© OECD, 2000.

© Software: 1987-1996, Acrobat is a trademark of ADOBE.

All rights reserved. OECD grants you the right to use one copy of this Program for your personal use only. Unauthorised reproduction, lending, hiring, transmission or distribution of any data or software is prohibited. You must treat the Program and associated materials and any elements thereof like any other copyrighted material.

All requests should be made to:

Head of Publications Division
Public Affairs and Communication Directorate
2, rue André-Pascal, 75775 Paris
Cedex 16, France.

Social Sciences for a Digital World

BUILDING INFRASTRUCTURE AND DATABASES FOR THE FUTURE

OECD

ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT

ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT

Pursuant to Article 1 of the Convention signed in Paris on 14th December 1960, and which came into force on 30th September 1961, the Organisation for Economic Co-operation and Development (OECD) shall promote policies designed:

- to achieve the highest sustainable economic growth and employment and a rising standard of living in Member countries, while maintaining financial stability, and thus to contribute to the development of the world economy;
- to contribute to sound economic expansion in Member as well as non-member countries in the process of economic development; and
- to contribute to the expansion of world trade on a multilateral, non-discriminatory basis in accordance with international obligations.

The original Member countries of the OECD are Austria, Belgium, Canada, Denmark, France, Germany, Greece, Iceland, Ireland, Italy, Luxembourg, the Netherlands, Norway, Portugal, Spain, Sweden, Switzerland, Turkey, the United Kingdom and the United States. The following countries became Members subsequently through accession at the dates indicated hereafter: Japan (28th April 1964), Finland (28th January 1969), Australia (7th June 1971), New Zealand (29th May 1973), Mexico (18th May 1994), the Czech Republic (21st December 1995), Hungary (7th May 1996), Poland (22nd November 1996) and Korea (12th December 1996). The Commission of the European Communities takes part in the work of the OECD (Article 13 of the OECD Convention).

© OECD 2000
Permission to reproduce a portion of this work for non-commercial purposes or classroom use should be obtained through the Centre français d'exploitation du droit de copie (CFC), 20, rue des Grands-Augustins, 75006 Paris, France, Tel. (33-1) 44 07 47 70, Fax (33-1) 46 34 67 19, for every country except the United States. In the United States permission should be obtained through the Copyright Clearance Center, Customer Service, (508)750-8400, 222 Rosewood Drive, Danvers, MA 01923 USA, or CCC Online: http://www.copyright.com/. All other applications for permission to reproduce or translate all or part of this book should be made to OECD Publications, 2, rue André-Pascal, 75775 Paris Cedex 16, France.

FOREWORD

This report brings together a selection of the contributions presented at the workshop on the infrastructure requirements of the social sciences, entitled "Social Sciences for a Digital World: Building Infrastructure for the Future", co-organised by the Social Sciences and Humanities Research Council (SSHRC) of Canada and the OECD, and hosted by the former in Ottawa on 6-8 October 1999. The workshop was attended by high-level experts and policy makers from OECD countries and international organisations.

The Ottawa Workshop was the first in a series of international workshops on the future of the social sciences to be hosted by OECD Member countries, and was approved by the OECD Committee for Scientific and Technological Policy as a follow-up to the conference on social sciences held in April 1998 in Paris. The second workshop in the series (Bruges, 26-28 June 2000) will focus on the contribution of the social sciences to knowledge and decision making for public policy design and implementation. The third (Tokyo, 29 November-2 December 2000) will focus on strengthening the role of the social sciences in relation to innovation, and the final workshop (Portugal, May 2001) will draw the main conclusions of the programme.

The report is published on the responsibility of the Secretary-General of the OECD.

TABLE OF CONTENTS

THE CHALLENGE OF BUILDING INFRASTRUCTURE IN THE SOCIAL SCIENCES: OPENING REMARKS

by

Marc Renaud
President, Social Sciences and Humanities Research Council of Canada

I would like to take this opportunity to welcome you all to this Workshop on "Social Sciences for a Digital World: Building Infrastructure for the Future", organised by the Social Sciences and Humanities Research Council of Canada and the OECD. Our discussions will be much facilitated by the participation of so many first-class experts from around the world, and I sincerely hope that we can come out of this meeting with a clear understanding of the common challenge we face and a concrete set of recommendations for the OECD and for our national governments. Last night, we had a chance to meet and talk with each other informally, and to hear Dr. Soete and Minister Manley describe what they see as the challenges and expectations confronting the social sciences. And now, it's time to get down to the job of defining specific issues and actions to move the social science agenda forward.

Where this workshop fits in the larger scheme of things

In broad-brush terms, it seems to me that social scientists share a common, passionate belief in the power of rational work to decipher human society and to help build a better world. It also seems clear that we are new to the "infrastructure game", in the sense that, unlike bench researchers, we do not have a long tradition of working with collective tools. In a way, the social sciences have been at the bottom of science's totem pole, much as biology was in the 1950s before Crick and Watson's discovery of DNA. However, the advent of global information technology – our functional equivalent of DNA – is ushering in sweeping changes, dramatically altering perspectives and opening up whole new vistas of opportunity.

Today, history offers us a unique opportunity to fashion the future. Rather than adopting tools invented by engineers and computer specialists, we ourselves can actually design, build and use the tools we need. But, for this dream to become reality, we have to act now. To fulfil our needs as researchers, but also to fulfil the needs of society, we have to build the necessary infrastructure which will serve to efficiently collect, transmit and make use of knowledge.

Today, more than ever, we need to work on global societal issues such as human development, the creation and redistribution of wealth, health, environmental change, new ways of understanding how communities function, the development of a sense of social and international responsibility, reducing social conflicts, and many others. These are problems experienced by all our countries, although to differing degrees. And globalisation is forcing us to look for solutions that are not confined within national borders. In this context, the social sciences must respond to a demand for more and more complex sources of knowledge from government and elsewhere, including the international organisations.

As one of these organisations, the OECD has a key role to play in promoting and facilitating the international co-operation that is indispensable if we want to rise to the challenge of solving the social and economic problems which face the world today. The OECD offers a forum in which governments can discuss, elaborate and implement social and economic policy. Through its committees, OECD Member countries can exchange experiences, seek solutions to their common problems and co-ordinate their domestic and international policies.

The work of the committees is largely based on a thorough understanding of social and economic trends. Much of the underlying research and analysis of fundamental issues is furnished by the OECD Secretariat's multidisciplinary teams of experts. Their analyses are underpinned by policy data collected by social science specialists in both OECD and non-member countries. The OECD – and the countries themselves – must also have access to international data.

The social sciences have an unprecedented window of opportunity to shape the tools – both intellectual and technological – that will structure the reality of the future. If we can put our minds to harnessing these new powers of technology effectively, we will be able to do comparative research on a hitherto unimaginable scale. We will be able to unravel many pressing questions in society – such as as why people in one place die earlier than those living in another, what helps some kids develop and succeed better than other kids, and so on.

The creation of the Canada Foundation for Innovation provides a case in point, showing the importance of being ready to respond when opportunities appear. When our federal government launched its now CDN 1 billion fund to provide support for research infrastructure at Canadian universities, hospitals and similar institutions, the natural sciences, health sciences and engineering were able to quickly mount specific proposals and attract the required partners. Leaving aside the fact that it is a greater challenge to find private sector partners for social science research, researchers in our disciplines were slower to respond, largely because we haven't spent much time thinking about and articulating our collective infrastructure needs, either now or for the future.

To further our own quest for knowledge and help build a civic, caring and creative society, we need to be ready for a future which is coming at us at dizzying speed. To meet the demands of that future, we are called upon to "reinvent" the social sciences in some pretty fundamental ways. We must be able not only to move beyond traditional disciplinary and geographic boundaries. We must also build new capabilities and reflexes of co-ordination, collaboration and communication – including among researchers, both nationally and internationally, and also between researchers and people from other walks of life.

It is my fervent hope that we will come out of this Workshop with:

- A clear understanding of common goals for maximising the outputs of the social sciences.
- A mapping out of all the possible ways we could go about meeting these goals.
- Some concrete ideas of specific actions we can start taking within individual countries and at the international level.

As I said, that's no mean task. But I think we're up to it. We have an incredible array of talent and expertise gathered in this room. Let's put it to full use. To make real breakthroughs, we have to be willing to take risks, including being willing to dream in colour and put forward some ideas that may initially sound somewhat crazy. Let's take those risks. Let's go for it!

Chapter 1

SOCIAL SCIENCES FOR A DIGITAL WORLD: BUILDING INFRASTRUCTURE FOR THE FUTURE

by

David Moorman
Policy Analyst, Social Sciences and Humanities Research Council of Canada

Introduction

Is it possible to conduct a truly global social survey? Can effective international standards for data collection and documentation be established? How can the connections between the social science research community and policy makers be improved? These were just a few of the complex questions addressed at the recent international workshop on "Social Sciences for a Digital World: Building Infrastructure for the Future". Hosted by the Social Sciences and Humanities Research Council of Canada and co-sponsored by the Canada Foundation for Innovation, the Organisation for Economic Co-operation and Development (OECD) and the National Science Foundation, this workshop brought together 57 representatives from 18 countries to explore common issues, problems and possibilities in the future development of social sciences research infrastructure.

Among the proposals that emerged from the workshop were:

- Create a digital and interactively evolving "Ottawa Manual" that would articulate standards for the collection and documentation of statistical data, and define best practices for organisations and agencies involved in providing resources for social science research.
- Implement a complex longitudinal World Social Survey that could capture vital social development information from a large number of countries simultaneously.
- Establish a large-scale international data archive that would take advantage of new communications technologies to combine a central depository with regional or national nodes and broad Web-based access.
- Develop internationally networked knowledge transfer centres that would provide convenient and timely public access to research results and infrastructure.
- Create an OECD Working Group to advance the exploration of technical and legal issues surrounding the development of data management practices.

During the workshop a clear consensus emerged that social science itself must become more visible, more involved and more respected if its practitioners are to address transnational social problems. Participants repeatedly emphasised that many of the issues identified require international comparative investigation and analysis if they are to be effectively understood. They were also well aware that most people, in the developing world in particular, lack access to modern information technologies and that this is creating a "digital divide" that must be overcome if broadly applicable cross-national research is to have a significant impact on the lives of ordinary people. Underlying the discussions was the realisation that while many social problems are becoming "globalised", there is very little political will to develop the research infrastructure necessary to address these problems.

The participants defined social science infrastructure as the collective structures that enable, enhance, embody and structure research. These include data itself, facilities for its effective and efficient management, research tools such as computer hardware and software, activities that facilitate communication both within the research community and beyond, and the internal structure of organisations that promote and facilitate research. At the most general level, workshop participants stressed that, in order to become more effective, social science infrastructure:

- Must be expanded to an international level.
- Must promote international sharing of data and resources from different countries and disciplines.
- Must bridge the gaps among data producers, users, policy makers and the public.

Focus of the Ottawa workshop

This gathering marked the first in a series of OECD international social sciences workshops, launched as follow-up to a conference on the social sciences held in April 1998 at the OECD headquarters in Paris. The series will be organised under the auspices of the OECD Committee for Scientific and Technological Policy and hosted by Member countries, with key research institutions playing a major role. The objective of the workshop series is to stimulate change and progress in the social sciences and their use in the policy-making process, to explore concrete ideas for improving the effectiveness of social science research, and to promote the sharing of knowledge and information between researchers from various disciplines and policy makers and agency managers from OECD Member countries. The series has been given an umbrella title "Reinventing the Social Sciences". A steering committee, comprised of leaders of the organising institutions, has been established to supervise the workshop programmes.

My position is that we need a double paradigm shift in the social sciences: a paradigm shift from publication-driven research towards change-driven research, and a paradigm shift from disciplinary-driven research agendas towards research driven by problems and their driving forces. Such a double paradigm shift, which should not lose sight of the most strict quality control, can in turn only be realised if there is a shift in science policy.

Luk van Langenhove, Deputy Secretary-General, Federal Office of Science Policy, Belgium, and President of the Steering Committee of the OECD International Social Sciences Workshops

The Ottawa workshop involved a series of plenary sessions on specific infrastructure issues such as the use of longitudinal databases, the complexities of data archiving and the research potential of qualitative data. The participants broke into three working groups to discuss areas of general interest:

- *Operational issues and new technologies*, including such issues as information management, data documentation and verification, metadata, linking databases, access to data, comparative analysis, Web-based collaboratories, digital libraries and archives, and multimedia presentation of information.

- *Development of new surveys and databases*, including current plans and practices, suggestions for new surveys at national and international levels, policy development, and the application of new technologies.

- *Communication and interaction*, focusing on relations between data producers and data users, linkages between research and policy development, research and management training, and public interaction.

The steering committee requested each working group to identify the most significant and pressing knowledge gaps related to building research infrastructure, determine which policy issues require immediate attention, and suggest specific steps that could be taken to improve international co-operation and co-ordination.

New directions, new resources and new analysis

Much of the discussion in both the working groups and plenary sessions revolved around the potential of new forms of research data, new resources for the collection, documentation and archiving of data, and new methods of analysis. Facilitated by rapid advances in information and communications technologies, a broad range of new research tools are allowing researchers to ask new questions and compile both quantitative and qualitative information in new forms. Several of the plenary speakers explored aspects of these new developments.

Participants discussed such issues as the need to develop new applications for geographic information systems describing the spatial and temporal relationships of communities and individuals within representative data models. Such complex geographic information systems, as well as longitudinal surveys, would capture data on dynamic life-cycle transitions in the areas of health, education, family structure, crime, income levels and consumer behaviour. Developments in these areas, along with a number of others, promise to revolutionise our understanding of human relations and our interactions with the natural world.

While discussing new digital resources for the social sciences, participants explored such developments as new computer software packages that match individual data files across large incompatible data sets; Web-based national and international surveys; multimedia presentations of environmental records that incorporate three-dimensional surveillance photographs, sounds and smells; new forms of qualitative data, including video and textual information and new search engines to facilitate analysis; as well as ways to link large-scale administrative and organisational databases.

Major issues for the future

Workshop participants identified four major areas that must be addressed in ensuring the development of adequate research infrastructure for the social sciences:

- Sharing research data and resources between disciplines and countries.
- Improving linkages between data sets.
- Improving the interactions between researchers, policy makers and the public.
- Funding of international research projects and of better training for the next generation of researchers.

Sharing research data and resources

A frequently discussed topic at this OECD workshop concerned data sharing and access. Participants emphasised that social science infrastructure involves more than just technology. Researchers also need to develop a culture of sharing data and other forms of research infrastructure both domestically and internationally. Such co-operation would improve the quality of social science research and analysis and broaden the scope of subjects that could be addressed. The participants repeatedly stressed that no data set ever reaches its full potential unless it is shared with other researchers. Sharing data leads to full exploitation of the information collected, allows for the replication of research results, reduces respondent burden and provides for re-analysis from alternative perspectives.

> It is precisely the diversity of the analyses (including modelling and simulation) performed on well-specified and well-documented data sets, by a large variety of researchers, who differ in their interests and orientations, who differ in their options for applied *vs.* basic social science, for empirical inquiry or for modelling, which offers the best opportunities to produce the kind of cumulative knowledge that, in a spiral process, progresses to better (re)conceptualisation, better observation and better measurements, to more productive and critical interaction between data, models and theories. These are the arguments for fostering a new generation of scientific enterprises which as such put new demands on the social sciences at all levels: scientific, organisational, financial as well as political.
>
> *Gaston Schaber, President, CEPS/INSTEAD, Luxembourg*

Workshop participants identified a number of barriers to the sharing of research data and other forms of information. These include:

- A fundamental tension between increasing access to data and protecting confidentiality.
- National legislation that sometimes prevents the export or sharing of data outside the country of origin.
- Differing requirements for the protection of confidentiality among countries.

- A broad range of technical barriers that would require international agreements on common practices and standards.

- In some countries, a lack of national institutions that could facilitate the sharing of research data.

- The absence of privacy laws in some countries, making others reluctant to share information.

The participants suggested a range of practical steps that could be taken immediately by research and statistical agencies to promote the sharing of data, steps that do not require either technical advances or international agreements. Recommendations included such measures as imbedding data-sharing clauses in research grant regulations; engaging professional associations in discussions on creating a data-sharing culture; establishing codes of ethics to combat the deliberate misuse of data; encouraging statistical agencies to develop advisory boards to promote international co-ordination and co-operation among researchers; encouraging universities to give greater recognition to academics who create data sets and other forms of research infrastructure.

Linking data

Linking the vast array of data sets compiled by researchers, statistical agencies, government departments and private organisations holds great promise for the research community and its ability to explore social and economic problems. Linked data sets increase sample size, broaden the range of variables that can be incorporated into analysis, and expand the scope of questions that can be asked. Rapidly improving computing power is helping to make it technically feasible to link even the largest administrative and organisational data sets.

There remain, however, a number of difficult obstacles to overcome in the linking of data sets. These include:

- A lack of common standards for database design.

- Database designs that favour specificity and flexibility over compatibility and linkability.

- Administrative bodies that gather data according to a varying of principles and modes of operation.

- Proprietorial ownership of data.

- Software limitations.

- Lack of standards for collection and documentation of data.

- Lack of expertise for the development of standards and best practices.

- Legal restrictions designed to protect confidentiality.

- Disciplinary divides and a lack of agreements for linking data.

Overcoming these limitations will require on-going technical developments, particularly computer software, new ways of designing databases to facilitate linkages, the establishment of

standards for collection and documentation of data, and a greater appreciation of the rich research possibilities in linking data sets.

Improving interactions between researchers, policy makers and the public

Participants agreed that steps must be taken to increase interaction and improve relationships among researchers, policy makers and the public. This is vital for the future of evidence-based economic and social policy decision making. In the broadest sense, the social science community must provide the tools and information to stimulate the vision and imagination of the public and their political leaders. Although the specifics generated considerable debate, there was broad agreement on a number of practical measures:

- Increase the involvement of researchers from developing nations in both domestic and international research projects.
- Make research results more widely available to journalists and politicians so that research knowledge can better inform evidence-based decision-making processes.
- Bring social science research directly into the classroom at all levels.
- Encourage co-operation and communication among professional societies.
- Increase researcher and public involvement in the design of surveys.
- Involve survey respondents in analysis through consultation and direct feedback.
- Involve data users in setting the research agendas of national statistical agencies.
- Improve access to research results for the public, and in particular for non-governmental organisations and voluntary agencies that cannot afford to commission research themselves.

> The culture with respect to access to official data has changed in many countries in recent years, with a view gaining prominence that the provision of good data for a wide community of users including Parliament, industry and commerce, academia, the media and the general public must be met alongside the needs of government. Analyses should be in the public domain and basic data made available for further analyses. This change is driven by the increasing recognition that not to use data or to use inadequate data has costs for society.
>
> *Denise Lievesley, Director, Institute for Statistics, UNESCO*

Research funding and training issues

One of the more difficult aspects of international co-operation and co-ordination involves the funding of research infrastructure across national borders. Outside the European Union, there are essentially no funding mechanisms for research of an international scope. Workshop participants concluded that in the absence of international agreements on specific programmes or projects, it is unlikely that any progress will be made in the near future. In addition, participants agreed that funding for social science research infrastructure development and training needs to be increased within

individual countries. Several suggestions were made for arguments that could be used to improve the situation:

- Emphasise common transnational or global research areas, such as sustainable development, immigration, changing social values and family structures.

- Point out that the social cost of restricting access to data can outweigh the potential funds to be raised by charging for access to data.

- Argue that research infrastructure is essentially a sunken cost, where the returns on investments already made can only be maximised through increased usage.

> Once a standing funding mechanism… is arranged, one could begin to ask: Which data? When? How? And for how long? Only those projects that produce comparable data and only those efforts that make such data available to researchers more generally should be considered. But, until a benevolent philanthropist or a group of national foundations comes along with multiple millions for international data infrastructure, the questions of which data, domains, disciplines, etc., is on hold.
>
> *Tim Smeeding, Director, Luxembourg Income Study, Centre for Policy Research, Syracuse University*

Training the next generation of social science researchers is a crucial challenge that will determine the quality, quantity and direction of educational efforts. During the workshop discussions, a number of areas were identified where improvements must be made to address the current situation:

- Increase training in quantitative analysis, particularly in areas that employ the new computing and communications technologies.

- Increase expertise in such areas as the development of best practices, common standards and documentation.

- Train research brokers or facilitators capable of acting as a bridge between researchers, policy makers and the public.

Next steps

In the course of the OECD Ottawa workshop, participants made a number of suggestions and recommendations for next steps in the development of research infrastructure and for improving international co-operation and co-ordination. Some of the more general recommendations include:

- Actively engage software and hardware designers in discussions on infrastructure development so that researchers can shape the technology to meet their needs.

- Push governments to support and promote the archiving and sharing of research data.

- Harmonise the data produced by statistical agencies in close co-operation with a wide variety of stakeholders.

- Convince researchers to assume responsibility for the evaluation of standards, ensure that research is kept up to date and of the highest quality, promote inter-disciplinary and international communication, and foster ethical practices.

- Make every effort to include researchers and agencies from developing countries in international decision-making processes regarding research infrastructure.

A number of ideas emerged for specific research projects or undertakings designed to take advantage of new technologies or to address particular problem areas in infrastructure development.

One is the creation of a longitudinal World Social Survey. Such a survey could be based on existing international models, but would be extended to all countries that wish to join. The questions surveyed would focus on social issues of common concern to all communities, such as family structure, employment, health and education, and on the impacts of the forces of globalisation. Respondents would, in effect, be "consultants to the world", and this would be highlighted as a way to encourage people to participate.

Other action proposals include:

- Create a digital and interactively evolving Ottawa Manual on the standardisation of data collection and documentation and on best practices in infrastructure development.

- Develop an international network of knowledge transfer centres that would provide public access to research results and research infrastructure.

- Create user-friendly and publicly accessible databases on issues of direct interest to journalists, policy makers and civil society.

- Establish an international data archive that would combine a central depository, regional or national nodes and broad Web-based access.

- Initiate an OECD-led investigation of Member countries' regulations on confidentiality, privacy, access, linking of data sets and other pertinent matters.

- Create an international licensing system for researchers who work with confidential research materials from countries other than their own.

- Establish an international working group, under OECD auspices, to explore common issues in infrastructure development and possibilities for international agreements.

Conclusion

Discussions at the OECD Ottawa workshop covered wide territory, yet several broad areas of consensus emerged. Participants agreed that effective exploration of transnational social and economic change requires social scientists to work at an international level, in an interdisciplinary fashion, conducting international research projects and sharing and integrating research resources across borders. There was strong consensus that this new mode of research means that social scientists must develop a "culture of sharing" that goes far beyond the exchange of research results.

The participants recommended that the OECD examine impediments to data access. Some delegates suggested the immediate creation of an international working group to explore the possibility of establishing agreements on such issues as standards for preservation, confidentiality, privacy protection and researchers' responsibilities. Others advocated a more cautious and limited approach that would focus on working within each country to change national policies. It is clear, however, that international governmental organisations such as the OECD, UNESCO as well as the national research-granting agencies have a key role to play in any measures taken.

Many participants expressed the desire that social scientists from developing nations, as well as young scholars, be involved in international research collaborations, while others made specific suggestions about how to involve data users in setting research agendas. Finally, workshop participants agreed that social scientists should formulate a set of best practices for social science infrastructure, including techniques for linking international data.

Throughout the plenary and working group sessions of the Ottawa workshop, the implications and challenges of moving social science infrastructure into the next century inspired the discussion. Social scientists and governments today face two major challenges: the need to understand and shape information technologies so they best serve society, and the need to find global solutions to global social and economic problems. Highlighting the importance of international research collaboration, identifying the range of issues and challenges which shape and delimit social science infrastructure and recommending ways to improve opportunities for social scientists to co-operate at an international level will help to address these needs. The workshop proved to be an important step in that direction. At its conclusion, participants agreed that if social scientists do not actively participate in these measures, international co-operation and co-ordination in the development of research infrastructure will simply not occur.

> To further our own quest for knowledge and help build a civic, caring and creative society, we must be ready for a future which is coming at us at dizzying speed. To meet the demands of that future, we are called upon to "reinvent" the social sciences in fundamental ways. We must be able not only to move beyond traditional disciplinary and geographic boundaries. We must also build new capabilities and reflexes of co-ordination, collaboration and communication – including among researchers, both nationally and internationally, and also between researchers and people from other walks of life.
>
> *Marc Renaud, President, Social Sciences and Humanities Research Council of Canada*

Chapter 2

REINVENTING THE SOCIAL SCIENCES: SETTING THE STAGE

by

Luk van Langenhove*
Deputy Secretary General of the Belgian Ministry of Science Policy and
President of the Steering Committee of the OECD International Workshops on the Social Sciences

Introduction

This is the first of four OECD International Social Sciences Workshops organised under the generic heading of "Reinventing the Social Sciences". As president of the Steering Committee of these workshops, I would like to raise two questions:

- What needs to be re-invented in the social sciences?
- Why should the OECD bother with the social sciences?

Taking the second question first, I would like to describe in greater detail the relationship between the OECD and the social sciences. The OECD has a long tradition of looking at the social sciences. Already in 1966, a report on *The Social Sciences and the Policies of Governments* was submitted to the second Ministerial Meeting on Science. In 1976, the OECD Committee for Scientific and Technological Policy examined the social science policies of three countries (France, Finland and Japan). Based on these assessments, a report was published in 1979, *The Social Sciences in Policy Making*, in which the following recommendations were made to Member governments:

- A more flexible and pluralist system of financing research should be devised.
- The social sciences research system should be developed in a more balanced way.
- The role of science policy bodies should be broadened so as to ensure the development and use of the social sciences.
- Communication between the government and the scientific community should be intensified.

* The views expressed in this chapter do not engage the responsibility of the Ministry of Science Policy.

- Contacts between governmental and non-governmental specialists in the social sciences should be intensified.

- Decision makers should be urged to take account of the results of social science research.

These recommendations remain valid today. One could ask why this is the case? One could also ask why, since that 1976 initiative, no further work on the social sciences has been carried out by the OECD for more than 20 years!

Not until 1997 did the Belgian Delegation to the CSTP group on the Science System propose to re-examine the role of the social sciences in the scientific system with the aim of furthering research in the social sciences. Many other delegations supported this request and in April 1998 a Workshop on the Social Sciences was held, focussing on the problems encountered by social science disciplines and proposing ways forward. Among the topics discussed at that workshop were:

- *The status of the social sciences*. The social sciences do not enjoy the same status as the natural sciences in the eyes of either the scientific community or the general public. This lack of recognition has serious consequences for both public funding and public legitimisation.

- *The influence of the social sciences.* Two conflicting attitudes can be observed: those who believe that the social sciences do not appear to be of use in solving the problems facing society *vs*. those who think that some social science disciplines, *e.g.* economics or management, have a considerable, albeit diffuse, influence.

- *Institutional rigidities and interdisciplinarity*. The division of labour among the various social science disciplines and sub-disciplines was widely recognised as hindering the status and influence of the social science research in general. In addition to disciplinary boundaries reflected in institutional rigidities, methodological arguments between disciplines entail a lack of unity in the study of man in society.

The proceedings of the 1998 Workshop on the Social Sciences were published under the title *The Social Sciences at a Turning Point?* (OECD, 1999).

In March 1999, the CSTP decided to organise a number of follow-up international workshops on the social sciences under the generic heading "Reinventing the Social Sciences". Today, we are gathered at the first of these follow-up workshops.

So, finally, after more than 20 years, the social sciences are again on the OECD agenda. It can be no coincidence that other international initiatives have occurred in the interim that also deal with the social sciences, of which the following are two important examples:

- In 1994, UNESCO established the MOST programme for the fostering of international policy-relevant social science research. In 1999, UNESCO published the first *World Social Science Report* (Kazancigil and Makinson, 1999). This groundbreaking report – the first in a series – aims to take stock of the current situation in the different social science disciplines and monitor and prospectively map out developments in all their dimensions and perspectives.

- In 1996, the Gulbenkian Commission on the Restructuring of the Social Sciences published its report *Open the Social Sciences* proposing radical measures for modernisation of the

discipline: these ranged from how to award university chairs, to setting syllabi and raising funds (Wallerstein, 1996).

The reasons underlying this renewed concern with the social sciences in international organisations and commissions could be put down to a general disillusion with the current state of affairs. This discontent stems from the claim that the social sciences are not contributing to improving our understanding of society, let alone to solving societal problems.

Some blame this situation on the fact that governments have not invested sufficiently in the social sciences. Others blame social scientists themselves, claiming that they are not delivering the right goods. Still others point to the state of flux of today's societies which makes understanding and forecasting ever more difficult.

It is, of course, true that the economic crises of the 1970s and 1980s forced governments in many countries to cut back on public spending. Research in general, and the social sciences in particular, have suffered from these cuts. In the United States, for example, expenditure on social science R&D in the university sector dropped sharply between 1973 and 1985 from 8.01% to 3.96% (Oba, 1999).

During that same period, a number of governments adhered to a neo-liberal ideology that focused on individuals and markets. As a result, paying attention to "social" aspects was not politically popular. Neither was the idea that the social sciences could contribute to policy making. The elections of Margaret Thatcher in 1979 and Ronald Reagan in 1981 have been widely seen as marking a change in climate in the relations between government and the social sciences (Bulmer, 1987).

Another reason for "discontent" can be found in the social sciences themselves. At the end of the 1960s, critical movements occurred within many of the social science disciplines, challenging the methodological and disciplinary foundations of the disciplines. These critical movements, on the one hand, made it clear that orthodox research traditions within the social sciences – modelled on those of the natural sciences – pose major problems in contributing to an understanding of the social realm (see, among others, Argyris, 1974; Manicas, 1987; Gergen, 1982; Wallerstein, 1996).

Alternative methods and approaches have been developed but have met with resistance in the discipline-organised academic setting. One fringe of the critical movement turned away from empirical research and – influenced by postmodernist theories – engaged in pure speculation and normative debates.

Another issue which has probably contributed to the renewed concern for the social sciences concerns the enormous changes that our societies are going through. In recent years we have witnessed the emergence of global-scale issues and global-scale problems. Globalisation is everywhere. The term "globalisation" describes the emergence of a global economy whose effects spread across the whole planet, but it also refers to a whole range of non-economic issues with worldwide implications.

Giddens once defined this phenomenon as "action at distance": our day-to-day activities are increasingly influenced by events happening on the other side of the world. And, conversely, local lifestyles have become globally consequential. As noted in the UN report *Our Global Neighbourhood*: "never before have so many people had so much in common, but never before have the things that divide them been so obvious."

Globalisation creates both stresses and opportunities for societies. In today's changing world, input from the social sciences in formulating innovative approaches to solving global scale issues is

vital. However, many people are frustrated by what they perceive as the inability of the social sciences to contribute to societal progress.

So, what needs to be reinvented in the social sciences? I foresee four major challenges?

First, there is a need to reinvent the social science infrastructure. A globalised world means that national infrastructures will have to be framed within international infrastructures. Traditionally, the social sciences developed around the concept of a nation-state and the expression of national culture (Martinotti, 1999). Increasingly, the social sciences will have to transcend the "local" if they want to have a role to play in understanding the "global".

Digitalisation changes the way in which the social science research will be performed and communicated. The Internet provides researchers with vast opportunities for transdisciplinary collaboration, data-sharing and Web-based archiving. Qualitative analysis is also benefiting enormously from new computer applications.

Second, public legitimisation of the social sciences needs to be reinvented. Although it is widely accepted that this legitimisation depends upon two aspects: *i)* the claim of making the world more intelligible; and *ii)* the contribution to problem solving and policy making in society.

One of the major problems of the social sciences would seem to be that they are not able to demonstrate tangible achievements in these two fields.

Third, up until the present time the social sciences have divided the world into disciplines and within those disciplines into variables. This was the only way to follow the model of the natural sciences. Today, we are aware of the limits of that approach (van Langenhove, 1996) and we know that other models are emerging within the natural sciences themselves.

Dealing with complexity is possibly the most challenging task facing us today. The social sciences need to be able to investigate the complex social reality of the world in which we live at all its different temporal, spatial and aggregation levels.

Fourth, one of the most pressing tasks is to reinvent the disciplinary structures. While most research is still performed from the vantage point of single disciplines, social problems are multidisciplinary in nature. As noted by Wallenstein (1998), the disciplinary boundaries no longer represent clearly different fields of study with clearly different methods, but the corresponding corporate structures and boundaries are still in place and are very effective in disciplining the practices of research.

Taken together, these four challenges can be interpreted as a call for change. Of course, I am not alone in making such claims and pleading for change. A serious problem arises from the fact that it would seem that the institutional organisations that embody the social sciences are unable to engage in such changes themselves.

Pressures against change are strong. Governments have a crucial role to play in this respect: not only do they manage a significant share of the available resources which they can use as "leverage", they are also the biggest "buyers" of results.

However, governments also need an incentive to change. This is where the OECD has a useful role to play: it is tasked with analysing and recommending feasible means of action that can respond to

various political concerns. Our task will be to prepare such recommendations based on the assessments and discussions that emerge from this and the upcoming workshops.

I would like to present my personal viewpoint on how the social sciences should develop. I know how ambitious – if not pretentious – it is to do so. But then this is the first in a series of "reinventing" workshops. So we can always dream!

My position is that we need a double paradigm shift in the social sciences away from *publication*-driven research towards *change* driven research and away from a disciplinary-driven research agenda towards research driven by social problems and the forces which give rise to them.

Such a double paradigm shift should not lose sight of the strictest quality control, and can only be realised through a shift in science policy (van Langenhove, 1999).

The social sciences need to be able to generate knowledge of relevance for all those who want to change a given situation. As such, social science research should aim to bring together researchers, the actors with a role to play in the phenomena studied and decision makers. However, the social sciences cannot claim to act as an agent of change "on behalf" of society: social scientists must work together with industry, government and civil society.

Key issues concern empowerment through the social sciences and participative research that includes all the stakeholders involved. In addition, there is a need for recognition that compartmentalising complex societal issues and their interrelations into simple disciplinary issues is counterproductive.

For those who think that this is a utopian ideal, I would like to point out that methodological tools exist that allow such a double paradigm shift, namely participative methods.

Participatory methods is an umbrella term which describes interactive approaches that actively involve a range of stakeholders, ranging from decision makers to laypersons. The rationale for participative social sciences research can be explained against the background of two questions:

- From whose perspective is research performed?
- How can social science research influence decision making?

The first question has to do with the values of the initiator which are of primary importance in defining a research issue that are of primary importance when initiating research. Who determines this and on what grounds?

While, in the case of basic research, it is mainly the researchers themselves who decide what to study, in applied research it is the body that commissions and/or pays for the research. For instance, an academic sociologist or a government can decide to initiate a research project on inter-group relations between immigrants and non-immigrants. Seldom, if ever, are the immigrants and non-immigrants implicated in that decision and, in the majority of cases, their role in the research process will be limited to passively responding to questions. Most will never even see the results. At best, the research results might influence a development path because they will be used in making decisions on, for example, how to improve inter-group relations in a community.

In the case of participative social science research, the people involved will have an active say in *i)* defining research goals; *ii)* conducting the research; *iii)* interpreting the results; and *iv)* translating

them into development paths. Such an approach to social sciences takes as its starting point a community of enquiry that uses theoretical and methodological expertise to influence the process of change.

In the view of some academics, this perspective may seem utopian since it ignore the distinction between experts and laypersons. I think that that is wrong because such working methods are already common in some disciplines, for instance psychological research (Reason and Heron, 1996) and in certain practices such as management performance audits in organisations (Argyris and Schön, 1974). There is no reason why it could not work for other societal issues.

All this, and much more, is up for debate in the four OECD workshops on "Reinventing the Social Sciences". Today, our focus is on infrastructure and the challenge of the digital world. This is, and will continue to be, a major driving force in changing the social sciences.

I hope that our four workshops will serve to stimulate the debate and to raise awareness that there is a vital need for change if the social sciences are not to become an intellectual backwater in the 21st century. As Charles Handy once said: "The best way is always yet to come, if we can rise from our past...".

REFERENCES

Argyris, C.(1980), *Inner Contradictions of Rigorous Research*, Academic Press, New York.

Argyris, C. and D. Schön (1974), *Theory in Practice: Increasing Professional Effectiveness*, Jossey Bass, San Francisco.

Bulmer, M. (1987), "The Social Sciences in an Age of Uncertainty", in M. Bulmer (ed.), *Social Science Research and Government*, Cambridge University Press, Cambridge.

Gergen, K. (1982), *Toward Transformation in Social Knowledge*, Sage Publications, London.

Kazancigil, A. and D. Makinson (1999), *World Social Science Report 1999*, UNESCO Publishing and Elsevier, Paris.

Manicas, P.(1987), *A History and Philosophy of the Social Sciences*, Basil Blackwell Publishers, Oxford.

Martinotti, G. (1999), "The Recovery of Western European Social Sciences Since 1945", in A. Kazancigil and D. Makinson (eds.), *World Social Science Report 1999*, UNESCO Publishing and Elsevier, Paris.

Oba, J. (1999), "The Social Sciences in OECD Countries", in A. Kazancigil and D. Makinson (eds.), *World Social Science Report 1999*, UNESCO Publishing and Elsevier, Paris.

OECD (1999), *The Social Sciences at a Turning Point?*, proceedings of the OECD Workshop on the Social Sciences, OECD, Paris.

Reason, P. and J. Heron (1996), "Co-operative Inquiry", in J. Smith, R. Harre and L. van Langenhove (eds.), *Rethinking Methods in Psychology*, Sage Publications, London.

Van Langenhove, L. (1996), "The Theoretical Foundations of Experimental Psychology", in J. Smith, R. Harré and L. van Langenhove (eds.), *Rethinking Psychology*, Sage Publications, London.

Van Langenhove, L. (1999), "Rethinking the Social Sciences? A Point of View", in *The Social Sciences at a Turning Point?*, proceedings of the OECD Workshop on the Social Sciences, OECD, Paris.

Wallerstein, I. *et al.* (1996), *Open the Social Sciences*, Stanford University Press, Stanford.

Wallerstein, I. (1998), "The Heritage of Sociology. The Promise of Social Science", Presidential address at the XIVth World Congress of Sociology.

Chapter 3

SOCIAL SCIENCES DATABASES IN OECD COUNTRIES: AN OVERVIEW

by

Jun Oba
Science and Technology Policy Division, OECD

Introduction

There has been increasing demand for various types of social science data. In many OECD countries, decision makers promote the use of empirical social science data to monitor social trends relevant to public policy. Greater availability of data on society could favourably affect research quality in the social sciences. Data collection efforts, such as social surveys, have increased and international co-operation has also been undertaken.

Many data archives set up in OECD countries make data available to researchers. In addition, recent developments in information and communication technologies (ICT) have considerably improved access to social science resources and increased the potential of social science research. The benefits of data archiving are obvious: researchers can share data, thus avoiding duplication and improving data quality through further analysis by other researchers. Nowadays, data archives disseminate data to users via the Internet. However, the system is not yet fully satisfactory, and many researchers find it difficult to deal with Web-based data retrieval systems. Research into new systems is under way. Various levels of training courses are available for data archivists as well as social scientists. Moreover, data archiving is time-consuming and requires many resources.

The restrictions to be imposed on future use of the data are an important issue. Owners and copyright holders are often reluctant for various reasons, such as privacy protection, to provide data, especially microdata, for secondary analysis. Confidential data is generally made anonymous, but this may hamper social science research, since researchers with complex theories and powerful statistical tools increasingly find that suppression of information significantly limits their analysis, and it becomes increasingly difficult to develop a microdata file that maintains confidentiality when longitudinal data are available for a number of years.

Access to publicly produced microdata varies greatly at national level. Until now, few countries have adequately liberalised researcher access, and many social scientists have denounced their insufficient access to public microdata. In answer, governments have undertaken initiatives to make

data available. However, official data which are considered confidential are often only available to officials of statistical offices, or, if they are accessible to researchers, it is only on the premises of the statistical office.

Archived data are not always comparable across countries because of the paucity of standards and comparative studies in social surveys and data collection. Consequently, most empirical research in the pertinent disciplines is not internationally comparable. There are pressing needs for harmonisation of data, but harmonisation of already acquired data is very difficult and can be an extremely lengthy process. Preferably, co-ordination of surveys and harmonisation of data should be undertaken prior to the survey. International organisations play a major role in standardising statistics. International social surveys and research programmes have also been undertaken.

