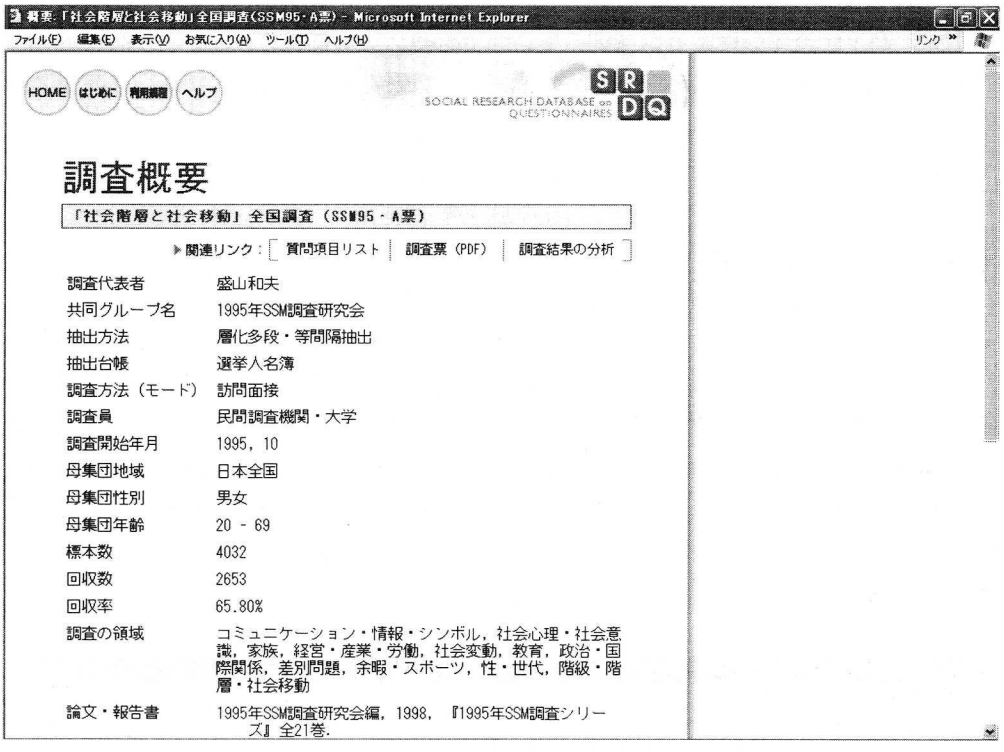

図3　調査概要のページ

このページの上部では、調査の基本的な情報が表示される。項目は、調査名称、調査代表者、共同グループ名、抽出方法、抽出台帳、調査方法（モード）、調査員、調査開始年と月、母集団地域、母集団性別、母集団年齢の下限と上限、その他の母集団の情報、標本数、回収数、回収率、調査の領域、論文・報告書である。これらのデータはデータベースで管理されており、該当しない項目、たとえば年齢の上限がない場合などの非該当部分は表示されない設定になっている。

下段において、質問文と選択肢の検索ができる。この場合の検索範囲は、表示されている調査だけに限定される。

また、このページ上部には「関連リンク」という項目がある。このリンクをつかえば、「調査概要」のページから1クリックで質問項目リストを表示することができる。また一部の調査データについては、この「関連リンク」を通じて調査票のPDFファイルの閲覧が可能である。また分析可能な調査データについては、この「関連リンク」から調査結果の分析画面に移動することできるようになっている。