Networking of archives is essential for efficient use of data, but it is necessary to deal with the variety of interfaces, data formats and information resources available in the world's data archives. New systems are being developed. There are also language barriers, since data are compiled in different languages. International collaborations are under way, including integration of data catalogues, development of new networking systems, standardisation of data formats, etc. They require international and interdisciplinary co-operation.

In OECD countries, governments are increasingly called upon to address social issues that require analysis of empirical, often internationally comparable, data relevant to a wide range of social science disciplines. Such analysis depends largely on the availability and quality of data. Many OECD countries have a large number of databases on societies which are generated by social surveys, social research, etc., and which are stored in data archives and other institutions. However, many are not available for secondary analysis or are reserved to small number of authorised people, especially in the case of publicly produced data. Recent developments in ICT have the potential to significantly improve access to such data if the restrictions on use are removed. International co-operation for collecting data and integrating databases is under way and could enhance comparative studies. In addition to data collection efforts, an electronically linked comprehensive Web-based database system from which researchers could extract relevant harmonised world data would greatly increase the potential of social science research.

The increasing need for data in the social sciences

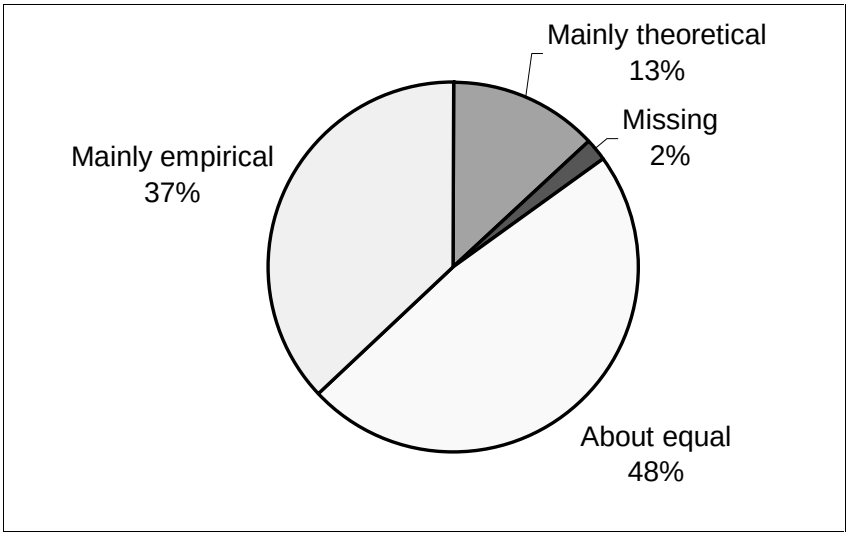
There has been increasing demand by social scientists not only for new data but also for exploitation of existing data for secondary analysis. In many OECD countries, decision makers promote the use of empirical data by the social sciences to monitor social trends relevant to public policy. Improved availability of data on society could favourably affect research quality in the social sciences.

Advances in social science methodologies require large amounts of data. Researchers increasingly require access to detailed microdata to conduct research in many areas. The powerful statistical techniques needed to analyse multilevel, longitudinal data cannot be used with aggregate data; access to microdata is essential (Bernard *et al.*, 1998).

However, valuable research data have been lost through destruction of materials and through media obsolescence. In order to preserve them and make them available to users, data archives have been created in OECD countries (see Annex).

According to a survey carried out in the United Kingdom in December 1997 at the London School of Economics (LSE), mostly of its research staff via e-mail, over 85% of respondents would benefit from improved access to data. Over a third of respondents said that their research interests were "mainly empirical". Half said that their interests were about equal (*i.e.* theoretical and empirical) (Figure 1). The survey also showed widespread interest in a wide range of data. Three-quarters of respondents expressed an interest in using data from inter-governmental organisations and the EU (Brackenbury, 1998).

Figure 1. **Type of research conducted by respondents**



Source: Brackenbury, 1998.

Recent developments in ICT have considerably improved access to social science resources. Researchers can now browse, identify and extract data in social sciences from various archiving institutes via the Internet. Data manipulation, such as merging data sets, is greatly facilitated by ICT. Machine-readable data offer diverse possibilities for research, for example by comparing various data sets across periods or societies. In addition, secondary analysis – the re-analysis of machine-readable data – is one of the best supplements to traditional teaching methods, especially for teaching research methodology and statistics (Guy, 1997).

Many social surveys have been undertaken for the purpose of social science research. A well-known example is the General Social Survey (GSS) launched in 1972 in the United States. It is an annual personal interview survey of US households conducted by the National Opinion Research Center (NORC) and aims to make timely, high-quality, scientifically relevant data available to the social science research community. Since 1982, GSS has had a cross-national component, and since 1985 the cross-national module has been part of the International Social Survey Programme (see below). The General Social Survey Data and Information Retrieval System (GSSDIRS) allows users full online access to the GSS 1972-98 data and documentation, as well as online analysis of data via the World Wide Web.

A survey (Gadd, 1998), conducted at the University of Plymouth in the United Kingdom, in 1998, shows the level of use of social data by socio-economic staff and research students interested in secondary analysis of data (Table 1). More than half were interested in census data. However, some data sets, notably British Social Attitudes, were not used despite interest expressed by respondents (37.2%).

To obtain longitudinal data, such surveys should continue, using consistent questions and samples. Longitudinal data are especially important for social science research. In addition, new surveys should be undertaken, since social science studies increasingly need more data focused on various aspects of society. Thanks to ICT developments, Web-based surveys could facilitate collection of data from large numbers of respondents at relatively low cost. However, such surveys may not yield satisfactory results, because Internet tools do not yet allow for conducting surveys on representative populations. Methodologies should be developed to produce scientifically reliable data from Web-based surveys, which have enormous potential.

Table 1. **Use of secondary statistical data sets in the University of Plymouth**

	Never heard of it	Not interested	Interested/ Not used	Interested/ Used
Census 1991	4 (4.2%)	39 (41.1%)	19 (19%)	32 (33.7%)
British Social Attitudes	8 (8.5%)	37 (39.4%)	35 (37.2%)	12 (12.8%)
General Household Survey	7 (7.6%)	41 (44.6%)	27 (29.3%)	16 (17.4%)
Quarterly Labour Force Survey	9 (9.9%)	42 (46.2%)	24 (26.4%)	15 (16.5%)
Family Expenditure Survey	11 (12.5%)	40 (45.5%)	25 (28.4%)	11 (12.5%)
Family Resources Survey	20 (22.2%)	42 (46.7%)	24 (26.7%)	3 (3.3%)
Eurobarometers	42 (47.2%)	27 (30.3%)	11 (12.4%)	8 (9.0%)
British Household Panel Study	28 (32.9%)	39 (45.9%)	13 (15.3%)	4 (4.7%)
Office for National Statistics Databank	47 (52.8%)	26 (29.2%)	12 (13.5%)	3 (3.4%)
National Child Development Study	14 (15.7%)	60 (67.4%)	9 (10.1%)	5 (5.6%)

Source: Gadd, 1998.

International co-operation for data collection, such as the International Social Survey Programme (ISSP), are under way (see below). Some research institutes undertake international social surveys, such as the *Centre d'Études de Populations, de Pauvreté et de Politiques Socio-économiques*/ International Networks for Studies in Technology, Environment, Alternatives, Development (CEPS/INSTEAD) in Luxembourg.

Data archiving issues

Role and benefits of data archives

Data archives are national resource centres for research, education and other purposes. They acquire data, put them into appropriate format for secondary analysis, store and disseminate them through various media, including the Internet. Data archiving can help researchers share data, thus avoiding duplication and improving quality through further analysis by other researchers.

In the OECD countries, many data archives in social sciences now exist, such as the Institute for Research in Social Science (IRSS) Data Archive of the University of North Carolina at Chapel Hill in the United States, founded in 1924, which is one of the oldest facilities of its kind in the world. It disseminates data sets to users, principally via the Internet. However, the Internet system is not yet satisfactory and much remains to be done. The Directorate for Social, Behavioural and Economic Sciences of the National Science Foundation (NSF) of the United States has announced a programme on Enhancing Infrastructure for the Social and Behavioural Sciences, with emphasis on creating Web-based collaboratories to enable real-time controlled experimentation, to share the use of

expensive experimental equipment, and/or to share widely the process and results of research in progress (NSF, 1999).

Granting agencies such as the NSF often require researchers to deposit their data in a public archive upon completion of the project.[1] Many government departments, national statistical agencies and international organisations redistribute data from their own Internet sites, largely free of charge. Individual researchers also place data obtained from their research on the Internet from their research institute's server. In order to locate needed data in different institutions, clearinghouses and gateways have been created.

The benefits of data archiving

Reinforces open scientific inquiry. When data are widely available, the self-correcting features of science work most effectively.

Encourages diversity of analysis and opinions. Researchers with access to the same data can challenge each other's analyses and conclusions.

Promotes new research and allows for the testing of new or alternative methods. There are many examples of data being used in ways that the original investigators had not envisioned.

Improves methods of data collection and measurement through the scrutiny of other work. Making data publicly available allows the scientific community to reach consensus on methods.

Reduces costs by avoiding duplicate data collection efforts. Some standard data sets, such as the Longitudinal Study on Ageing and the National Longitudinal Surveys of Labour Market Experience, have produced literally hundreds of papers that could not have been produced if the authors had had to collect their own data.

Provides an important resource for training in research. Secondary data are extremely valuable to students, who then have access to high-quality data as a model for their own work.

However, data archiving is time-consuming and sometimes expensive. Although many investigators are more than willing to make their data available to others, they are frustrated by the task of preparing data for outside use, particularly in terms of creating complete documentation (ICPSR, 1999). The data collected during social science research projects and social surveys are often not comparable across countries and periods because of the lack of equivalent classifications or consistent formats. Such data are often unsuitable for secondary analysis. In addition, they are not always accessible to researchers because of administrative reasons, data protection regulations, etc. Even so, data providers often charge high fees.

Many data archives continue to be open to researchers throughout the world, and international co-operation has progressed. Some data archives, such as the Inter-university Consortium for Political and Social Research (ICPSR) in the United States, are strongly international. Established in 1962 at the Institute for Social Research, University of Michigan, ICPSR is a membership-based, not-for-profit organisation serving member colleges and universities in the United States and abroad. ICPSR provides access to a large archive of computerised social science data, training facilities for the study of quantitative social analysis techniques and resources for social scientists. The data holdings cover a broad range of disciplines, including political science, sociology, demography, economics, history, education, gerontology, criminal justice, public health, foreign policy, and law. The consortium encourages social scientists in all fields to contribute to and utilise ICPSR's data resources. ICPSR includes among its members over 325 colleges and universities in North America and several hundred additional institutions in Europe, Oceania, Asia and Latin America. Member institutions pay

annual dues that entitle faculty, staff and students to acquire the full range of services provided by ICPSR. Individuals at non-member schools can also order data for a fee.

Professional development assistance for archive staff has been made available by the principal archives and international organisations. Some archives offer training courses to archivists and data librarians, such as the Summer Training Program in Quantitative Methods of Social Research of ICPSR and the Essex Summer School in Social Science Data Analysis and Collection. Such courses are also useful for forming networks among researchers for possible future international co-operation. The International Association for Social Science Information Services and Technology (IASSIST) provides guidelines for professional development and opportunities for learning through a recommended curriculum for training staff of social data information centres located in national archives, academic libraries, computing centres, research institutes, governmental agencies, or private corporations.

Data collection and storing

A data archiving institute collects, stores and classifies data for provision to a broad research community for further analysis. On the basis of its mission and criteria, every archive selects a limited number of data sets, as data archiving is a costly operation. Mochmann and de Guchteneire (1998) refer to four categories of criteria: *i)* scientific criteria dealing with relevance, size and scope of the data set; *ii)* technical criteria such as internal format, data size and media for transfer; *iii)* administrative criteria such as documentation, privacy protection and ownership; and *iv)* financial criteria.

A data archive has many data collection channels, such as academic research, national surveys and governmental statistical offices. Data archives generally do not buy data sets. Researchers view their contribution to the archive as a further means of making their research public. For donating institutes, the data archive may function as an external backup of their own holdings. For these and other reasons, most data sets arrive at the data archive free of charge. Many archives can, however, reimburse the expenses involved in the transfer of the data (Mochmann and de Guchteneire, 1998).

In order to store data in the archive, the data should be processed so as to be accessible to users. Ideally, the data set should be accessible using a standard statistical package, such as SPSS or SAS. Thanks to the development of ICT and data-analysis techniques, social science data use increasingly complicated data structures. For these reasons and others, it is crucial to success to plan for a data collection project at the outset. The cost can thus be significantly reduced. ICPSR (1999) recommends that documentation should be as much a part of project planning as questionnaire construction or analysis planning and that a project plan should, at a minimum, involve decisions on the following topics:

- *File structure.* What is the data file going to look like and how will it be organised? What is the unit of analysis? Will there be one long data record or several smaller ones?
- *Naming conventions.* How will files and variables be named?
- *Data integrity.* How will data be converted to electronic form, and what checks will be in place to find illegal values, inconsistent responses, incomplete records, etc.?
- *Code-book preparation.* What will the code-book look like and how will it be produced? What information will it contain?

- *Variable construction.* What variables will be constructed following the collection of the original data? How will these be documented?
- *Documentation.* What steps will be taken to document decisions taken as the project unfolds? How will information be recorded on field procedures, coding decisions, variable construction, and the like?
- *Integration.* To what extent can the various tasks mentioned above be integrated into a single process? (This point is critical: to the extent that one can use a single computer programme or an integrated set of programmes to carry out these tasks, they are made simpler, less expensive and more reliable.)

A data set has documentation (metadata) to help users to locate and navigate the relevant data. The documentation consists of catalogue, code-book, user guide, etc. It is crucial for the archiving system, and without it many valuable resources would be under-utilised or even lost to the research community. Recent developments in the structure of metadata have been led by an international committee of data archivists and data librarians. The committee has proposed a structure, or DTD in SGML terminology, that will make it possible for data producers and archives to produce consistent metadata for use by data librarians in searching mechanisms (Musgrave, 1998).

The content and format of code-books vary considerably among data collections. The typical code-book contains the following information (Guy, 1997):

- A description of how the data were collected, including the sample design.
- The variables contained in the data.
- In the case of surveys, the survey instrument or questionnaire used to solicit responses from the respondent and the coded values of each question.
- The location and format of the variables within the raw data file.

The validity of data is sometimes questioned. Unless data come from government statistical institutions, whose data are generally deemed reliable (of course, they should also be subjected to secondary analysis to test their validity), the checking of the data an archive receives is important to ensure that the archive is reliable.

Dissemination and data retrieval

The normal way to disseminate data sets from the archive is via the Internet or recordable media such as CD-ROM and diskette. Data sets from an archive may be available in several formats for different types of machines. To obtain data via the Internet, users browse the Web and identify the data they need (Figure 2), and they may then extract data sets via the Internet or order them from the archive. A catalogue with summaries of data sets stored in the archive is useful.

Figure 2. **Extract of research results from the Roper Center's data catalogue**

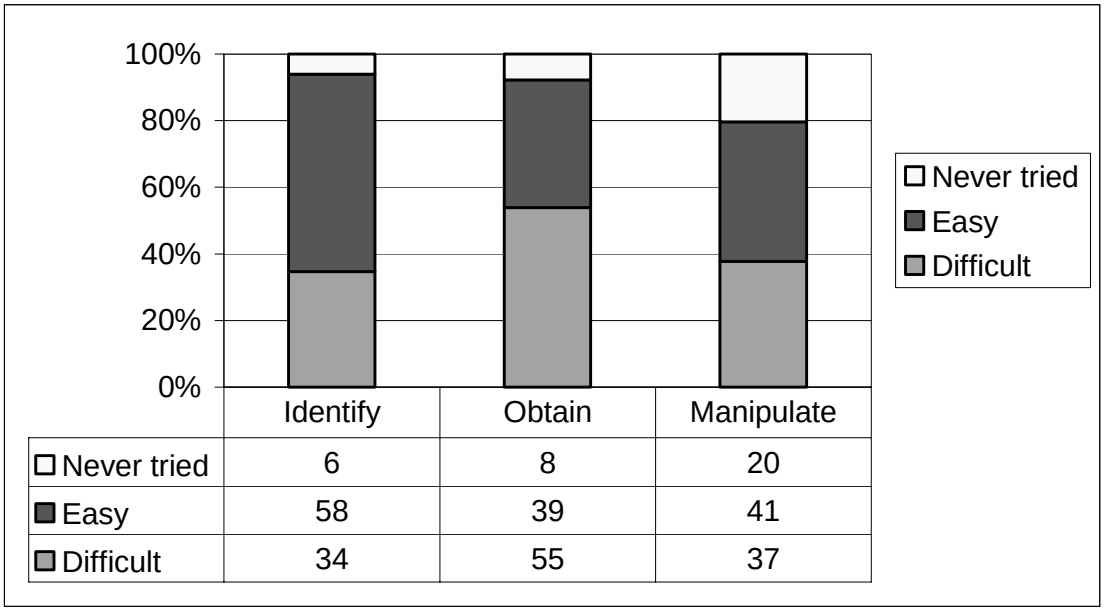
Study Number: USAIPOGNS1998-9812046

Title: Gallup News Service Poll # 9812046: Economy/Religion
Survey Firm: Gallup Organization
Survey Sponsor:
Field Dates: December 4-6, 1998
Sample: National adult
Sample Size: 1070
Sample Notes:
Variables: 74
Major Topics Covered:
Clinton job performance (1); Clinton impeachment (2); opinion of political leaders (4); Republicans vs Democrats in Congress (2); future of Clinton (2); 2000 presidential election (2); biggest threat to the country (1); power in a few big companies (1); government involvement in mergers (2); Middle East conflict (3); Social Security (8); economy (2); employment (2); Christmas gift spending (7); the Bible (2); new Bible reading affects life (7); religion (7); 'born-again' Christian (1).

Source: The Roper Center for Public Opinion Research, http://www.ropercenter.uconn.edu/.

However, many researchers find it difficult to identify and, more importantly, to obtain the data that they need. The above-mentioned survey at LSE showed that over half of the respondents, who are probably the major users of data, find it difficult to obtain data for their research, and that over a third found it difficult to identify data for research purposes (Figure 3).

Figure 3. **Ease with which respondents acquired data for research**



	Identify	Obtain	Manipulate
Never tried	6	8	20
Easy	58	39	41
Difficult	34	55	37

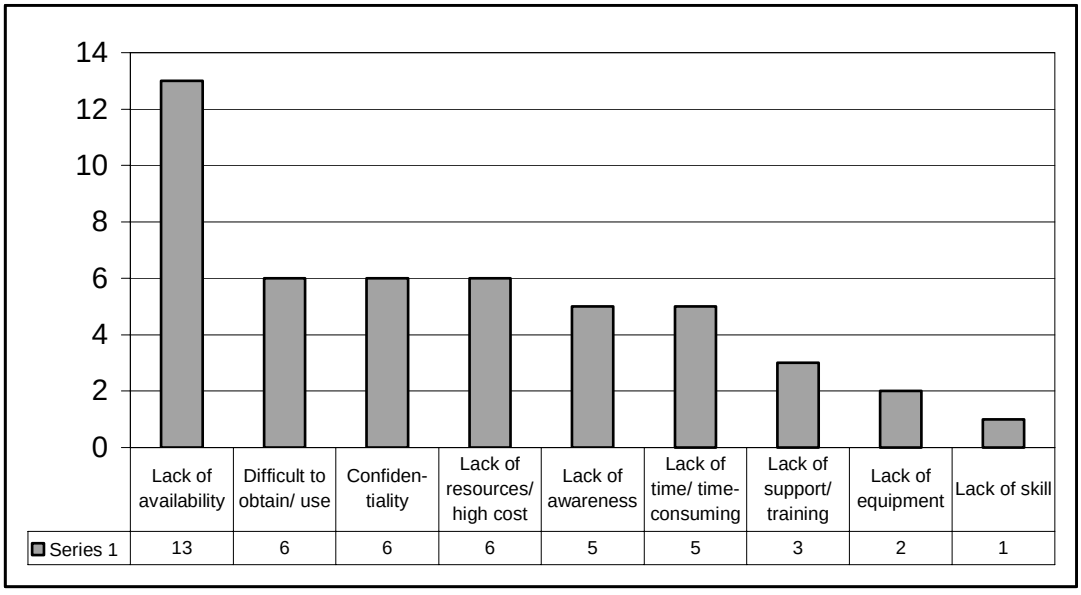
Source: Brackenbury, 1998.

The survey at the University of Plymouth (see above) shows low levels of use of data archives and other data services. The Data Archive was used only by 12.3% of respondents who expressed an interest in secondary data analysis, but 37.8% had never heard of it and 48.0% knew of it but had not

used it (Gadd, 1998). The LSE survey highlights several bottlenecks preventing more widespread use of data (Figure 4). It concluded that data librarians should be aware of external data services (*e.g.* academic data centres, commercial online services) as well as local data services (*e.g.* print and CD-ROM collections) and recommends that access to data should be improved, especially by developing easy-to-use Web-based data delivery and a wider range of networked CD-ROM data (Brackenbury, 1998).

A number of Web research machines, such as Alta Vista, Lycos and Yahoo!, may be useful for identifying data archives and locating data. However, such generic search machines, in spite of – or even because of – their natural-language algorithms, are little suited for searching scientific disciplines and must be regarded as of limited use in scientific queries (Ohly, 1998). A proficient documentation-browsing engine is needed which enables users to navigate throughout the data and to view the lists and definitions of each data set so that the required data may be easily identified, extracted and put into an appropriate format for secondary analysis. The system should respond quickly, be relatively simple and easy to browse.

Figure 4. **Factors hindering data use in research**

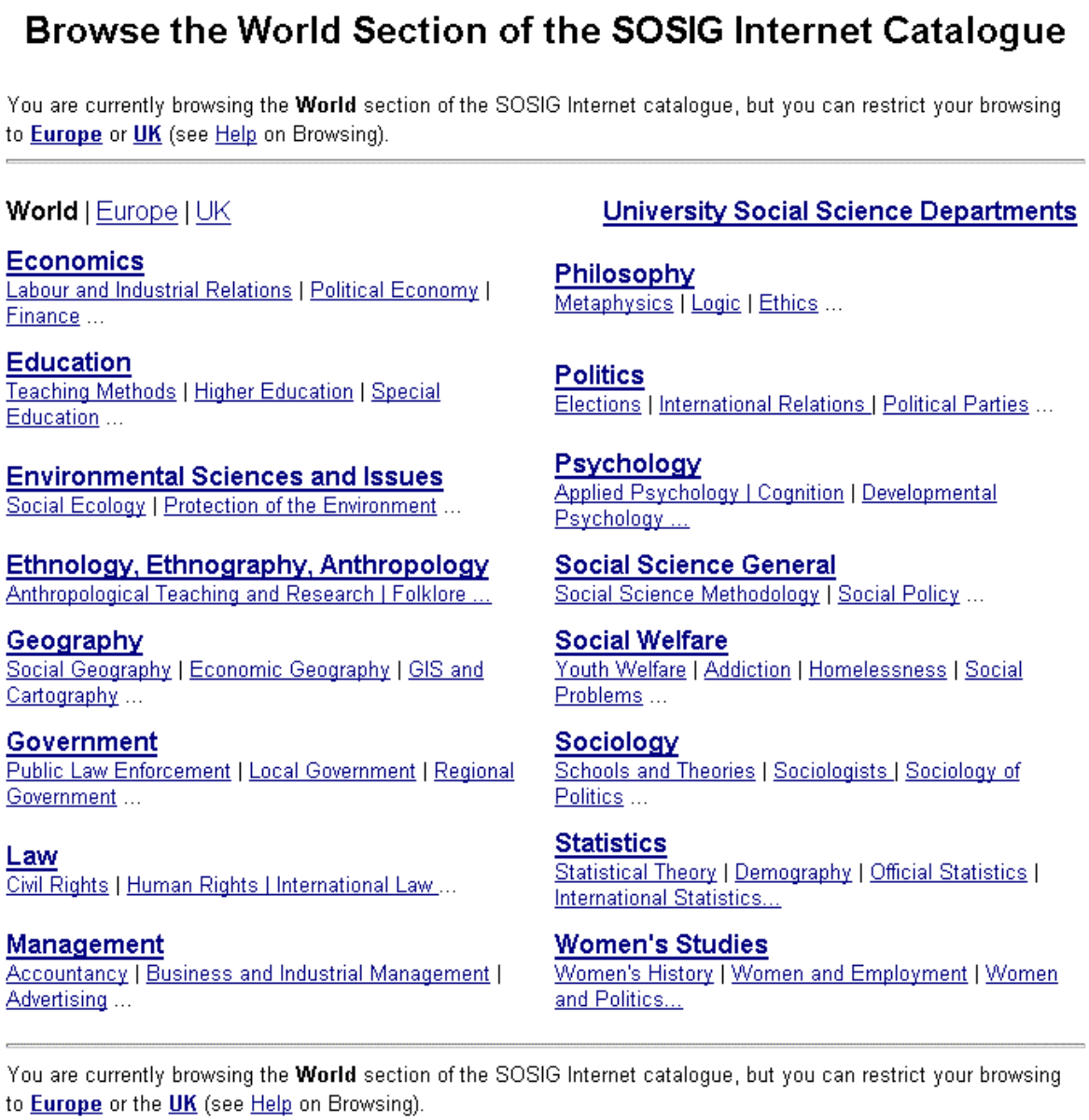


Source: Brackenbury, 1998.

In recent years, Internet gateway services have been developed for the social sciences, which allow researchers to identify data archives containing the data necessary for their research. The Social Science Information Gateway (SOSIG) in the United Kingdom, for example, was established in 1994 to provide social science researchers, academics and librarians fast, easy access to relevant, high-quality networked resources. The gateway provides access to Internet resources via an online catalogue in which resources are classified and described by an information professional (Figure 5), or via search by keywords. The SOSIG team locates, assesses and describes high-quality networked resources not only in the United Kingdom but also around the world. It adds value and saves time and effort for researchers and users by providing the means of browsing and searching resource descriptions and connecting directly to resources of interest. The catalogue currently contains over 3 500 descriptions of resources under over 160 subject headings ranging from anthropology to statistics (Hiom, 1998/SOSIG, http://www.sosig.ac.uk/).

However, these services are not yet well known or sufficiently used by researchers. The above-mentioned survey conducted at the University of Plymouth (Gadd, 1998) shows that these resources are largely under-exploited. According to the survey, only BIDS is widely used and found useful (51.3% of respondents). More than a half of all respondents (52.2%) have never heard of SOSIG. It should be noted, however, that very few find these gateways of no use (0.6% for SOSIG). There still remains the problem of making researchers aware of these services.

Figure 5. **SOSIG browsing page**



Browse the World Section of the SOSIG Internet Catalogue

You are currently browsing the **World** section of the SOSIG Internet catalogue, but you can restrict your browsing to **Europe** or **UK** (see Help on Browsing).

World | Europe | UK

University Social Science Departments

Economics
Labour and Industrial Relations | Political Economy | Finance ...

Education
Teaching Methods | Higher Education | Special Education ...

Environmental Sciences and Issues
Social Ecology | Protection of the Environment ...

Ethnology, Ethnography, Anthropology
Anthropological Teaching and Research | Folklore ...

Geography
Social Geography | Economic Geography | GIS and Cartography ...

Government
Public Law Enforcement | Local Government | Regional Government ...

Law
Civil Rights | Human Rights | International Law ...

Management
Accountancy | Business and Industrial Management | Advertising ...

Philosophy
Metaphysics | Logic | Ethics ...

Politics
Elections | International Relations | Political Parties ...

Psychology
Applied Psychology | Cognition | Developmental Psychology ...

Social Science General
Social Science Methodology | Social Policy ...

Social Welfare
Youth Welfare | Addiction | Homelessness | Social Problems ...

Sociology
Schools and Theories | Sociologists | Sociology of Politics ...

Statistics
Statistical Theory | Demography | Official Statistics | International Statistics...

Women's Studies
Women's History | Women and Employment | Women and Politics...

You are currently browsing the **World** section of the SOSIG Internet catalogue, but you can restrict your browsing to **Europe** or the **UK** (see Help on Browsing).

Source: SOSIG, http://www.sosig.ac.uk/.

Table 2. **Use of Internet services in the University of Plymouth**

	Never heard of it	Heard of/ Not used	Used/ No use	Used/ Some use	Used/ Very useful	Unsure
British Education Index (BIDS)	38 (23.8%)	36 (22.5%)	3 (1.9%)	30 (18.8%)	52 (32.5%)	1 (0.6%)
SOSIG (social science)	84 (52.2%)	53 (32.9%)	2 (0.6%)	14 (8.7%)	6 (3.7%)	2 (1.2%)
Nursing and Medical Information Gateway (OMNI)	129 (80.1%)	29 (18%)	-	1 (0.3%)	-	2 (1.2%)
Biz/ed (business and economics)	143 (88.3%)	17 (10.5%)	-	1 (0.6%)	-	1 (0.6%)
Devline (development studies)	143 (88.8%)	17 (10.6%)	-	-	-	1 (0.6%)

Source: Gadd, 1998.

Among generic search engines, some offer catalogues of information resources in specific fields. For example, Yahoo! provides access to social science information resources in many languages. Its US Web site (http://dir.yahoo.com/Social_Science/) includes "social sciences" in its sub-categories, with an information catalogue ranging from anthropology to women's studies. This service is available in several languages.

Even when researchers discover a resource which seems relevant, they cannot easily find out whether the information is reliable and recent. In addition, data retrieval is time-consuming and often expensive.

Researchers' abilities in terms of ICT skills are very broad. While ICT, especially the Web systems, can significantly widen access to worldwide data and information resources, it may also reduce accessibility for those who lack sufficient computer skills. According to the survey at the University of Plymouth, even among those interested in secondary data analysis, 85% said that they were likely to use journals or books containing statistical tables, but only 60% said that they were likely to use a computer programme or CD-ROM, and the same percentage said that they were likely to use the Internet. Moreover, 61% said they were likely to need support services to find relevant sources, while 46% said they were likely to need them to help analyse secondary data. Finally, 31% confirmed that they would take staff development training in this area (Gadd, 1998). There is substantial need for training courses, especially for those lacking in computer skills, as well as more user-friendly systems. The above-mentioned Essex Summer School in Social Science Data Analysis and Collection, for example, offers courses for such people.

On the other hand, a simple search system which is easy for every user to use will not satisfy more experienced users who demand a more sophisticated system. Striking the right balance between the two kinds of requirements is problematic. Moreover, users' computer systems have differences in capacity. Some researchers use a simple and slow dial-up connection, while others use material that allow for very high rates of data transmission and powerful data analysis. This implies a range of different approaches in order to satisfy these different patterns of use.

There are several projects for developing more advanced systems. The NESSTAR project, for example, funded by the European Commission under the information engineering sector of the Telematics Applications programme and maintained by the Data Archive in the United Kingdom, is developing a set of generic tools that will make it easier to locate data sources across national boundaries, to browse detailed information about the data, to tabulate and visualise the data, and to download the appropriate subsets of data in one of a number of formats.

Restrictions

Restrictions to be imposed on future use of data raise very important questions which should be settled when a data set is transferred to the archive. Technically, it is possible for any user in a network to access and download all data sets in the archive. Some archives make their data accessible to the general public, subject only to restrictions imposed by legislation or by the investigators (for example, DDA in Denmark). In the case of the PACO database (see below), for instance, its database (containing harmonised data and documentation) can be accessed by outside users, but not the PACO Panel Archive (containing the original data). Mochmann and de Guchteneire (1998) enumerate a set of possible clauses which data archives can offer:

- No restrictions.
- Free for academic public research.
- Publication based on the data requires the donor's consent.
- Accessible only after written permission of donor.
- Use and publications to be brought to the attention of donor.

Restrictions imposed by the data owner are one of the constraints on the rapid delivery of data across the network. Data owners are often reluctant to provide data for secondary analysis, especially microdata, for reasons such as privacy protection. Because these restrictions need to be handled efficiently, a system which sets up a rapid means of authorising access to more restricted parts of the system is essential. At the same time, once authorisation is granted, it is essential that it is secure and that authorised users can be reliably authenticated and that unauthorised users are unable to breach the security of the system (NESSTAR, http://dawww.essex.ac.uk/projects/nesstar.html).

Data have various owners. Some archives, such as the UK Archive, do not own data but hold and distribute them under licences signed by data owners. Data ownership should be clarified to users, but it is not easy to deal with; Mochmann and de Guchteneire (1998) point to use of the concept of copyright instead of ownership for facilitating the problem. They explain: "The original researcher who created the data set normally owns a copyright to the data set. This copyright can be handled similarly to the copyright on books or works of art. The data archive may acquire its own copyright to a data set. Storing a data set in the archive will almost always involve changes to the original materials. Data are reformatted, documentation is added, variables are recoded to standards, etc. This process will add a claim on copyright to the data set for the part which the archive contributed. So, on any archived data set both donor and the archive will have a claim."

In the United Kingdom, the Office for National Statistics (ONS), whose data are provided to the academic community by the Data Archive, upholds certain requirements imposed by holders of data copyright. For example, ONS has delegation of copyright from Her Majesty's Stationery Office, and consequently, Crown copyright data should not be passed on without appropriate recognition; where appropriate, a royalty fee may be due. In practice, for *bona fide* academic research purposes (defined as when the researcher has total control and responsibility for the research outputs and their publication, without any conditions imposed by the sponsor), royalty fees are waived, although the data must still be recognised formally as Crown copyright material. However, when academics move outside pure academic research, royalties become due (Sylvester, 1998).

As a formality, some archives have registration process for users. In the case of the SSDA in Australia, users are simply asked to sign and return an "Undertaking Form" in which they agree to acknowledge the original depositor and distributor in any work based on the data. However, Cole and

McCombe (1998) report that, based on feedback from users of MIDAS (see Annex) in the United Kingdom, the complex registration process for many of the copyright/commercial data sets held in the archive acts as a major deterrent to use, particularly for teaching purposes.

Archives open to the public try to preserve the confidentiality of respondent data and to ensure that confidentiality is protected when data are released. However, a considerable number of research projects in the social sciences need to use data on identifiable persons. Mainly for that reason, severe restrictions have been imposed on publicly supported data collections such as census data, in spite of demands for open access from researchers. Once data are released to the public, it is hardly possible to monitor use to ensure that other researchers respect respondent confidentiality. Thus, it is common practice in preparing public-use data sets to alter the files so that information that could imperil the confidentiality of research subjects is removed or masked before the data set is made public (ICPSR, 1999).

If ICPSR receives data sets with identifiable variables, such as name or social security number, it removes them as part of the first level of processing. Increasingly, consideration is being given to returning to investigators data sets received with specific identifiers. This is because ICPSR practice is to preserve originally submitted data; legal action could be taken should it be known that ICPSR maintains a copy of such a data set. ICPSR staff consult with principal investigators to help them design or modify a public-use data set so that it maintains (insofar as possible) the anonymity of respondents (Dunn and Austin, 1998).

Cost is another problem. Normally, data are not sold by archives. High royalty fees may be prohibitive for scientific use and this approach has not been widely introduced. Some archives charge a service fee for each request, others do not. Charging a small fee forms a threshold which avoids abundant misuse of the services (Mochmann and de Guchteneire, 1998).

Access to publicly produced data

Each government has a long history of collecting data on its society in order to facilitate decision making. These data derive from social surveys, censuses, administrative records, etc., and remain the major source for social science research. Guy (1997) reports that more than 70 agencies in the United States produce statistics. The official data are stored separately in each ministry or gathered in a national archive or a governmental statistical office.

Many government statistics are derived from administrative systems. The main advantage of such data sources is that they tend to be relatively cheap to use and are often timely. The disadvantage is that the statistics derived from such administrative systems are intrinsically governed by the scope, objectives and processes of that system. This means that they may not always use definitions and coverage that would be preferable from a statistical perspective. Changes in the workings of the particular system can lead to discontinuities in the statistical series across time (Government Statistical Service, http://www.statistics.gov.uk/aboutgss/uksub.htm).

Official data are published in the form of white papers, governmental statistical reports, etc. They are now often made accessible via the World Wide Web, such as through the Web site of the National Center for Education Statistics of the US Department of Education. Some of the official data are resold by private companies, often with value-added metadata.

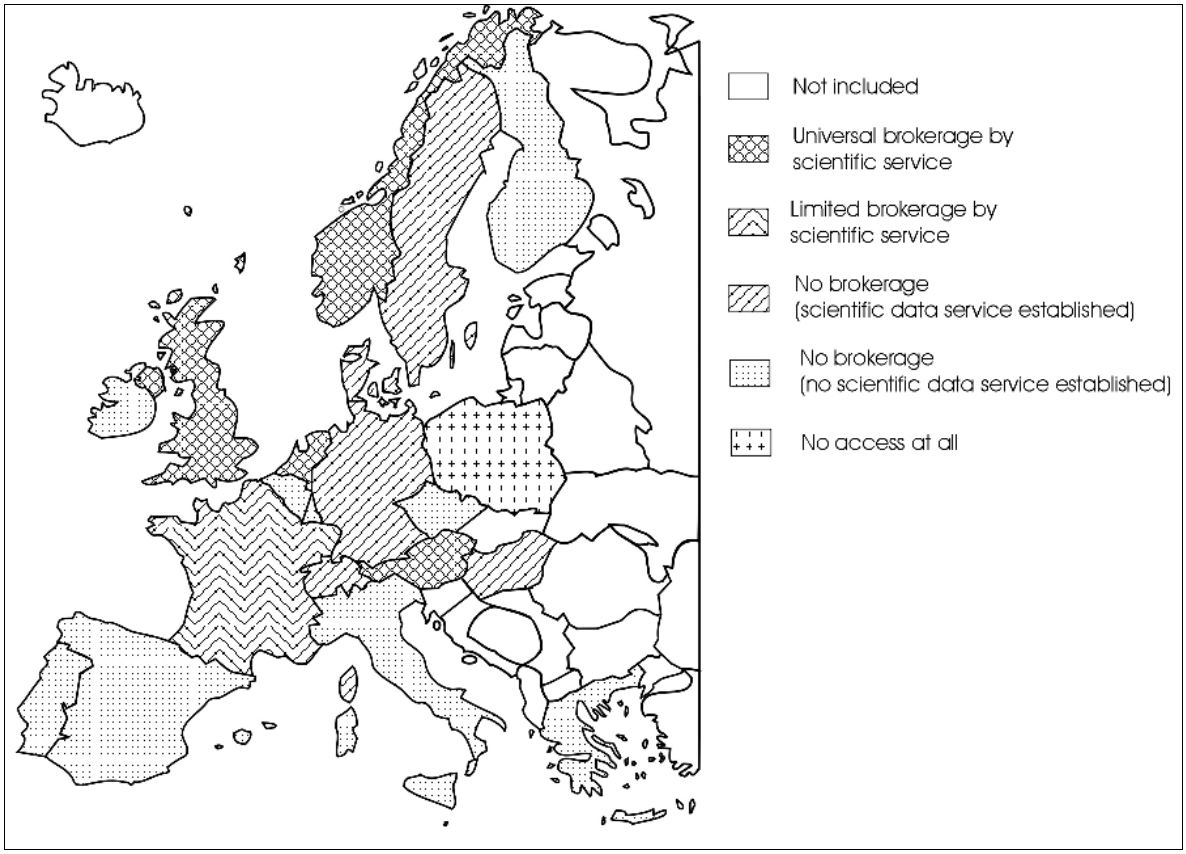
However, these governmental data are often under-exploited because they are not disclosed to the public or made available for reasons such as protection of confidentiality. In addition, although

accessible, they are often not suitable for secondary analysis and require processing before being analysed.

It should be noted that microdata derived from academic surveys are widely made available via social sciences data archives, but that access to government microdata, such as census data, is severely limited owing to the confidentiality requirements often imposed by data protection laws. Kraus (1998) notes that, in contrast to academia, with its established tradition of data sharing and the early reconciliation of the inherent conflict between data protection and data access, it is usually more difficult for government agencies to allow third parties access to the collected microdata. Also, confidence of the interviewees is a costly good and data protection a crucial tool to ensure this confidence.

Access to publicly produced microdata varies greatly across countries. Kraus (1998) presents an overview of the situation of access to official microdata in Europe, including information on the role of national social science services (Figure 6 and Table 3). Flora (1997) notes that in Europe only a few countries, above all Britain and Norway, have liberalised access adequately (and provided it at reasonable cost) to satisfy the needs of current academic social research. Many social scientists have denounced the insufficient access to public microdata and have demanded free access for academic purposes. Kraus (1998) points out that access to key sources such as population censuses and establishment surveys is most restricted in Europe, and proposes making national science organisations intermediaries between government agencies and individual users in order to arrange data protection measures and meet users' research needs.

Figure 6. **Availability of country-wide services of official microdata in Europe**



Source: Kraus (1998).

Table 3. **Access to official microdata in Europe: some basic characteristics of regulations as of 1998**

Country	Provision of access	Application procedure and access restrictions	Data dissemination practice and support by social science infrastructure
Austria	Optional	- Application to statistical office - Specification of research purpose - No access to samples of population censuses	- National social science service available (WISDOM) - Supports use of official microdata (disseminates ready-to-use files after approval by statistical office)
Belgium	Optional	- No standard procedure - Access difficult	- No national science service in operation - Mediation through Belgium Ministry for Scientific Research
Czech Republic	Optional	- Order to statistical office; user-tailored files	- No national science service in operation
Denmark	Optional	- Application to statistical office - Specification of research purpose and data needs - Use limited to security area of the statistical office - Accepted users are to be sworn in (Special situation because of the extensive use of register data instead of survey data)	- National social science service available (DDA) - Supports users in their applications (no dissemination of data allowed)
Finland	Optional	- Application to statistical office (research proposal and data needs) - Linkage of information from registers possible	- National social science data archive in phase of foundation (note of the author: Finnish Social Science Data Archive was founded in 1999)
France	Optional	- Non-members of CNRS: Order to INSEE - Members of CNRS: LASMAS - No access to samples of population census	- National social science service available, but brokerage established outside the service - CNRS/LASMAS provides centralised support for members of CNRS
Germany	Optional; but limited to applicants resident in Germany	- Application to statistical office - Specification of research purpose, duration and of required variables [Special protection measures mandatory (admission lists, etc.)]	- National social science service available (GESIS) - Only general support, but no data brokerage allowed (Prohibition to disseminate anonymous microdata to (non-German) researchers abroad is derived from the general principle that German law is not valid outside of Germany)
Greece	Optional	- Request to statistical office; decision taken by a specialised committee of the office.	- No national social science data archive available
Great Britain	Mandatory	- Application to ESRC Social Science Data Archive; - Specification of variables - Access to microdata of population census limited to members of ESRC organisations	- National economic and social science services available - Provision of ready-to-use files by ESRC data archive - Wide variety of services for enhanced use of official microdata (MIDAS etc.)
Hungary	Optional	- Order to statistical office - Public use files and user-tailored files - Access to samples of population censuses	- National social science service available (TARKI), but no brokerage established
Ireland	Optional	- Order to statistical office - Standard files - No access to samples of the population census	- No national social science service available

Table 3 *(cont'd)*. **Access to official microdata in Europe: some basic characteristics of regulations as of 1998**

Country	Provision of access	Application procedure and access restrictions	Data dissemination practice and support by social science infrastructure
Italy	Mandatory	- Application to statistical office; - Specification of research purpose and duration - Standard files - Evaluation of request by a special committee based on published criteria (No access to samples of population censuses)	- No national social science service available (initiative running)
Netherlands	Optional	- Application to statistical office - Specification of research purpose and duration - Standard files (WSA) resp. user-tailored files (CBS) - Evaluation of application by special committee [Special situation: increased use of administrative sources (population census discontinued)]	- National social science service available, but brokerage established outside this service - Special agreements between statistical office (CBS) and Dutch Research Council - Dissemination of standard files by an infrastructure unit of the national research council (WSA)
Norway	Optional	- Application to national social science services (NSD) - Specification of research purpose and duration - Standard files - Evaluation of request by NSDl - Access to small samples of the population census possible	- Dissemination of ready-to-use standard files by NSD
Poland	No access	- Law on official statistics, issued 29 June 1995, Art. 38	- No national social science service established
Portugal	Optional	- Application to statistical office - Specification of research purpose - User-tailored files - Access, in principle, to all surveys	- No national social science service available
Spain	Optional	- Order to statistical office - Provision of public use files - Access to small sample of population censuses possible	- No national social science service available (note of the author: the *Archivo de Estudios Sociales* was recently created)
Sweden	Optional	- Application to statistical office Specification of research purpose, duration and data needs User-tailored files (Special situation: extensive use of administrative data instead of surveys)	- Direct dissemination (national social science service available, but no brokerage established) - In special cases work in secure area of the office
Switzerland	Optional	- Application to statistical office - Specification of purpose and variables (Access to small sample of population survey possible)	- Direct dissemination (national social science service available, but no brokerage established)

Source: Kraus (1998).

Official data are often reserved to officials of statistical offices or, if they are accessible to researchers, they must be used within the statistical office. In the European Union, the Commission Decision of 21 April 1997 (97/281/EC) stipulates that data considered confidential shall be made accessible within the Commission only to officials of Eurostat, other staff of Eurostat and other natural persons working on the premises of Eurostat under contract, and shall be used by them only for purposes defined in the framework of the Basic Regulation.[2]

As for statistical confidentiality, Council Regulation (EC) No. 322/97 of 17 February 1997 on Community Statistics (http://europa.eu.int/eur-lex/en/lif/dat/1997/en_397R0322.html) stipulates that data used by the national authorities and the Community authority for the production of Community statistics shall be considered confidential when they allow statistical units to be identified, either directly or indirectly, thereby disclosing individual information, and that, to determine whether a statistical unit is identifiable, account shall be taken of all the means that might reasonably be used by a third party to identify the said statistical unit (Article 13). Access to confidential data for scientific purpose is also laid down in the Article 17, but is fairly limited as mentioned above.

In the United Kingdom, a comprehensive agreement was concluded in 1996 between the Data Archive at the University of Essex and the Office for National Statistics (ONS), to facilitate access by academics to official data. It covers all the data sets provided by ONS and makes the entire process easier and quicker for those wishing to access ONS data via the Data Archive, which acts as an ONS agent, (Sylvester, 1996). In addition, ONS launched in 1998 "StatBase", on behalf of the Government Statistical Service, to provide access to a comprehensive set of key statistics drawn from the whole range of UK official statistics available across government.

In Germany, the Federal Statistical Office, with the support of the Ministry of Education, Science and Technology, has made major efforts to provide the scientific community with easier access to the microdata of various surveys. However, the law still prohibits the dissemination of data via third parties, such as the National Social Science Services (GESIS), as well as the dissemination of microdata among interested researchers abroad (Kraus, 1998).

The Data Liberation Initiative (DLI), a Canadian pilot project initiated in 1996, provides Canadian academic institutions with affordable access to Statistics Canada data files and databases for teaching and research. Under the DLI, universities are able to acquire data for a set annual subscription fee (CAD 3 000-12 000). This eliminates the need for *ad hoc* consortia and grant-funded purchases. It includes public-use microdata files (anonymous records from surveys such as the General Social Survey, the Census, and the Survey of Labour and Income Dynamics). It should be noted that the use of the files is limited to instruction and scholarly research and that the files may not be used for commercial purposes.

Belgium's AGORA programme, initiated by the Federal Office for Scientific, Technical and Cultural Affairs (OSTC) in 1999, is to provide scientific support, upon request, to federal institutions to facilitate exploitation of their information by outside users. It helps federal institutions to harmonise and reprocess their data and facilitates researchers' access to the data. It also aims to collect non-administrative socio-economic data which are useful for research that may enlighten federal policies (note from OSTC).

Some efforts have been made by governmental statistics authorities to make microdata available while protecting confidentiality, thereby often suppressing some data. However, this kind of anonymous data may hamper research in the social sciences since researchers with more complex theories and more powerful statistical tools increasingly find that suppression of information in traditional public-use files significantly limits their analysis. It becomes increasingly difficult to develop a microdata file that maintains confidentiality when longitudinal data are available for a number of years. Bernard *et al.* (1998) propose "research data centres" across major urban centres and university campuses throughout Canada, which would provide as secure physical locations for confidential data as the offices of Statistics Canada.

International co-operation

Harmonisation of data

Today, archives store a very large number of data sets derived from various social surveys and social research across the world. However, the data are not always comparable across countries; consequently, most of the empirical research done in the pertinent disciplines is not internationally comparable. This is true in part because many, probably most, researchers are not internationally oriented to start with (Kaase *et al*., 1997). Because of the paucity of standards and comparative studies in social surveys and data collections, the harmonisation of acquired data is a very difficult task. Mochmann (1998) notes:

> "Owing to the lack of standardisation in social research, the database has to be harmonised…. This is a conceptually, technologically and methodologically demanding task. Since contemporary structures of social research still are geared to national needs, these additional challenges can hardly be tackled with given resources. In addition, interests differ from country to country and is very difficult to define priorities or to decide what should be harmonised." (Mochmann, 1998)

Co-ordination of surveys and harmonisation of data should preferably be undertaken before the survey in terms of targets, sampling size, data format, etc., to ensure better possibilities for comparative research because *post hoc* harmonisation efforts such as the Panel Compatibility Study (see below) necessitate much more work and are generally less reliable. Flora (1997) reiterates the need to strengthen efforts to harmonise social surveys *ex ante* in Europe in decentralised but co-ordinated efforts, and proposes as a first step to determine the overlap between existing surveys and longitudinal data. However, harmonisation of one country's data with those of other countries may break the continuity of the data and make longitudinal research impossible in that country. For this reason, data collection across countries should preferably be harmonised in such a way that they may still be connected to data at national level.

There have been many efforts by international organisations, such as the UN, UNESCO and the OECD, and the academic community to standardise data classifications, such as ISCED at UNESCO in the field of education and the OECD's *Frascati Manual* in the field science and technology including social sciences. Kraus (1998) notes that in the last two decades there has been a remarkable convergence of enumeration programmes, concepts and methods for standardisation of statistics, although he points to considerable remaining differences. This standardisation enables various forms of data comparison among countries. There is also much cross-national co-operation in data collection by international organisations, such as the European Commission's Eurobarometer public opinion surveys. However, data sets obtained by international organisations do not always cover all the member countries owing to the lack of sufficient co-operation or because they are not suitable for comparative research because of country specifications tolerated in the data collection methodology.

The Luxembourg Income Study (LIS), co-ordinated by CEPS/INSTEAD, is a co-operative research project with membership in 25 countries. It aims to test the feasibility of creating a database consisting of social and economic household survey microdata from different countries, and to provide a method that allows researchers to access the data under the privacy restrictions set by the countries providing the data. It promotes in particular comparative research on the economic and social status of populations in different countries. In 1994, in association with LIS, the Luxembourg Employment Study (LES) was initiated to construct a data bank of labour force surveys from countries with quite different labour market structures. These surveys provide detailed information on areas such as job search, employment characteristics, comparable occupations, investment in education, migration, etc.

The LES team has harmonised and standardised the microdata from the labour force surveys to facilitate comparative research. Attention has focused on including countries with very different structural characteristics regarding the level of unemployment, the participation of women in the labour force, the importance of trades unions, and the existence of special labour market contracts (Luxembourg Income Study, 1998).

The International Social Survey Programme (ISSP) is an annual programme of cross-national collaboration on surveys covering topics that are important for social science research. It brings together existing social science projects and co-ordinates research goals, thereby adding a cross-national, cross-cultural perspective to the individual national studies of the 31 member countries. It was established in 1984 to: *i)* develop jointly topical modules dealing with important areas of social science; *ii)* include the modules as a 15-page supplement to the regular national surveys (or a special survey if necessary); *iii)* include an extensive common core of background variables; and *iv)* make the data available to the social science community as soon as possible. ISSP researchers especially concentrate on developing questions that are meaningful and relevant to all countries and can be expressed in an equivalent manner in all relevant languages.

	SSP surveys
1985	Role of Government I
1986	Social Networks
1987	Social Inequality I
1988	Family & Changing Gender Roles I
1989	Work Orientations I
1990	Role of Government II
1991	Religion
1992	Social Inequality II
1993	Environment
1994	Family & Changing Gender Roles II
1995	National Identity
1996	Role of Government III
1997	Work Orientations II Fielding & Archiving
1998	Religion II
1999	Social Inequality III

The PACO (Panel Comparability) project of CEPS/INSTEAD takes a centralised approach to creating an internationally comparable database integrating microdata from various national household panels over many years. The PACO database contains harmonised and consistent variables and identical data structures for each country included. The PACO database increases the accessibility and use of panel data for research and facilitates comparative cross-national and longitudinal research on the processes and dynamics of policy issues such as labour force participation, income distribution, poverty, problems of the elderly, etc.

At European level, Eurostat works towards an integrated European statistical system. Eurostat has created common classifications, methods and organisational structures for compiling comparable statistics on EU member states. All data collected from the national statistical institutes are checked by Eurostat, compiled in the appropriate form and, where applicable, harmonised with European Statistical System standards. Eurostat now aims to include central and eastern European countries. The

EUREPORTING project, which aims to create a science-based European System of Social Reporting and Welfare Measurement, is under way with financing from the European Commission (see Annex).

In 1997, as a science-based initiative, the European Social Survey (ESS) was launched by the European Science Foundation with a view to building an integrated social science database for European comparative research. ESS is expected to provide systematic and regular data on topics of major interest to the European social science community and, as a facility, to encourage comparative analysis of political, social and economic trends by measuring systematically, at regular intervals, citizens' attitudes and behaviours relating to a core set of political, social, and economic domains (The European Science Foundation, 1998). Flora (1997) proposes, at European level, a new kind of European observatory, based on networks of social scientists and research institutes, focusing on more specific policies and institutions, and trying to integrate a variety of sources and data.

Networking of archives

Many data archives maintain a variety of national and international co-operative efforts to ensure efficient use of information resources. Generally, the Web site of each archive allows users to reach many other data-collecting institutions by virtual links. Sometime users can obtain data sets from other institutions through their preferred archive. Internet gateway services have also been made available to guide researchers to many of the world's data archives and data service institutions.

One problem is that networking means dealing with the variety of interfaces and formats for the various data and information resources available in the world's data archives. To do so, a special software package which supports many data formats (SIR, SAS, SIR, etc.) is required. The ROADS Project in the United Kingdom aims to provide an infrastructure to support the development of subject-based information gateways. Its purposes are: *i)* to produce a software package which can be used to set up subject-specific gateways; *ii)* to investigate methods of cross-searching and interoperability within and between gateways; and *iii)* to participate in the development of standards for the indexing, cataloguing and searching of subject-specific resources. The ROADS software (version 1) is freely available for production use and is being used, for example, by the Social Science Information Gateway (SOSIG) (http://www.roads.lut.ac.uk/roads-software.html).

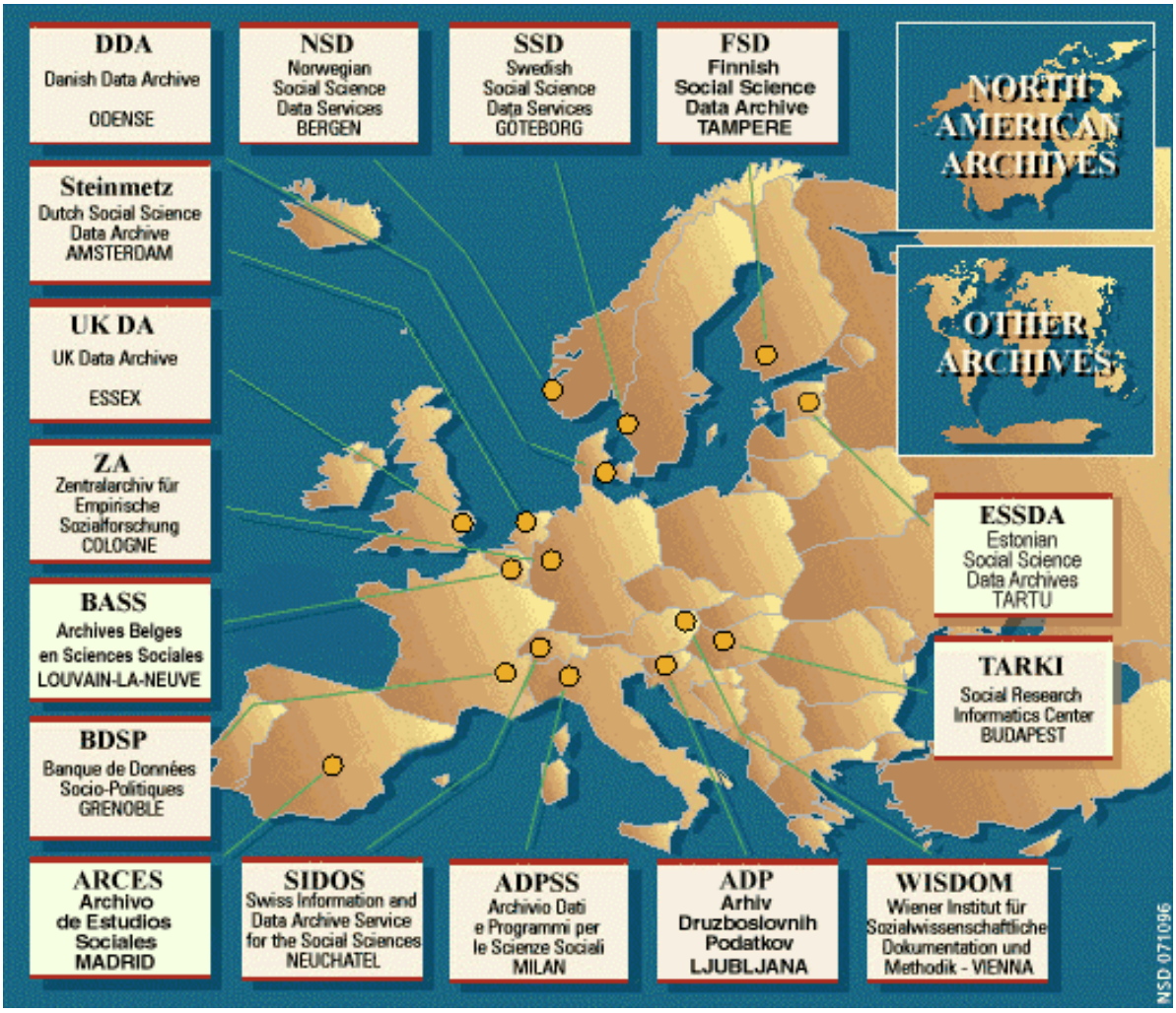
Another problem is that most databases are written in the language of the data owner, especially in the case of social survey outputs. This could be a major barrier for comparative research projects, even though the data are made available to all researchers worldwide via the Web system. There have been some efforts to translate data sets, such as the Roper Center's Japanese Data Archive in the United States, which is a collection of Japanese survey data covering some 1 500 reports and about 200 data sets (http://www.ropercenter.uconn.edu/jpoll/home.html).

GESIS (*Gesellschaft Sozialwissenschaftlicher Infrastruktureinrichtungen*) in Germany's *Wissenschafts-gemeinschaft Gottfried Wilhelm Leibniz* (WGL) is an infrastructure association which supplies fundamental social science services mainly for German-speaking researchers, at national and international level, in terms of both theory and practice. GESIS consists of *Informations Zentrum Sozialwissenschaftern (IZ), Zentralarchiv für Empirische Sozialforschung an der Universität zu Köln* (ZA) and *Zentrum für Umfragen Methoden und Analysen* (ZUMA) (http://www.social-science-gesis.de/index-e.htm).

International data organisations were founded by national archives to promote co-operation on the multiplicity of archival tasks and to foster a worldwide network of data services for the social sciences. In 1976, the Council of European Social Science Data Archives (CESSDA) was created by European archives and has extended collaboration to the rest of the world. It aims to promote the

acquisition, archiving and distribution of electronic data for social science teaching and research in Europe (Figure 7). It encourages the exchange of data and technology and fosters the development of new organisations in sympathy with its aims. The organisational network is being electronically linked via the Internet. The CESSDA homepages allow easy access to the catalogues of member organisations and provide a central news forum about its activities and other relevant information.

Figure 7. **CESSDA participants in Europe**

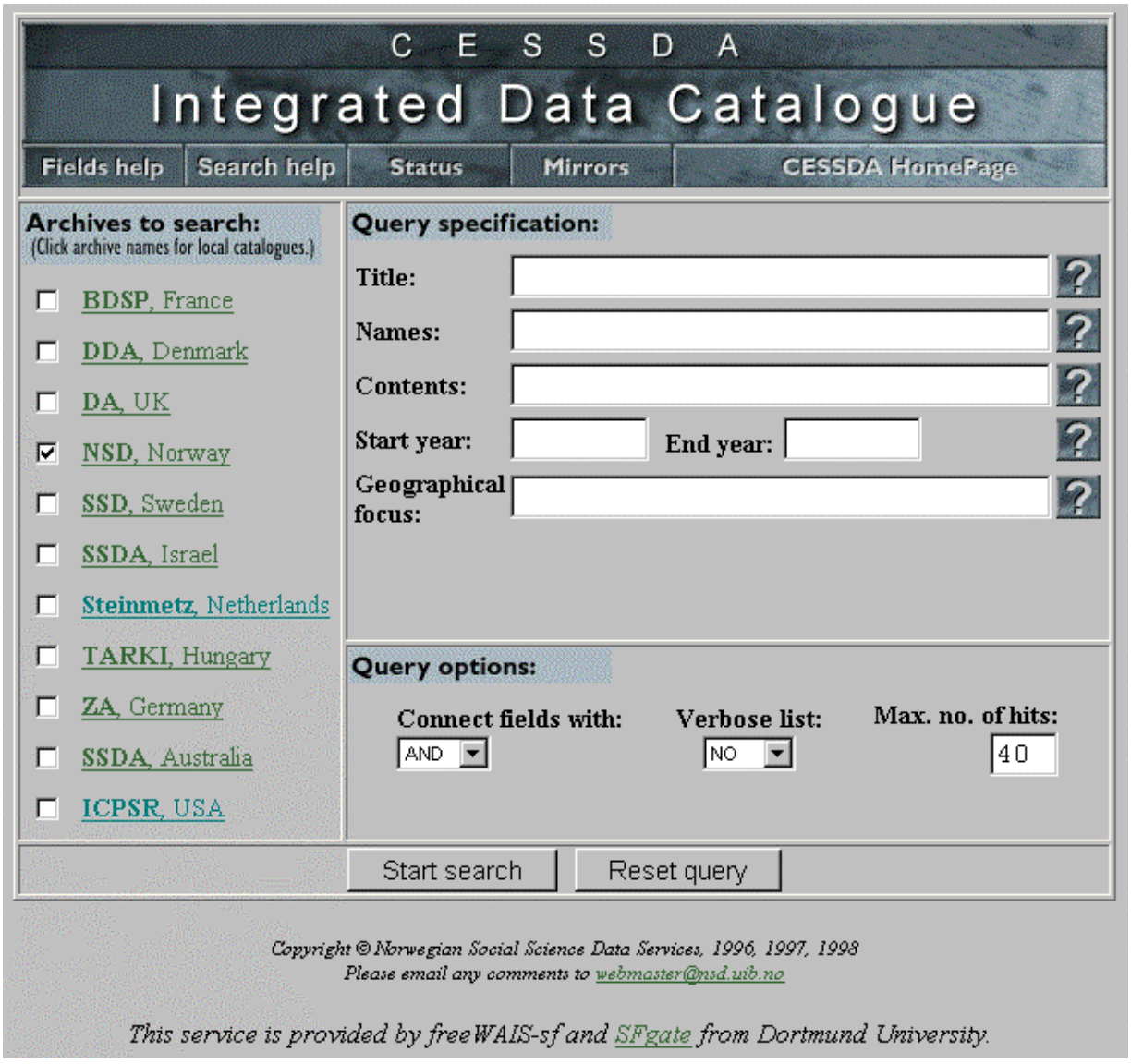


Source: CESSDA, http://www.nsd.uib.no/cessda/.

CESSDA's Integrated Data Catalogue (IDC) offers easy searching of its participants' data archive catalogue information. Presently, eleven catalogues can be searched in an integrated way via the Internet (Figure 8). The search can be done by entering keywords in relevant fields. Five fields have been agreed by the CESSDA partners: title of study, name(s) of the principal investigator(s), contents (abstracts describing the studies or keywords), start/end year and geographical focus (country). One problem is that some of the participating archives have their data catalogues in their national language. In the first stage, only the English translation of study titles is required, but translation of the rest of the material will be an ongoing process. However, translation involves a considerable amount of work. The possibility of integrating a (multiple language) thesaurus in the

search system is being evaluated as another solution to minimise the need for translation (Hassan *et al.*, 1996).

Figure 8. **CESSDA Integrated Data Catalogue**

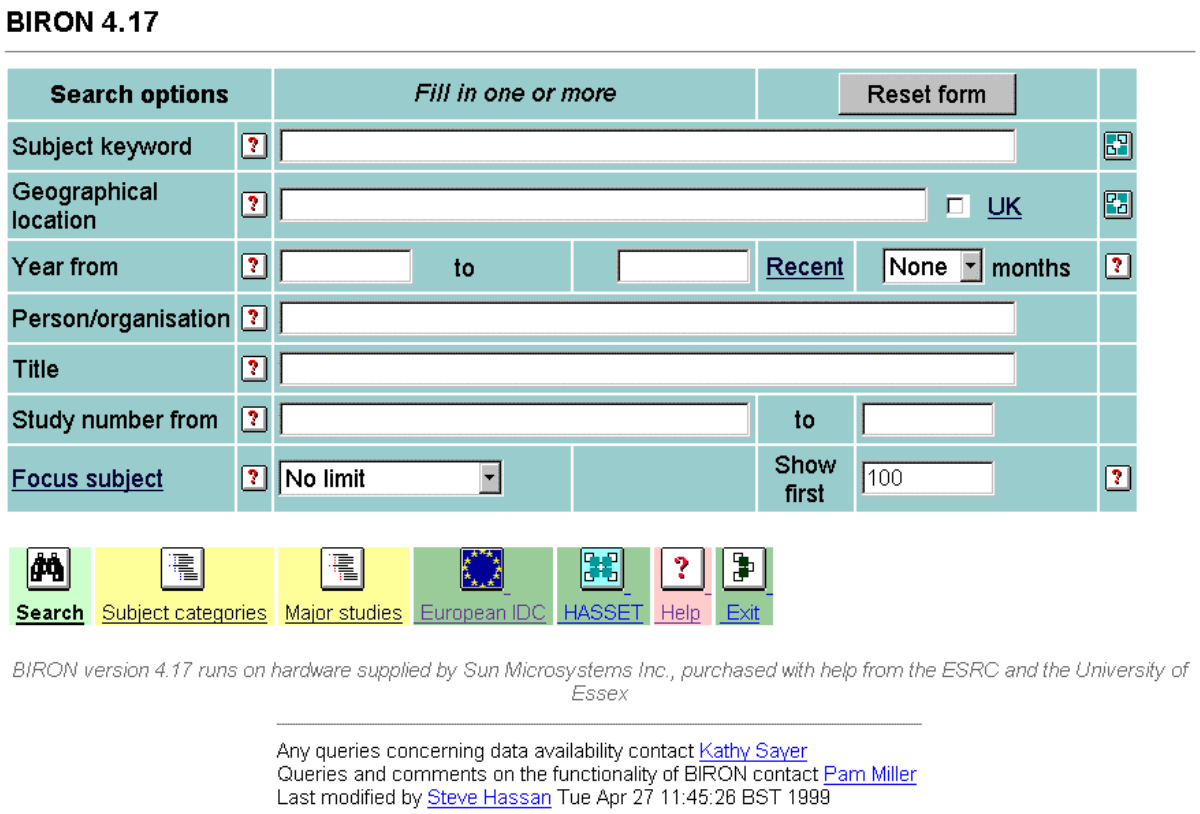


Source: CESSDA, http://www.nsd.uib.no/cessda/.

The IDC system has the advantage of being extremely simple to use as well as offering an easy way of extending the search into the WAIS[3] catalogues of other data archives worldwide. For a more complex search, users may link directly to archives to see their complete catalogue. The Data Archive in the United Kingdom, for example, with its more complicated BIRON (Bibliographic Information Retrieval On-line) system, offers greater accuracy for research (Figure 9). IDC's search mechanism is much less finely tuned and does not use the power of the authority lists which BIRON employs to ensure that names and other access points are fully cross-referenced. Similarly, although keywords are listed in the IDC, the thesaurus function, which links non-preferred and preferred terms and allows

searching up and down hierarchies is not operational for the IDC (The Data Archive, http://dawww.essex.ac.uk/).

Figure 9. **BIRON 4.17 catalogue**

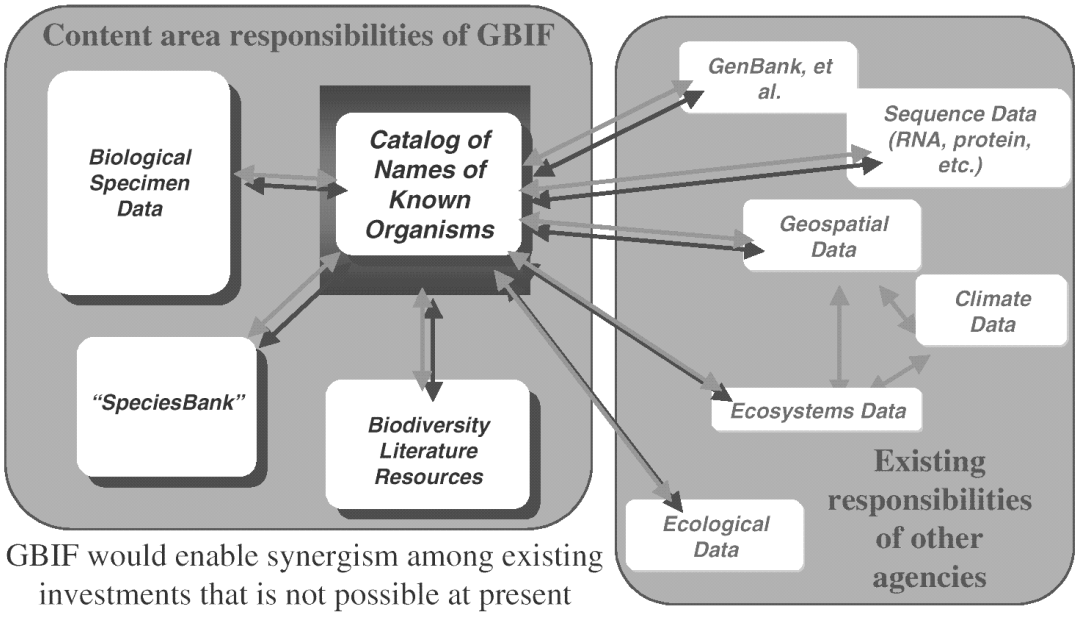


Source: The Data Archive, http://dawww.essex.ac.uk/.

Building on these experiences, more advanced systems are under construction, such as the NESSTAR system, which is an enhanced version of IDC, and ILSES (Integrated Library and Survey Data Extraction Service Project). ILSES is a project of a number of Dutch, German, French and Irish institutes, which was accepted by the European Commission under the Fourth Framework Programme. It aims to develop a service that enables individual users to access and retrieve documentary information and empirical data related to large-scale surveys. In addition, it offers tools and procedures for the normalisation, cataloguing and controlled distribution of (distributed) holdings of documentary and data resources.

One example in another field of science which could be drawn on is the Global Biodiversity Information Facility (GBIF), studied in the OECD's Megascience Forum. It is proposed as a distributed system of interlinked and interoperable modules (databases, software and networking tools, search engines, analytical algorithms, etc.), which will enable users to navigate and put to use vast quantities of biodiversity information via Internet (Figure 10). The GBIF knowledge base and informatics will provide infrastructure support to information networking efforts, including clearinghouses.

Figure 10. **Global Biodiversity Information Facility**



Source: OECD, 1999.

The International Federation of Data Organizations for the Social Sciences (IFDO) was formed in 1977 for the larger community of social sciences. Its IFDOnet provides structured information on social science topics and the up-to-date World Wide Web infrastructure of data archives. Fundamental documents on archive management are offered as well as reflections on data analysis issues. The site supports international comparative analysis on sustainable development with study packs for graduate students. IFDOnet is supported by UNESCO with the affiliated MOST[4] Clearing House. At present, 29 of the world's archives are participating in IFDO.

The International Association for Social Science Information Services and Technology (IASSIST), an association of social scientists, information specialists and computer specialists, offers assistance to archives and their staff worldwide. The association provides guidelines for professional development and opportunities for learning, develops guidelines for creating social data information centres, and with other social data organisations works towards the development of guidelines and standards for the preservation of social data. It promotes global linkages between social science data information centres, including developing centralised data collections, policies for data sharing and co-operative acquisitions, and technological linkages for data and information exchange.

Finally, although networks of archives, researchers and data librarians are being developed and each system increasingly allows instant access to available resources, the quality and usability of data for comparative research is still, to a large extent, dependent on what was originally collected and documented by principal investigators. Again, the harmonisation of data as well as networking should be enhanced.

Conclusion

The need for empirical data in the social sciences will certainly grow. Data archiving will continuously increase its importance for the research community. The computing environment changes rapidly and new techniques for managing, documenting and analysing data are being developed, and the Web offers considerable scope for promoting wider and more effective use of data and information resources to meet demand. However, archiving is a major task. Far greater efforts are needed to ensure better preservation of data.

Until now, much effort has been expended to archive the best social science resources and to integrate catalogues of archives from many countries, such as CESSDA's Integrated Data Catalogue. However, problems such as the variety of archiving systems and structures in data archives and language barriers have not been completely overcome, and full integration has not been achieved. Further developments in ICT will surely enable the creation of more integrated systems in the future. For this reason, international and interdisciplinary co-operation among social scientists, data archivists and computer specialists is essential.

Data harmonisation efforts should be increased with a view to better comparative research which would be of great potential value for policy making. Such harmonisation requires organisational networks and exchanges among researchers, as countries' experiences differ. These networks can be formed, for example, at international conferences and in training courses. International programmes for data harmonisation should be developed as well.

In addition, access to existing data, especially publicly produced data, should be facilitated and new data should be produced on the basis of defined priorities. The anonymity of microdata is crucial, especially for government data. Needless to say, the development of an archiving infrastructure as well as education and training of specialists in archives are also important. Training programmes for researchers, especially young ones, should be encouraged.

Finally, the protection of confidentiality and of intellectual property rights (IPR) is important. Special attention should be given to legislation on privacy and data protection. Data archives must protect the confidentiality of those whose responses form the basis of their resources. However, advances in social science research methodology, which leads to much more sophisticated structured data, are making safeguarding data confidentiality more difficult. Careful attention should be paid to protecting privacy without reducing the value of data. Similarly, the right balance should be struck between the protection of IPR and the research community's free access to data.

ANNEX*

1. International efforts

International governmental organisations

International Monetary Fund (IMF)(http://www.imf.org)

The IMF provides key global economic indicators and publishes them twice a year in its *World Economic Outlook.* The projections and analysis contained in the *World Economic Outlook* are an integral element of the IMF's ongoing surveillance of economic development and policies in its member countries and of the global economic system, including world economic activity, exchange rates, and conditions in international financial and commodity markets. Selected series available in the publication are also available at the World Economic Outlook (WEO) Database Web site.

Organisation for Economic Co-operation and Development (OECD)
(http://www.oecd.org/statlist.htm)

The OECD provides comprehensive and comparative economic and social data, in fields such as economy, health, education, science and technology, social policy, etc. It co-operates with other international organisations in harmonising data in many fields. The indicators are published in print and machine-readable media. A limited number of data sets can be accessed via Internet.

Statistical Office of the European Communities (EUROSTAT)
(http://europa.eu.int/en/comm/eurostat/eurostat.html)

Established in 1953, Eurostat collects its data, using uniform rules, from the national statistical institutes of the countries concerned. All data are checked by Eurostat, compiled in the required form and, where applicable, harmonised with European Statistical System standards. The process of harmonising statistical data also extends to all the European Union's partners: members of the European Economic Area (EEA), the United States and Japan. Its statistical work is presented to the public via electronic media or printed publications.

---

* The information below has been collected principally via each organisation's homepage. The items are classified in alphabetical order in each section.

United Nations (UN)

InfoNation (http://www.un.org/Pubs/CyberSchoolBus/infonation/)

InfoNation is an easy-to-use, two-step database that allows users to view and compare the most up-to-date statistical data for the member states of the United Nations, including geography, economy, population and social indicators. In the first menu, users are invited to select up to seven countries they then can proceed to the data menu where they are able to select statistics and other data fields.

Social indicators (http://www.un.org/Depts/unsd/social/main.htm)

Social indicators covering a wide range of subject fields, such as population, human settlements, housing, health and literacy, are compiled by the Statistics Division, Department of Economic and Social Affairs of the United Nations Secretariat, from many national and international sources in the global statistical system. These indicators are issued in general and special print or machine-readable publications of the Division.

United Nations Educational, Scientific and Cultural Organization (UNESCO)

UNESCO Statistical Database (http://unescostat.unesco.org/database/DBframe.htm)

UNESCO collects data on education, culture and science. At present, access is provided to a wide range of education statistics maintained in the database. Access to statistics on science, culture and communication is to be developed in the future. Some of the information collected by UNESCO is organised and edited especially for publication in the *UNESCO Statistical Yearbook*.

World Education Indicators (http://unescostat.unesco.org/Indicator/Indframe.htm)

Based on the indicators published in successive editions of the *UNESCO World Education Report* and using available statistics from the UNESCO Statistical Database, an initial set of 16 most commonly used indicators have been introduced in the Web site, and another 20 are to be added shortly. An important feature of this indicator Web site is that users interested in knowing more about the conceptual and methodological aspects may directly access international technical references in the form of a conceptual framework for education indicators as well as technical specifications for individual indicators.

World Bank (http://www.worldbank.org/html/extdr/data.htm)

The World Bank provides data on the world economy through its publications and Web site. Its most general statistical publication is *World Development Indicators*. Along with its companions, the *World Bank Atlas* and the *World Development Indicators CD-ROM*, it offers a broad view of the record of development and the condition of the world and its people. It also provides a continuing survey of the quality and availability of internationally comparable indicators. The 1999 *World Development Indicators* presents 600 indicators in 83 tables, organised in six sections: world view, people, environment, economy, states and markets, and global links. The tables cover 148 economies and 14 country groups, with basic indicators for a further 62 economies.

World Health Organization (WHO) (http://www.who.int/whosis/)

WHO provides data on global health through publications and its "WHO Statistical Information System (WHOSIS)". WHOSIS describes - and to the extent possible provides access to - statistical and epidemiological data and information presently available from WHO and elsewhere in electronic or other forms. It includes "Basic Heath Indicators", "WHO Statistical Information", "WHO Mortality Statistics and Estimates", etc., and allows the user to search by keywords through the entire WHO Web site, and globally throughout the WWW.

Non-governmental bodies

Council of European Social Science Data Archives (CESSDA)(http://www.nsd.uib.no/cessda/)

CESSDA promotes the acquisition, archiving and distribution of electronic data for social science teaching and research in Europe. It encourages the exchange of data and technology and fosters the development of new organisations in sympathy with its aims. Its Integrated Data Catalogue (IDC) offers easy searching of its members' data archive catalogue information via Internet.

International Federation of Data Organizations for the Social Sciences (IFDO) (http://www.ifdo.org/)

IFDOnet provides structured information on social sciences subjects and the up-to-date WWW infrastructure of Data Archives. Fundamental documents on archive management are offered, as well as different reflections on data analysis issues. The site supports international comparative analysis on sustainable development with study packs for graduate students. IFDOnet is supported by UNESCO and developed in co-operation with the affiliated MOST Clearing House.

International Association for Social Science Information Services and Technology (IASSIST) (http://datalib.library.ualberta.ca/iassist/index.html)

IASSIST is an organisation dedicated to the issues and concerns of data librarians, data archivists, data producers, and data users. It aims to help bridge the interests and concerns of three distinct communities – social researchers and scientists, information specialists who preserve social data and computing specialists – to advance the interests of these data professionals, to promote professional development of this new career, and to take an active role in the promotion of global exchange of information, experience, and standards. The association has a membership of individuals, many of whom are professionals in social data information centres.

International Statistical Institute (ISI) (http://www.cbs.nl/isi/test.htm)

ISI is an autonomous society having as its objective the development and improvement of statistical methods and their application throughout the world.