図4は、質問文の検索の欄に「お仕事」と入れた場合の検索結果のページである。

収録されている全113件の調査票を対象に検索がおこなわれ、該当する件数と該当し

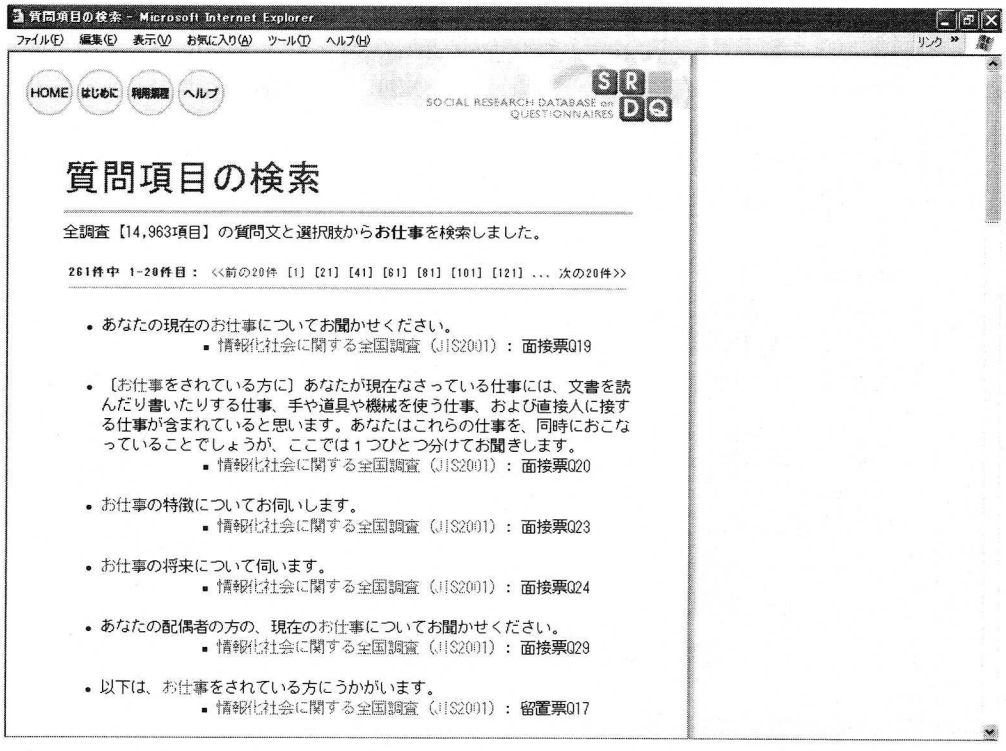

図4　検索結果のページ

た質問文、その質問文が含まれる調査名と問番号が表示される。検索語は、赤字で表示される。該当した質問文の直後の行にある調査名（一番上の例では、「情報化社会に関する全国調査(JIS2001)」）をクリックすると図3に示したような「調査概要」のページが表示される。問番号（一番上の例では「面接票Q19」）をクリックすると、該当する質問文（面接票Q19）が赤字で示され、それにつづく質問文を見ることができる。画面を上にスクロールさせれば、Q18からさかのぼって、それ以前の質問文を見ることができる。つまり検索結果のページから、この例の場合だと「情報化社会に関する全国調査(JIS200)」の全質問文を見ることができる仕組みになっている。

4.1.4 調査データ分析画面

つぎに、調査データをWebブラウザ上で分析する「分析」画面の説明にうつる。

「トップページ」（図2）の左下に「分析可能な社会調査データ」表示されている。この「分析可能な社会調査データ」からいずれかの「分析」をクリックすると図5のような画面があらわれる。

図5　調査結果の分析ページ

「調査結果の分析」というタイトルの下に分析するデータ名「情報化社会に関する全国調査（JIS2001）」が青字で表示される。その下の行には、このJIS2001調査の「調査概要」（図3参照）と「質問項目リスト」へのリンクがある。その下の「分析メニュー」がSRDQでつかうことができる分析手法である。分析は先述したとおり、SPSSのWebAppを使用している。WebAppでは、SPSS Baseとオプションに含まれるすべての機能を利用することができる。しかしそのためにはカスタマイズが必要であり、現在のSRDQにおいては、「分析メニュー」にあげられている度数分布表、ケースの要約、クロス集計、Webキューブ、T検定、一元配置の分散分析、線型回帰、相関係数、偏相関係数、因子分析の10種類の統計手法とその中の一部のオプションが利用可能になっている。

ここでは、クロス集計表の分析の方法を紹介する。

図5の「分析メニュー」から「クロス集計」を選ぶと図6のような画面があらわれる。

左の枠内に分析につかうことができる変数が表示されている。この枠内を下にスクロールしながら、分析したい変数を選んで指定する。指定は、行変数、列変数を指定すると普通のクロス表になり、さらに層変数を指定すると3重クロス表が作成される。行変数、列変数、層変数とも複数指定することができ、それぞれの枠内で上から順に組み合わせて実行される。この順番を変更したいときには、すぐ右側の「up」と「down」の

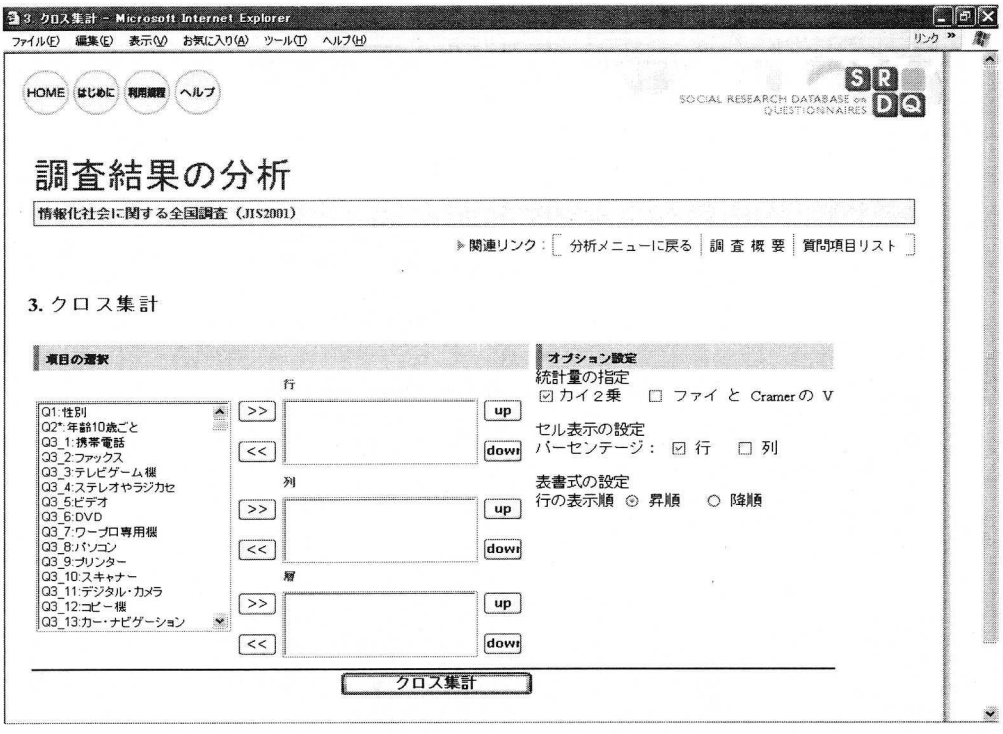

図6 クロス集計

ボタンをつかえばよい。

さらに右側の「オプション設定」の欄では、統計量として「カイ2乗」値と「ファイとCramerのV」の値を指定することができる。各セルには、「行」をチェックすれば行パーセントが、「列」をチェックすれば列パーセントが表示される。また行の選択肢の表示順を昇順、降順のどちらにするかを指定することもできる。なおデフォルトでは、カイ2乗、行パーセント、昇順が選択される設定になっている。

行変数に「Q1:性別」、列変数に「留置票Q1:インターネットやパソコン通信をご存じですか」を選び、「カイ2乗」、「行」パーセント、「昇順」を指定した場合、分析結果は図7のようになる。

いちばん上にある「クロス集計へ戻る」をクリックすると、図6へ戻る。つぎの「処理したケースの要約」からは有効ケース数とそのパーセントが分かる。つぎの枠がクロス集計表の結果である。そしてその下に「カイ2乗検定」の結果が表示される。他の分析もふくめて、分析の方法、出力の形式は、パソコンで使われるSPSSとほぼ同じ形式である。また、このページにも「関連リンク」という項目があり、メタデータのページに1クリックで移動することができる。