Inter-university Consortium for Political and Social Research (ICPSR), Institute for Social Research at the University of Michigan (http://www.icpsr.umich.edu/)

Established in 1962, ICPSR is a membership-based, not-for-profit organisation serving member colleges and universities in the United States and abroad. ICPSR provides access to the world's largest archive of computerised social science data, training facilities for the study of quantitative social analysis techniques and resources for social scientists using advanced computer technologies. The data holdings cover a broad range of disciplines, including political science, sociology, demography, economics, history, education, gerontology, criminal justice, public health, foreign policy, and law. ICPSR encourages social scientists in all fields to contribute to and utilise ICPSR's data resources. ICPSR includes among its members over 325 colleges and universities in North America and several hundred additional institutions in Europe, Oceania, Asia and Latin America. Member institutions pay annual dues that entitle faculty, staff, and students to acquire the full range of services provided by ICPSR. Individuals at non-member schools can also order data for an access fee. IFDO member.

Paco Data Archive (http://www.ceps.lu/paco/pacopres.htm)

The archive, which has been set up by PACO (see below), includes original household panel data sets from ten countries. The Panel Archive contains original (not harmonised) variables, but the original data have been shifted from different platforms and formats into a common format: SPSS system files for Windows on the PC. National documentation on the original panel studies has been collected at the PACO data centre.

Resource Centre for Access to Data on Europe (r·cade) (http://www-rcade.dur.ac.uk/)

r·cade is an interdisciplinary resource centre set up by the Economic and Social Research Council (ESRC) to help researchers and analysts identify and acquire data on European social sciences. It has been set up in the United Kingdom at the Centre for European Studies at the University of Durham. It provides efficient access to key statistical data about Europe. Official statistics are available from Eurostat, UNESCO, ILO and UNIDO.

Research institutes for international studies

Centre d'Études de Populations, de Pauvreté et de Politiques Socio-Économiques / International Networks for Studies in Technology, Environment, Alternatives, Development (CEPS/INSTEAD) (http://ceps-nt1.ceps.lu/index.htm)

CEPS/INSTEAD is a research centre and network based in Luxembourg which carries out national and international socio-economic studies and develops comparative socio-economic microdata sets, with the aim of providing instruments for analysing, programming and simulating socio-economic policies. In 1997, the Centre was selected by the EC's DG XII for support under its TMR programme ("Access to Large-scale Facilities").

International Programs Center (IPC), US Census Bureau (http://www.census.gov/ipc/www/)

IPC conducts demographic and socio-economic studies and strengthens statistical development around the world through technical assistance, training, and software products. Its work is commissioned and funded by US federal agencies, international organisations, non-governmental organisations, private businesses, and other governments. For over 50 years, IPC has assisted in the collection, processing, analysis, dissemination, and use of statistics with counterpart governments throughout the world. Its International Data Base (IDB), a computerised data bank, contains statistical tables of demographic, and socio-economic data for 227 countries and areas of the world.

International programmes in social sciences

European Social Survey (ESS) (http://www.esf.org/sp/ESSa.htm)

ESS is to seek to define a "large-scale facility" for social science research, a research instrument measuring systematically, at regular intervals, citizens' attitudes relating to a core set of political, social and economic issues. Its data findings would be accessible to researchers through a co-ordinated network of national data archives and other facilities. It would complement existing data collections from other sources such as national bureaux or the European Bureau of Census (Eurostat).

European System of Social Indicators (EUSI)
(http://www.zuma-mannheim.de/data/social-indicators/eusi.htm)

EUSI aims to provide a theoretically as well as methodologically well-grounded selection of measurement dimensions and indicators, which can be used as an instrument to continuously observe and analyse the development of welfare and quality of life as well as changes in the social structure at European level. Particular attention is to be paid to coverage of the European dimension (identity, cohesion); the incorporation of new dimensions of welfare and social change, such as social exclusion and sustainability; the search for new and better indicators within the domains covered; exploitation of the best available databases and cross-country comparability. This project is a subproject of EUREPORTING.

Integrated Research Infrastructure in the Socio-economic Sciences (IRISS-C/I)
(http://www.ceps.lu/iriss/iriss.htm)

IRISS at CEPS/INSTEAD (IRISS-C/I) is a research grant programme of CEPS/INSTEAD, funded by the Training and Mobility of Researchers (TMR) programme ("Access to Large-scale Facilities") of the European Commission (DG XII). It offers access to CEPS/INSTEAD's national and international comparative micro-databases on individuals, households and firms, and promotes problem-oriented socio-economic analyses based on international microdata relevant to society and industry. It aims at bringing together individual researchers from different countries and disciplines in an informed socio-economic research environment.

International Social Survey Programme (ISSP) (http://www.issp.org/)

ISSP is a programme of cross-national annual collaboration on surveys covering topics important for social science research of which 31 countries are members. It brings together existing social science projects and co-ordinates research goals, thereby adding a cross-national, cross-cultural perspective to the individual national studies.

Luxembourg Employment Study (LES) (http://lissy.ceps.lu/LES/les.htm)

LES was set up in 1993 in association with LIS (see below). Its aim is to construct a data bank containing labour force surveys from the early 1990s from countries with quite different labour market structures. Such surveys provide detailed information on areas such as job search, employment characteristics, comparable occupations, investment in education, migration, etc. The LES team has harmonised and standardised the microdata from the labour force surveys in order to facilitate comparative research.

Luxembourg Income Study (LIS) (http://lissy.ceps.lu/access.htm)

LIS is a co-operative research project with members in 25 countries. Its goals are: *i)* to test the feasibility of creating a database consisting of social and economic household survey microdata from different countries; *ii)* to provide a method allowing researchers to access these data under the privacy restrictions set by the countries providing the data; *iii)* to create a system that will allow research requests to be quickly processed and returned to users at remote locations; and *iv)* to promote comparative research on the economic and social status of populations in different countries.

The Panel Compatibility Project (PACO) (http://www.ceps.lu/paco/pacopres.htm)

The PACO project is a centralised approach to creating an international comparative database integrating microdata from various national household panels over a large number of years. The PACO database contains harmonised and consistent variables and identical data structures for each country included. The PACO database increases the accessibility and use of panel data for research and facilitates comparative cross-national and longitudinal research on the processes and dynamics of areas such as labour force participation, income distribution, poverty, problems of the elderly, etc.

Towards European System of Social Reporting and Welfare Measurement (EUREPORTING) (http://www.zuma-mannheim.de/data/social-indicators/eurepwww.htm)

The EUREPORTING project's long-term objective is create a science-based European System of Social Reporting and Welfare Measurement. The project aims at advancing existing and developing new and innovative concepts, methodological approaches and empirical tools; better utilising existing databases and developing new ones where necessary; applying those concepts, tools and databases to urgent problems of social exclusion and socio-economic inequality; establishing a network of co-operating research teams, which could be quickly mobilised for short-term information and could establish the core for longer-term endeavours, *e.g.* a "European Social Report". EUREPORTING is funded for three years (1998-2001) by the European Commission and is divided into three sub-projects: Developing a European System of Social Indicators (EUSI), Stocktaking of Comparative Databases in Survey Research and Access to Comparative Official Microdata.

Other programmes

Development of a European Service for Information on Research and Education (DESIRE) (http://www.desire.org/)

DESIRE is a major international project aimed at building large-scale information networks for the research community, involving collaboration between project partners working at ten institutions from four European countries - the Netherlands, Norway, Sweden and the United Kingdom. The project's focus is on enhancing existing European information networks for research users across Europe through research and development in three main areas of activity: Caching, Resource Discovery and Directory Services. On-going DESIRE phase 2, which began in July 1998, is focused on distributed Web indexing, subject-based Web cataloguing, directory services, and caching. It is financed by the EU's Telematics Application Programme.

Integrated Library and Survey-data Extraction Service Project (ILSES) (http://www.gamma.rug.nl/ilses)

ILSES is a project of a number of Dutch, German, French, and Irish institutes, accepted by the European Commission under the Fourth Framework. It aims to develop a service that enables individuals users to access and retrieve documentary information and empirical data related to large-scale surveys. In addition, it offers tools and procedures for the normalisation, cataloguing and controlled distribution of (distributed) holdings of the documentary and data resources.

2. Data archives and statistical offices in OECD countries

Australia

Social Science Data Archive (SSDA), Australian National University (http://ssda.anu.edu.au/)

Located in the Research School of Social Sciences at the Australian National University, SSDA was set up in 1981 to collect and preserve computer-readable data relating to social, political and economic affairs and to make the data available for further analysis. IFDO member.

Australian Bureau of Statistics (http://www.abs.gov.au/)

Austria

Wiener Institut für Sozialwissenschaftliche Dokumentation und Methodik (WISDOM) (http://www2.soz.univie.ac.at/wisdom/)

WISDOM is the Austrian social science data archive. It archives all relevant data in various fields of social sciences. CESSDA/IFDO member.

Österreichisches Statistisches Zentralamt (http://www.oestat.gv.at/)

Belgium

Belgian Archives for the Social Sciences (BASS), Université catholique de Louvain (http://www.logi.ucl.ac.be/logi/)

BASS's mission is to acquire, to prepare and make available to users and to archive demographic, economic and political data and to answer queries from researchers at the university. IFDO member.

National Institute of Statistics (http://statbel.fgov.be/)

Canada

Carleton University Data Centre (http://www.carleton.ca/~ssdata/)

The Data Centre was originally part of the Social Sciences Data Archive, located in the Department of Sociology and Anthropology. Its function was to serve as a depository for data and encourage the use of the data by faculty and students. It is now a part of MaDGIC (Maps, Data and Government Documents Centre) of the university. IFDO member.

Leisure Studies Data Bank (LSDB), University of Waterloo (http://library.uwaterloo.ca/uweds/lsdb/lsdb.htm)

LSDB is a non-profit research facility which maintains machine-readable data on leisure-related topics. LSDB provides all researchers with ready access to diverse data sources and provides assistance in their use and analysis. LSDB was founded in 1972 by faculty members in the Department of Recreation and Leisure Studies so that students could gain access to data of national and provincial significance. IFDO/IASSIST member.

UBC Numeric Data Services, University of British Columbia (http://www.datalib.ubc.ca/)

UBC Numeric Data Services is part of the Humanities and Social Sciences/Government Publications Division of the UBC Library system. It maintains raw electronic data for research in economics, demography, public opinion, geography, and other social sciences. Only current UBC students, staff and faculty members may access these data for non-commercial use only. IFDO member.

Statistics Canada: (http://www.statcan.ca/)

Czech Republic

Czech Statistical Office (http://www.czso.cz/)

Denmark

Danish Data Archives (DDA) (http://ddd.sa.dk/dda/)

The DDA is a national data bank and documentation and technical service facility for researchers and students in Denmark and abroad. The DDA is an independent unit of the organisation of Danish State Archives. It was established in 1973 by the Danish SSRC (Social Science Research Council) and developed into a national social science resource facility. CESSDA/IFDO member.

Danmarks Statistik (http://www.dst.dk/)

Finland

Finnish Social Science Data Archive (FSD) (http://www.fsd.uta.fi/)

FSD is a national resource centre for social science research and teaching. It began operations at the beginning of 1999 as a separate unit in the University of Tampere. FSD provides a wide range of services from data archiving to information services. The main task for FSD is to increase the use of existing data in the social sciences by disseminating data throughout Finland and internationally. The unit is funded by the Ministry of Education. CESSDA member.

Statistics Finland (http://www.stat.fi/tk/home.html)

Banque de données socio-politiques (BDSP), Institut d'Études Politiques (http://www-cidsp.upmf-grenoble.fr/)

Located at the Institut d'Études Politiques, Grenoble, BDSP was created in 1981. It stores and distributes numerical machine-readable data. Its data collections include political science (election studies, political opinion polls, etc.), sociology (organisational and industrial sociology, socialisation, mass media, communication, welfare conditions, etc.), history and demography. CESSDA/IFDO member.

Laboratoire d'Analyse Secondaire et de Méthodes Appliquées à la Sociologie - Institut du Longitudinal (LASMAS-IdL) (http://www.iresco.fr/labos/lasmas/accueil_f.htm)

LASMAS-IdL, a CNRS (*Centre national de la recherche scientifique*) laboratory in France, is a resource centre of surveys for social science research. It archives and disseminate pertinent data. Based on agreements with various producers of data (INSEE, *Centre d'études et de recherche sur les qualifications*, Ministry of Education, Ministry of Culture), LASMAS-IdL acquires data from surveys and diffuses them to CNRS researchers.

Institut National de la Statistique et des Études Économiques (INSEE) (http://www.insee.fr/)

Germany

Eurodata Research Archive, Mannheim Centre for European Social Research (http://www.mzes.uni-mannheim.de/eurodata/eurodata.html)

The Eurodata Research Archive is a unit of the Mannheim Centre for European Social Research (MZES), founded at Mannheim University in 1989. Eurodata's basic function is to support comparative European research at MZES. The archive covers all of Europe and focuses on socio-economic and political data. Its core relates to aggregate data at national and regional level. The archive participates in research projects for developing comparative databases.

Zentralarchiv für Empirische Sozialforschung an der Universität zu Köln (ZA) (http://www.za.uni-koeln.de/)

The Central Archive (ZA) archives primary material (data, questionnaires, code plans) and results of empirical studies, prepares them for secondary analysis and makes them available to the interested public. The range of the ZA encompasses all technical areas in which procedures of empirical and historical social research are used. CESSDA/IFDO member.

Statistisches Bundesamt Deutschland (http://www.statistik-bund.de/)

Greece

National Statistical Service (http://www.statistics.gr/en/index.htm)

Hungary

Social Research Informatics Centre (TÁRKI) (http://www.tarki.hu/)

TÁRKI was founded in 1985 as a service organisation by Hungary's major social science research units to set up an information network for social research in Hungary and to develop methodological co-ordination. Its archive currently stores data of more than 300 sociological and socio-statistical surveys. In addition, TÁRKI has its own survey system with a trained network of interviewers, so that it can quickly carry out reliable surveys on samples of general population or on selected target populations. CESSDA/IFDO member.

Hungarian Central Statistical Office (KSH) (http://www.ksh.hu/)

Iceland

Statistics Iceland (http://www.statice.is/)

Ireland

Irish Central Statistics Office (CSO) (http://www.cso.ie/)

Italy

Archivio Dati e Programmi per le Scienze Sociali (http://www.nsd.uib.no/cessda/adpss.html) (*Note:* this is not the Web site for this institute, but a relevant document can be found there).

ADPSS is part of the Istituto Superiore di Sociologia, an institute sponsored by the four universities of Milan, the University of Pavia and local agencies. Although it performs other services in the field of social research computing, as well as teaching and consulting in methods of social research, ADPSS specialises in the collection of ecological data files and survey files of particular interest for urban and political studies. The main data holdings are large ecological data files. In addition, ADPSS has a Data Archive of Italian and International Surveys with a selected number of survey files (ready for SPSS analysis) mainly used for secondary analysis and for teaching purposes. CESSDA/IFDO member.

Istituto nazionale di statistica (http://petra.istat.it/)

Japan

Social Science Japan Data Archive (SSJDA), Institute of Social Science, University of Tokyo (http://www.iss.u-tokyo.ac.jp/pages/ssjda-e/)

SSJDA is located in the Institute of social Science's Information Centre for Social Science Research on Japan. This archive is a collection of data from social and statistical surveys, which are available for secondary use in academic research projects. The data archive itself is currently available only in Japanese. Summary information on the data archive is provided in English.

Statistics Bureau of the Management and Co-ordination Agency (http://www.stat.go.jp/)

Korea

Korean Social Science Data Centre (KSDC) (http://www.ksdc.re.kr/index.htm)

KSDC, established in November 1997, manages and collects social data sets for academic research in the social sciences. It maintains and makes available both national and international data sets related to social sciences, in such areas as census, economics, social issues, law, public health and elections. It also plans surveys and statistical analyses related to important issues and critical events in contemporary Korea.

National Statistical Office (http://www.nso.go.kr/)

Luxembourg

Service central de la statistique et des études économiques (Statec) (http://statec.gouvernement.lu/)

Mexico

Instituto Nacional de Estadística, Geografía e Informática (INEGI) (http://www.inegi.gob.mx/)

Netherlands

Netherlands Institute for Scientific Information Services (NIWI)
(http://www.niwi.knaw.nl/welcome.htm)

NIWI is an institute of the Royal Netherlands Academy of Arts and Sciences (KNAW). NIWI formally started on 1 September 1997, out of the merging of six institutes which provided scientific information. NIWI provides scientific information in various scientific fields. including social sciences and history. The Netherlands Historical Data Archive (NHDA) and the Steinmetz Archive (social sciences) (STAR), both located at NIWI, are members of IFDO and of IFDO-CESSDA, respectively.

Scientific Statistical Agency (WSA) (http://129.125.158.28/wsahomuk.html)

WSA was instituted by the Netherlands Organisation for Scientific Research (NOW) in 1994 to improve the accessibility and availability of data sets, with a focus on social-scientific research. Its aims include improving the availability and supply of data and in particular microdata sets, promoting and developing new uses of data sets, initiating and sustaining contact with user groups, taking stock of users' wishes concerning data sets and advising them on the feasibility of their wishes, consulting on and implementing microdata protection measures, providing information, price setting, if necessary setting up a user registration, and in general acting as an intermediary between data set providers and data set users.

Centraal Bureau voor de Statistiek (http://www.cbs.nl/)

New Zealand

New Zealand Social Research Data Archives (NZSRDA)
(http://www.massey.ac.nz/~NZSRDA/nzsrda/archive.htm)

NZSRDA, located in the Faculty of Social Sciences, Massey University, was established in 1992. Its purpose is to collect, clean, document and preserve computer-readable data relating to social, political and economic affairs and to make the data available for further research and analysis. It currently holds 33 social sciences survey data sets, including attitudes to religion, social issues surveys, elections, etc. IFDO member.

Statistics New Zealand (http://www.stats.govt.nz/)

Norway

Norwegian Social Sciences Data Services (NSD) (http://www.uib.no/nsd/)

Established in 1971, NSD is a national resource centre, under the Research Council of Norway, which serves the research community. It is funded mainly by the Research Council of Norway, Norwegian universities and different ministries. NSD provides a wide variety of resources across all social science fields, such as survey data, regional data, data about political parties, historical data, etc. It has set up a special privacy issue unit responsible for contacts between the research community and the Norwegian Data Inspectorate. CESSDA/IFDO member.

Statistics Norway (http://www.ssb.no/)

Portugal

Instituto Nacional de Estancia (INE) (http://infoline.ine.pt/si/index.html/)

Spain

Archivo de Estudios Sociales (ARCES), Centro de Investigaciones Sociológicas (CIS) (http://cis.sociol.es/primera.html)

ARCES is a service of the Centro de Investigaciones Sociológicas, an autonomous state agency dedicated to the study of Spanish society, primarily through survey-based research. ARCES serves as a data archive for social surveys and makes available information and access to social data in different archives throughout the world. CIS has contributed, through over 1 200 public opinion surveys, to a better understanding of the new social and political realities. All the surveys carried out by the CIS are kept in its Data Bank, where they are available to any natural or legal, public or private person requesting them. CESSDA member.

Instituto Nacional de Estadística (INE) (http://www.ine.es/)

Sweden

Swedish Social Science Data Service (SSD), Göteborg University (http://www.ssd.gu.se/home.html)

SSD serves all Swedish universities especially those with full educational programmes at graduate level. Its main task is to increase the availability of machine-readable data for secondary research in historical and social science disciplines. CESSDA/IFDO member.

Statistics Sweden (SCB) (http://www.scb.se/)

Switzerland

Swiss Information and Data Archive Service for the Social Sciences (SIDOS)

SIDOS is a service of the Swiss Academy of Social Sciences and Humanities (ASSH). It facilitates access to available data so at to promote secondary analysis and reanalysis of data for scientific purposes as well as to give support for training students in social science methodology. SIDOS has a research data archive which is part of the world archive network, as well as documentation on research in the social sciences. CESSDA/IFDO member.

Swiss Federal Statistical Office (http://www.admin.ch/bfs/)

Turkey

State Institute of Statistics (http://www.die.gov.tr/

United Kingdom

The Data Archive, University of Essex (http://dawww.essex.ac.uk/index.html)

Founded in 1967 at the University of Essex to preserve and distribute electronic data for research and teaching, the Data Archive is a specialist national resource containing the largest collection of accessible computer-readable data in the social sciences and humanities in the United Kingdom. It is funded by the Economic and Social Research Council (ESRC), the Joint Information Systems Committee (JISC) of the Higher Education Funding Councils and the University of Essex. CESSDA/IFDO member.

Manchester Information Data sets and Associated Services (MIDAS) (http://midas.ac.uk/)

MIDAS is a national research support service located at Manchester Computing at the University of Manchester, specialising in on-line provision of strategic research and teaching data sets, software packages, training and large-scale computing resources for the UK academic community. MIDAS is a free service for higher education throughout the United Kingdom and is funded by JISC, ESRC and the University of Manchester. Its data sets include the 1991 Census of Population Statistics, government and other large continuous surveys, macroeconomic time series data banks, etc.

QUALIDATA, University of Essex (http://www.essex.ac.uk/qualidata/)

QUALIDATA was set up by ESRC to rescue the most significant research material of previous years. QUALIDATA is not an archive: it is an action unit. Its aims are to: locate, assess and document qualitative data and arrange for their deposit in suitable public archives; to disseminate information about such data; and to encourage their reuse. The Qualidata catalogue (Qualicat) is available via the Internet, where researchers can search and obtain descriptions of qualitative research material, its location and accessibility.

Government Statistical Service (GSS) (http://www.statistics.gov.uk)

Office for National Statistics (ONS) (http://www.ons.gov.uk/ons_f.htm/)

United States

Data and Program Library Services (DPLS), University of Wisconsin-Madison
(http://dpls.dacc.wisc.edu/)

DPLS is the central repository of data collections used by the social science research community. Its mission is to promote academic research by facilitating the use of secondary research materials. The collection includes major surveys from other distributors, US government data, and locally produced archival data sets. Longitudinal surveys, macroeconomic indicators, election studies, population studies, socialisation patterns, poverty measures, labour force participation, public opinion polls, education and health data, and local/national/international governmental statistics are among the subjects represented. Data sets are free to download via the Online Data Archive, but registration is required. IFDO member.

Institute for Research in Social Science (IRSS) Data Archive, University of North Carolina at Chapel Hill

IRSS maintains the country's third-largest archive of computer-readable social science data. Holdings include national and international economic, electoral, demographic, financial, health, public opinion and other types of data to meet a variety of research and teaching needs. It maintains a computer-searchable catalogue of the more than 2 800 studies and series in its data holdings. Since 1965, IRSS has operated the Louis Harris Data Center, the national depository for survey data collected by Louis Harris and Associates, Inc. More than 1 000 Harris Polls from as early as 1958 are archived at the Center, with over 60 000 questions asked of more than 900 000 respondents. IFDO member.

Institute for Social Science Research (ISSR) Data Archive, University of California, Los Angeles (http://www.sscnet.ucla.edu/issr/da/)

The ISSR Data Archive was established in December 1977 as a service unit supporting quantitative research, primarily secondary analysis, in the Institute for Social Science Research and the greater social science community at UCLA. Data are primarily acquired through UCLA's membership in the Inter-university Consortium for Political and Social Research (ICPSR). Other sources for data include government agencies such as the National Center for Health Statistics, Bureau of the Census, National Center for Education Statistics, and the Bureau of Labor Statistics. Surveys conducted by the Institute for Social Science Research, such as the Los Angeles County Social Survey, are also acquired by the Archive. IFDO member.

National Opinion Research Center (NORC), University of Chicago (http://www.norc.uchicago.edu/homepage.htm)

NORC is a non-profit corporation affiliated with the University of Chicago which conducts survey research in the public interest for government agencies, educational institutions, private foundations, non-profit organisations, and private corporations. It collects data to help policy makers, researchers, educators and others address the crucial issues facing the government, organisations, and the public. Those seeking information about NORC data sets may wish first to consult the ICPSR Web site. IFDO member.

The Roper Center for Public Opinion Research (http://www.ropercenter.uconn.edu/)

Founded in 1947 and located at the University of Connecticut, the Roper Center is the largest library of public opinion data in the world. The Center's mission focuses on data preservation and access, education and research. The Center brings together findings of surveys conducted by thousands of research organisations in the United States and abroad, typically in machine-readable form, for further secondary analysis, and facilitates access to these data by interested researchers everywhere. The Center is committed to the further development of a rich comparative database containing comprehensive research collections on the United States and other countries of the Americas (including Canada, Mexico, Venezuela, Brazil, and Argentina); the leading industrial democracies (including Japan, Britain, France, and Germany); and eastern Europe and the nations of the former Soviet Union. IFDO member.

Social Sciences Data Collection (SSDC), University of California, San Diego (http://ssdc.ucsd.edu/index.html)

SSDC is a collection of numeric data in the social sciences maintained by the Social Sciences and Humanities Library (SSH) of University of California, San Diego. The collection is available online to UCSD faculty and students. Also available through SSDC are online data analysis, full-text and map resources provided by the SSH Library.

Bureau of Labor Statistics (http://stats.bls.gov/blshome.htm/)

Census Bureau (http://www.census.gov/)

Federal Interagency Council on Statistical Policy (FEDSTATS) (http://www.fedstats.gov/)

National Center for Education Statistics (NCES) (http://nces.ed.gov/)

National Center for Health Statistics (http://www.cdc.gov/nchswww/default.htm)

3. Clearinghouses for the social sciences

Argus Clearinghouse (http://www.clearinghouse.net/index.html)

The Argus Clearinghouse has 13 categories for search, including Social Sciences and Social Issues. This category includes guides on topics pertaining to the study of people and their activities, customs, origins and institutions, as well as the larger social context and social concerns under consideration or debate. The sources included are evaluated by experts, using as criteria objective and subjective information content, design, schemes used, meta-information.

Scholarly Internet Resource Collections (INFOMINE), Library of the University of California, United States (http://infomine.ucr.edu/Main.html)

INFOMINE's objective is the introduction and use of Internet/Web resources of relevance to faculty, students, and research staff at university level. It is being offered as a comprehensive showcase, virtual library and reference tool containing useful Internet/Web resources, including databases, electronic journals, electronic books, bulletin boards, online library card catalogues, articles and directories of researchers, among many other types of information. It has been built by contribution from over 30 University of California and other university and college librarians.

Social Science Information Gateway (SOSIG), United Kingdom (http://www.sosig.ac.uk/welcome.html)

The SOSIG Internet Catalogue is an online catalogue of high-quality Internet resources. It offers users the chance to read descriptions of resources available over the Internet and to access those resources directly. The Catalogue points to thousands of resources; each one has been selected and described by a librarian or academic, making this the Internet equivalent of an academic research library for the social sciences. The SOSIG service receives funding from ESRC, JISC and the European Union.

Social Sciences Virtual Library (SSVL), College of Liberal Arts and Sciences at the University of Florida, United States (http://www.clas.ufl.edu/users/gthursby/socsci/)

The Virtual Library Web catalogue is run by a loose confederation of volunteers. Sites are inspected and evaluated for their adequacy as information sources before they are linked from SSVL. Its catalogue offers access to data archive institutes in the world selected by SSVL.

Sociosite, Faculty of Social Sciences at the University of Amsterdam, Netherlands (http://145.18.241.97/sociosite/index.html)

SocioSite is a project based at the Faculty of Social Sciences at the University of Amsterdam. It is designed to give access to information and resources which are useful to sociologists and other social scientists and intends to provide a comprehensive listing of world social science resources on the Internet. It includes a comprehensive (but not exhaustive) list of the world's data archives.

NOTES

1. See “SBER Archiving Policy”, http://www.nsf.gov/sbe/sber/common/archive.htm.

2. Council Regulation (EC) No 322/97 of 17 February 1997 on Community Statistics (see below).

3. WAIS (Wide Area Information Servers) is a user interface program used in IDC.

4. MOST is a research programme, designed by UNESCO, to promote international comparative social science research, with the emphasis on supporting large-scale, long-term autonomous research and transferring the relevant findings and data to decision makers.

REFERENCES

Bernard, Paul *et al.* (1998), "Final Report of the Joint Working Group of the Social Sciences and Humanities Research Council and Statistics Canada on the Advancement of Research Using Social Statistics", http://www.sshrc.ca/english/policydocs/discussion/statscanreport.pdf.

Brackenbury, Simon (1998), "Hates Statistics, Hates Computers. Promoting Networked Statistics at the University of Plymouth and the London School of Economics. Priorities for Establishing a Data Support Service. Results from a User Requirement Survey at the London School of Economics, December 1997", paper presented to the IRISS Conference, Bristol, March.

Dunn, Christopher S. and Erik W. Austin (1998), "Protecting Confidentiality in Archival Data Resources", *ICPSR Bulletin*, Fall 1998.

European Science Foundation (1998), "Blueprint for a European Social Survey (ESS)", Strasbourg.

Fienberg, Stephen E. (1994), "Sharing Statistical Data in the Biomedical and Health Sciences: Ethical, Institutional, Legal, and Professional Dimensions", *Annual Review of Public Health*, Vol. 15, Annual Reviews, Inc., Palo Alto, CA.

Flora, Peter (1997), "A System of Socio-economic Reporting on Europe – Memorandum for the Fifth Framework Programme of the EU", *EURODATA Newsletter* No. 5, MZES, Mannheim.

Gadd, David (1998), "Networked Statistics Research Support Survey", *Networked Statistics Bulletin*, Issue 27, May, University of Plymouth.

Government Statistical Service, "The Work and Structure of the Government Statistical Service", http://www.statistics.gov.uk/aboutgss/uksub.htm.

Guy, Laura (1997), "A World of Knowledge: Government Information in Action – The Data and Program Library Service: 30 Years of Providing Access to Electronic Information", http://dpls.dacc.wisc.edu/wla/.

Hassan, Steve, Gunnar Sylthe and Repke de Vries (1996), "The CESSDA Integrated Data Catalogue", IASSIST '96 Conference Paper, Minneapolis, May.

Hiom, Debra (1998), "The Social Science Information Gateway: Putting Theory into Practice", paper presented to the IRISS Conference, Bristol, March.

Inter-university Consortium for Political and Social Research (ICPSR) (1999), "ICPSR Guide to Social Science Data Preparation and Archiving", http://www.icpsr.umich.edu/ICPSR/Archive/Deposit/dpm.html.

Kasse, M. *et al*. (1997), "Survey of Strengths, Gaps and Weakness in European Science: Economics and Social Sciences – A Report for the European Science and Technology Assembly (ESTA) and for the European Heads of Research Councils (EUROHORCs)", Bonn.

Kraus, Franz (1998), "Towards a Data Infrastructure for Socio-economic Research on Europe: Improving Access to Official Microdata at the National and the European Level", *EURODATA Newsletter* No. 7, MZES, Mannheim.

Luxembourg Income Study (1998), "Luxembourg Income Study", http://lissy.ceps.lu/LES/les.htm.

Mochmann, Ekkehard (1998), "European Co-operation in Social Science Data Dissemination", http://www.ifdo.org/archiving_distribution/eudd_bfr.htm.

Mochmann, Ekkehard and Paul de Guchteneire (1998), "The Social Science Data Archive Step by Step", http://www.ifdo.org/archiving_distribution/archive_sbs_bfr.htm.

Musgrave, Simon (1998), "Resource Discovery and Use: Improved Web Tools to Find and Browse Data", paper presented to the IRISS Conference, Bristol, March.

National Science Foundation (1999), "Enhancing Infrastructure for the Social and Behavioral Sciences - Program Announcement NSF 99-32", http://www.nsf.gov/pubs/1999/nsf9932/nsf9932.htm.

OECD (1999), "Final Report of the OECD Megascience Forum Working Group on Biological Informatics", http://www.oecd.org/dsti/sti/s_t/ms/prod/BIREPFIN.pdf.

Ohly, H. Peter (1998), "Structuring of Social Science Information Resources in the Internet: Knowledge-Organization Aspects in Building a Clearinghouse", paper presented to the IRISS Conference, Bristol, March.

Sylvester, Yvonne (1996), "The ONS and the Data Archive", *The Data Archive Bulletin*, No. 63, September.

Chapter 4

PLANNING LARGE-SCALE INFRASTRUCTURE: LONGITUDINAL DATABASES

by

Gaston Schaber
President, CEPS/INSTEAD

Introduction: setting the frame

The original title of this paper was intended to be "The ***Challenge*** of Longitudinal Studies". The word "challenge" eventually disappeared from the title, probably because its meaning was unclear. But this fact made me think about how to cope with the undetermined meaning of the title. Why a challenge? Where does the challenge lie? Who is challenged or should feel challenged, and by what? Is there only one challenge? For only one group of people? Does it go in multiple directions? This paper may provide some of the answers to these questions.

First, the immediate challenge in the social sciences is to develop large-scale research infrastructures, in particular through the creation of longitudinal databases for comparative research purposes (Schaber, 1993).

The infrastructure concept

Stating that there is no formal or comprehensive definition of research infrastructure for the behavioural and social sciences, the Commission on Behavorial and Social Sciences and Education of the US National Research Council (1998) defines two main categories of infrastructure:

- *Category 1:* Multidisciplinary centres which provide an opportunity to bring together a critical mass of experts interested in common problems; a novel variation of the traditional centre is the "virtual centre" made possible by the Internet.

- *Category 2:* This category covers the much broader category of research instrumentation and equipment, and contains at least five sub-categories:

  – Platforms and observational systems (*e.g.* neural imaging equipment, observational coding systems).

- Computational systems (*e.g.* supercomputers, mass storage devices, visualisation systems).
- Laboratory and analysis systems (*e.g.* electron microscopes, statistical software, image processing).
- Communication and network systems (*e.g.* vBNS, Internet).
- Information systems and databases (*e.g.* digital libraries, large surveys).

The last sub-category is considered by some to constitute the intellectual infrastructure, with a major component being the methodological developments (and computer programs) essential to any sophisticated analysis of the data.

At the end of the definition, the Commission of the US Research Council states: *i)* that international databases face additional problems of data comparability arising from differences in national data collection methods and definitions; and *ii)* that no international body, or union of national bodies, exists to support such an international comparative infrastructure.

The key challenge is to develop the required intellectual infrastructure, in an internationally comparative perspective, with limited funding.

The increasing need for representative micro-data

Quoting from the background document prepared for the workshop by Jun Oba of the OECD Secretariat (OECD, 1999, see Chapter 3 of this volume):

> [There is] "an increasing demand by social scientists not only for new data but also for exploitation of existing data for secondary analysis. In many OECD countries, decision makers promote the use of empirical data by the social sciences to monitor social trends relevant to public policy. Improved availability of data on society could favourably affect research quality in the social sciences.
>
> Advances in social science methodologies require large amounts of data. Researchers increasingly require access to detailed microdata to conduct research in many areas. The powerful statistical techniques needed to analyse multilevel, longitudinal data cannot be used with aggregate data; access to microdata is essential (Bernard *et al.*, 1998)"

The position and status of data

Data are central to any infrastructure venture but they do not control the intellectual work involved. The scientific aim of the project is not simply to bring together and archive data or statistics from national data sets and make them available to researchers around the world (although this would already be an important step forward). As stated in the co-ordination report (Schaber, 1998) of the Targeted Socio-economic Programme Committee to the Committee for Research in Science and Technology (CREST) at the EU Commission, infrastructure building in a genuine research perspective has to give high priority to fundamental activities such as conceptual and methodological work, with a view to developing indicators and producing sound theories, models and data sets of high quality and with a sufficient degree of comparability to serve as the basic tools for systematic analysis of socio-

economic development and critical changes, and continuous monitoring of economic and social policies.

Defining infrastructure-building in this *comparative* way, means identifying the corresponding challenges, some of which are related to data, others to structural and cultural specificities.

Data-related issues

Comparative research on the social and economic aspects of specific countries and societies is hampered by the unavailability of appropriate microdata. Microdata are usually produced only at the national level and in a national perspective and, even where they are available to researchers at the national level, such data are often difficult to handle without inside expertise. Where such data are available for different countries, they are difficult to pool together for analysis of larger entities such as Europe.

Structure- and culture-related issues

In the absence of adequate and systematically organised documentation, the diversity of cultures, languages, social institutions, regulations, distribution systems, government programmes, etc., is such that it is extremely difficult to move from domestic to foreign, from national to European, and from European to international socio-economic data for any specific area or topic.

Before asking the question of whom, *i.e.* which institutions or organisations, would be best able to rise to these challenges, we should attempt to translate the challenges into tasks; this will perhaps make it easier to sketch out the profiles of the people or entities that do (or should) carry out these tasks.

Defining the tasks

The following tasks will need to be undertaken in order to move the social and economic sciences forwards. Implementation of these task will give the social sciences a comparative dimension and a sufficient degree of coherence to form a research infrastructure for a larger scientific community, whether at the European Union level, the OECD level, the international level or, in a somewhat more distant future, the global level.

At the basic level, irrespective of the type of study undertaken, these tasks will include:

- Bringing together relevant existing microdata sets from reliable sources, whether produced by national statistical offices, administrations or research institutions (*this can often make a huge difference in terms of access and co-operation*).

- Documenting the data sets at the technical and institutional levels, so that users are aware of the methodologies used and the precise meaning of the information in the broader and more complex context of national institutions, regulations, etc.

- Achieving comparability through harmonised and consistent variables and files (with identical variable names, labels, values and data structures).

- Developing appropriate technical devices and arrangements to ensure that the microdata sets are available to and easily accessible by researchers, taking into account that, in accordance with specific national regulations on data protection and privacy protection, some of the data may only be accessed under controlled conditions.

- Creating training programmes and generous opportunities for the younger generations of economic and social researchers to use these databases transnationally and in a comparative perspective, both when working alone and in groups.

At a more advanced level of social and economic analysis, there is a need for:

- *Longitudinal microeconomic and micro-social surveys* on persons, households, firms and spatial units, which take into account the time dimension by focusing on the same units of observation over a number of years. This type of approach is indispensable for observing change and analysing the dynamics of change. It imposes stringent demands on quality control in regard to data (data production and processing, database construction, data documentation and data comparability). (This point covers the most labour-and cost-intensive aspects of the whole undertaking.)

- *Time-and-space-referenced integrated information systems* enabling demographic, social, economic and ecological data of mutual relevance for comparative research and policy analysis to be combined in a meaningful way. The informational status of these heterogeneous sets of data and measures needs to be documented, and their respective values for combined analysis of highly complex inter-relations assessed. (This is still rare, as can be seen below relating to US panels and geo-referencing, *e.g.* on the basis of census data.)

- *Building models and theories* able to deal comparatively with both short- and long-term societal changes. This needs to be done on a higher level than enlightened journalism, and thus requires use of the microdata sets referred to above (see Box 1 for a recent example of a modelling exercise). These databases will need to be large-scale, detailed, longitudinal and well documented both technically and institutionally, with relevant complementary information at the macro and meso levels, and, above all, they need to be comparable. The list of requirements is long, and the successful implementation of the databases will require a great deal of work.

Box 1. **A modelling exercise using the CEPS/INSTEAD comparative databases**

At an international conference held in Luxembourg and organised by Tim Smeeding, Director of the LIS project at INSTEAD, on “Children's Well-being in Wealthy and in Transition Countries”, Pierre Hausman, senior researcher at CEPS/INSTEAD and his collaborators contributed a simulation exercise in which they applied the Luxembourg system of family allowances to the US states of California and Pennsylvania, using the CEPS/INSTEAD comparative databases. According to the criteria used in this study, the mean percentage of children living in poor families in France or in Luxembourg is about 7%, while the mean percentage for the United States is 23%. The simulation exercise shows that, if the Luxembourg system of family allowances were applied *a)* in California, the percentage of children living in poverty would be only 8.8% (which means a poverty reduction of 75%), and *b)* in Pennsylvania, the corresponding value would be 8.2% (a 93% reduction in poverty). With such a Luxembourg-like system, the United States would hardly have more children living in poverty than the two European countries. The United States would, of course, have to spend five times more on their children than they do today – an expense that would have to be compared to the lifetime price they pay for having 23% of their children growing up in poverty.

A new research policy and new facilities

The implementation of these requirements and tasks, and the cross-national development of the corresponding research infrastructures (indispensable for assessing social and economic development and performance, social and economic policies, social and economic cohesion, both within and between countries), go far beyond the confines of the academic tradition of social sciences. In the same way, they fall outside the framework of individual academic careers and go beyond the confines of national data-producing offices and administrations.

Such a project will require a research policy that fosters existing and emerging data enterprises, large-scale research facilities and long-range research networks, conceived and equipped to work in close co-operation with university institutions, on the one hand, and official data producers in statistical offices and public administrations, on the other. Furthermore, it should aim to: *i)* help create the necessary databases and methodologies for comparative research; *ii)* provide training facilities for new generations of socio-economic researchers; and *iii)* offer the means of communication necessary for new modes of cross-national and intercontinental co-operation.

In this perspective, we hope to achieve the following objectives:

- The use of academic computer communication networks and electronic storage of scientific use files will make it common practice for widely dispersed researchers, either alone or in groups, to work on shared databases, much as is already the case for scientists in the exact sciences.

- Organising socio-economic data in such a way will allow research to be carried out more economically by reducing unnecessary repetition or duplication of basic tasks.

- In particular, this new organisation will lead to the production of more accurate research results and to the accumulation of research material in a consistent corpus of comparative knowledge to be critically reviewed and amended over time.

- To meet the needs of the scientific community for a broader range of analyses to be performed on substantial data sets accessible by all researchers.

It is precisely the diversity of the analyses (including modelling and simulation) performed on well-specified and well-documented data sets by a large variety of researchers who differ in their interests and orientations, in their options for applied *vs.* basic social science, for empirical inquiry or for modelling, that offers the best opportunities for producing the kind of cumulative knowledge that, in a spiral process, will lead to better (re)conceptualisation, better observation and better measurement, and thus to more productive and critical interaction between data, models and theories, etc.

These are the arguments for fostering a new generation of scientific enterprises which will place new demands on the social sciences at all levels: scientific, organisational, financial and political.

Actual and potential contribution of longitudinal databases to large-scale infrastructure

Existing longitudinal databases

Canada

I will limit myself here to a reference to the Final Report of the Joint Working Group of the Social Sciences and Humanities Research Council and Statistics Canada on the Advancement of Research using Social Statistics (December 1998). This remarkable report, prepared by Paul Bernard and collaborators, was distributed as background information for the workshop (Bernard *et al.*, 1998). As Paul Bernard stated at the preparatory meeting in Paris, the Canadian set of newly developed surveys and panel studies is not only the most recent, it has serious chances of being the best: "as the Canadians have learned from the mistakes and misfortunes of the pioneering predecessors from elsewhere". Personally, I tend to agree.

Europe

In the case of the European Union, the picture is not impressive. Europe as a genuine entity is only in the making, as are the specifically European surveys and longitudinal studies. Of course, most of the EU member states do take part in a number of surveys which then are put together internationally. However, specific EU initiatives are rare; in fact, I could mention here only the European Community Household Panel, which is of quite recent origin.

A relatively exhaustive and reliable overview of studies and databases for Europe which could be related to the US and Canadian achievements is not yet available. Intermediate reports are in the process, but these will basically reflect the fragmented, nationally defined efforts undertaken to date. To my knowledge, the main reference is the European Commission supported project on "EuReporting (1998) : Towards a European System of Social Reporting and Welfare Measurement". These important, but heterogeneous, elements will have to be elaborated in very innovative ways if the intention is to reflect the reality of a nascent European Union.

United States

The United States have the oldest and richest collection of nation-wide longitudinal databases. The following list was presented in a report prepared by the Subcommitteee on Federal Longitudinal Surveys (1986):

- Survey of Income and Programme Participation
- Consumer Price Index
- Employment Cost Index
- National Longitudinal Study of the High School Class of 1972
- High School and Beyond
- National Longitudinal Survey of Labour Market Experience
- Social Security Administration's Retirement History Study
- Social Security Administration's Disability Programme Work Incentive Experiments

- National Medical Care Expenditures Survey
- National Medical Care Utilisation and Expenditures Survey
- Longitudinal Establishment Data File
- Statistics of Income Data Programme

This impressive set of longitudinal studies maintained by the United States over the years represents a research infrastructure in itself. However, in the perspective of this international workshop, these remarkable endeavours are purely "domestic" – although the term domestic in this case refers to a country the size of a continent. Domestic or not, it will take some time before other countries, or even the European Union, will be able to parallel this performance.

In the meantime, in relation to the concerns of this meeting, *i.e.* large longitudinal data sets and infrastructure-building, US research has made considerable progress in recent years, not only in making data available to the research community, but also in developing large co-operative networks enabling better use of these data resources. These resources can be used as stand-alone databases or in combination, thus enriching the databases in creative ways through geo-referencing and other linkages (using the resources of other data sets, irrespective of whether these are produced administratively or in universities).

A model in this respect is the work accomplished over the last few years by teams such as that run by Greg J. Duncan and partners, working on integrating federal statistics in the field of child development research (Brooks-Gunn *et al.*, 1995). The aim of this work is to improve research on child and adolescent development by improving specific national data collection projects in regard to the richness and quality of the data obtained, the representativeness of the samples, the expansion of information on outcome domains, resources and explanatory variables – improvements which will serve both the policy and the academic research communities.

Their theoretical discussions and empirical analyses have led these researchers to make a number of highly interesting suggestions, some of which would involve only minor and fairly inexpensive changes but would bring large analytical benefits; others are more expensive but would open up invaluable opportunities for analysis.

The work focuses on twelve national data collections:

1. Long-term Longitudinal Surveys

(1.1.) Panel Study of Income Dynamics (PSID)
(1.2.) National Longitudinal Survey of Youth (NLSY)
(1.3.) National Longitudinal Survey of Child-Mother Data
(1.4.) National Educational Longitudinal Survey of 1988 (NEL88)
(1.5.) National Survey of Children (NSC)

2. Short-term Longitudinal Surveys

(2.1.) Survey of Income and Program Participation (SIPP)
(2.2.) National Survey of Families and Households (NSFH)
(2.3.) High School and Beyond (HS&B)
(2.4.) Consumer Expenditure Survey (CEX)
(2.5.) National Crime Victimization Survey (NCVS)

3. Cross-sectional Surveys

(3.1.) Decennial Census, Public Use Micro-Sample (PUMS)
(3.2.) National Health Interview Survey-Child Health Supplement 1988

The work described above is remarkable and deserves international attention – but it does not permit comparative studies. I would now like to present a project for a genuinely comparative database integrating mature and younger longitudinal household surveys for cross-national research. This project has been set up by CEPS/INSTEAD, which does not only hold a microdata archive with well-documented, anonymised, original microdata files from many countries, but is also a data producer in its own right (Luxembourg Household Panel and Firm Panel) and a comparability producer (LIS, LES, PACO) (Box 2).

Box 2. **The CEPS/INSTEAD comparative databases LIS, LES, PACO**

The "Luxembourg Income Study" (LIS), since 1983

The LIS is an international comparative cross-sectional database on income distribution which presently contains 80 micro-data sets from 26 countries, covering the period 1968 to 1997: Australia, Austria, Belgium, Canada, the Czech Republic, Denmark, Finland, France, Germany, Hungary, Ireland, Israel, Italy, Luxembourg, the Netherlands, Norway, Poland, Portugal, Russia, the Slovak Republic, Spain, Sweden, Switzerland, Chinese Taipei, the United Kingdom, the United States. The database is accessed globally via electronic mail networks by over 300 users in 28 countries (the jobs being sent are closely checked for reasons of data protection. In addition to harmonised data, LIS users are offered extensive documentation at both the technical and institutional levels.

The "Luxembourg Employment Study" (LES), since 1993

LES is an international comparative database integrating microdata from a set of Labour Force Surveys from the 1990s. The database presently includes data sets from 15 countries, with about 90 variables, and new data sets are continuously being added. The countries currently include: Austria, Canada, the Czech Republic, Finland, France, Hungary, Luxembourg, Norway, Poland, the Slovak Republic, Slovenia, Spain, Sweden, Switzerland, the United Kingdom, the United States. The following countries will be entered in the near future: Australia, Belgium, Bulgaria, Denmark, Israel, Italy, Lithuania, the Netherlands, Portugal, Romania.

The "Panel Comparability Project" (PACO), data archive and database, since 1990

PACO is an international data archive and database integrating national longitudinal household panels from west and east. The PACO data archive presently includes 70 original panel data sets from 17 countries, with original (not standardised) variables transformed into a common format (SPSS system files for Windows on PC). The PACO database contains scientific use files and presently includes 40 data sets with fully comparable standardised variables from household panels of the following countries: France (Lorraine), Germany, Hungary, Luxembourg, Poland, the United Kingdom, the United States. The PACO network is still developing at CEPS/INSTEAD and will include panels from additional countries (Belgium, Russia, Sweden, etc.). Information in the PACO files is available: for households and individuals on the micro-level; for single years; in longitudinal form – and is completed by extensive technical and institutional documentation. The scientific use files are available on CD-Rom, according to specified confidentiality rules and pledges.

Web site: http://www.ceps.lu/paco/pacopres.htm

The project follows on from the PACO experience in open systems able to integrate both long-standing and nationally produced household panels, the European Community Household Panel, and other data sets.

The idea of developing a comparative database of household panels (the Panel Comparability Project – PACO) emerged two years after the creation of the Luxembourg Income Study (1982), with one important difference: there were already a number of one-time surveys on income distribution waiting to be put together into LIS, whereas in the case of longitudinal household studies, there was only the US Panel Study on Income Dynamics (PSID), the model from which the German, the Dutch, and the Luxembourg/Lorraine twin-set household studies were derived (since 1984). Useful, exploratory comparative work was undertaken during the mid-1980s, initiated by Duncan and Schaber, which led to a seminar at CEPS/INSTEAD on 15-19 June 1987. However, the project elaborated did not receive funding because the potential funding agencies considered (rightly) that the European partners had not yet reached the critical mass necessary for matching the US contribution in a balanced comparative transatlantic venture.

In 1990, when the United Kingdom joined the CEPS informal panel network, a first comparative panel project obtained funding from the European Science Foundation (from 1990 to 1993). From 1993 to 1996, the comparative project – now known as PACO – was funded, or more precisely co-funded, by the Human Capital and Mobility Programme of the European Commission, with basic financial support from the Luxembourg Government.

Since July 1996, PACO has been funded by the Luxembourg Government, with additional funds being sought – and sometimes found – under respective EC programmes for adding more countries and more yearly waves and for integrating additional variables into the database.

The work of the PACO division at CEPS/INSTEAD aims to consolidate and enlarge the internal documentation system on panels, ranging from original questionnaires of existing panel studies to documentation on these panels and to literature on panel issues such as methodology, data processing, data protection, etc.

CEPS has recently initiated the creation of a consortium of household panel producers and managers in order to implement an infrastructure project that should enable European and intercontinental socio-economic research to be carried out on the living conditions of persons and households. This project aims to:

- Create an international comparative micro database containing longitudinal data sets from national household panels (in Europe, the United States and Canada) and from the European Community Household Panel (ECHP).

- Link key information from existing macro/institutional data sets to complement the comparative database and provide support through utilities for panel analyses.

- Improve analysis and understanding of social change and its implications for individual people and households and for social institutions and policy making.

The database will comprise comparable variables transformed to fit a common format and will use standardised international classifications, where these are available. Information will be provided: for households and individuals on the micro level; for single years; and as longitudinal information, all of which elements will be linked to macro- and institutional data. The base will contain harmonised and consistent variables and identical data structures for each country: *i.e.* 14 EU countries, Poland,

Hungary, the United States and Canada. The data will be stored as system files for the statistical packages SPSS, SAS and STATA, and will contain identical variable names, labels, values and data structures. Each country file will be adequately anonymised and rated as a scientific use file. On the basis of dissemination rules agreed upon between the consortium and the data owners, the database will be made available on CD-ROM and distributed to the scientific community under appropriate conditions of confidentiality and data protection.

The tasks defined by the consortium include:

- Developing and (re)defining rules on standardisation.
- Building up and/or enhancing/reconverting comparable panel databases.
- Creating documentation and a user guide for the databases thus created.
- Collecting and preparing key information from macro-, meso- and institutional data.
- Improving information on and access to original country panel data.
- Enhancing the ECHP, etc., for scientific use.
- Enhancing the data processing techniques for using panel data.
- Setting up an Internet information system on household panel studies.
- Running exemplary panel analyses, *e.g.* on labour market dynamics and other basic issues.

The comparative database will allow the consortium to conduct research into the following issues:

- Housing
- Demography
- Marriage history (current and retrospective)
- Fertility biography (current and retrospective)
- Education, vocational training
- Health
- Biographical labour force participation information (retrospective)
- Labour Force participation (current)
- Unemployment
- Income (gross and net), taxes, social security
- Time use
- Subjective variables
- Organisational variables

The following countries currently participate in the project: Austria; Belgium; Canada; Denmark; Finland; France; Germany; Greece; Hungary; Ireland; Italy; Luxembourg; the Netherlands; Poland; Portugal; Spain; Switzerland; the United Kingdom, the United States.

Issues relating to substance

In thinking about infrastructure-building and becoming acquainted with the new, high-speed, multimedia communication and information technologies which will enable us to enhance our still underdeveloped socio-economic research infrastructures, we should give some thought to substantive issues – unless we want to run the risk of having state-of-the-art technologies … and not much to say.

The report by Paul Bernard *et al*. highlights a number of priority areas for research in social statistics:

- Child development.
- Youth in transition.
- Families in flux.
- Growing old.
- Education, skills and literacy.
- Distribution of wages and work.
- Social and community supports.
- Social impacts of science and technology on families/children and on well-being.
- Evolving workplace and technology use.
- Welfare, income and poverty.

I wholeheartedly support these priorities, to which I would only suggest adding, in the same vein:

- Intergenerational relations in a changing demographic context.

Nevertheless, I would like to add a short list of issues and domains which we should not shy away from. In spite of the fact that the present framework does not allow for an explicit presentation, I will at least mention them. The socio-economic sciences will have to face these issues sooner or later; they should take them into account sooner – at the level of their projects for infrastructure-building. These issues and domains include:

Structural change in a comparative perspective

This issue will require examining structural change at all levels (local, regional, transregional, international), and linking in an effective way economic, social, demographic, ecological and, of course, political components.

In our advanced/industrialised countries, we tend to consider other countries as being "in transition" (in comparison with our own countries). Taking this attitude could prevent us from seeing that we too may be in a stage of transition, perhaps at a different level but one which is equally critical.

Inequality, poverty and socio-economic performance

This theme was recently (re)defined by work at the World Bank. It should be as much a concern for the OECD countries as the "transition" theme outlined above. It should not be seen simply as a

"third-world issue": even in our wealthiest countries and their wealthiest cities, there are areas where the conditions of living and the levels of development of large sectors of the population are similar to, or even below, those observed in third-world countries.

From the outset, we need to develop the concepts, methods and strategies for collecting comparable data which will allow us, within a broader intellectual framework, to deal simultaneously and in an inter-related way with trends towards wealth and poverty in rich and poor countries. The best way to prepare for this task might be to start with the study of cities.

This brings me to the last item which I would like to see on the agenda for innovative socio-economic research.

Stability or instability, and sustainability of complex socio-economic urban systems

Contemporary society is becoming massively concentrated in cities and urban areas, where it generates the conditions and the push for equally massive wealth, progress and change. However, the urban world created by contemporary society leads not only to progress, social development, economic growth and corresponding opportunities; it also leads to growing inequality, poverty, deterioration of living conditions, social and economic exclusion, and thus to growing disparities, tensions and risks of unrest, violence and disruption.

It is only recently that cities have become a focal concern for a number of national governments and for major international institutions and organisations. Encouraging initiatives have been taken by the latter, such as:

- The United Nations programme on "Human Settlements" (UN, 1997).
- The European Commission "Urban" programme (EU, June 1994).
- The US National Science Foundation's "Urban Research Initiative" (URI) (NSF, 1998).

The UN and the EU programmes are still under development; they have been designed from the outset to be comparative, although they will not necessarily be longitudinal (which, of course, would be the ideal to aim for). The US initiative is not in itself internationally comparative, but it does insist on the importance of longitudinal data and neighbourhood studies. It is unfortunate that this recent programme will have only a very short life.

Cities have been studied by many scientific disciplines for decades, and a great deal of information is already available, although the available data are often not very enlightening since they are highly fragmented and unrelated, reflecting the barriers and the diverging or conflicting interests which separate the actors and analysts involved, whether engineers and other professionals, scholars, business people, politicians, or simply people.

Much remains to be done from the point of view of the development of research infrastructures requiring large, comparative and longitudinal databases.

I should add here that, in order to handle the last three issues listed, we will need to develop the complex integrated information systems described above. Such systems will have to pull together in meaningful ways, across space and over time, demographic, social, economic and ecological data of

mutual relevance for comparative research and policy analysis, and will have to make combined use of three different micro-units of observation:

- *People (individual persons, households/families)*, to enable us to learn more about the social tissue.

- *Firms and businesses* are rarely dealt with at the micro level, but can teach us more about the economic tissue.

- *Neighbourhood*, the third micro-unit, deals with spaces or areas (geographically identifiable, even if the distinctions are rather blurred), where living conditions and situations are given and events take place, and where, for better or for worse, people grow up, develop and interact.

If our descriptions and analyses are to reflect reality, then these three levels of observation will have to be taken simultaneously into account.

The necessary longitudinal data sets have not yet been developed, but they will contain social, economic and other information at a micro level which requires anonymisation and protection. For reasons of data and confidentiality protection, "safe centres" will have to be created where data owners, data processors and data analysts can work together under legally defined conditions to produce adequately anonymised scientific information and scientific use files for research purposes and policy analysis.

Funding of infrastructures and comparability production

Infrastructure-building in the socio-economic sciences resembles to only a limited extent infrastructure-building in the exact sciences. In the exact sciences, crossing borders (whether physical, cultural or political) is achieved at low or near zero cost. In the socio-economic fields, however, such an endeavour runs into data- and culture-related obstacles which are costly to overcome. Building research infrastructures which cross country borders is not a national concern and even less a national priority. And this is even more true for the production of comparability.

I would like to make reference here to the Mannheim Memorandum, authored by Prof. Peter Flora and signed by colleagues from EU countries. [1] The Mannheim Memorandum pleads for the creation of a system of socio-economic reporting on Europe, and for the development of appropriate infrastructures (Flora, 1997).

The Mannheim Memorandum addresses only the European Commission and the European Union member states, *i.e.* an international configuration that is smaller than the OECD membership. Nevertheless, the considerations and proposals of the 19 signatories might be of some relevance here. Once the European reporting system has been outlined, and this evidently requires appropriate data and transnational research infrastructures, the painful question of funding will arise. The signatories acknowledge that basic efforts should be made at the level of the member states themselves, but they also consider that the over-arching Union and Commission should fulfil their complementary responsibilities in supporting such an endeavour. And they conclude by stating:

> "This proposal is not to suggest that the Commission as such should create and run a European system of socio-economic reporting or that the Commission should finance third parties to do so for the European Union, but to suggest, according to

> the principle of subsidiarity, that the Commission contribute an adequate share at the level of the member countries and their respective scientific communities to the development of a truly European research infrastructure, able to produce and manage relevant, representative, and comparable information and databases in a bottom-up approach, involving the countries as well as the Community, starting not from zero, but from scientific research institutions, establishments, facilities which on their own have already achieved larger or smaller parts of such an infrastructure-building endeavour." (Flora, 1997)

I would also like to make extensive reference to a significant paper by Timothy Smeeding distributed as background material to participants at this workshop (Smeeding, 1999).

In his introduction, Smeeding states very clearly that successful infrastructure projects require at least four elements: top-flight researchers to address the research question; survey vehicles to implement data collection; liberal data access so that a wide cadre of researchers can use these data; and most of all, decision-making and fund-granting institutions to sponsor cross-national collaborative activities.

In his paragraph on "Funding Mechanisms: The Achilles Heel", he goes on to say:

> "Decision criteria for different types of infrastructure are important. But what is most sorely needed is an international funding mechanism that would allow social scientists to collaborate on research-based data harmonisation efforts, while at the same time providing the impetus for national central statistical offices and other national data collectors to implement various infrastructure efforts. Questions of how to decide among different proposals and their gestation are premature. The idea of a plausible standing mechanism needs to wait for an initial experimental mechanism!" (Smeeding, 1999)

He notes that Europe is only slightly ahead of the Americans on this front; in its "Training and Mobility of Researchers" programme, the European Union has provided funds for large-scale facility (LSF) projects such as the "Integrated Research Infrastructure in the Socio-economic Sciences (IRISS) project run by CEPS/INSTEAD (the home of LIS and PACO). The EU Commission has granted funds to IRISS to bring together groups of researchers for periods of from two weeks to three months to work together using cross-national databases applied to key social problems. While this initiative is a good start for researchers from the EU countries and the EU-associated countries, it excludes non EU-associated researchers from funding. This calls for complementary solutions.

To provide some basic information for readers who are unfamiliar with the EU programmes, the current support provided by the Commission to Large-scale Facilities does not serve for the development of research infrastructures as such, *e.g.* the infrastructure that CEPS/INSTEAD is committed to developing, in part through its IRISS project. The Commission supports access for researchers to existing large-scale facilities, but does not support the development of infrastructures as such, since the Commission considers this to fall under the responsibility of the EU member states. Although we are grateful that the EU Commission has over the last few years included the social and economic sciences in its programmes, and that it is supporting access to the very young, large-scale facilities in these fields, we continue to think that the Commission should, according to the principle of subsidiarity, contribute its share to the development of infrastructures in the socio-economic sciences – as it did, a decade ago, for the hard sciences, in particular for physics. It should be borne in mind that this workshop deals not only with access to large-scale facilities and existing infrastructures, but also with the actual *development* of the transnational infrastructures.

In regard to funding mechanisms, one way to proceed is for several large funders of research and data collection, such as the National Science Foundation, or the National Institutes of Health in many nations, to unite with their international counterparts (*e.g.* the European Community) to discuss co-operative support of data infrastructure and research tied to important cross-national scientific issues, *e.g.* population, ageing, migration, educational attainment, and economic status. These funders should seek agreement for a co-operatively funded effort to build a flexible international data infrastructure to meet the specific needs of researchers.

An alternative approach is for a large international organisation to make the investments necessary to create such an infrastructure by subsuming existing initiatives (*e.g.* LIS, PACO, PSID, SOEP, ECHP), and providing the funds to staff the project and make the data available. Given today's technological advances, instead of simply being marginal cost users of under-funded public goods, these organisations could then become public good funders for their member states – at a modest cost. All that is required is for a far-sighted administrator in one of these groups to make the case for such an organisational investment.

Smeeding envisages – summarily and without great expectations – a third way of obtaining funding: through globally oriented businesses which would agree upon a longitudinal database, *e.g.* a consumer expenditure survey, which would provide market-based information on people. This suggestion is interesting as it stands, but would be worth discussing from a scientific point of view *only* if one of the outcomes would be the production of some sort of public good, independently assessed for quality and accessible to the scientific community.

Smeeding comes to the following conclusion:

> "The international data infrastructure in the social and behavioural sciences is not ready for major questions about its shortcomings, because there is no international body, or union of national bodies, which supports such an infrastructure to begin with! Small co-operative ventures such as the EC's IRISS programme and its United States' partner, the National Science Foundation, are a step towards research co-operation. A similar longer-term project to add to the database infrastructure is also needed." (Smeeding, 1999)

I should specify that the co-operation venture mentioned by Smeeding between the EC and the US National Science Foundation is not between the EC and NSF; it is in fact between CEPS/INSTEAD and the US National Science Foundation in relation to our own IRISS project, which is co-funded by the European Commission. We would be delighted if the EU recognised the potential of this very new collaborative venture and decided to take advantage of it in order to become active at a higher institutional level.

Turning back to Smeeding for his final considerations:

> "Once a standing funding mechanism (*e.g.* a long-term financial partnership or a single larger funder such as the OECD) is arranged, one could begin to ask: Which data? When? How? And for how long? Existing under-funded projects should be strong candidates for such funding. Clearly, this financial body or partnership should consider projects which closely link research and data collection. Only those projects which produce comparable data, and only those efforts which make such data available to researchers more generally should be considered. But, until a benevolent philanthropist or a group of national foundations come along with multiple millions for international infrastructure, the question of which data,

> domains, disciplines, etc., is on hold. Meantime, the prototypical international data infrastructure projects which have begun over the past 15 years remain underdeveloped, under-utilised, and financially threatened." (Smeeding, 1999)

I have nothing to add to Smeeding's developments, except to say that I completely agree with them.

Conclusions – or rather claims

At the beginning, the title of my paper on infrastructure-building and longitudinal databases contained the word "challenge". That got lost in the planning process. But I rather liked it, because its meaning was sufficiently unclear to be used in more than one sense. In fact, the challenge is multidirectional and applies to everyone of us:

- Data and database producers, whether in central statistical offices, administrations, independent advanced research centres or universities.
- Producers of longitudinality (a rare species) and producers of comparability (even rarer).
- Users of all kinds, whether fundamental or applied researchers, policy analysts or policy planners.
- Funders (the rarest species of all).

We are all challenged, in our different ways, by the following requirements:

- Data caring.
- Data sharing.
- Data comparing.

And, these three actions have to be completed by:

- Training efforts which focus on the younger generations of researchers.
- Fundraising efforts and pressures at the level of national governments and scientific institutions as well as at the level of the international and global bodies.

NOTES

1. The 19 signatories are: J. Berghman, Tilburg; J.Bradshaw, York; J. Commaille, Paris; R. Erikson, Stockholm; P. Flora, Mannheim; R. Hauser, Frankfurt; B. Henrichsen, Bergen; M. Kaase, Berlin; F. Legendre, Paris; Y. Lemel, Paris; D. Lievesley, Colchester; B. Marin, Vienna; A. Martinelli, Milan; G. Martinotti, Milan; H.H. Noll, Mannheim; G. Schaber, Luxembourg; N. Westergaard-Nielsen, Aarhus; M. Pérez Yruela, Cordoba; and W. Zapf, Berlin.

REFERENCES

Bernard, Paul *et al.*, (1998), "Final Report of the Joint Working Group of the Social Sciences and Humanities Research Council and Statistics Canada on the Advancement of Research Using Social Statistics", http://www.sshrc.ca/english/policydocs/discussion/statscanreport.pdf.

Brooks-Gunn, Jeanne, Brett Brown, Greg J. Duncan and Kristin Anderson Moore, (1995), "Child Development in Context of Family and Community Resources: An Agenda for National Data Collection", in *Integrating Federal Statistics on Children: Report of a Workshop*, National Academy Press, http://books.nap.edu/books/0309052491/html/27.htm.

Committee on National Statistics (CNSTAT) (1995), *Integrating Federal Statistics on Children: Report of a Workshop*, National Academy Press, http://www.nap.edu/books/0309052491/html/.

EuReporting (1998), Project financed by the European Commission in the framework of the TSER Programme for a three-year period starting March 1998. *i)* Sub-project on "European System of Social Indicators", *ii)* Sub-project on "Stocktaking of Comparative Databases in Survey Research"; *iii)* Sub-project on "Access to Comparative Official Microdata". http://www.zuma-mannheim.de/data/social –indicators/eureporting/

European Union (1994), "The URBAN Audit", http://www.inforegio.cec.eu.int/urban/audit.

Flora, Peter (1997), Mannheim Memorandum, *EURODATA Newsletter*, No. 5, pp. 2-7, http://www.mzes.uni-mannheim.de/eurodata/newsletter/no5/nl5_Peter_Flora.html.

National Science Foundation (1998), "The Urban Research Initiative (URI)", NSF homepage http://www.nsf.gov/cgi-bin/getpub?sbe981.

OECD (1999), "Social Sciences Databases in OECD Countries: An Overview", background paper prepared for the Ottawa Workshop on Infrastructure Needs for Social Sciences (Chapter 3 of this volume), OECD, Paris.

Schaber, Gaston (1993), "Developing Comparative Databases", position paper prepared for the Second CIESIN Users' Workshop in Atlanta Georgia, organised by the Consortium for International Earth Science Information Network, 11-13 October, http://www.ceps.lu/ceps_i/cidownld.htm.

Schaber, Gaston (1998), "Co-ordination Report to CREST on Research Infrastructures in the Socio-economic Sciences", European Commission, Directorate General XII, Science, Research and Development, Targeted Socio-economic Programme Committee, 1 October, http://www.ceps.lu/ceps_i/cidownld.htm.

Smeeding, Timothy M. (1999), "Problems of International Availability of Microdata: The Experience of the Luxembourg Income Study (LIS) Infrastructure Project", prepared for the DIW/GESIS Meeting on Co-operation between the Academic Community and Official Statistical Bodies: Practice and Prospects, Wiesbaden Germany, 1 June (paper distributed at the Ottawa Workshop).

Subcommittee on Federal Longitudinal Surveys and the Federal Committee on Statistical Methodology (1986), "Statistical Policy", Working Paper 13, OMB, May, http://www.bts.gov/swart/cat/sw13.htm.

United Nations Centre for Human Settlements, Global Urban Observatory (1997), *Monitoring Human Settlements with Urban Indicators, Draft Guide 1997*, http://urbanobservatory.org.

US National Research Council, Commission on Behavioral and Social Sciences and Education (1998), *Investing in Research Infrastructure in the Behavioral and Social Sciences*, National Academy Press, Washington, DC, http://www.nap.edu/catalog/6176.htm.

Chapter 5

SHARING AND PRESERVING DATA FOR THE SOCIAL SCIENCES

by

Denise Lievesley
Director, Institute for Statistics, UNESCO

Introduction: the importance of data

There is increasing recognition throughout the world of the importance of the exploitation, in social research and analysis, of the rich data resources of official agencies in particular, but also from academic and commercial sources. This chapter concentrates on the value of providing access to data in electronic form (alongside conventional published material) and, where legal constraints permit it and operational procedures can be devised, giving access to individual-level data as well as the aggregate data. Not all of these data will result from surveys and censuses - access to data which are by-products of administrative processes can be just as valuable for research, although this is only now being appreciated.

We need to *create a culture in our institutions and in society more generally in which data sharing is the norm.* Even within the academic social science sector in the developed world, there are pockets where the culture of data sharing is not accepted – perhaps because primary researchers are concerned that they get no formal credit for data sharing and that it brings risks, particularly of other academics obtaining faster or better publications from their secondary research. The institutional reward system for research needs to be examined to identify and remove such barriers.

Both data producers and data brokers need to forge better links with professional associations and in particular with journal editors to promote a culture of data openness in which criticism of the original data collection or analysis should be factual and temperate, and to ensure a responsible use of data.

Increasingly, the funders of social and economic research and professional organisations are requiring grant recipients and those who publish data to place data and accompanying documentation in the public domain, inviting scrutiny as well as secondary use of data (partly due to the pressure on research budgets which means that more effective use has to be made of existing data). In the future, this may also facilitate overview studies, the social and economic fields having been slow to advance in overview studies such as those taking place in clinical fields. It has been argued that Cochrane-type evidence-based systems are required in relation to research on society (Smith, 1996). Thus we may see

the development of mandatory reporting standards comparable to those set out in the CONSORT statement for those involved in clinical trials (Begg, 1996).

But the main reason for developing data policies must be to ensure that: *i) deliberate replication* is encouraged but *ignorant duplication* does not happen; and *ii)* to exploit investments in data. An example of a data policy is that of the UK Economic and Social Research Council which asks of all applicants whether new data collection could be avoided and requires them to check if secondary data are available. If they argue successfully that new data collection is required, they can receive finance to enable them to document and archive the data. Researchers must discuss data access early and offer the data to the UK Data Archive.

It is crucial to communicate and explain the rights of data providers and ensure they get proper recognition – but also to argue within our political and institutional settings for low cost or free access for the use of data in scientific research and teaching because it is in the public interest for data to be used in this way.

Data access in the developing world

Despite the growing recognition of the value of data, there are many areas of the world in which data are simply not available for research purposes, often because of a poor infrastructure and limited expertise in data handling but also often because of the weak links between government officials and academics. This must be of concern since the effect is a widening gap between the developed and the poorer countries of the world, exacerbating disparities in society. The innovative use of knowledge is key to social and economic development, as identified by Drucker in 1969. The development of endogenous capabilities is an effective deterrent to the brain drain. The UNESCO report on the World Summit for Social Development (1996) states that: "Alongside the action to enhance national and regional capabilities for higher education and scientific and technological training, it is also essential to promote both basic and applied scientific research and the dissemination of its results."

Support must be given to enhancing access to information and communication technologies and to ensure that wider use is made, by researchers in the countries concerned, of the data which have been generated with the very precious resources of the country. It is extremely worrying that the lack of preservation facilities and expertise means that the small trickle of precious data is often lost, and the lack of accepted systems for the involvement of academics means that the existing data are not exploited. The lesson that electronic data are not a finite exhaustible resource and that their value is increased, not diminished, by their use has not yet been learnt in much of the developing world. Efforts by existing data archives to transfer their "know-how" and technology to poorer nations are to be applauded, but better co-ordination and financing is required. The need for access to relevant data banks and archives, alongside "centres of excellence where researchers and scientists could be trained" were identified in UNESCO's Audience Africa programme in 1995.

Benefits of data access to the scientific research community

The benefits of providing access to electronic data for the scientific community are numerous:

- The advantages of materials in digital form are immense and include unfettered access, flexibility and enhanced capabilities since access to numerical data in electronic form permits a level and depth of analysis which cannot be undertaken with published material (Marsh *et al*., 1991). Even with text, digitisation enhances manipulation and searching as

well as the ability to copy and amend information very easily, and it facilitates types of analysis which would be very resource-intensive, and perhaps even impossible, otherwise.

- High-quality, policy-relevant research is facilitated by giving researchers access to data which are often collected at great expense to the public purse and in terms of public participation. These data may be of a type which it would not be possible for the scientific community to collect directly, such as population census material or products of administrative systems.

- A fundamental principle of scientific scholarship is that research results should be accessible to others to enable them to refute, clarify or extend the findings. Developments in science (including social science) should be incremental, building on what has been discovered or explained before.

- Frequently, data are used for purposes unanticipated at the time of their collection. Making data available for secondary analysis can encourage a variety of different perspectives to be brought to bear and the resulting analyses can be richly diverse.

- Electronic data are an important resource for teaching purposes - this is expanded in the section below on the use of data in teaching.

- While the electronic nature of material raises resource issues, it also brings great opportunities in the ways that it can be delivered to the user communities, permitting the integration of information from different sources (for example, enabling the creation of themed data rather than data separated by source), the incorporation of quality indicators and other metadata into data sets, the creation of spatially referenced data sets, the merging of data from different points in time, and the extraction of different levels of data geared to different uses.

The benefits to data providers

So, let's turn to the benefits to data providers of making their data available. Many of the greatest benefits can be classified as *altruism*, which might include:

- Contributing to the development of knowledge.

- Ensuring that data are exploited.

- Encouraging multiple perspectives on the same data.

- Facilitating comparative research.

- Enabling new researchers to be trained in data analysis and interpretation. Singh and Crothers (1999) have identified the increasingly explicit demands for skills from employers who require the immediate usability of graduates they employ.

The importance of such altruism depends critically on the environment (the *data culture*) within a country, including the existence of legislation on open government, the status and influence of the scientific community, and the appropriate data protection and confidentiality legislation. This perhaps goes some way to explaining the large discrepancies encountered in data access across different countries.

The culture with respect to access to official data has changed in many countries in recent years, with a view gaining prominence that the provision of good data for a wide community of users including Parliament, industry and commerce, academia, the media and the general public must be met alongside the needs of government. Analyses should be in the public domain and basic data made available for further analyses. This change is driven by the increasing recognition that not to use data or to use inadequate data has costs for society. Perhaps the growing communication among social scientists across the world has also resulted in pressures being put upon some governments to improve data access for research purposes.

A further important reason for encouraging secondary analysis is to reduce respondent burden. Compliance costs are a concern, particularly in surveys of businesses or elites and in small countries. So, secondary analysis of existing data rather than fresh data collection is to be encouraged.

The benefits of forging links between users and producers of data

An OECD report as long ago as 1979 on *The Social Sciences in Policy Making* identified the value of contacts between government and non-government social science specialists – yet this relationship remains weak in many countries. This lack of mutual understanding between academic social scientists and public servants is deplored in the UNESCO report on the World Summit for Social Development (1996).

Reducing the divide between data producers and academic researchers in the social sciences will be of benefit to both groups since the quality of data collection can be improved by feedback on the use of the data and since data analysis, to be informed, must take account of the methodology and conceptual framework which underlie the collection. In the case of official data, collaboration between these groups needs to be carefully managed so that the integrity of official statistics and the independence of academic research are not undermined.

In the United Kingdom, we suffered for many years from having an academic community which was dismissive of official data producers who were viewed as merely carrying out rather routine activities whilst, in turn, academic research was perceived by the public servants to be self indulgent and unrelated to real life. Amongst some parts of the social sciences the use of data was seen as ‘mere empiricism’ and was considered inferior science. Fortunately such divisions have broken down as official statisticians have become more user focussed and as the incentives for academic researchers to carry out policy relevant research have increased. The promotion of the use of data by the scientific community can greatly assist in the process of getting financial and institutional support for the collection of high quality data, especially as increasingly the social scientists are carrying out publicly accessible data analysis.

An important benefit to data generators of making their data available for research purposes is to forge links with scientific users in order to:

- Take advantage of users’ expertise and to consult this “expert group”.
- Create a community of knowledgeable data users.
- Establish a community of supporters who will fight with data providers to ensure that important data collections are protected.

- Obtain feedback on use, especially relating to policy relevant research.
- Carry out collaborative research with users.

Concerns of data providers

Although I have concentrated so far on the benefits to data producers of making their data available, they often have concerns which may include:

- Defensiveness, concern about deliberate or accidental misuse of data.
- Anxiety that users may use data in a way that is politically embarrassing.
- Concern about the resources needed to document data, and that users will find errors.
- Need to ensure equity of access for all users.
- Need to use the data for revenue generation.
- The fact that the use of the data by others may undermine their own competitive advantage or intellectual capital.

Responsibilities of data users

Some of the legitimate concerns of data providers can be alleviated by ensuring that users understand that they have responsibilities and obligations in return for using data. An important responsibility of users of secondary data is to give credit to the data provider. It is rare for publications to be quoted by others without giving full credit to the source of the material, yet it still happens that electronic data sources are used without proper acknowledgement.

It can be essential for legal, ethical and other reasons that users respect the conditions of access agreed with data providers. This may entail the implementation of controls over use and perhaps the collection of charges for data.

Confidentiality

It is especially critical that users and data brokers are sensitive to confidentiality issues. This means being fully informed about:

- Relevant statistical legislation, both general and specific.
- The interpretation of that legislation.
- Data protection legislation.
- International and regional as well as national legislation.
- Pledges on confidentiality, access or anonymity made to respondents.

- The sensitivity of the particular information.
- The data culture.

All of these can influence the availability of data for use by scientific researchers. While recognising that these are important issues – data providers have an important responsibility to respect the rights of respondents and great damage can be inflicted on future data collections if breaches of anonymity occur or are perceived to have occurred – there is sometimes a lack of political will to find solutions to minimise the problems and to permit data access. On the other hand, great efforts have sometimes been made to overcome the obstacles to data sharing. These include, for example, the revision of statistical legislation in the Republic of Ireland.

Protection of confidentiality

In many situations the protection of the anonymity of the respondents will be critical. Extensive literature is being built up on ways to do this while providing access to the electronic material (see especially the proceedings of the Eurostat conference series on this topic).[1] However, it is vital to balance the needs of today's researchers with those of the future who may require access to the individual identifiers. Thus, provided that the full unanonymised versions can be held in a secure environment with no risk of disclosure, it is preferable to retain them for historians of the future. (However, we need to consider the dangers inherent in retaining some data in identifiable form, as recently outlined in Seltzer, 1998). The version provided to current researchers may be protected by employing:

- Various methods to reduce risk of disclosure such as grouping, rounding, suppression, adding noise.
- Legal contracts to alert researchers to the fact that they must not use data to identify individuals.
- "Safe access" whereby users do not have direct contact with the raw data.
- Privileged access in which only specified authorised users may analyse the data.
- Delayed access until it is no longer sensitive.

Requirements of users

The complexity of different data sources, particularly in countries with a wide range of organisations collecting data such as those with decentralised statistical systems, can confuse new users of data. The user must be able to locate data easily, preferably through systems of "one-stop-shopping". Two levels of documentation are required: *i)* a rich description of the data resource's content, structure and provenance, including where possible contextual information to enable researchers in the future to understand the resource; and *ii)* an abbreviated subset, conforming to appropriate cataloguing standards, which is included in an online catalogue through which users can locate and then gain access to the material in question. Comprehensive, informative and accessible metadata are vital. These should cover concepts, methods and procedures, and should convey as much information as possible on quality to enable the user to judge fitness for purpose, particularly as those not involved in collecting the data have more limited ways in which to estimate quality (for example, the measurement of *process quality* is not usually feasible for them).