図7　クロス集計の出力

4.2 データ公開とプライバシー保護

4.2.1 情報漏洩の統計学的な制御・制限

調査データを調査主体の管理を離れた公共のものとするとき、被調査者のプライバシーを守りながらデータの情報をできるだけ活用することが必要になってくる。もちろんSRDQのようなインターネットを通じた調査データの公開についても調査データを公共のものとする活動にふくまれるから、被調査者のプライバシーを守る方法についての対策が必要である。そのための方法論が情報漏洩の統計学的な制御・制限（Statistical Disclosure Control あるいは Statistical Disclosure Limitation）である。このような方法論は、近年、ミクロ統計データの公開がすすむにつれて研究がさかんになっている（Doyle et al. 2001; 松田ほか 2000; 竹村・廣松 2002; Willenborg & Waal 1996, 2001）。

情報漏洩をふせぐための統計学的な制御・制限には複数の方法がある。識別されると被調査者のプライバシーに関わるような変数をデータ・ファイルから削除する方法やクロス集計表で実現値の少ないセルを削除・統合してしまうような方法があり、これらは Suppression とよばれている。

あるいはリコードをおこなってプライバシーの漏洩をふせぐ方法もある。たとえば年齢の変数が2歳きざみだったものを5歳きざみにリコードする（global recoding）、世帯収入のような変数で最上位、最下位のカテゴリーをひろげる処理（それぞれtop recoding、bottom recoding）が提示されている。これらがリコードによってプライバシー漏洩をふせぐ方法である。

また原データの適切な統計的標本を抽出し、その抽出標本を公開することで原データに含まれている情報の漏洩を制御する方法もある（Sampling）。この場合、もし原データそのものが標本ならば、そのデータは自ら情報漏洩を制御することもあるとかんがえられる。

4.2.2 事実上の匿名化

たしかに調査データの公開にあたって被調査者のプライバシーの漏洩をふせぐ統計学的方法論について研究は蓄積されつつある。しかしデータを公開する以上、完全な匿名化は不可能であるというのもまぎれもない事実である。なぜなら、もしデータに対して完全な匿名化をおこなうならば、そのデータからは性別、学歴、職業、収入といった情報が削除されてしまうからである。社会科学において必要不可欠なそういった情報が削除されたデータでは公開する意義はない。また一般の人々も性別、学歴、職業、収入に関連した情報を知りたいであろう。

完全な匿名化が不可能である場合、そこで重要になるのが事実上の匿名化という概念である。ドイツ連邦統計局では、1980年代後半から1990年代前半にかけて、具体的な統計データから被調査者がどれくらい識別可能かについて実験をおこなっている（Müller, Blien, Wirth, 1995）。さきに紹介したような情報漏洩の統計学的な制御・制限に関する研究では、人工的に生成されたデータにもとづいて被調査者識別の危険性が議論されている。しかしドイツ連邦統計局によるこの実験結果では、実際の統計データの場合、人工的なデータよりも被調査者の個体識別の危険性ははるかに小さいと報告されている。

統計データから被調査者の識別をおこなうためには、その統計データのほかに、全数調査などにもとづいた識別用のファイルが必要である。そして統計データと識別ファイルの照合をおこなった結果、はじめて被調査者の個体識別が可能になる。たとえば、公開された統計データからA地域に年収3000万円の男性がいるという情報が得られた場合をかんがえる。この男性が誰であるかを識別するためには、まずA地域在住者の年収に関する個人情報が掲載された識別ファイルを入手しなければならない。そして公開された統計データから得られた情報とこの識別ファイルを照合した結果、はじめてA地域における年収3000万円の男性が誰であるか識別される。しかし一般的にかんがえて、このように被調査者の識別のために統計データと識別ファイルを照合するには、識別ファイルの入手やデータとファイルの照合など、照合者は莫大なコストを支払わなければならない。したがって実際の統計データにおいて被調査者の個体識別の危険性は小