A key aspect to the information provided to users is information on charges. Where possible, charges should be seen to be reasonable and equitable but, most importantly, they must be predictable so that, for example, academic researchers are able to apply for relevant funds for their research. Charges for data can be a serious disincentive, particularly to young researchers and for exploratory or experimental work unless there is access to good research funds and the funding agencies recognise the need for research resources.

Timeliness is also important. It is a myth that academic research can get by with old data; much scientific research requires up-to-date material. Institutional frameworks need to be established in order to involve users in key decisions especially in the selection of material for preservation and distribution and in the design of access systems.

Institutional framework

Data producers can allay many of these concerns by working in strategic alliances with responsible data intermediaries who will understand the needs of academic users while respecting the obligations users have towards the data providers. Data dissemination centres can act as intermediaries between scientific users and data producers. They can be desirable facilities because many queries and problems which users raise are, in fact, unrelated to the data and consequently, if not handled elsewhere, take valuable time of the data providers. In other words, supporting users is time consuming. The close liaison between such centres and the data providers is essential, however, since the user support staff must develop a good understanding of the data.

The dedicated specialist work of data dissemination centres can provide value-added services through their understanding of the diverse needs of users, by integrating different data sets, adding contextual information, reformatting for data delivery, extracting subsets of data and documentation, and by developing services to enable users to visualise, browse and select data.

The use of data in teaching

The use of “real”, as opposed to synthetic, data sets in teaching adds interest and relevance to courses and, if the data are updated on a regular basis, ensures that the courses are pertinent to current issues. Access to the rich resources of data centres or archives should make it unnecessary to use data created especially for teaching purposes and ought also to reduce reliance on a small number of rather old data sets which have been used extensively on many courses. Students who gain their experience of data analysis from the use of specially constructed data rarely have a good understanding of the complexity of data analysis in the “real” world. An appreciation of this complexity through the use of unadulterated data can give appropriate training for applied careers. Students also have the opportunity to understand the rationale for collecting data and can develop critical faculties to judge the strengths and weaknesses of particular data and of statistical indicators. For a thought-provoking article on the teaching of social science, see Singh and Crothers (1999) in which they identify the need to “relate data and theory to the specification and assessment of policy options”.

Data can be chosen to be of particular relevance to the subject being taught and thus can bring both substantive and statistics courses alive. Data archives throughout the world contain rich sources of demographic, behavioural and attitudinal data with which to address many substantive topics. Such data may be used in conjunction with the main publications arising from the analysis conducted by the study’s originators. Students could be asked to replicate research already conducted, to extend this research or to examine the data from an entirely different perspective.

Several data sets are good sources for comparative analysis, most notably the Eurobarometer Survey conducted across all EU countries simultaneously using the same questionnaires, and the International Social Survey Programme in which an identical core of questions are included in surveys in 22 countries. A large number of data sources permit the exploration of change over time, including change in the demographic or other aspects of the structure of society, change in behavioural patterns or in attitudes. Longitudinal data can result from a number of different designs: in the United Kingdom, users can have access to fresh cross-sectional designs such as the British Social Attitudes Survey, panels such as the British Household Panel Survey, rotating panels such as the Labour Force Survey and cohorts such as the National Child Development Survey. Such material can be used for an exploration of the various methods of collecting data over time and their implications for the analysis and interpretation of both point-in-time estimates and estimates of change. There are also rich veins of data to be tapped to assist in teaching about various methodological issues connected with the collection of data. The potential is substantial and a short paper available from the UK Data Archive provides a few ideas for teaching methodology using data.[2]

Access is usually provided to both data and documentation and, increasingly, the documentation is being recognised as a valuable resource in its own right. It can be used to train students in data collection methods and to provide model surveys which students might copy or adapt. A number of key journals now require the authors of any article which uses data to make their data available for others to re-analyse and try to replicate their findings. These developments will change the ethos whereby data are seen to be the property of an individual researcher and will shift the culture to one of data sharing. This is bound to have benefits for using data in teaching since students will not merely read an article passively but will be able to explore the data themselves.

It is interesting to note an increase in the use of data in teaching short courses which, because of changes to the computing environment, are more likely to include “hands-on” work. Several summer school courses now focus on the analysis of particular data sets or on secondary data analysis more generally. There has also been a marked increase in the inclusion of “real” data in software demonstration and teaching material.

Developments have been particularly noticeable in teaching materials for schools. An exemplar project is run in Norway and involves schoolchildren in carrying out an election or referendum a few days in advance of the national election or referendum. Children learn about data collection and analysis, as well as about national politics, and the results of the surveys are widely covered in the press and on television. A children’s census for the Millennium will be carried out in 2000-01 by the Royal Statistical Society Centre for Statistical Education, together with the UK Office for National Statistics.

Data preservation

Many developed countries have some sort of strategy in place to address the risk of losing valuable data sets, but in some countries this issue has only been tackled in an *ad hoc* fashion, which has already lead to the loss of some important and irreplaceable data sets. It can be valuable to closely integrate data preservation strategies with data access systems thus ensuring that not only are data retained but also that they are available to be exploited widely.

The significant increase in the creation of electronic material poses challenges for the preservation of our heritage. Electronic material, unlike paper which can be kept for quite long periods of time without the need for intervention, needs proactive preservation husbandry if it is to continue to be useable over time. A preservation strategy will require resources and expertise. This is an especially

important issue in situations in which there is no paper version of the material at all; this is increasingly the situation as data are being collected using computer-aided methods.

It is vital that electronic information is preserved in a way which permits them (data and documentation) to be accessible over time despite changes to hardware and software so that they remain useable in the future by scientific researchers - including the data producers themselves. The preservation of electronic material requires expertise, equipment, operational systems, and trust and credibility.

A preservation system needs to ensure:

- The physical reliability of digital data.
- The security of the data from unauthorised access.
- The usability of the data.
- The integration of the data, where appropriate, into information and dissemination systems.
- The maintenance of effective data and documentation.
- The protection of the authenticity of data content.

Archiving does need to be distinguished from backing up. The two are rather confused in the minds of many people. Thus, IT staff in particular would seem to believe that they have an archival strategy in place when they really mean a back up and retrieval system for live data. The latter is a necessary but not sufficient condition of archiving.

In terms of the *physical reliability* of data, the longevity of the physical media on which the data are stored is critical and this depends upon the storage conditions and the way the media are used (for example re-writing media can lead to faster corruption of the data). Even under good storage conditions, media are more fragile than manufacturers claim.

Good archival practice requires that:

- A series of checks are built into the data at the time they are first preserved so that regular sweeps through the data recalculating such checks enables corruption to be identified.
- Rigorous checks take place immediately to ensure that what has been written is what was intended.
- If corruption is identified, a follow up should take place - often poor-quality media are received in batches so it may be necessary to check other material.
- Media should be regularly refreshed (*i.e.* the data rewritten to new copies of the same type of media) - this will depend upon the manufacturer’s claims, plus experience and information from experts.
- Media should be kept under review and migration to new media considered if they prove more robust and cost effective.
- Multiple copies of the electronic versions should be kept so that retrieval can take place if the media are found to be corrupt.

- These copies should be kept at different locations.
- It is important to ensure that the finding aids and audit trails, internal management systems, etc., are also archived.
- Techniques which may make it more difficult to retrieve old material should be avoided.

In terms of the *security* of the data, most agencies will have policies which ensure that only authorised users within the organisation can have direct access to the full unanonymised data. It will be necessary to examine how this security system will be extended to cover archived data. At the minimum, it will require access to be controlled and the data-writing capacity to be restricted to specified carefully controlled accounts. One of the security aspects which is specific to archiving is which version of the data to preserve. As already discussed, a decision is required as to whether anonymised versions only are to be preserved, thus reducing the risk to security of the information, or whether full versions, including copies of links to identifiers, are to be retained.

A major problem in archiving data is *technological obsolescence*. There are some who believe that this is an even more important problem than the fragility of the media. As a matter of principle, software used for archival purposes should be under the organisation's control so that the situation cannot arise that the software owner goes out of business or makes significant changes to their product without consultation.

There are three main strategies with respect to the format for data preservation:

- Maintain the data in software-specific form but migrate all of the data to the next version of software as the software changes.
- Maintain all the data in the original software-specific form and only migrate data (if necessary through several changes of software which will have to be retained in an archive themselves) when they are required.
- Transfer the data into a standard form which it is not necessary to migrate unless major incompatibilities with hardware occur and then can be carried out with one sweep through all the data.

The relative costs and risks of these three options will depend upon many factors, including the range of software being used, the length of time data must be retained, the proportion of data being accessed and the frequency of access. The need for more research into methods of digital preservation are described in Hendley (1998). Data are increasingly being created in software-specific formats which cannot be disentangled without risking the usability of the data. On the other hand, many users want to access the data in a software-specific format and thus there is a tension between the archival service and providing immediate and easy access to the data.

Other archival issues

Huge advances in the production of relatively inexpensive hierarchical storage management systems mean that data management may be integrated with data archiving. They facilitate logical structuring of data collections in order to exploit automated management tools. It is essential that the preservation of documentation be taken seriously. In order to improve the storage of documentation, its delivery to users and to take advantage of new technological ways to integrate data and documentation, the optimum strategy will be to hold documentation in digital form. Within

organisations it is important to clarify who will take responsibility for bringing the relevant documentation together and to try to ensure that this happens as early as possible in the process of data collection since it is more difficult to reconstruct documentation at the end of a project's life. Documentation will usually be created in digital form - usually as a word-processed document - but for some of the legacy material digital versions may not be available or locatable and in this case documentation may have to be scanned or re-keyed. Digital documentation may be stored in image format. There are great advantages in consolidating on a single format for data documentation. Most archives seem to be converging upon portable document format (.pdf) files which can be read by Adobe Acrobat and some other software. This format is worth investigating since it maintains the documentation structure and appearance. However, some organisations are choosing to use optical character recognition software to convert documentation to text-based versions. This would give the potential to use SGML (standard generalised mark-up language) for documentation which would facilitate the building of easy access systems and to integrate data and documentation to a greater extent. If this is too resource-intensive, perhaps only priority documents or documents for priority data sets could be dealt with in this way.

Sometimes the decision as to which versions or parts of data sets are to be preserved is a complex one and it can be particularly problematic in relation to dynamic data (whether to select a snapshot and also preserve change files), data which are revised (one approach is to preserve those versions for which paper publications were produced), and very large databases (for which maybe only a sample should be selected for preservation).

In this chapter, I have focused upon data but there are also issues in relation to archiving statistical or economic models, specialist software required for the analysis of particular data and geographical information to enable users to understand the spatial structure of the data. For a comprehensive review of the topic of electronic preservation, see Waters and Garrett (1996). A framework for creating and preserving digital information is provided in Greenstein and Beagrie (1998).

Selection of data for preservation and access

Because of the resource implications of the long-term preservation of data, it is simply impossible to retain it all and it becomes necessary to develop selection criteria. Hard choices need to be made and these should be informed by the goals of the institution and its values, but they should be open for criticism (Hazen *et al.*, 1998). Such criteria will need to balance both current and, as far as they can be predicted, future needs of the research community. Thus, judgements need to be made on the long-term scholarly value of the material, a decision which is inevitably subjective and so advice should be sought from researchers from within the discipline area and historians. One of the difficulties of selecting material for preservation purposes is that the data creators often see the completion of their research or the publication of current figures as the end of the process and thus the costs of maintaining data will be viewed as outweighing the benefits. In order to ensure that data are preserved for the future, the difficult issue of who pays for long-term preservation must be tackled and also more attention needs to be paid to the benefits to data producers of ensuring that their data remain accessible. Data which are unique, which have been expensive to create and which are difficult to reproduce may feature highly.

As an example, the acquisitions policy of the UK Data Archive is driven by users' views which are ascertained through both direct approaches and via the gathering of intelligence about data sources and needs. Data sets are acquired according to the following list of priorities:

- At the specific request of a researcher.
- To ensure that the Archive holds the strategic social science and economic data sets (*e.g.* the Family Expenditure Surveys, unemployment statistics).
- When it is judged that the data set will be substantially used based on the current or recent experience of similar data sets, especially if there is already a clearly defined user group.
- To create time series (*e.g.* the UK Archive will search for a missing data set in a series).
- To build up complete and coherent holdings on specified themes or geographical areas (examples include the extensive holdings of Northern Irish data and the various surveys of disability conducted in the late 1980s and early 1990s).
- According to agreed strategies with fellow archives whereby each specialises on particular topics in order to make better use of limited resources.
- In recognition of an unmet need for data relating to a particular area.
- In recognition that research using a data set has been exceptionally influential in its field.
- In order to build a collection of data sets of value for teaching purposes, *e.g.* small, easy-to-use and clearly documented data ["Ready access to shared electronic files can transform the classroom" (Maindonald, 1998)].
- When the original data have been accessed from the Archive but value has been added, say, by combining the data with other information:

Of course, data sets may sometimes be lodged in an archive even though they have not been solicited. This happens particularly when researchers retire or when institutions close. If such deposits are judged to be overly resource-intensive to process and archive, and are of a low priority in relation to the selection priorities, it may be necessary to reject them. Before data are accepted, the data must be accompanied by at least the minimum documentation which secondary analysts need to make informed use of the data, and legal permission to distribute the data must be provided by someone with the authority to do so. In particular, it is essential to ensure that copyright is not being breached by the data distribution; this can be particularly problematic in situations where the data are being deposited in an archive by someone other than the originator or sponsor. Because of the resource implications of preservation, it is probably sensible to reject studies for which usage is predicted to be low, *e.g.* small, very localised or peripheral studies, or for which the depositor wishes to impose too stringent a vetting process or too restrictive a set of conditions before allowing access.

In developing a selection policy, there is a tension between establishing stringent quality rules and ensuring they are applied vs. a more *laissez-faire* approach in which the secondary analyst is expected to judge whether data are fit for their purpose. Different data sets impact very differently upon the resources required to acquire and preserve; in particular, it can be very time-consuming to rescue old data sets. This is particularly the case when the originator is not available to assist with queries. In fact, this argues for the deposit of data as soon as the processing and basic analysis has been completed – before the creator moves on or forgets relevant information. Finally, it is worth realising that the use of data sets cannot easily be predicted and can be very affected by unpredictable issues such as publicity for a key publication. Usage can to a great extent be manipulated, for example by holding data workshops to promote use or featuring the data in publicity material.

Preservation policies

As illustrated above, preserving data costs money because it requires specialist staff and equipment. One of the greatest problems is where this money can be raised. The dominant short-term focus of research funding has exacerbated this problem. Occasionally, extra finances can be generated by showing that the costs of having to obtain alternative data because of the lack of a preservation strategy are high, and in these cases data preservation can be seen to be cost-effective and to avoid re-inventing wheels. Similarly, there may be excellent reasons for retaining access to information in order to achieve better corporate intelligence. Data creators who attach little value to long-term preservation of data resources are less likely to adopt procedures or standards which facilitate their preservation and thus the costs are increased. Raising their awareness of these issues is the first step towards creating a culture in which long-term preservation is given a higher priority. Many organisations suffer from the lack of an overall strategic policy with respect to archiving electronic data. In particular there is often confusion about the authority to dispose or retain material and very little understanding of the financial impacts of each. A critical issue is to separate an archival policy for the future from the need to deal with legacy material which is already at risk.

Conclusion

An evaluation of the particular institutional, legal and financial circumstances within a country may identify barriers to a desirable co-operation between scientific users of data and the researchers or officials who have responsibility for collecting them. It is in the interests of both groups to overcome such barriers and I hope I have indicated some ways in which this can happen.

Despite the universal acknowledgement of the importance of access to data for research and teaching purposes and the need to address the long-term preservation of such data, there is a notable absence of appropriate facilities in many countries. A concerted effort is needed on behalf of the international community to ensure that resources are made available to disadvantaged social scientists.

NOTES

1. International seminars on Statistical Confidentiality. For enquiries, contact Eurostat, Jean Monnet Building, Plateau du Kirschberg, Luxembourg (http://europa.eu.int/comm/eurostat/).

2. Contact the UK Data Archive, University of Essex, Wivenhoe Park, Colchester, United Kingdom (http://dawww.essex.ac.uk/).

REFERENCES

Begg, C. *et al.* (1996), "Improving the Quality of Reporting of Randomised Controlled Trials", *Journal of the American Medical Association* 276.

Drucker, P. (1969), *The Age of Discontinuity: Guidelines to Our Changing Society*, Harper and Row.

Greenstein, D. and N. Beagrie (1998), *A Strategic Framework for Creating and Preserving Digital Resources*, Library Information Technology Centre, London.

Hazen, D., J. Horrell and J. Merrill-Oldham (1998), *Selecting Research Collections for Digitisation* European Commission on Preservation and Access.

Hendley, T. (1998), *Comparison of Methods of Digital Preservation*, Library Information Technology Centre, London.

Maindonald, J. (1998), *New Approaches to Using Scientific Data*, Australian National University Occasional Paper GS98/2.

Marsh, C. *et al.* (1991), "The Case for Samples of Anonymised Records from the 1991 Census", *Journal of the Royal Statistical Society* A Vol. 154, Part 2.

OECD (1979), *The Social Sciences in Policy Making*, OECD, Paris.

Seltzer, W. (1998), *Population and Development Review* 24(3).

Singh, M. and C. Crothers (1999), "Training the Social Scientist: What are the Indispensable Skills and Tools?", *UNESCO World Social Science Report*.

Smith, A. (1996), "Mad Cows and Ecstasy: Chance and Choice in an Evidence-based Society", *Journal of the Royal Statistical Society* A Vol. 159.

UNESCO (1996), *UNESCO and the World Summit for Social Development*, UNESCO, Paris.

Waters, D. and J. Garrett (1996), "Preserving Digital Information", report of the Task Force on Archiving of Digital Information commissioned by the Commission on Preservation and Access and the Research Libraries Group, http://www.rlg.org/ArchTF/.

Chapter 6

NEW TECHNOLOGIES FOR THE SOCIAL SCIENCES

by

William Sims Bainbridge*
Senior Science Advisor, National Science Foundation

Introduction

The social-scientific promise of new computing and communications technology is extremely bright, but possibly more so in the private sector than the public sector. The new information technologies will be of moderate direct benefit to the forms of social science that serve the needs of government. Such developments as digital libraries will greatly facilitate the storage and dissemination of data, and in some areas new analytical techniques will also be beneficial. Collection of data for government use may not be improved very much, except possibly in areas related to information technology itself, which will be of increasing interest for setting policies. However, the new information technologies will be of absolutely decisive importance in two other areas of indirect interest to government. First, these technologies will revolutionise many fields of fundamental research across the social sciences. Second, collaboration between social scientists and information technologists will create many valuable new goods and services that strengthen the economy and serve the needs of citizens. Thus, this paper will concern changes in priorities and conceptualisations, as well as the impact of the new technologies.

Cyberspace

William Gibson, the man who coined the term *cyberspace*, presented a disturbing picture of the computerised future in his books *Neuromancer*, *Count Zero* and *Burning Chrome* (Gibson, 1984, 1986a, 1986b). Just a few years from now, according to Gibson, the chief role of government will be mismanaging gloomy housing projects for the unemployed underclass, while information technology corporations dominate economics, politics and culture. This image is not entirely new. Seventy-five years ago, in his allegorical drama *R.U.R.*, the Czech writer Karel Capek coined the term "robot" to describe the mechanical labourers who were created by capitalists to replace the unwanted working class (Capek, 1923).[1] Gibson and Capek wrote fiction, but policy makers and scientists in the real world have become very concerned about the so-called *digital divide* that separates social, ethnic,

* The views expressed in this chapter do not necessarily represent the views of the National Science Foundation or the United States.

racial and national groups in their access to Internet and related technology. In a series of reports, the US Department of Commerce has warned that the digital divide is widening, although the data are open to multiple interpretations (US DOC, 1999). On 28 July 1998, Vice President Gore (1998) advocated regulations that would force Internet service providers to subsidise school access to the Net, specifically because the Commerce Department research "reveals that this 'digital divide' between households of different races and income levels is growing".

In the private sector, William H. "Bill" Gates has established the Gates Learning Foundation "to bridge the 'digital divide' between those who have access to computers and the Internet and those who lack such access" (http://www.glf.org/LearningFoundation/). Awareness of the central importance of advanced technology is not limited to any one ideological camp. Conservative political columnist George Will recently wrote:

> The chastening fact for candidates and political analysts is that familiar political certainties may need revising. Watch a few hours of television, and you will see a slew of advertisements for products and services that did not exist a decade ago. Half of all Americans aged 18 to 29, and half of all with household incomes of USD 75 000 or more, go on line for information every day. The *Wall Street Journal* recently reported that in just the past two years high-tech industries in Texas have created more jobs than exist there in oil and gas extraction. Science and commerce have imparted far more direction and velocity to social change than politics ever has. (George Will, 1999)

From observations such as these, I draw the lesson that social scientists must aggressively carry out research on the nature and consequences of socio-technical changes brought about by the information revolution. But, more than that, they must participate actively in this revolution themselves, so that they can influence the direction it takes, for the benefit of the sciences and for the public good.

Research on the digital divide will often use tried and true methodologies such as random sample surveys, for example in order to be able to compare the people who do have Net access with those who do not (Novak and Hoffman, 1998). National sample surveys will have a significant role to play in the future, as they have in the past. For example, the General Social Survey (GSS) is one of the prime data infrastructure projects supported by NSF since 1972. One GSS item that bears on the theme of this paper was asked in both 1985 and 1996: "The federal government has a lot of different pieces of information about people which computers can bring together very quickly. Is this a very serious threat to individual privacy, a fairly serious threat, not a serious threat, or not a threat at all to individual privacy?" The results show a significant increase in concern over time. In 1985, 31.8% felt that this was a very serious threat, but in 1996, fully 39.5% felt this way. The fraction calling government databases a fairly serious threat rose from 30.8% to 34.8%. Looking at this the other way, the percentage who felt this was not a serious threat or no threat at all dropped from 37.4% to 21.0%.

Traditional government-supported surveys serve primarily the first of three purposes for government-supported social science. These are:

- Providing factual data to guide the establishment and development of government programmes to solve or mitigate social problems.

- Answering fundamental questions about the nature of social reality, thereby contributing to human intellectual understanding and the gradual awareness of truth.

- Creating new goods and services that benefit human beings and contribute to the vigorous growth of the economy.

Let us briefly consider the infrastructure needs of these three in turn.

Social indicators

First, consider factual data to guide government programmes. Accurate social indicators are very expensive to collect. For example, the massive Current Population Survey carried out each month by the US Bureau of the Census provides data from which the Bureau of Labour Statistics calculates the unemployment rate. Similar methods are used by Canada, Mexico, Australia, Japan and all of the countries in the European Union, although some nations instead get data from employment office registrations or unemployment insurance records.

Unemployment data have many uses, such as helping to guide the fiscal and monetary policies of nations over the business cycle. But, at the risk of oversimplifying, one could say there are two schools of thought about how they might contribute to dealing with the problem of unemployment: either you want to provide government benefits to the unemployed, in which case the unemployment rate is crucial information. Or you want to create new jobs, in which case the rate is of very little interest at all. At least in the United States, the emphasis has shifted to the creation of new jobs, and that takes us to the need for new kinds of information often connected to growing industries employing advanced technology.

To some extent, new technologies will facilitate collection of data which the government already knows it needs, and make possible and necessary the collection of new data. For example, in the United States, highly accurate data on land and housing sales are supposed to be kept at the county level, but the procedures and physical means for recording the information vary widely. One can imagine a future electronic system that not only made the work more efficient for the county but that automatically fed into a national database system for use by several government agencies, as well as by economists, geographers and other social scientists. Other nations may be much closer to this goal than the United States.

In order to track the development and impact of information technologies, government agencies are beginning to develop new indicators. For example, the National Science Foundation (NSF), the Department of Commerce, and the Office of Science and Technology Policy recently co-operated in holding a conference on digital commerce. It "assessed research to date on the scale, direction and significance of the emerging digital economy; engaged the private sector in suggesting the research, tools, indicators and data collection that may help to inform investment decisions and policy development; and promoted an open research agenda for better understanding the growth and socio-economic implications of information technology and electronic commerce" (http://www.digitaleconomy.gov/).

The NSF's Science Resources Studies (SRS) division has also been supporting work to develop indicators of the impact of information technologies. One accomplishment has been "Social and Economic Implications of Information Technologies: A Bibliographic Database Pilot Project". This small digital library provides "searchable listings of research publications, data sets and Web sites that can help us understand the social and economic implications of information, computation and communication technologies" (http://srsweb.nsf.gov/it_site/index.htm). SRS has also launched a new data collection effort called the Information Technology Innovation Survey to provide descriptive data for government policy makers on significant innovative activities in US corporations. This project will

also analyse patterns of innovation within industry sectors and amalgamate the available research findings into a unified database for investigation of relations between product and process innovations.

Perhaps the most cost-effective way of increasing the benefits of social science data to government is to make them more readily available. The people, after all, are the real government of a democratic nation, and they should be able to see the information possessed by public servants, whenever possible. Combined with judicious organisational reform, today's information technologies can accomplish a real revolution in this area.

For example, in June 1996, the entire General Social Survey was placed on the Web, with funding from NSF. In terms of the history of the Web, that was a long time ago – half way back to the beginning – but even the original version of the site had all the essential features of a fully-fledged survey digital library. There was a complete code book of all questions that had been included in the survey since 1972, with a search engine, multiple hierarchical indexes, hypertext links from each item to abstracts of all publications that employed it, simple response frequency tables, a decent online analysis facility and provisions for downloading the raw data for more sophisticated statistical analysis. Since that pilot project demonstrated the practicality and utility of such systems, NSF and other agencies have invested in further efforts of that kind.

The GSS Web site was created parallel to the initial phase of the Digital Library Initiative, which was supported by NSF, the Defense Advanced Projects Agency (DARPA) and the National Aeronautics and Space Administration (NASA). The second phase has added the Library of Congress, the National Library of Medicine, the National Endowment for the Humanities, and the Federal Bureau of Investigation (http://www.dli2.nsf.gov/). Among the major projects supported in Phase II is Harvard University's Virtual Data Centre, "an operational social science digital library". More recently, the NSF Directorate for Social, Behavioural and Economic Sciences (SBE) has supported Richard Rockwell and the Inter-university Consortium for Political and Social Research in a series of improvements to the ICPSR Web site that include better tools for the discovery of data and integration of major data resources outside of ICPSR fully into the searchable metadata databases.[2] These two projects are complementary, and together they move us towards the day when researchers, students and policy makers will be able to efficiently locate the existing survey data they need on line.

Several efforts are in progress to improve access to government statistical data. SBE has made a major award to Steven Ruggles and the University of Minnesota to create and disseminate an integrated international census database composed of high-precision, high-density samples of individuals and households from seven countries. The abstract of the NSF award boasts, "It will be the world's largest public-use demographic database, with multiple samples from each country enabling analyses across time and space".[3] Over the years, we have helped Ruggles and his team to create the Integrated Public Use Microdata Series (IPUMS) of American census data. The original phase of the Digital Library Initiative supported six large projects, but I like to refer to the GSS Web site as the seventh digital library of that generation, and IPUMS as the eighth.

Recognising the importance of maximising scientific value of government data, while preserving confidentiality rights, NSF has worked in partnership with the US Bureau of the Census to establish centres where data can be analysed under controlled conditions, in Boston, Los Angeles, Berkeley (California), and Durham (North Carolina).[4] Linking data about specific individuals across data sets is technically challenging, ethically sensitive and scientifically highly fruitful. Thus, a new award has been made to John Abowd, John Haltwanger and Julia Lane to develop a number of aspects of this crucial work. The NSF-funded National Computational Science Alliance (NCSA) will be providing technical assistance in the area of Web-based collaboration tools.[5]

In areas as diverse as regulation, taxation and education, government needs to understand the patterns of growth and influence of the new information technologies. The biennial Study of Public Attitudes Toward and Understanding of Science and Technology is an NSF-supported data infrastructure survey that polls about 2 000 Americans and has recently begun including questions about Internet use. In 1997, only about 13% of respondents were able to give a satisfactory verbal answer to the question, "What is the Internet?" (National Science Board, 1998a). About 15% said they used the World Wide Web at least an hour each week. However, of the sub-group of respondents identified as attentive to science and technology issues, fully 30% used the Web regularly, and the figure was 37% for respondents with graduate or professional degrees (National Science Board, 1998b). We would expect the figures to be higher today, and we await publication of results from the 1999 survey.

Although the survey is very long by telephone standards, it has room for only a very few items on any given topic, and perhaps for that reason the scientific questions that can be addressed with the data are limited. Even in a representative democracy, it is difficult to see how the views of the general public about science are translated into public policy, or how vague public attitudes shape the career choices of those few individuals who might become scientists. Among the more sophisticated aspects of the survey is that it attempts to identify the core respondents who are most knowledgeable about and attentive to each major aspect of science or technology, and in 1997 only 14% (just 288 respondents) were judged to be generally attentive to S&T issues.

A broad public survey like this could be far more valuable if it were combined with a series of ancillary studies that zeroed in more closely on sub-groups of the population for whom science was more relevant, asking a much greater number and variety of questions, not only about knowledge and attitudes but about actual experience with technology and how beliefs about science fit in with other ideological factors. Only then would it be possible to connect the data to the substantial body of theory and social scientific findings about worldviews, technical careers, and public policy making.

Ideally, national or regional research centres should be set up specifically to carry out surveys of both the general public and of sub-groups within it, about all aspects of science and technology, but focusing especially on the diffusion of the newest computing and communications technology. These centres could be of the conventional type, situated in a particular location to accomplish a single task. Or they could be a more modern network of multi-disciplinary research projects linked by the new communications technologies into a distributed collaboratory. This takes us to the territory of fundamental research.

Fundamental scientific research

Recently, NSF has funded a variety of moderately large research projects looking at particular aspects and impacts of information technology. While these are not linked together in a collaborative network, if we were able to support a greater number it might make sense to organise them into a virtual centre. Rob Kling studies the shaping of knowledge networks in scientific communication. Sara Kieseler examines the impact of information technology on scientific collaborations within and among disciplines. Hal Varian leads a team that combines economics, social legal studies and computer science to explore the changing dynamics of intellectual property rights management. Sharon Dawes is evaluating several empirical cases involving groups of agencies in New York State engaged in programmatic or administrative innovations that depend on the sharing of knowledge and information across multiple organisations. Noshir Contractor is using computational modelling as well as empirical data analysis to understand the co-evolution of knowledge networks and 21st century organisational forms. Alaina Kanfer is studying the NSF-funded National Computational Science

Alliance to learn the extent to which knowledge really can be distributed electronically rather than requiring a team of scientists concentrated in the same physical location.[6]

While focused on topics of great interest to government, these projects are based in universities and intimately connected with the social scientific search for fundamental knowledge. Scientific progress benefits to only a limited extent from data collected primarily for the benefit of government bureaucrats. Indeed, there is a law of diminishing returns which says that each additional year of conventional data adds progressively less knowledge. A second law of intellectual exhaustion says that government bureaucrats tend to focus only on a very few areas of human life. Years ago, I believe it was Albert Biderman who pointed out that government collects much administratively convenient data that may not be valid measures of anything anyone is really interested in. But even valid data about important long-term social trends can be cost ineffective if collected too often, and may contribute nothing to the growth of scientific knowledge outside a very limited intellectual area.

The social sciences are a vast realm, covering anthropology, archaeology, economics, political science, social psychology, sociology and numerous sub-disciplines and inter-disciplines. Some of the most exciting questions are in fields remote from social problems. The appropriate research methodologies are often cutting edge rather than traditional, for example gene sequencing for anthropology, Internet-based auction experiments for economics and Web-based surveys for sociology. In order to promote the public good of scientific knowledge, governments must invest in many kinds of research infrastructure outside the cramped territory of social problems.

Three major NSF awards establish facilities to store or analyse samples of spoken language. Mark Kornbluh received funding from the Digital Library Initiative to create the National Gallery of the Spoken Word.[7] Brian MacWhinney is building a distributed, Web-based data archiving system for transcribed video and audio data on communicative interactions.[8] Francis Quek is carrying out research on human gesture, speech and gaze that involves video analysis, speech/signal processing and Web-based multimedia databases.[9]

Research of this kind can be of indirect value to national security agencies since computer voice recognition permits automatic scanning of telephone and radio transmission for words related to such topics of concern as the planning of terrorist activities. However, a vast number of valuable fundamental research projects can be carried out using this technology in such fields as linguistics, political science, psychology and sociology.

NSF has begun supporting work to develop collaboratories to study social and economic interaction through online, real-time experiments. A prominent example is the project headed by David Willer, to build, maintain and evaluate a software and data library for experiments, simulations and archiving for the social and economic sciences.[10] NSF is also supporting a number of smaller projects exploring the potential not only of online experimentation but also of Web-based surveys. However, we have not as yet established a survey research centre or other major project that takes telephone administration of questionnaires the next logical step to the Internet.

The reliability and validity of Web-based surveys are a subject for research and debate, rather than assumption. Surveys can be targeted to specific populations and even to selected lists of individuals over the Internet, and need not be a convenience sample of visitors to a Web site. We must keep in mind that random samples are only one of several methods for rendering results generalisable.

An alternate methodology is experimentation, with random assignment of research subjects to alternative treatment and control groups. A different, but equally valid, approach is to state theories rigorously and formally, then to put them at risk of disconfirmation by any data that properly

operationalise the concepts. A third alternative is repeated replication of the same study with different populations believed to differ significantly in terms of relevant parameters, in sophisticated variants of meta-analysis (Johnson, 1991). A fourth also involves replication, but employs somewhat different research techniques and is sometimes called triangulation. Finally, if inexpensive Web-based surveys produce a really interesting finding with some regularity, then a good argument can be made for including the best measures in an expensive survey carried out by more traditional means.

If the 21st century is to really be an information society, then culture should become the prime topic. By concentrating excessively on issues of social class and large-scale stratification, sociology has been applying 20th century methodologies to 19th century questions. By limiting its prime area of interest to pre-literate cultures, cultural anthropology has been applying 18th century methodologies to prehistoric questions. Either both must move into the new Millennium, or a new, autonomous social science of culture will arise, employing rigorous scientific techniques to the measurement of cultural structure and explanation of the dynamics that create them.

In recent years, I have noticed the emergence of what I call new social sciences. They cover some of the empirical territory already covered by sociology and social psychology, but they draw upon little of the traditions of those fields. The scientists who practice them tend to have degrees in computer science, information science, communications or even library science. Perhaps because they carry none of the political baggage of traditional social sciences, they are readily employed by corporations not only to study the productivity of information technology, but also to analyse work in terms of information science models. Theoretically, they tend to begin with a *tabula rasa* lacking most of the traditional social science ideas, as illustrated by the recent collection *Simulating Organizations*, co-edited by a former NSF programme officer and by a former member of an NSF advisory panel (Prietula *et al.*, 1998). For better or for worse, traditional social sciences are at risk of being replaced by these newly ambitious competing disciplines.

New products and services

This is a good point to emphasise that this paper expresses my own views rather than the official position of the National Science Foundation. Any discussion of the future of new technologies in social sciences is bound to be speculative. Given the revolutionary nature of the technology and of some of the social forces it has unleashed, any official pronouncements are likely to be too short-sighted. One reason is that this revolution is likely to upset the current balance of commercial, political and cultural vested interests. Another is that imagination is a very inaccurate predictor of the future. On the other hand, without free imagination and a willingness to take risks, we cannot create the best future for ourselves and our descendants.

With those provisos, we can examine the untapped potential of the social sciences to innovate new products and services. For too long, social and behavioural scientists have contributed far less than their fair share to economic growth, and thus they have collected less than their share of the profits. Phrases like “the information society”, “the global village” and “the digital economy” may be rhetoric, but in truth we really may have reached a new watershed in technology and the form of society. If we wish to benefit and to reduce the problems associated with change, we have to adopt the working assumption that a really significant transformation is indeed in progress. This will often mean that government should invest in innovative research centres and in programmes to develop new kinds of infrastructure.

An excellent example is the recent development of geographic information systems (GIS). Much of the progress has been in physical geography, and there is much room for improvement of

applications suitable for socio-economic data. One of the pioneers in the commercial development of GIS is Jack Dangermond. His Environmental Systems Research Institute should be an inspiration to social scientists who possess both technical expertise and an entrepreneurial spirit (http://www.esri.com/).

For a number of years, NSF supported the National Centre for Geographic Information and Analysis, and both phases of the Digital Library Initiative funded the GIS-related Alexandria DL. The current version is called the Alexandria Digital Earth Prototype Project, and it will assemble geo-referenced data collections as well as develop new computerised environments called Iscapes (or Information Landscapes) based on the Earth metaphor.[11] Parallel to these efforts, SBE has funded the Centre for Spatially Enabled Social Science, headed by Michael Goodchild.[12] Note that the rapid development of GIS has required parallel efforts in the public and private sectors, including government support for science and technology development centres.

In a manner not dissimilar to GIS, computerised techniques are being developed to handle three-dimensional representations of physical objects. Of course such systems already exist, from computerised tomography in medicine to industrial design such as employed in the aviation industry. But new systems developed to facilitate specific areas of scientific research will have many spin-off benefits in industry.

Timothy Rowe is working on a digital library project to apply high-resolution x-ray computerised tomography techniques to vertebrate morphology, which will have implications for physical anthropology.[13] Gregory Crane's Perseus Project DL is a well-established source of comprehensive information about classical civilisations, and it is adding substantial material about the architecture of archaeological sites.[14] An interdisciplinary team of computer scientists and archaeologists, headed by David Cooper and Martha Joukowsky, is developing systems for representing and analysing artefacts and sites, using as its test-bed the marvellous remains at Petra in Jordan.[15] All this fundamental research should have implications for future industries, for example in all areas related to the production of works of art. One could well imagine a time in the not-too-distant future when sculpture will be composed on computers much in the way that music has been for centuries, as a set of abstract specifications which can be rendered into physical objects at many different points in time and space, at varying scales and from various materials.

Logically, the social and cognitive sciences should be central to the information society. To some extent this means continuing old traditions of social indicator research, but adding new indicators that are sensitive to important changes. But it also means the application of social science principles and research methodologies to the development of new information and communication systems – socio-technical design. Rather than letting engineers and computer scientists build a new system in isolation for its potential users, social and behavioural scientists will participate actively in the design and development process from the very beginning. Indeed, often the original idea for a new product or service will come from a social scientific discovery or insight.

Before social science can make its maximum contribution to the well-being and prosperity of future generations, it must radically revise its research priorities to the point that many new disciplines must be created, existing ones transformed significantly, and some terminated as genuinely obsolete. In addition, we must escape the decades-old assumptions about what is possible that currently imprison our imaginations, often without our awareness. An area in which old assumptions are already under serious challenge is the music industry.

Music

The music industry is a major source of wealth and employment. Tastes and styles vary enormously, and the creation of music is a highly complex form of social interaction. However, major journals in the social sciences almost never publish articles about music, and the few academic jobs available for musicologists are in conservatories and music departments whose main emphasis is the teaching of performance. Music shares with astronomy the honour of being a subject of rigorous scientific research in the ancient world, notably in the mathematical study of scales and chords. Outside acoustics and the psychology of hearing, today one finds hardly any quantitative studies in a field dominated by historians, critics and folklorists. The NSF programme in Cultural Anthropology has on rare occasions funded projects that incidentally examined the music of a people. For example, Deborah B. Gewertz and Frederick Errington of Amherst College were supported to study urbanisation among the Chambri ethnic group of Papua New Guinea, and preference for music styles was one measure of adaptation to global mass culture.[16]

American popular music of the 19th century is the focus of one of the new digital library projects. The Lester S. Levy Collection at Johns Hopkins University contains 29 000 pieces of American music, chiefly printed scores of songs, covering the period 1780 to 1960. Already, much of that collection is available on line in the form of bitmapped images, with a simple tool for searching the metadata but not the music or text of the songs themselves.[17] The DLI award will enhance the online collection with advanced search capabilities, sound renditions of the music and full-text lyrics. A key part of this effort is the development of optical music recognition software – comparable to optical character reading of scanned text documents – that will have wide applicability beyond this particular collection. Another digital library, headed by Samuel Armistead, also involves music, the multimedia archive of Folk Literature of Sephardic Jews, which contains many ballads.[18]

If culture will be the fundamental resource of the information society, then we need a wholly new discipline, a quantitative cultural anthropology, to chart it and to generate a range of valuable new systems for managing and even creating it. The scientists who will create this field can begin work immediately in collaboration with enterprises already on the Web. For example, the Web-based All Media Guide "is an ongoing project to review and rate all music (whether in-print CDs or out-of-print on vinyl) and list (and rate) all feature movies and provide their credits and related information" (http://allmusic.com/amg_root.html). It uses a variety of means to categorise recorded music, including maps showing historical linkages between genres and a hypertext encyclopaedia of composers, performers and genres. More than 200 "freelance music experts" have contributed text and data to the Web site, but there is no sign that quantitative methods of analysis have played a role. One approach, already employed by Amazon.com, is to identify for any given music CD, what other CDs were purchased on line by the people who bought it (http://www.amazon.com). Given a sufficiently large database of purchases, it is possible to derive a cultural map of music, showing how individual artists cluster into schools and lineages that themselves cluster into genres.[19]

Furthermore, the Internet may already be in the process of revolutionising the music industry. Time will tell, but the fact that artists may now sell self-recorded tracks, collections and CDs directly over the Web may reduce the influence of the recording companies, possibly shifting more of the rewards to the musicians themselves. On 21 September 1999, popular rock artist David Bowie began to distribute his latest work on line from his own Web site, as well as in the conventional manner through stores (http://www.davidbowie.com/). He envisioned a time in the near future when distribution would be entirely Web-based: "It's impractical at the moment because so few people have the bandwidth and programs necessary to download an album. But, mark my words, this is where the consumer industry is going, We are not going back to record companies and through shops. Within five years, it will have morphed so spectacularly that no-one will recognise the music business."[20]

One factor sustaining the power of music distribution companies is the protection provided by copyright, yet many thousands of people are sharing music recordings on line every day in violation of copyright. Issues related to intellectual property rights have become the centre of vigorous debate and the recent report from the National Research Council, *The Digital Dilemma*, considered the music industry to be an early test case and harbinger of things to come (Computer Science and Telecommunications Board, 2000).

Even as governments are attempting to defend traditional intellectual property rights against corrosive forces associated with the new technology, a good case could be made for ending music copyrights altogether. Of course, performers would still have the right to be paid for their highly skilled labour in performing music live, but the absence of government copyright protection would diminish the ability of a few performers and of the distribution corporations to monopolised recorded music. Copyrights are a form of government regulation, and ending them would merely be another case of deregulation.

One of the traditional arguments in favour of copyright protection is that it encourages creative production of value, but this is an assumption that should be tested empirically. The NRC report says, "Research should be conducted to characterise the economic impacts of copyright" (Computer Science and Telecommunications Board, 2000, p. 17). This is a fruitful area for government-funded research, perhaps justifying the establishment of centres for research on the social, economic and legal issues concerning property rights in the information age. An example of the kind of research they might do is the project being carried out by Hal Varian's interdisciplinary research team at the University of California, Berkeley, examining the complex interaction between pricing, copyright and physical copy protection technologies.[21] Ultimately, such research should include exploring the possibility that copyright has a net deleterious economic effect, increasing profits for a few artists and for major distribution companies, but decreasing wages for most artists who cannot benefit from mass distribution and who do not have the money needed to defend copyrights in courts of law.

Research could test the counter-hypothesis that copyrights in the music industry discourage creativity by concentrating power in the hands of distributors who increase their profits through promoting the celebrity of their artists rather than by improving the quality or diversity of the music itself. If the balance of power swings back to artists from the recording companies, then the social structure of music may also be transformed. This may give valued social roles and satisfactory incomes to large numbers of locally popular artists. Their live performances would be promoted by their personal Web sites that also bring in some income by selling music tracks at a sufficiently low cost that the motivation to pirate them is weak. Their intellectual property rights may be sustained by the informal social network of the community rather than by government regulation.

Memorials

Music is a good example of a well-established industry that is possibly facing transformations that might require greater social-scientific attention. It is important to realise that wholly new industries may arise out of the information revolution, that also may call for re-direction of some government funding and social scientific energy. At this point in time we cannot predict what these new industries will be. Therefore, any example I could offer would be controversial, but it is essential to recognise that revolutions have revolutionary consequences.

An interesting feature of the popular *Star Trek* universe is that mass-media popular culture is absent from its fictional future world. Several characters play musical instruments and the preferred styles of music are classical, whether European or belonging to some other high culture. Perhaps

precisely because the characters are living very future-oriented lives, they turn to historical sources like Mozart for their aesthetic recreation. Presumably, the copyrights have all expired. Instead of passively watching television programmes and movies, they programme their own "holodeck" virtual reality dramas in which they play active roles, often with historical settings. Government is certainly not in the science fiction business, but government-encouraged research is currently developing the technology to realise the *Star Trek* prophecies.

As several of the digital libraries described above demonstrate, the new information technology has the capacity to revolutionise our use of data about the past, and thus potentially transform our relationship to previous generations and to earlier periods of our own lives. At the extreme, the technology may bring the past alive.

Many of the social sciences employ historical data, but some excellent examples of how the new technologies have been applied are closer to the humanities. The "Valley of the Shadow" at the University of Virginia "is a hypermedia archive of thousands of sources for the period before, during and after the (American) Civil War for Augusta County, Virginia, and Franklin County, Pennsylvania," communities that fought on opposite sides in this bloody struggle (http://jefferson.village.virginia.edu/vshadow2/choosepart.html). Steven Spielberg's Shoah Visual History Foundation is creating a vast hypermedia archive including digitised video interviews to "ensure that the voices of more than 50 000 Holocaust survivors and witnesses will speak to people around the world for generations to come" (http://www.vhf.org/). On a much lower technical level, people have begun putting up Web sites for deceased members of their family. For example, the Empty Arms Web-ring has created fully 584 sites to commemorate deceased children (http://www.emptyarms.org/).

At the same time, many kinds of online communities have been created to facilitate virtual interaction between living people (Rheingold, 1994). Now, there are beginning to appear both commercial and non-profit Web sites that combine virtual community with preservation of the past. For example, ClassMates.com is an online database intended to help people find fellow graduates of the schools they have attended (http://www.classmates.com/). The B.C. Harris publishing company, which has printed alumni directories for more than three decades, is now trying to connect alumni into online communities (http://www.bcharrispub.com/). Around the world, many schools have added alumni communication facilities to their Web sites, and those Web sites also tend to offer some historical information. For example, the Web site of the Choate School not only has a history of the school, but also a facility for e-mailing alumni directly, plus hot links to their personal Web sites (http://www.choate.edu/). In addition, some private citizens have created yearbook-like Web sites to link graduates of their particular class, sometimes with biographies and other historical information.[22]

A recurrent theme in William Gibson's science fiction novels is that deceased people can live again in cyberspace. Perhaps the first detailed literary development of the idea that entire communities could be preserved in cyberspace for all eternity is the 1953 novel, *The City and the Stars*, by the English author, Arthur C. Clarke (Clarke, 1953). It is certainly far beyond current technical capability to transfer the neural net of a human brain into a computer, and this may never be possible. But already it is possible to record some aspects of a person's memory and personality, and to simulate some aspects of a person's behaviour.

Much work is going on in the areas of voice recognition and computerised speech, and on the computer-generated images of living people called avatars. Already, a few simple embedded systems use human voices. On the Washington DC subways, the digitised recording of a real woman's voice warns people that the doors are closing. If the door sensors detect that people are blocking the doors, she tells them to get away from them. This is the absolutely minimal example of artificial intelligence.

Now let us suppose that the real woman whose voice was recorded dies. A very small aspect of her has become immortal on the Washington subway – her nagging.

Interactions with future computers will often take the form of conversations, and it will be technically feasible to give them the voices, mannerisms and even some of the memories of real people. Much behavioural science research on human cognition, memory, speech and physical movement will be required to create robots and information systems that have personalities. The social sciences will be needed to provide the context of family and community beyond the individual – employing representations of social networks, geographic information systems, databases of government census data, and digital libraries of all kinds of documents from the lives of individuals and organisations.

For many years, government agencies and private corporations have invested in the development of expert systems or decision support systems that automate the practical knowledge of professionals in a given field and provide advice from it as needed by practitioners (Bainbridge *et al.*, 1994). The best-known examples are probably in the medical field. One of the chief limitations of expert systems is the assumption that experts agree with one another, and when they do not the best systems fall back on fuzzy logic or probability statements to provide rough guidance when definitive answers are impossible. But this limitation is irrelevant if the aim is to capture the knowledge of a single individual, with all its arbitrariness and historical locatedness. The automatic kitchen of the future really could be given the cooking knowledge of a deceased member of your family, let's say Aunt Bessie. If you wish, the voice with which your kitchen speaks could be hers.

One of the original digital libraries, the Informedia project at Carnegie Mellon University, emphasised automatic transcription, abstracting and search of audio-visual materials, such as television news programmes. One of the central ideas was an educational system in which the student would ask questions – speaking them aloud – of a famous person such as President John Kennedy. The computer would search a vast repository of all of Kennedy's speeches, and then display the segment that best responded to the question. Intended for instructional purposes, this is tantamount to a conversation – albeit stylised – with a deceased person. A new component of Informedia called Experience on Demand, funded by the Defence Advanced Research Projects Agency, is developing methods for "capturing, integrating and communicating experiences across people, time and space". Its goal is to create a "complete searchable record of personal memory and experience" (http://www.informedia.cs.cmu.edu/html/enter/html; http://www.informedia.cs.cmu.edu/eod_html/Eod7-10/sld001.htm).

To the extent that information technology products and services of the future will involve simulated human personalities, the psychology of personality may have to redirect its research efforts. In recent decades, there has been much research exploring how many dimensions are required to measure variations in the ways that people can be described, chiefly in terms of their interpersonal style. Some have said three dimensions, others have claimed as many as sixteen, and today the debate is centred around five to seven. But all this research has been *nomothetic*, seeking to identify general principles that hold statistically across large number of individuals. Future research may need to be more *idiographic*, measuring ways in which individuals are unique (Pelham, 1993; Shoda *et al.*, 1994). In addition, the focus of measurement may have to expand, adding fresh interest in the skills that different people possess as well as in their preferences. Merely to have Aunt Bessie's recipes is to have the benefit or her culinary knowledge and to be able to savour the foods that she liked and consider adding them to your own preferences. I do not know whether you will also want your robot kitchen of the future to display her moods. How far we push the technology will be an interesting research question in the social scientific study of culture.

Conclusion

In an article which I shared with participants at the Paris meeting that prepared for the Ottawa conference, I argued in favour of the creation of an Internet-based network of collaboratories combining survey, experimental and geographic methodologies to serve research and education in all of the social sciences (Bainbridge, 1999). Some elements of this International Network for Integrated Social Science already exist, chiefly specialised survey data archives, but we do not yet have centres for Web-based experimental and survey research, nor meta-centres dedicated to integrating the methodologies. In this paper, I have suggested that the social science infrastructure of the near future must also include a number of centres and laboratories for the development of social science oriented towards information technology, perhaps with a new emphasis on areas that will generate economically valuable applications.

Some of these centres would apply somewhat new methods to the archiving and analysis of the traditional forms of data that government needs for its own purposes, augmented by new kinds of data about the diffusion and impact of information technology itself. But many others would be laboratories to apply the technology to fundamental research in social science, and to apply social science to the further development of the technology. This will require substantial investments, and the research awards mentioned here are all fairly large, from around USD 500 000 to over USD 5 000 000. The Digital Library Initiative is a perfect example of how such an effort can be conducted co-operatively by several agencies of government. The new International Digital Library effort is an excellent vehicle to begin global co-operation (http://www.dli2.nsf.gov/; http://www.dli2.nsf.gov/intl.html).

If we do our work well, social science will draw successfully upon the new technologies to achieve all three goals: providing factual data to guide public decision making, answering fundamental questions about the nature of social reality, and creating new goods and services that will be of tremendous importance to the citizens of our nations. Conceptualising the digital divide in terms of access to receive information by means of the new technologies is far too one-sided. Individuals and groups also need equal access to provide information. Integration of social science into all aspects of the development and design process can help ensure that people in all sectors of society will play an active role in creating information technology that serves genuine human needs.

NOTES

1. In several Eastern European languages, the word “robot” specifically refers to “work” and conjures up images of the exploited working class.
2. NSF award 9977984.
3. NSF award 9907416, http://www.ipums.umn.edu.
4. NSF awards 9311572, 9610331, 9812173, 9812174 and 9900447.
5. NSF award 9978093.
6. NSF awards 9872961, 9872996, 9979852, 9979839, 9980109, and 9980182.
7. NSF award 9817485, http://www.ngsw.org/.
8. NSF award 9980009.
9. NSF award 9980054.
10. NSF award 9817518.
11. NSF award 9817432, http://www.alexandria.ucsb.edu/adept/10.html.
12. NSF award 9978058.
13. NSF award 9874781.
14. NSF award 9817484, http://www.perseus.tufts.edu/.
15. NSF award 9980091, http://www.brown.edu/Departments/Anthropology/Petra/.
16. NSF award 9417980.
17. NSF award 9817430, http://levysheetmusic.mse.jhu.edu/.
18. NSF award 9874771, http://philo.ucdavis.edu/SEFARAD/.
19. An example of methodology applied to a literary genre rather than to music is given in Bainbridge, 1986.
20. David Bowie, quoted by Edna Gundersen in “Turning to Face the Challenge”, *USA Today* online edition, 4 October 1999.
21. NSF award 9979852.
22. A good example is the Web site for the 1954 graduation class of the elementary school of Old Greenwich, Connecticut, http://users.erols.com/bainbri/og.htm.

REFERENCES

Bainbridge, William Sims (1986), *Dimensions of Science Fiction*, Harvard University Press, Cambridge, Massachusetts.

Bainbridge, William Sims (1999), "International Network for Integrated Social Science", *Social Science Computer Review*, Vol. 17, pp. 405-420.

Bainbridge, William Sims, Edward E. Brent, Kathleen M. Carley, David R. Heise, Michael W. Macy, Barry Markovsky and John Skvoretz (1994), "Artificial Social Intelligence", *Annual Review of Sociology*, Vol. 20, pp. 407-436.

Capek, Karel (1923), *R.U.R*, Doubleday, Page, Garden City, New York.

Clarke, Arthur C. (1953), *The City and the Stars*, Harcourt, Brace and Company, New York.

Computer Science and Telecommunications Board, National Research Council (2000), *The Digital Dilemma: Intellectual Property in the Information Age*, National Academy Press, Washington, DC, pp. 76-95.

Gibson, William (1984), *Neuromancer*, Ace, New York.

Gibson, William (1986a), *Burning Chrome*, Arbor House, New York.

Gibson, William (1986b), *Count Zero*, Ace, New York.

Gore, Al (1998), "Statement by Vice President Gore on Department of Commerce Report about the Growing Digital Divide", 28 July, http://ed.gov/PressReleases/07-1998/wh-0728.html.

Johnson, Blair T. (1991), "Insights About Attitudes: Meta-analytic Perspectives", *Personal and Social Psychology Bulletin*, Vol. 17, pp. 289-299.

National Science Board (1998a), *Science and Engineering Indicators 1998*, Table 8-9, National Science Foundation, Arlington, Virginia, http://www.nsf.gov/sbe/srs/seind98/start.htm.

National Science Board (1998b), *Science and Engineering Indicators 1998*, Table 7-24, National Science Foundation, Arlington, Virginia, http://www.nsf.gov/sbe/srs/seind98/start.htm.

Novak, Thomas P. and Donna L. Hoffman (1998), "Bridging the Digital Divide", 2 February, http://ecommerce.vanderbilt.edu/papers/race/science.html.

Pelham, Brett W. (1993), "The Idiographic Nature of Human Personality", *Journal of Personality and Social Psychology*, Vol. 64, pp. 665-677.

Prietula, Michael J., Kathleen M. Carley, and Les Gasser (eds.) (1998), *Simulating Organizations*, MIT Press, Cambridge, Massachusetts.

Rheingold, Howard (1994), *The Virtual Community*, Harper Perennial, New York.

Shoda, Yuichi, Walter Mischel and Jack Wright (1994), "Intra-individual Stability in the Organization and Patterning of Behaviour: Incorporating Psychological Situations into the Idiographic Analysis of Personality", *Journal of Personality and Social Psychology*, Vol. 67, pp. 674-687.

US Department of Commerce (USDOC) (1999), *Falling Through the Net: Defining the Digital Divide*, http://www.ntia.doc.gov/ntiahome/fttn99/contents.html; see also http://www.digitaldivide.gov/.

Will, George F. (1999), "Bradley's Amble", *Washington Post*, 9 September, p. A21.

NEW TECHNOLOGIES FOR THE *NEW* SOCIAL SCIENCES: DATA, RESEARCH AND SOCIAL INNOVATION

A COMMENT ON WILLIAM S. BAINBRIDGE'S "NEW TECHNOLOGIES FOR THE SOCIAL SCIENCES"

by

Paul Bernard
Université de Montréal

Introduction

William S. Bainbridge is perfectly right to argue, in "New Technologies for the Social Sciences", that we should make better use of advanced information and communication technologies to capture and analyse social life, and particularly those forms of social life that are evolving under the influence of these very technologies. I would, however, argue that social scientists still have to take the full measure of this transformation. Society will increasingly become self-organising, and it will increasingly make use of knowledge – especially social sciences knowledge – in this process. Our disciplines are thus called upon to provide new knowledge, and to provide it in new and more active ways. Consequently, our research and communication infrastructure has to become a vast feedback loop, linking the social sciences to their public. In other words, new technologies should be put in the service of new, more reflexive social sciences.

One can summarise Dr. Bainbridge's argument on the connection between new technologies and the social sciences (here and in an earlier paper he refers to) in the following three statements:

Social scientists should use the latest information technologies more extensively in order to better capture social reality

Two of the innovations he proposes have to do with data collection: Web-based surveys; and Web-based collaboratories. Two others have to do with the storage of data and their transformation in view of analysing them: Web-based data archives; and social sciences geographic information systems. The key challenge here is to make these innovations more than new gadgets; to shape them so that they serve the purposes of the active knowledge society.

Social reality itself is increasingly shaped by the use of new technologies and the cultural changes that these foster

For instance, to extend Dr. Bainbridge's example in the field of music and give it a "social network" twist, I am thinking of a suggestion, made a while ago, that classical music lovers might be interested in listening, for one evening, to what YoYo Ma has put on his own sound system (not a compilation, but a real-time sharing of tastes); on the following night, one might want to hear the most popular (or the most idiosyncratic) pieces of music as identified by the set of people who listened to YoYo Ma's selection the night before, and so on; allow for some exchange of e-mail, and you have a whole new social structure being formed, with the interplay of convergence and divergences that is at the core of any enriching set of social interactions. James Fallows, the editor of US News and World Report, has argued along similar lines about the way the daily newspapers of the near future will be put together: by subscribing to personalised electronic media that are more flexible than the standard paper publications, sets of readers will essentially devolve to editors the task of selecting and putting in context for them, from all the available information, the material that they consider valuable. This editing/brokering function of hitherto unexplored material, as well as devolution itself, are key to the new information society, as we will see later.

As a consequence, new technologies themselves should be used by social scientists to access this emerging reality

In other words, the digital society can best be analysed using digital methods and thinking in a "digital framework". Indeed, as I will try to show below, the distinction is less and less clear between social scientists and their methods on the one hand, and on the other hand participants in the information society who try to find their way in the richness and complexity of all that is available. Just as there is a notable increase in searches for health tips on the Web for purposes of "self-healing", and of finding one's way in the maze of medical services, social intelligence (reliable or not!) also is increasingly used for decision making. This is going on not only among policy makers, groups and citizens concerned with public issues, but also among individuals and families who face choices with respect to where they can find a residence, where they can send their kids to day-care or school, where they can look for jobs, which consumer goods they should get (perhaps taking into account ecological and social information), and so on.

Dr. Bainbridge draws our attention, then, to the importance of the new opportunities, and indeed the new capacities for social interaction created by these technologies. And, this applies equally well to the social processes we want to measure and analyse, and to the research processes with which we try to perform these measurements and analyses.

This being said, I will argue that the social sciences have to move radically in the direction of a more active involvement of social actors in the research process itself, rather than connect with their public only at the two endpoints of the research process: when data are gathered; and when research results are "disseminated".[1] The performance of social scientists in this respect has been quite uneven. Qualitative researchers have usually been better at this sort of involvement than quantitative ones. Statistical agencies are often required by law to pay attention to the dissemination end of things. Academics often have trouble finding a place in their *curriculum vitae*-driven schedules for such an involvement with the public. Non-profit organisations have often occupied that niche, while private sector research organisations usually sell their information to those who can afford it.

To argue my case for the public's involvement, I will take advantage of an unnoticed convergence between the perspectives of social scientists who approach society and social processes from very different perspectives.

The shape of the knowledge society

Fukuyama (1999), generally considered as a rather conservative analyst, recently argued that moving towards the information society has involved going:

- from the tight bonds of pre-industrial communities …
- to the bureaucratic co-ordination of individuals in industrial societies …
- and on to innovation and self-organisation in information societies.

This leads not only in the direction of decreasing, or at least transforming the role of the state; it also concerns all manners of large organisations, which Fukuyama sees more or less as dinosaurs. These organisations will have to devolve responsibilities, while keeping their ability to guarantee that the delegated mandates will be fulfilled. This devolution will not be so much to atomistic markets as to self-organising groups and networks. Only these will be in a position to take timely advantage of the abundance and complexity of information required to steer a course and achieve normatively determined goals.

Obviously, this requires that this information be produced, interpreted, synthesised and transformed, largely by the very action of these self-organising entities. The social sciences can play a key role in there. But they will not be in a position to play it "from above"; if they assume an overarching position, they will always be one or many steps behind this self-organising activity. Social scientists can provide valuable information, but in the very same process that they collect it (which does not mean that analysis can be dispensed with). They become part of the social innovation process, that part of the process which can be more reflexive because it is generally better equipped conceptually and methodologically.

Block (1996) comes from the other side of the political spectrum. As he examines the social origins of the current myth of the "Vampire State", he shows that most advanced societies currently operate under conceptions of economic activity which make them oblivious to some of the most important sources of economic prosperity. They focus to a very large extent on what makes stockholders and markets happy, thus disregarding the many possibilities offered by institutions that take into account the interests of all stakeholders, and that foster a delicate balancing (actually, a dialectics!) of co-operation and competition, within and among organisations. In Canada, the fine field study by Osberg *et al.* (1995) demonstrated the extraordinary economic importance for enterprises of what the authors call "soft technologies" for promoting motivation and co-ordination.

Knorr Cetina (1996), a sociologist of science, argues that society is not only impacted by scientific discoveries and their technological applications, but that its very texture has become infused with knowledge processes. It increasingly operates according to rules that mimic those of science: society gradually comes to resemble a laboratory, where people construct new worlds, use paradigms, feel their way, develop emotional relationships to knowledge objects and to technical objects, and so on.

Michael Gibbons *et al.* (1994) have argued that scientists are developing a new way of doing things, which they label Mode 2. While Mode 1 involved single disciplines, peer-review orientation

and distance from applications, Mode 2 is just the opposite: problem- and design-driven, open to the appreciation of outsiders as well as peers, and interdisciplinary. Moreover, the input of Mode 2 experts derives an important part of its value from their past association with people from different backgrounds with whom they have worked. A geologist with some exposure to anthropology might come handy, for instance, in prospecting work that involves going to territories where aboriginal populations live.

Paradoxically, one of the major reasons why Mode 2 scientific activity has arisen in the biomedical and natural sciences fields is the increase in the supply of scientific manpower, which could not all be used to fill the ranks of university faculty. There is a striking parallel, I think, with our numerous former students who work in the private social sciences industry (social sciences have already been much more active than Dr. Bainbridge assumes in creating new goods and services). Even though the academic practitioners of the social sciences usually hate to admit it, they are the close cousins of the public opinion polls and focus groups specialists, of the marketing experts, of the organisational consultants, and so on.

Two key questions arise as one examines this broad variety of contexts for the practice of the social sciences. First, how can their independence be maintained in spite of economic and political pressures? How can they walk the fine line between scientific validity and social usefulness? And second, how can the knowledge thus generated be made public? This is an indispensable requirement for two basic reasons: knowledge can only be validated if it is submitted to open criticism; and social engineering usually fails when it relies on magic bullets and manipulations, especially so in the self-organising society.

The role of the social sciences in the information society

If we are thus moving towards a self-organising information society, where the interests of a variety of stakeholders have to be accommodated, where epistemics increasingly becomes the texture of society, where knowledge itself is generated and used according to Mode 2, what should the role of the social sciences be? Basically, our disciplines should:

- Help society to understand its ongoing historical transformation.
- Provide information and shape it conceptually, so that innovation and self-organisation can take place.
- Identify organisational forms for flexible self-organisation.

This research and action agenda requires that:

- Active research methods and processes be substituted for passive ones.
- All parties to the information/knowledge society be involved in the interchanges about generating and applying knowledge.
- The social sciences themselves become self-organising.

All three aspects of this research agenda will be powerfully shaped by the research infrastructure that we have at our disposal and that we decide to use. Indeed, infrastructure not only enables and enhances research, but also orients it by putting at the disposal of researchers tools that are either dedicated to certain types of research, or that at least are more adapted to the development of some

research avenues than others. These are the issues we will briefly address in the rest of this chapter, examining successively the three following questions. First, to what extent do infrastructures accommodate an active stance in social sciences research? Second, how can we shape research infrastructures so that they support a broad involvement of the public? And third, how should the research infrastructures of the social sciences be governed in order to reach these goals?

Active and passive research methods and processes: a discussion of some of Dr. Bainbridge's suggestions

As I mentioned earlier, Dr. Bainbridge has suggested a number of innovative uses for information technologies in the social sciences, uses around which we would shape our research infrastructure in the future. While I applaud the initiative, I think we run, with each of them, some of the risks inherent in any top-down approach, where the full richness of social life cannot be adequately encompassed. I will provide here brief indications of these potential weaknesses with respect to all four of his suggestions.

Innovations in data collection: Web-based social surveys

Besides being timely, flexible and inexpensive, Web-based surveys offer the advantage of reaching out to the most technically innovative segments of the population. But they will obviously tend, for still quite a while, to disregard people on the other side of the digital divide: according to Grossman (1999): "the Internet could become an engine of inequality". In the United States, for instance, three-quarters of households with incomes greater than USD 75 000 have computers, as opposed to one-third of those with incomes between USD 25 000 and USD 35 000, and one-sixth of those with incomes less than USD 15 000. And when one takes into account the whole world, including the less developed countries, the divide obviously become quite a bit more impressive.

While some of the sampling biases involved in Web-based surveys can be corrected using weighting systems, the divide is simply too large to be dealt with technically. What risks being neglected here are forms of social innovation that take place outside the Web (*e.g.* daily coping with market and bureaucratic constraints among ordinary people), as well as the diffusion processes of technical and social innovations.

Innovations in data collection: Web-based collaboratories

These collaboratories would allow us to run, relatively inexpensively, large-scale social experiments, thus improving prospects for research and training. Besides the usual danger of overconfidence in the transferability of experimental results to real social contexts, however, such large-scale infrastructures would tend to increase the distance between experimenters and research subjects, who largely would not have direct contact with one another. Maybe we should not be so sure that we can dispense with all the informal interaction surrounding ordinary experimental situations. In Canada, the three major granting Councils (medical, natural sciences and engineering, and social sciences and humanities) have even adopted a common set of directives with respect to research involving humans, where the notion of research subjects has been replaced by that of research participants, playing a much more informed and active role in the research process. This of course does not rule out collaboratories; but it does introduce a note of caution about the ways in which participants should be able to "get back" at the researchers.

Innovations in the storage and transformation of data: Web-based data archives

Such archives are obviously indispensable in order to systematise data and improve comparability and access, while insuring protection of confidentiality. They involve two major risks, though. I can only sketch them here, and they would require thorough critical examination in current and future efforts at setting up archives, especially the most ambitious, international comparative and longitudinal ones.

First, the very welcome insistence of archives on making the data comparable across time and societies can have the perverse effect of "freezing" data in systems of concepts and categories that become outmoded with respect to changing social reality. Changing professional nomenclatures, classifications of industries, or even civil status categories in order to keep abreast of transformations in work, economic activity and family relationships is difficult enough when it is done in a single country; it becomes a nightmare when many counties are involved, each with its own cultural traditions and sets of interest groups. And yet, reality does change and does require adaptations in our measurement strategies, on a scale and at a pace that is simply unknown in the non-social world. Taxonomies used for describing and understanding the natural world remain fixed for long periods of time. Our classification schemes, by comparison, rapidly become inadequate, if not obsolete, oftentimes because they are used in social interaction outside the realm of social sciences, in social and political debates (think, for instance, of controversies surrounding the measurement of "race" and "visible minorities", of "sexual orientation" and "same sex couples").

Second, the construction of archives can lead to implicitly delegitimising data that is not easily amenable to standardised treatment. I am thinking especially here of non-quantitative data, which is not what archive builders usually have in mind first and foremost (the Qualidata archive at Essex notwithstanding). While quantitative data are often required to throw light on many issues, they have two important drawbacks. For one thing, quantitative data tend to permanently be out of tune with reality, because the data collection process rests, by definition, on conventional understandings of reality (embodied in questionnaires and forms). Everything social that is emerging, still difficult to express in more or less universally shared vocabulary, or simply not taken in charge by category-creating and legitimising institutions (public administration, the banking system, the media, etc.) is unreachable through conventional methods.

Moreover, focusing exclusively on quantitative data means neglecting the considerable amount of "different" digital data (discursive, textual, iconographic, even musical) that is generated in everyday activity and that can point us to a lot that is not accessible through questionnaires and forms. For instance, crises such as floods, earthquakes and tornadoes lead to the production of enormous amounts of data, if one counts in administrative and financial records, demographic counts, communications among authorities, media coverage including talk shows, and so on (to which one can, of course, add the results of sample surveys, qualitative in-depth interviews, focus groups, as well as observation data from relief centres and systematic inquiries about which social networks people used to solve the problems inherent in such crises). Much of the "spontaneously generated" material is frustrating because it has not been gathered for social sciences purposes and it provides incomplete information as to the context in which it was produced; but it has the immense advantage of being raw, uncontaminated by research procedures. It also represents a challenge because it is so massive that we don't quite know how to begin the task of roughly processing it in order to find leads for further refined (qualitative) exploitation; new methods revolving around lexical statistics do begin to offer some hope in this respect, though.

Innovations in the storage and transformation of data: social sciences geographic information systems

Geographic information systems provide extremely suggestive presentations of data: the multifaceted context in which our lives unfold is captured in a way that is so rich and so easy to grasp, because of the spatial representation, that hypotheses seem to immediately come to mind about what is going on; for this very reason, maps are also an excellent way for social scientists to get their ideas across to the public. Moreover, geographic information systems offer the possibility of bringing together data that have been developed in very different scientific, administrative and policy environments: one can learn a lot, for instance, from an overlay of data on various types of economic activity, on pollution sources and levels, and on the health situation.

The major danger with geographic information systems is of course ecological fallacy, the drawing of conclusions at the individual level from data that is defined at the aggregate level (territories). One should simply stand clear of the fallacy and use aggregate data for what they can really provide: an oft-neglected description of the local context in which people live, earn a living, send their kids to school, interact with one another informally and in local political institutions such as school boards, municipal and community councils, and so on (studies of concentrated urban poverty have demonstrated, for instance, the importance of such local contexts). Yet, social scientists should not neglect the capacity of individuals and groups to make their own sense of reality and develop their own strategies *vis-à-vis* their multifaceted local context; this context shapes the constraints and opportunities available, but social actors can in turn figure and strategize. In other words, we should use geographic information systems without getting entangled, as Dennis Wrong once put it, in an "oversocialised" vision of human beings.

Databases as instruments of social innovation

Not only should our research infrastructure avoid tuning out the active involvement of social actors in shaping their destinies, but it should bring about the active involvement of all parties involved in social sciences research in the generation and application of knowledge. It is not only a question of social justice, of attenuating the effects of the digital divide. It is also because social change cannot be implemented from above, without involving the people concerned.

Moreover, broadly connecting people to social science research may powerfully contribute to enriching our research agenda. Ommer (1999), for instance, has elaborated on the theme of "the centrality of marginality" in advanced societies, drawing attention to the fact that populations in marginal resource communities (for instance, fishing villages) know first-hand about where our food come from and are consequently in a position to be "stewards" of our natural resources and of our ecological balance. Similar reasoning could be applied to "hackers", who are marginal with respect to mainstream organisations in the world of information and communications technologies, and yet are actively sought after by these very organisations because they can play a key role (in security, but also in other matters) in the core of the knowledge-based economy.

In order to shape this active involvement of all parties concerned in social sciences research, we need to build active feedback loops. I will illustrate this proposition in the case of databanks and archives, which we can really turn into instruments of social, as well as research, innovation.

Plans for such an infrastructure should involve a first feedback loop between the producers/shapers of data, and the researchers using these data. This is where value is really added to data. The infrastructure can provide:

- A rich set of interconnections between data and results, so that researchers (and eventually research users) can find their way to what they need by pulling on any thread.

- An accumulation of procedural (usually unpublished) knowledge:

  - about the potential and the shortcomings of data;

  - about original ways of using the data: derived variables, longitudinal sub-samples, etc., which can tremendously accelerate and improve research.

- A constant process of critical sharpening of concepts underlying the data collection (concerning question formulation as well as sampling procedures).

- An opportunity to transform our classification schemes according to constant changes in the way we work, produce, form families, live the various stages of our lives, like youth or old age, and so on.

We also need a second feedback loop between the research producers (including the producers/shapers of data), and the users of research, decision makers, policy analysts, but also lower level and local implementers of policies, interest groups, non-governmental organisations and community groups. There are no magic bullets in the self-organising information society (if there ever were). Social innovation requires that all players be brought on side, inasmuch as possible, for the very basic reason that everything social takes on an historical character and involves conscious actors who define strategies; neglecting this ability of people to figure things out and to act autonomously often dooms plans for change that were carefully laid, but in closed circuit. The "connectedness agenda" which now adorns much of political discourse should involve much more than technology and getting cables to the people. It should mean helping people develop a familiarity with information, an ability to find what they need, to shape it to their purposes, to put it in context, to criticise it. Training and knowledge brokering will play crucial roles as prerequisites of the self-organising information society.

These two feedback loops are parallel as well as complementary. In both cases, the idea is:

- To take data into places where it has never been:

  - into research projects and conceptual debates that are innovative;

  - and into social uses, for political discussion and for administrative purposes.

- To bring back into the databases "travel" or field notes.

On the basis of these "experiments" in research and in civic discussion, data collection and analysis procedures can be improved, so that they fit our needs for knowledge and better capture social reality. In all of this, I reiterate, qualitative researchers are not the adversaries of quantitative data users, but rather allies, doing at the very least the exploratory work that the latter often cannot perform quantitatively.

Self-organisation in the social sciences

If the social sciences are to serve the purposes of the self-organising society, then they themselves should be self-organising. In a sense they are, of course, at least as academic disciplines. But, as I have

tried to show, they now have to undertake a new and different dialogue with society, and their self-organisation should lead them towards fulfilling the mandate I have been sketching above.

I advance only one proposition in this respect here: the key to our self-organisation will be *hybrid organisations*, where the interests of all the stakeholders are represented and where some adequate balance can be maintained between the demands of social, economic and political usefulness, and the requirements of scientific validity. Actually, I should not be talking about a balance, but rather about a dialectics: the social sciences can only be useful to the extent that they enjoy a large measure of scientific autonomy, and can thus reveal heretofore unknown facts about social reality; but at the same time, they can only preserve, and indeed increase social, economic and political support for their scientific endeavour if they keep being relevant to the needs of people. This is exactly what Fukuyama says, not only of the social sciences enterprise, but also of the whole information society.

Many examples of such hybrids in the social sciences have sprung up recently, such as:

- Information brokers/editors/synthesisers, who are themselves hybrids of researchers and communication specialists.

- Organisations with hybrid mandates of political usefulness and scientific credibility, such as statistical agencies.

- Granting councils moving towards oriented research, while leaving researchers free to define the best ways of pursuing relevant knowledge; peer review and partnerships are now being mixed in interesting and innovative ways.

- Research forums, with bipartite or tripartite leadership (researchers, users and statistical agencies), where leading-edge information and knowledge is exchanged.

- Banks of Frequently Asked Questions on various Web sites, where the information generated is a joint product of those who ask the questions and of those who answer them.

- Science Shops or Community-University Research Alliances where research users in the community team up with researchers to pull information together in order to generate new information for the purposes of deliberation and problem-solving.

One could go on and on, and the list will only lengthen over the years, if the *new* social sciences use new technologies in order to rise to the challenge of the self-organising information society.

NOTE

1. Note, incidentally, how the verbs “gather” and “disseminate”, as well as the word “data” connote passivity on the part of the non-researchers.

REFERENCES

Bainbridge, William S. (1999), *International Network for Integrated Social Science*, National Science Foundation, p. 31.

Block, Fred L. (1996), *The Vampire State and Other Myths and Fallacies about the U.S. Economy*, The New Press, New York.

Fukuyama, Francis (1999), *The Great Disruption: Human Nature and the Reconstitution of Social Order*, Free Press.

Gibbons, M., C. Limoges, H. Nowotny, S. Schwartzman, P. Scott and M. Trow (1994), *The New Production of Knowledge. The Dynamics of Science and research in Contemporary Societies*, London, Thousand Oaks, Sage Publications, New Delhi.

Grossman, Wendy M (1999), "Cyber View: On-Line U", *Scientific American*, July, p. 41, citing "The Virtual University and Education Opportunity", published by the College Board in Princeton, NJ.

Knorr Cetina, Karin D. (1996) "Epistemics in Society. On the Nesting of Knowledge Structures into Social Structures", in W. Hijman, H. Hetsen and J. Frouws (eds.), *Rural Reconstruction in a Market Economy*, MansholtStudies 5, Mansholt Institute, Wageningen, pp. 55-73. 1998 published in *Sociologie et Sociétés*, "Sociology's Second Wind", 30(1), pp. 37-50. 1997 Dutch translation by Hans Harbers (ed.), *Kennis en Methode* 21(1), pp. 5-28.

Osberg, Lars, Fred Wein and Jan Grude (1995), *Vanishing Jobs : Canada's Changing Workplaces*, James Lorimer, Toronto.

Ommer, Rosemary (1999), "Building Communities for the Knowledge-based Economy: The Role of Research in Resource-based Communities Making the Transition", paper presented at the Canada Foundation for Innovation Conference, Ottawa, November.

Chapter 7

FINAL REPORT OF THE JOINT WORKING GROUP OF THE SOCIAL SCIENCES AND HUMANITIES RESEARCH COUNCIL AND STATISTICS CANADA ON THE ADVANCEMENT OF RESEARCH USING SOCIAL STATISTICS

by

Paul Bernard, Chair, Betty Havens, Peter Kuhn, Céline Le Bourdais, Michael Ornstein, Garnett Picot, Martin Wilk and J. Douglas Willms, assisted by Hélène Régnier

Introduction

Canada's economy and society are in a period of rapid and difficult change. Timely and objective analysis of economic and social conditions is required to understand this transformation, to provide a basis for broad and informed debate on public policy, and to establish a foundation for intelligent policy formation. The need is particularly acute because Canada's social policy has not kept pace with the dramatic changes in its economic policy over the past two decades. Governments at all levels have acknowledged the importance of redesigning social policy so that it meets the needs of all Canadians, and leads us towards more civil and economically sustainable communities.

In one respect, Canada is well equipped to meet these needs. We now have a number of excellent and timely social surveys covering a variety of topics. Theoretical developments and advances in research design have led to the creation of longitudinal surveys which track individuals over extended periods of time. These new research tools provide information on the dynamics of poverty, the effectiveness of training programmes, the consequences of job loss, the influence of childhood experiences, and several other topics pertinent to redesigning social policy. Together, these form the basis for establishing a well-integrated system of "social statistics", a term we use to encompass information describing a wide range of human activity and the social, economic, educational and cultural features that affect our daily lives.

However, as a nation we have very little capacity to conduct social policy research, evaluate social programmes or monitor progress towards achieving social aims. The federal government recently acknowledged the need to strengthen its research capacity, and established the Policy Research Initiative to recommend and oversee the implementation of an interdepartmental research agenda. Similarly, provincial governments, the private sector, and non-government organisations recognised this need and have attempted to revitalise the policy research community.