さいとかんがえられるのである。

このように被調査者の識別に莫大なコストがかかる場合であれば、その調査データにおいて事実上の匿名化が実現しているとかんがえることができそうである。しかし、このことを逆にかんがえると、被調査者の識別が比較的小さなコストで実現可能なデータにおいて事実上の匿名化は破綻しているとみなされるべきだといえる。あまりコストをかけなくても被調査者の識別が実現するデータとは、変数が小さなサブ・カテゴリーにわかれているデータや、細かい地域情報がふくまれているデータである。このようなデータは、どちらも識別ファイルの入手・照合を容易にするという点で被調査者識別のコストを下げている。

低いコストで被調査者が識別可能になることをふせぐ、いいかえるならば公開データにおける事実上の匿名化をまもるためには、以下にあげる対策が最低限必要であろう。まず、変数が小さなサブ・カテゴリーにわかれているデータに関しては、リコードをおこなってサブ・カテゴリーを統合したデータを公開するという対策が必要である。これは先述した global recoding や top recoding、bottom recoding といった方法である。また、細かい地域情報、たとえば被調査者の住所・氏名に直接むすびつく ID や地点番号などを除外したデータ、さきほど登場した用語でいえば Suppression をおこなったデータを公開データとするといった対策も必要である。SRDQ で公開される調査データに関しても、公開データの事実上の匿名性を守るためにこれら最低限の対策が施されている[9]。

4.3 このシステムの意義

では SRDQ による調査データの公開には、どのような意義があるのだろうか。ここでは、研究者に対する意義、調査対象者に対する意義、そして教育における意義の 3 点をとりあげておく。まず研究者に対しての意義としては、すぐに知りたい分析結果が得られるシステムは研究に取り組むうえで便利だという点をあげることができる。また SRDQ は一部ではあるが英語でも利用できるようになっているので、もちろん日本語の読めない海外の研究者も手軽にデータ分析が可能である。

いつでも誰でもデータ分析できる SRDQ の特徴は、これは調査対象者にとっても価値がある。これまでデータ・アーカイブによるデータ公開はもっぱら研究者が対象だった。しかし調査結果を知りたいとのぞむ調査対象者はおおく、調査結果のデータをオンライン分析が可能な形式で公開することは、調査対象者に対する調査報告としても意義がある。

教育面でもあたらしい意義がある。データ分析のための特別なパッケージがいらず、Web ブラウザだけで調査データの分析が可能できるので、学生が社会調査法の学習に取り組みやすくなる。SPSS や SAS などのインストールされたパソコンがおかれた教室だけでなく、自宅で統計分析の練習をおこなうことも可能である。さらに公開されて

いるデータがアメリカのGSS調査ではなく、「情報化に関する全国調査」ように日本でおこなわれた社会調査データであれば、学生の関心も高まるだろう。これは、高等教育の枠組みをこえて、中等教育や生涯教育の場における社会学教育でも有用性を発揮する可能性を秘めている。

5. 今後の課題

あたらしい調査データ公開システム、SRDQは研究だけではなく、教育上も意義をもったシステムであり、また、調査データを対象者に公開するという目的にも役立つものである。しかし同時に、動きはじめたばかりのシステムでもあり、これから乗り越えていかねばならない課題もいくつかある。そこで本稿のしめくくりに、SRDQにとっての今後の課題についてふれておきたい。

まず公開データの利用者が参照できるマニュアルを整備する必要がある。これは現在SRDQ上でHelp機能として備えている、システムを利用するために最低限必要な操作マニュアルだけでは十分といえない。これまでのデータ・アーカイブがおもに研究者を対象として調査データを公開してきたのに対して、SRDQは調査対象者をはじめとする一般の人々の利用をも想定しており、そのためにWebだけで利用できるシステムを作製している。そのような一般の人々が調査データを自分自身で集計して、その結果を理解するためには、最低限度の統計学的知識が必要とならざるをえない。そこで市民に向けた統計学的知識の提供ならびに利用者から寄せられるデータの分析手法や分析結果に関する質問・疑問への対応といった利用者に対するコンサルティング機能が、SRDQの場合は重要で必要不可欠である。どのような形式でそのようなコンサルティング機能を提供するか、これを今後の課題のひとつとしてあげることができる。教育上の利用も含めて、理解しやすいテキストの作成やWeb上の掲示板のような機能をつかって質問や疑問に答えていくことがかんがえられる。

そして、このシステムが発展するためには、質の高い社会調査データのさらなる収集と公開が必要である。いくらシステムがもっている機能に有用性があっても、そのシステムを通じて提供される調査データが貧弱なものであれば、このSRDQというデータ公開システムの価値はないであろう。現在SRDQでは、無作為標本抽出法にもとづいた全国調査である「情報化に関する全国調査（JIS）」や「社会階層と社会移動（SSM）全国調査」の調査データを公開している。今後もこれらに劣ることのない質の高い社会調査データの収集と公開をおこなっていくこと、これもSRDQの課題である。

最後に調査データの公開とプライバシーの問題である。インターネットを通じた調査データの公開や分析は、社会科学の分野におけるインターネットの社会的影響（DiMaggio et al. 2001）の産物である。そのためSRDQによる調査データの公開も、IT革命によるあたらしい知的所有権やプライバシーの問題に抵触せざるをえない。知的所有権に関

しては、データ・アーカイブの作成者の権利がどの程度で認められるかという問題がある。SRDQ は、その膨大で有益な出力に対して、わずかな労力で作成されているといえるが、それでもデータ作成の人的労力はかなりのものであり、またこれは相当程度の社会調査の知識がないとできないものである。このような労力に対して、特に学術上の業績としての評価がともなわないと、誰もデータ・アーカイブの作成にたずさわらないであろう。プライバシーに関しては、調査データには調査対象者に関するおおくの情報がふくまれているという問題がある。SRDQ はそのようなデータを加工し、研究者による科学的分析を保障する程度には原データの情報をふくみつつ、研究者にかぎらずだれでもアクセス可能で、なおかつデータからのプライバシー漏洩による損害を引き起こさない対策をとらなければならない。また学術上は、調査データの 2 次分析、オンライン分析が認められているが[10]、調査対象者へのインフォームド・コンセントはこれまで通りのかたちでかまわないかどうかも議論の対象となるであろう。これらの点について、これから議論をふかめていく必要があるだろう。

〔付記〕本稿は平成 15-16 年度科学研究費補助金研究成果公開促進費 158052「質問紙法にもとづく社会調査データベース」（代表者：直井優）による研究成果の一部である。

注

1) IFDO の URL は http://www.ifdo.org/（2004 年 10 月 1 日現在）である。
2) CESSDA の URL は http://extweb.nsd.uib.no/cessda/（2004 年 10 月 1 日現在）である。
3) SORD の URL は http://www.sgu.ac.jp/soc/sordhp/（2004 年 10 月 1 日現在）である。
4) SSJ データ・アーカイブの URL は http://ssjda.iss.u-tokyo.ac.jp/（2004 年 10 月 1 日現在）である。
5) ただし調査概要情報を検索・閲覧するときは、閲覧者確認のためメール・アドレスの入力が必要である。
6) SDA の URL は http://csa.berkeley.edu:7502/（2004 年 10 月 1 日現在）である。
7) NESSTAR の URL は http://www.nesstar.com/（2004 年 10 月 1 日現在）である。
8) DDI の仕様書は現在も討議が進行しているため確定していないところがあるが、その一部はたとえば ICPSR の「Data Documentation Initiative(DDI)」で告示されている。このサイトの URL は http://www.icpsr.umich.edu/DDI（2004 年 10 月 1 日現在）である。
9) ここで参考のためにも、GSS データのオンライン分析用データについてふれておく。現在 GSS データは、さきほど紹介した SDA をつかってインターネット上でオンライン分析可能なかたちで公開されている。しかしその公開データには、リコード等のかたちで被調査者のプライバシー漏洩をふせぐ対策がおこなわれた形跡はない。公開データにおける情報漏洩の統計学的な制御という観点からみた場合、無作為抽出標本にもとづく

GSSデータは標本データであるというだけで情報漏洩の制御がおこなわれていると判断することは不可能ではない。ただしGSSデータのインターネット上の公開にあったって、そのような積極的な判断が下されたかどうかは不明である。

10) 2003年11月に設立された社会調査士資格認定機構の定める「社会調査倫理綱領」では、その第4条において、このことが認められている。

参考文献

DiMaggio, Paul, Eszter Hargittai, W. Russell Neuman, & John P. Robinson, 2001, "Social Implications of the Internet," *Annual Review of Sociology,* 27: 307-336.

Doyle, Pat, Julia I. Lane, Jules J. M. Theeuwes, & Laura V. Zayatz eds., 2001, *Confidentiality, disclosure, and data access : theory and practical applications for statistical agencies,* Elsevier Science.

松田芳郎・濱砂敬郎・森博美編, 2000,『統計調査制度とミクロ統計の開示』(講座ミクロ統計分析 第1巻) 日本評論社.

Müller, Walter, Uwe Blien, & Heike Wirth, 1995, "Identification Risks of Microdata: Evidence form Experimental Studies," *Sociological Methods & Research,* 24(2): 131-157.

佐藤博樹・石田浩・池田謙一編, 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.

Scheuch, Erwin K., 1990, "From a Data Archive to an Infrastructure for the Social Sciences," *International Social Science Journal,* 123: 93-111.

竹村彰通・廣松毅, 2002「統計制度：電子社会における経済と社会の計測」奥野正寛・竹村彰通・新宅純二郎編『電子社会と市場経済：情報化と経済システムの変容』(ライブラリ電子社会システム 2) 新世社：309-336.

Willenborg, Leon, & Ton de Waal, 1996, *Statistical disclosure control in practice* (Lecture notes in statistics ; 111) Springer-Verlag.

Willenborg, Leon, & Ton de Waal, 2001, *Elements of statistical disclosure control* (Lecture notes in statistics ; 155) Springer.

The Open Data Archive of Social Research data : SRDQ

Taichi MATSUKAWA, Koichi HIGUCHI and Akira KAWABATA

The data archive exists for the collection, editing, process, storage, and dissemination of social research data. Looking back through history, a social research data archive was set up in America shortly after World War II. This archive is now held at the "Roper Center." After that, data archives were created throughout North America and Europe. Currently, the majority of these data archives are members of international organizations like IFDO and CESSDA for easy linkup.

The data stored in these archives fall into two categories: research data, which concentrates on statistical analysis, and metadata. Metadata is documentation pertaining to research and data, and includes technical information such as data collection methods and sampling methods, and general information such as, the subjects of research, the period of research, and the sponsors.

Broadly speaking, data archives are significant for two reasons. These are: (1) the research advantage, which enables existing data to be used effectively, and (2) the educational advantage of learning social research methods using good-quality research data.

Existing data archives, and in particular Japanese data archives, however, have traditionally had the following problems. One is that tabulated results and the results of analysis that would be of interest to researchers cannot be easily obtained. The research data is not sent to us upon application to an archive, but it takes from several weeks to one or two months, sometimes even longer. Another problem is that the collected metadata is limited to brief, generalized information such as the research title and research subject. Even when searching the research data in which a researcher is interested, he/she can only search using this limited information. For this reason, we cannot say that the features of today's excellent databases can be used to the full.

It was for this reason that we created the new data archive, SRDQ: Social Research Database on Questionnnaires. The two main advantages of SRDQ are as follows: Firstly, the data can be analyzed on the website. No special software is necessary; the data can be analyzed using an ordinary web browser. Another special feature of SRDQ is that metadata can be fully collected and then searched. This way, all pertinent items to an archive's user are logged, not just the outline topics like research title and main theme, and these can be searched freely. Further, depending on the research, the questionnaires can also be browsed in PDF format, as can the contents of the written reports and the reports themselves in PDF format.

SRDQ is also planning to address some issues. They are: (1) to offer the manual which a user without statistical knowledge can refer to, (2) to further collect good-quality social research data and (3) to make social research data open to the public without conflicting with a subject's privacy.