There are at least three significant barriers that need to be overcome if we are to develop our research capacity in social statistics. The first barrier is the sheer lack of trained researchers. More than a decade of federal and provincial government restructuring and downsizing has significantly reduced the number of researchers working internally. This has occurred during a period of decline of the training in statistics and research methods in social science departments of our universities, with the exception of economics. The second significant barrier is access to data. Paradoxically, the detail provided by new data sets, which makes them so valuable, prevents them from being made public, as it could potentially enable users to identify specific individuals. The Statistics Act sets out strict criteria for maintaining confidentiality, and Statistics Canada necessarily has a very strong position on ensuring that the law is followed. The third barrier is that there are very weak links between the work of social scientists and the potential users of the knowledge they generate. Even though there is a tremendous appetite for social statistics about education, employment, health, literacy and other pertinent social issues, many of the important findings of social scientific research have not been adequately conveyed to the policy community or to a wider public through the popular media.

The mandate of our working group, convened jointly by Dr. Ivan Fellegi, the Chief Statistician, and Dr. Marc Renaud, President of the Social Sciences and Humanities Research Council, is to make proposals to encourage quantitative research on major social and economic issues using large-scale data. This report recommends the funding of three components, each designed primarily to surmount the barriers described above. Together, they comprise a *Social Statistics Research System*, which would complement the integrated system of social statistical data for which Canada is already considered a world leader.

The first component is aimed at increasing the number of researchers engaged in quantitative research on social and economic issues. It includes three separate proposals: Research and Training Groups that would bring together researchers from different disciplines and institutions to conduct quantitative research and training in priority areas; a Training Programme, including a summer school, that would provide specialised training in advanced statistical methods complementary to apprenticeships and graduate programmes, and support for data librarians to increase the use of social statistics in undergraduate training; and a Fellowship Programme, including M.A., PhD and Postdoctoral Fellowships, aimed at providing support for young researchers pursuing careers in social statistics, and Senior Fellowships that would enable some of Canada's leading social scientists to devote more of their time to research, and provide leadership in the training of the next generation of researchers.

The second component entails two proposals: Research Data Centres and remote access capabilities that would provide access to detailed micro-data, while maintaining the strict rules for the preservation of confidentiality required of Statistics Canada under the Statistics Act; and support for the enhancement and expansion of the activities of the Data Liberation Initiative.

The third component is the development of a Social Statistics Communication Programme that would implement a communications strategy to inform and build the public constituency for quantitative social science research. A key component of this programme are research forums that would support research networks, provide an arena for the presentation of research findings and enhance communication among researchers, the policy community and the media.

The report also recommends that Statistics Canada and the Social Sciences and Humanities Research Council negotiate a memorandum of understanding that defines the goals and organisation of the proposed Social Statistics Research System. A co-ordination structure is also recommended so as to take advantage of opportunities to strengthen networks and increase synergy among researchers.

The full implementation of this System would take place over a period of five to eight years, beginning 1 July 1999. The cost, after an initial start-up period of two to three years, is estimated to be about CAD 10 620 000 per year. This would be new money, added mainly to the budget of SSHRC, and to a lesser extent to the budget of Statistics Canada. We believe that an intervention of this magnitude is necessary to establish the infrastructure required to dramatically change the scope of quantitative social science research in Canada.

Canada has experienced, and continues to experience, dramatic social, economic and technological changes during the last two decades. Our economic policy has adjusted to these changes, particularly in response to pressures stemming from global international markets, and the need to abate inflation and reduce the national debt. These changes in economic policy have been accompanied by major changes in the labour market and in family structure. There is a general sense among many Canadians that the major problems we face are not economic but social. Governments at all levels have acknowledged the need to redesign our social policy so that it fits better with our current economic policy.

To meet this need, we require research on a wide range of social, economic, educational and cultural issues. The renovation of social policy must build on a basic understanding of the life course, and of the complex relationships among factors at different levels, such as families, neighbourhoods and communities. Conducting such research requires a well-integrated system of social statistical surveys and the capacity to analyse the data.

Beginning with data from the census and surveys on trade, finances, prices, the labour force and other areas, Statistics Canada has developed a national statistical system. Historically, its aim has been to provide aggregate indicators of current economic and social conditions, including descriptive statistics on economic growth, health, education, justice, productivity, the labour force and so on. During the 1970s, two major changes shaped its evolution. First, research in the social sciences demonstrated that an understanding of many social phenomena, such as the nature of criminal activities and victimisation or the effects of poverty, required separate, focused surveys. Second, the policy community recognised the importance of research that could help us understand how particular life events, combined with peoples' habits and lifestyles, affected their long-term social outcomes. Research began to stress the importance of *social context* – the families, neighbourhoods, schools and organisations in which people live and work – in shaping and constraining individuals' actions. But the descriptive data available from cross-sectional surveys were inadequate for monitoring changes in social outcomes or understanding the causal mechanisms that led to desirable outcomes. This required longitudinal surveys, in which data were collected from the same sample of respondents on at least two and preferably more occasions.

Statistics Canada responded to this challenge by instituting a new generation of social surveys during the 1980s and early 1990s. They include, among others, the General Social Survey (GSS), the Graduate Follow-up Surveys, the Survey of Labour and Income Dynamics (SLID), the National Longitudinal Survey of Children and Youth (NLSCY), the National Population Health Survey (NPHS), and the Displaced Worker Survey (DWS). Over the next few years, additional major new surveys are planned, including the longitudinal Workplace and Employee Survey (WES), the longitudinal Youth in Transition Survey (YITS) and the Survey of Financial Security (SFS). Meanwhile, Statistics Canada also strengthened its efforts in collecting administrative data and taxation data, and its capabilities to link data from various sources. Canadian researchers also played a major role in the development and administration of international surveys, such as the International Adult Literacy Study (IALS) and the Third International Study of Mathematics and to Science (TISMS). Several other surveys are currently being developed, and efforts are being made to integrate the survey data so that they can better characterise Canadian life. Together, these surveys have taken

the national statistical system in a new direction, and have helped distinguish Canada as a world leader in the collection of social statistical data.

Despite this great strength, Canada has relatively little capacity to analyse the data from these surveys. The most difficult barrier to surmount is that there are simply too few researchers engaged in quantitative research. The problem is especially acute in areas requiring advanced statistical methods. The capacity to conduct research within government institutions has declined over the past two decades, mainly due to restructuring and downsizing. Although these institutions appreciate the need to bring analyses of social statistics to bear on public policy debates, they are not strongly enough connected to the universities, where the majority of social scientists conduct their research and where the training of new researchers occurs. Within the universities, training in statistics and research methods has also declined severely during this period. There are now very few faculty teaching courses in advanced statistical methods, and in most social science fields, the course requirements for M.A. or PhD degrees no longer include formal training in statistics or quantitative research methods. The notable exception is economics, but in this field research capacity has been and continues to be eroded by the migration of qualified researchers to the United States.

This problem is exacerbated by the fact that the data generated by the new surveys are relatively complex: the surveys are usually longitudinal and have a multilevel structure, such as students nested within schools, or workers within firms. Thus their analysis requires powerful computing and statistical techniques. There have been tremendous theoretical developments in this area, but there are very few social scientists in Canada who have training in their application.

Another significant barrier is access to data. In the past, much useful research was conducted on public-use data files made available to university and other researchers by Statistics Canada, especially after the launching of the Data Liberation Initiative (DLI) by the academic community, with support from Statistics Canada, the Humanities and Social Sciences Federation of Canada, and SSHRC. A difficult problem is posed by the need to make detailed micro-data – the exact responses to questions in surveys – available to researchers, without compromising the confidentiality promised to survey respondents. The powerful statistical techniques which are appropriate for the analysis of multilevel, longitudinal data cannot be performed with aggregate data; access to micro-data is essential. Moreover, even simple descriptive problems often require access to micro-data. Part of the problem, ironically, is that the detail on individuals provided by longitudinal micro-data could make it possible for a researcher to identify specific individuals. This would violate requirements to maintain the confidentiality of individuals' responses, as set out in the Statistics Act.

The third barrier to the development of our research capacity in the social sciences concerns communication. Many social scientists necessarily devote their energy to addressing narrowly defined research questions that will advance their academic discipline. Success in academic careers depends to a large extent on publication in scholarly journals, and usually there are relatively few incentives or resources available for preparing more popular pieces that would convey research findings to a broad audience. Moreover, the demands of an academic career to excel in research and teaching make it difficult for researchers to forge strong links with the policy community, and often the time span of academic research is too long to meet the immediate needs of policy makers. Consequently, the transfer of knowledge from research to policy and practice is not as rapid nor as strong as it could be. Perhaps for the same reasons, social scientists have not been particularly successful in popularising their findings through the public media.

However, in the longer term, the strengthening of social science research requires widespread public recognition of the benefits of social policy research. We need to demonstrate that empirical research provides a basis for informing the development of public policy, evaluating social

programmes and monitoring our progress towards achieving social aims. There is no substitute for a specialised dialogue among researchers and policy makers. As this will not occur spontaneously, an organised effort is required to increase public awareness of research findings.

A Joint Working Group of the Social Sciences and Humanities Research Council (SSHRC) and Statistics Canada was convened by Dr. Ivan Fellegi, the Chief Statistician, and Dr. Marc Renaud, President of the SSHRC. Its mandate is to present a set of proposals to encourage quantitative research on major social and economic issues using large-scale data. The effort of the Group takes place in the context of other initiatives aimed at improving research using social statistics. The federal government recently established the Policy Research Initiative (PRI), with a mandate to recommend and oversee the implementation of an interdepartmental research agenda designed to remedy gaps in the knowledge required for policy development in the federal government. The PRI has partnered with SSHRC to undertake a project on societal trends, with the objective of gaining insight into major social changes, to understand their implications for policy research, and to identify key areas for future research. In addition, SSHRC is developing programmes of strategic research in the areas of social cohesion, the knowledge-based economy and society, and population health. These are also priority research areas for the PRI.

This report recommends the funding of three components designed primarily to increase the number of researchers engaged in quantitative social science research, improve access to micro-data without compromising confidentiality, and inform and build the public constituency for social statistics. Together, these components comprise a Social Statistics Research System which would complement the national statistical system which is already well developed. These components are described separately in the next three sections of this report. The final section sets out a time-line and estimated budget for implementing the proposals, and recommends a process for co-ordinating the inter-related activities that would follow.

Building research capacity

The first component of our proposed Social Statistics Research System aims to increase the number of researchers engaged in quantitative social science research and to improve the skills of those already working in this field. It contains three separate proposals: Research and Training Groups, a Training Programme, including a summer school, and a Fellowships Programme that includes M.A., PhD, Postdoctoral and Senior Research Fellowships. These are described below.

Research and Training Groups

> *We recommend establishing a programme of Research and Training Groups with four interrelated functions: the conduct of quantitative social research; the training of researchers; the broad dissemination of research findings; and the provision of feedback about the usability of data sources. We envision medium-sized research groups, preferably cross-university and cross- disciplinary, working on major research problems, focused on data from a number of different sources. The programme for Research and Training Groups would have an adjudicating committee, including researchers from abroad, that would implement a strategic, multi-stage and proactive adjudication process.*

Central to the development of a strong and diversified body of research is a community of scholars who work together and encourage and criticise each others' work. Research communities that focus on particular research issues are especially critical to graduate students and researchers

beginning their careers, as they enable them to accumulate experience and judgement which is not easily learned from books or through university courses. The SSHRC has experience with a variety of different kinds of organisational forms, and it has found that the most productive environments are those where researchers meet face-to-face and work on a common intellectual theme.

The centrepiece of our initiative, therefore, is a programme to support research and training groups. We envisage middle-sized groups, typically with eight to 15 researchers. This number provides a critical mass while avoiding the co-ordination problems of larger groups. Each Research and Training Group would focus on a topic that is sufficiently broad to merit a co-ordinated programme of research, training and dissemination activities; and sufficiently important to attract a significant number of researchers. Ideally, the Groups would be multi-disciplinary, and entail partnerships among disciplines within a university, and between two or more universities. These Groups would focus their efforts on research pertaining to a particular theme. Annex A of this chapter describes some promising themes where there is already an abundance of available data.

There are already social research groups in Canada, in areas such as ageing, education, immigration and child development, with funding through SSHRC's Major Collaborative Research Initiatives (MCRI) and National Centres of Excellence. The proposed Research and Training Groups, however, would focus on quantitative social science research and place much more emphasis on developing the skills of new researchers and enhancing the skills of an entire research team. Good research teams provide valuable informal learning as a by-product of conversations, meetings and formal seminars, but the Research and Training Groups would be required to address the need for training more explicitly. They would identify training needs and describe concrete steps to address them. They might create and improve formal courses, or prepare background research papers on research methods or aspects of survey design. Efforts to establish apprenticeships that would bridge the gap between formal graduate courses and applied research would be of primary importance.

The Groups will also be expected to develop and carry out plans for the dissemination of their research results, not only to specialised audiences of academics, but also to policy makers and the general public. This would involve more than the episodic flurry of communications activity; it would entail sustained efforts at reaching out, culminating, in several cases, in the formation of partnerships with groups outside academia. Finally, these Research and Training Groups would provide feedback to those collecting data, in particular to Statistics Canada, about the quality of data being collected, its relevance to central research problems and its limitations. We expect that this would improve future data collection efforts and sustain researchers' commitments to this activity.

Research and Training Groups that work in similar or related areas would, over time, become involved in virtual networks – regionally, nationally, and internationally. These networks would allow researchers who were not geographically close to interact with the Groups, or even to become members. Research and Training Groups would be encouraged to pursue their training, dissemination and feedback objectives by putting to use other components of the Canadian Social Statistics Research System that we propose: the summer school and the specialised workshops to be offered by the Training Centre, the research forums, which would offer the opportunity of meeting with a variety of research users, and the communication activities to be developed at SSHRC.

We expect that most of the Research and Training Groups would focus on the analysis of large-scale Statistics Canada surveys. SSHRC and Statistics Canada would encourage research that uses complex longitudinal and multi-level data, and comparative research, especially using multi-country databases. Research and Training Groups might also plan and conduct their own surveys, either free-standing, such as studies of Canadian elections, or surveys that involved Canadian participation in an international social survey. Conducting such surveys outside of Statistics Canada,

or in partnership with it, would increase opportunities for training students how to plan, design and carry out a social survey. This would benefit Statistics Canada and the survey research industry.

Applications for Research and Training Groups would be peer reviewed, with successful applications funded for an initial period of five years, and renewable. Typically, the project budgets would support student research assistants and postdoctoral fellows, computers and software, travel, and in some cases, data collection and archiving. Funding could also be granted to support non-student research professionals; in particular these Groups would likely require professionals capable of preparing and documenting the complex data files required to conduct sophisticated quantitative analysis, of providing ongoing support to graduate students who undertake data analysis, and of solving problems with computers and software.

We recommend that the Research and Training Groups programme have a single separate adjudicating committee, with international members who have expert knowledge of quantitative social science research, including policy-related research. The committee would be charged with developing an adjudication process that was comprehensive and formative; that is, one that encouraged the development of new proposals and links with other groups and the policy community. We feel a separate adjudication committee is necessary as it would be able to understand the special requirements of quantitative social sciences research, and balance the criteria, preferences and objectives to be served by the Research and Training Groups. It would also serve to situate the activity of the Groups within the larger context of the Canadian Social Statistics Research System, and attend to the ongoing development of Groups, especially when in their early stages of their development.

Improving skills: Training Programme and summer school

> *We recommend the establishment of a Training Programme, including a national summer school, covering topics in quantitative methods and the use of Canadian micro-data. The programme would serve policy analysts, university researchers, and graduate and undergraduate students. It would also provide support to data librarians.*

To complement the Research and Training Groups, there is a need for a programme of short courses in social data analysis, similar to the American and British efforts at Ann Arbor and Essex, but more decentralised and adapted to Canadian data and research concerns. The goal is not to offer a substitute for university courses at the graduate and undergraduate levels, but to provide a series of intermediate-level courses in data analysis, and more advanced courses pertaining to the analysis of longitudinal and multilevel data. These courses would stress the application of statistical techniques, and would use Canadian micro-data.

A Training Programme of this nature is required for several reasons. One is that students at the senior undergraduate and graduate levels who take training in statistics are often unable to make the leap from the theoretical training offered in university courses to applying the techniques to real data. In many cases, the examples offered in textbooks pertain to social issues in the United States, and employ small, contrived data sets. Many students need some basic tutoring on how to manage data, such as how to retrieve data from a CD-ROM, merge files, select cases and sort data. They also tend to need instruction on rather simple topics which are seldom covered in their training, such as scaling variables, creating composite variables, handling missing data and using design weights. Some basic courses that used one or two Canadian data sets could generate considerable enthusiasm for pursuing a career in social statistics. Also, when students take such courses alongside other university researchers and policy analysts, they learn first-hand about the kinds of problems that the analysis of micro-data can address. A considerable amount of incidental learning occurs and important working relationships are formed, many of them inter-disciplinary.

Researchers who are at an advanced level also require training in the more complex statistical methods appropriate for analysing longitudinal and multilevel data sets. This is essential if we are to exploit the richness of the recently developed surveys. The audience for these courses would be mainly university researchers, and policy analysts working in government agencies and other organisations. These courses also need to be provided centrally, because there is not usually a critical mass of researchers wishing to pursue such advanced training at any one university.

Finally, because data librarians play an essential role in facilitating access to and use of the existing databases, efforts should be pursued to increase their training. This would contribute to undergraduate education as well, because data librarians can help professors use Canadian data in their courses and assist students in getting started on projects that require some data analysis.

The logical centrepiece of an effort to improve training in quantitative data analysis in Canada is a summer programme, modelled somewhat after the ongoing and successful summer programmes held at Michigan and Essex. The idea is to offer a series of courses in one or more locations, normally over a three- to four-week period. Ideas for a curriculum are provided in Annex B. However, this programme would differ from the UK and US programme in three important respects: *i)* it would focus on Canadian social issues and use Canadian data; *ii)* it would provide a social and intellectual environment for graduate students, faculty members and public sector researchers with an interest in quantitative research on Canadian issues; and *iii)* it would provide a means for follow-up activities, such as establishing an electronic network and the exchange of research papers. It would also give researchers an opportunity to establish contacts with personnel at Statistics Canada who are directly involved in managing particular data bases.

There is no reason to restrict these activities to the summer. The Training Centre should provide a platform for related educational and consultative activities throughout the year. These could include workshops to develop plans for new surveys, seminars to support research on a given topic or with a given data set, and workshops on statistical techniques and survey methods. Departments, foundations, businesses, unions and NGOs could use the facility for consultations and seminars on pressing research questions. Activities of this kind already take place on an *ad hoc* basis, but a Centre with a mandate to facilitate such training would dramatically reduce the effort presently required for researchers to conduct such training programme

We propose to begin with a programme organised by Statistics Canada and using their teaching facilities in Ottawa. Aside from providing a good, ready-made teaching environment, Statistics Canada's ability to operate both in French and in English is extremely attractive. From the beginning, Statistics Canada would work closely with academics through an advisory committee. As the programme becomes established, some of the courses could be taught at various universities. In the longer term, the programme itself could be moved to a university base in one or more universities, while retaining a partnership with Statistics Canada. Also, there should be an effort to reach an accreditation agreement with some universities, as this would likely be helpful for many students.

We recommend that funding be provided to initiate the programme. Funding is required for the salary of a director, office expenses, travel, instructors, and travel and honoraria of the advisory committee. The main source of funding for the Centre would be tuition fees and sponsorships for specific activities, such as workshops on a particular data set. In the longer term, this programme should strive towards becoming self-financing. Attracting graduate students to a summer programme will require SSHRC support for travel, accommodation and tuition. Support of this kind might also be made available to junior faculty members. It would be appropriate to approach MRC and NSERC, who could support graduate students in areas that would benefit from this programme.

Fellowship programmes

> *We recommend that SSHRC establish a targeted programme to support students and young researchers at the M.A., PhD and Postdoctoral levels in social statistics, and a programme of Senior Research Fellowships in Social Statistics.*

To develop a stronger research community in the longer term, we must ensure that young scholars are attracted to pursue a career in social statistics. The field needs a distinct and higher profile, visible to graduate students and post-docs as they decide on the orientation of their research. For this purpose, they should have access to specifically targeted financial support.

M.A., PhD and Postdoctoral Fellowships. The fellowships programme would be distinct from the regular SSHRC fellowships programmes, but its eligibility conditions, regulations and the size of awards would be the same. Applications would be restricted to quantitative social research, though the means of demonstrating this will vary according to the level of the award. M.A. candidates would be expected to take graduate courses that provide them with theoretical and analytic skills, and to write a review paper or thesis on a quantitative topic in social statistics. The application for PhD and Postdoctoral Fellowships would include a proposal to conduct a quantitative research project.

Given the targeted nature of this programme, evaluation of these applications requires the establishment of a separate expert adjudication committee. Also, there should be an explicit commitment to achieving a gender balance. We therefore recommend that all applications for these fellowships be adjudicated directly at the SSHRC.

Senior Fellowships in Social Statistics. The committee recommends that two Senior Fellowships be granted every other year to university scholars who are making a significant contribution to research and training in the area of Canadian social statistics. The Fellowships would be for CAD 100 000 per year for a period of five years, and would require that recipients devote at least 80% of their time to research and training. Renewal would depend on the recipient maintaining a high level of scholarly research. The Fellowships would be prestigious, and would likely increase the momentum of the scholars' academic careers. These Fellowships are modelled after those awarded by the Canadian Institute for Advanced Research (CIAR). The CIAR Fellowships have been highly successful in keeping some of our best scholars in Canada, and enabling them to develop significant research programmes.

We recommend that the adjudication committee be the same as for the other fellowships proposed above. Scholars could be nominated for a Senior Fellowship by their university or they could apply directly to the SSHRC. The application would entail a brief description of how their proposed research programme would contribute to social statistics and to the training of young scholars. Adjudication would be based on the applicants' records of scholarly activity, and their proposed research and training programme

Internships. We have also identified a need for more young researchers to obtain experience in policy research, in view, among other things, of eventually finding employment for their newly developed skills in quantitative social sciences. Although we see the immense potential benefit of a programme of internships, we could not develop precise ideas in this respect in the absence of a more general programme of internships at SSHRC. We can therefore only recommend that the new internship programme proposed in the Council's Innovation Scenario take into account the special need for social statistics expertise identified in this report, and that appropriate resources be provided for it. Moreover, we think that in view of the development of the Social Statistics Research System

proposed here, Statistics Canada should more fully promote its own internship programmes and encourage other government departments to do the same.

Access to data

If Canadians are to fully benefit from the substantial data resources that exist, researchers and analysts must have adequate access to these data. Research is required to convert data into usable information, and adequate access is necessary to facilitate and promote research. Data access has many dimensions. It can mean ensuring that users have adequate knowledge of and access to basic statistics from the data sources. It can also mean access to public-use micro-data files required for more complex analysis, and at yet another level, access to the detailed unscreened (confidential) micro-data. Archiving is also an access issue, as data sources may not be accessible in the future if care is not taken to properly archive them when they are created.

The Data Liberation Initiative (DLI) significantly improved the access to public-use household survey data sources. The Joint Working Group viewed this initiative as extremely positive and recommendations are made in this section regarding ways in which the initiative might be extended. Recommendations are also made regarding support for the archiving of data created by research groups and initiatives outside of Statistics Canada.

The major issue dealt with in this section, however, relates to the access by researchers to data which are confidential under the Statistics Act. The principal issue is how to create such access while ensuring that the respondents' confidentiality is protected as required under the Statistics Act. Confidentiality is the cornerstone of the statistical system. It is also true that in order to fully exploit the new data sources being created, researchers need to have some form of access to the detailed data. A number of alternative means of achieving these goals are discussed here, but the main proposal relates to the creation of Research Data Centres.

Research Data Centres

> *The committee recommends that SSHRC and Statistics Canada jointly create a national system of Research Data Centres, where researchers can access detailed micro-data for research purposes, while maintaining the confidentiality provisions of the Statistics Act.*

Increasingly researchers require access to detailed micro-data to conduct research in many areas, including health, child development, income and economic security, labour adjustment and workplace change. Paradoxically, as the quality and complexity of the data have increased, access to these micro-data may decline as a result of confidentiality requirements. Confidentiality protection is crucial. Not only is there a commitment of confidentiality of responses made to citizens when they provide information, but the security of data is vital to maintaining the public and political confidence required to carry on the broad range of data collection activities effectively.

Since the early 1970s, Statistics Canada has made micro-data (individual records) available while protecting confidentiality through the production of "public-use" micro-data files for household surveys. To prevent the release of records in which the information from individual respondents can be identified, some data must be suppressed: this may involve the omission of some elements of the individuals' records or "collapsing" response categories, so it is not possible to identify respondents with unusual characteristics. But this solution is no longer adequate for two reasons. First, many of the new longitudinal surveys have a data structure that prevents public-use micro-data files from being created at all, for it becomes increasingly difficult to develop a micro-data file that maintains

confidentiality when longitudinal data are available for a number of years. Second, researchers with more complex theories and more powerful statistical tools increasingly find that the suppression of information in traditional public-use files significantly limits their analysis. If the data access issue is not addressed, not only will a number of new and innovative data sources be under-utilised and valuable research lost, but researchers may increasingly turn to data from other countries, particularly from the United States, where the access problem has been more adequately addressed. Findings from American research data are often not applicable to the Canadian situation.

One means of addressing this issue is to have researchers sworn in under the Statistics Act as "deemed employees" of Statistics Canada, allowing them to conduct research using the detailed micro-data files. The Statistics Act provides for such deemed employee status under conditions that are discussed later. Even after being sworn in under the Act, however, there is still the issue of physical access to the data. It is proposed that a number of locations be created across Canada where such access would be possible. These offices would be called Research Data Centres, and would be modelled on a similar programme established by the US Census Bureau. Following is an overview of the Research Data Centre proposal. The details of the proposal can be found in Annex C.

Ultimately, we envision several research data centres distributed across major urban centres and university campuses throughout the country. A centre would provide a secure physical location for the confidential data. Legally it would be an office of Statistics Canada and a Statistics Canada employee would be on site. The centres would operate a research programme administered by SSHRC and Statistics Canada. Researchers who had a proposal accepted under the programme, and were sworn in under the Statistics Act, would have access to confidential micro-data files maintained at the Centres. A committee that operated under the auspices of SSHRC and Statistics Canada would carry out the adjudication of the programme submissions, which would be on the basis of merit. The Selection and Review Committee(s) would consist largely of academic researchers, with some representation from other research communities and Statistics Canada. The research merits of the proposal would be paramount in the selection process (combined with resource issues). Researchers at universities, research institutes, government agencies and other research organisations could submit proposals.

The Statistics Act governs access to confidential data, and its conditions must be incorporated into any programme of Research Data Centres. Researchers may become deemed employees of Statistics Canada and access confidential data providing that they are "perform(ing) special services for the Minister (*i.e.* Statistics Canada)". Since Statistics Canada clearly does not have the human resources necessary to conduct all the research necessary to exploit the data sources, arranging for other researchers to provide such service is one means of achieving this goal. In this environment, the research done would be similar to that which Statistics Canada itself would normally conduct. While this may sound restrictive, in fact Statistics Canada carries out a wide range of quantitative research; therefore, this is not likely to be an issue.

The output of the research programme associated with the Data Centres would consist, in the first instance, of research papers produced by the researchers. The papers would undergo the normal peer-review process, which would be managed by the Selection and Review Committee. It is anticipated that the vast majority would then be included in the Research Paper Series associated with the research programme. Since the research programme would be run jointly by SSHRC and Statistics Canada, these papers could not comment directly on policies or programmes, as that is outside the mandate of Statistics Canada (Annex C provides more detail.) Beyond this first stage, the researchers would be free to publish a revised paper (or the original research paper) in an academic journal or any other publication. There would be no concern in this case as to whether the product contained policy comment; the authors would be free to include any material they deemed necessary. It is proposed that the copyright for the original research paper would be vested with the researchers. Statistics Canada

would retain the right to publish the original research paper and to vet all publications for confidentiality.

As long as the site had a level of security comparable to that maintained by Statistics Canada, a Research Data Centre could be located at a university or a non-governmental research institution. For convenience, however, some Centres could be attached to existing Statistics Canada regional offices. The advantage of non-Statistics Canada locations is that they would allow the institution serving as a home base to develop a very strong empirical research capacity. Empirically-oriented researchers would be attracted to an institution with a Data Centre, allowing the university or institute to develop a strong programme. However, it may be appropriate to begin this programme with a pilot project at a Statistics Canada regional office. It is proposed that Research Data Centres initially include surveys and administrative data files with the exception of those concerning businesses; issues arise in the access to the latter that would need to be addressed before they could be included in the programme.

There are at least two alternative funding approaches for Research Data Centres. The first, modelled on the Data Liberation Initiative (DLI), is for the Centres to be entirely "block funded", with the costs shared by some combination of SSHRC, Statistics Canada, the university or institute at which the centre is located and other organisations, such as government departments and research institutes. If, as we anticipate, primarily academic researchers would use the centres, the majority of the block funding would be SSHRC-based. In this model, there would be no direct cost to the researchers using the Centres. The selection process, conducted by the selection and review committee, rather than applications for funding, would regulate access to the centres. A variant of this approach would be to have part of the annual cost of the Centres covered by block funding, and the remainder covered by fees paid by researchers using the Centre. University-based researchers whose projects were approved by the committee would receive a SSHRC grant, distributed through the review committee to cover this cost; researchers from other organisations would pay similar fees. In any case, care should be taken to facilitate access to data for graduate students and post-docs.

Remote access. In addition to deemed employees accessing confidential data at Research Data Centres, there are at least two complementary potential solutions to the data access problem. One is to develop a remote access capability in Statistics Canada. Researchers would submit their computer jobs to the agency, and they would be run on confidential micro-data files by Statistics Canada staff. The output would be vetted by Statistics Canada staff to ensure that confidential data were not released. This approach is complementary to, and not a substitute for, the Research Data Centres. It would be useful for researchers with smaller projects or who were not physically close to the Research Data Centres. The remote access would not, of course, allow the flexibility and quick interaction required by many research projects that the Research Data Centres would provide. A Statistics Canada committee has been struck to develop a proposal in this area, and an outline of their proposal is provided in Annex D.

Seeking approval from respondents to share the data with Data Centre users. The Statistics Act includes a provision for the sharing of data with users. If the respondent's permission is sought at the time of data collection, confidential micro-data may be shared with selected users. Data sharing would be possible with an Institute created by SSHRC, or some other body. Statistics Canada would then provide the raw micro-data (for respondents who agreed to share, most of whom do) to the Institute, which would of course agree under a contract arrangement to maintain confidentiality.

This approach is quite straightforward. The selection and vetting process would then be entirely in the hands of the Institute; Statistics Canada would play little role other than setting up the conditions to ensure data confidentiality. The shortcoming of this approach is that it can only be applied to data collected in the future, and to some data sets. The data sets that have been created to

date could not fall under such an arrangement. For that reason, this Joint Working Group believes that the approach should be seriously considered for the future, but will not solve the current issues. In the longer run, this approach is a potentially excellent solution to sharing selective data sets, and should be seriously considered once the initiatives in this report have reached some level of maturity.

Archiving and documenting data

> *We recommend that SSHRC allow researchers who created data files before documentation and archiving was an allowable expense in SSHRC grants to apply for funding to conduct such documentation and archiving on the more important older data files. We also recommend that SSHRC and/or DLI consider the creation of a national archiving system for quantitative social science data sets created outside of Statistics Canada.*

Organisations other than Statistics Canada often produce large quantitative research databases. SSHRC-funded initiatives can generate such files. The more important of these data sources should be archived for future use, and the databases should be created and maintained so that they are readily accessible and transportable. If this is not done, researchers other than those initially involved with the data have no access to it, particularly some time after the file was created. There is currently a lack of support for archiving important research databases created some time ago. In SSHRC- funded projects involving the collection of data, the cost of providing documentation and creating files for dissemination purposes has become an allowable expense. The difficulty is with databases assembled without SSHRC support or with only partial support over a period of time and those that pre-date the funding allowances for archiving. Examples of such data files include The Ageing in Manitoba data set, The Canadian Study of Health and Ageing, and The Canadian Fertility Studies. Our understanding is that at this time it is not possible to apply for funding to place the data in a form that can be disseminated. We therefore recommend that SSHRC expand one of its existing programmes or develop a new programme to provide for such funding. Applicants would have to make the case that the data are not presently in a form that allows them to be easily distributed or archived, provide plans for preparation of the data, a description of the final products and demonstrate that the data are of sufficient importance and interest that they will actually be used if archived and disseminated.

The Data Liberation Initiative

> *We recommend that the Data Liberation Initiative be invited to submit proposals with a view to pursuing and extending its mission of making data and research information easily available for academic research and for undergraduate training; this could include the training of data librarians, and the extension of networks of co-operation among them and with undergraduate instructors, as well as the preparation, dissemination and exchange of teaching materials.*

The Data Liberation Initiative (DLI) was launched a few years ago by the Humanities and Social Sciences Federation and Statistics Canada, together with a consortium of post-secondary institutions and federal departments. It has been an immense success, allowing researchers in universities and colleges much easier access to public-use data files. Many more researchers are developing an interest in quantitative data. A new generation of data librarians have been appointed and trained in most institutions, and they have learned to work in close co-operation with one another to increase the availability of both data and the information required to analyse them. Prior to the DLI, data centres existed in very few institutions, and now they are being created in many more. Software has been developed to provide meta-data on available data sources, and to help with their extraction: students

and relatively inexperienced researchers can now easily and rapidly bring quantitative data to bear on their analytical ideas; this nurtures their initial interest and paves the way for further involvement with social statistics.

Throughout this period, many data librarians have also engaged in collaborations with social science instructors who teach undergraduate students the logic, methods, and practice of data analysis. Many of these instructors have incorporated in their teaching much more hands-on experience with data by students. Moreover, they can send their students, graduate as well as undergraduate, to the data librarians to get help in accessing the data and the accompanying documentation.

The Data Liberation Initiative now is under review as it approaches the end of its initial five-year phase. It is a free-standing organisation, with its own Board and administration, and it is not the role of our Joint Working Group to make their case in their stead. This being said, we need to say how important it is that they be supported by Statistics Canada and SSHRC, morally as well as financially. The original mission of disseminating public-use data to academic researchers should of course be preserved; but other roles that have been assumed by the DLI should be recognised, pursued and augmented. First, the training of data librarians, who have become indispensable intermediaries between data sources and users, should be expanded, and the standards of the profession should be sanctioned by employers and employees alike. Also, the networks of co-operation among them should be further extended and strengthened.

Second, the DLI is ideally placed to assume a larger and more systematic role in improving undergraduate education in methods and social statistics, in co-operation with post-secondary institutions, as well as with the Humanities and Social Sciences Federation, Statistics Canada and SSHRC. It can promote exchanges among undergraduate instructors and with data librarians about improved ways of teaching, through written and electronic exchanges as well as in workshops, for instance at the annual Congress of the Social Sciences. It can also improve the availability of teaching materials for these instructors in the form of raw or semi-processed data, documentation about data and concepts, articles and research reports, and suggestions for course assignments.

Third, the work of the DLI, centrally as well as in the local data centres and in networks, may be seen as the precursor for an eventual Canadian data archive. Such archives already exist in most advanced countries and they have proved invaluable research instruments. While the creation of such an archive at the national level is not presently planned, that may change and the work of the DLI in producing better meta-data and documentation is most likely to move us in that direction. Therefore, it should be strongly supported. Indeed, the proximity of data librarians to the users will help to keep the material readily usable by the latter. We can anticipate that university-based data centres will even succeed in preserving and disseminating the precious procedural information developed in the course of using quantitative data in research. In other words, a data archive might grow from the bottom up over the first few years of maturation of the Canadian Social Statistics Research System; SSHRC and Statistics Canada should pay attention to potential developments and be prepared to lend them support.

Communicating research findings

Ultimately, support for an active programme of quantitative social analysis depends on public support and interest. Indeed, supplying relevant research and bringing it to the attention of people who could use it in debate and decision making is key to increasing demand for such research. But at best the process of recruitment and incentives around academic research does little to reward the abilities and temperament required to arouse public interest. With notable exceptions, policy makers find contemporary social science either not directly relevant to their problems or too narrowly focused.

And researchers, in turn, are often reluctant to engage in work that appears unduly oriented towards short-term practicalities. The links between social science research and the media are also relatively weak in spite of the tremendous appetite of the public for social statistics about education, employment, health, social assistance, ageing and other social issues. In both cases, obstacles crystallise, in particular, around the dispersion of the information, the inaccessibility of the language, and differences in operational timetables.

While some progress has been achieved in certain research areas, we need to be more proactive in the promotion of dialogue between social scientists and the potential users of their findings. What is needed is a change in culture, which can only come about after a rather extended period of interaction; and such interaction will in turn require that opportunities be systematically created and exploited for this purpose. This is why we propose the creation of research forums, as well as the organisation of a Social Statistics Communications Programme at SSHRC.

Research forums

> *We propose the establishment of a programme for the support of research forums in order to bring together social researchers based in universities, governments and non-governmental research organisations with a broad range of policy analysts and decision makers. The forums would sponsor conferences and related activities at which academic and government researchers would debate and would benefit from each other's research, data and experience.*

Too often today, academic research on social and economic questions of interest to policy makers in federal, provincial and municipal governments is never seen by those potential audiences. At the same time, academic researchers remain unaware of emerging policy debates to which their expertise can usefully contribute. As a result, a huge pool of expertise for the greater understanding of Canadian social trends and policy issues remains untapped. To bridge these gaps, we propose a programme of research forums.

Our model for research forums is the Canadian Employment Research Forum (CERF), which was formed in 1991 to foster Canadian research and bring together policy researchers in Ottawa and academic labour economists. As recently as ten years ago, empirically-oriented Canadian labour economists worked primarily with US data. Many Canadian faculty members and graduate students knew more about the workings and evaluation of the US welfare and unemployment insurance systems than their Canadian equivalents. At a very modest cost, CERF has effectively and rapidly eliminated this problem and has created a vibrant, interactive community of academic and government researchers. More information about CERF's activities is provided in Annex E.

We believe that CERF's formula can be adopted in other areas of inquiry and applied to a broader range of activities designed to bring together researchers and policy analysts from a wide range of institutions. A successful forum requires a critical mass of researchers in a well-defined but not too specialised area of research. The main activity of the forums would be to organise conferences, with the goals of informing policy makers and policy analysts of research results, providing researchers with a better understanding of policy debates, and more generally offering a setting to facilitate contact among the different sectors of the Canadian Social Statistics Research System. We also see this as an important opportunity for bringing graduate students and beginning researchers into a research milieu. Forums could also provide a convenient and effective vehicle whereby researchers in both the academic and government sector could discuss advantages and limitations of Statistics Canada's surveys for addressing current policy questions, and transmit these insights to Statistics Canada.

We propose that SSHRC conduct peer-reviewed competitions for research forums in quantitative social science. Each forum would be funded for a period of five years and would be renewable as long as the objectives of the programme were met. We envisage the creation of one or two forums per year in the initial five years. For simplicity, these applications could be assessed by the same committee that assesses Research and Training Groups. To qualify as a forum, a group of researchers would form a steering committee of approximately a dozen members who would be co-applicants. In addition to the university-based researchers, the appointment of researchers from government and the non-profit sector to the steering committee should be encouraged. Participation and, where possible, financial support from other organisations, including universities, government departments and the non-governmental organisations concerned with public issues, should be viewed positively in assessing applications.

Applications for new forums should include a plan for an inaugural conference in the first year, plus an indication of its direction for the following years. Of course, existing groups with activities of the kind we have described may apply for support. The funding is mainly required for a forum's key activity – organising conferences – and will include some secretarial and logistic support, travel (both for university faculty and graduate students) and small honoraria for the preparation of conference papers that involve a major departure from a researcher's ongoing research programme. To circulate conference papers and maintain communications, funds should cover the maintenance of a Web site; and some support should also be available for publicising the conferences. Because a successful forum will depend heavily on the largely volunteer efforts of academics in organising conferences, building and running the organisation, some teaching release for Canadian academic members of the steering committee should be available. Travel funding is also required for attendance at meetings of the steering committee.

Social statistics communications programme

> *We recommend that SSHRC establish a social statistics communications programme, specifically aimed at increasing public awareness of quantitative social science research. In co-operation with Research and Training Groups, with research forums and with Senior Fellows, and in consultation with the communication services of universities and of Statistics Canada, this programme should work towards raising the profile of social statistics research finding, and towards furthering debate around the publication of such statistics.*

Social statistics are a privileged area for communications in the social sciences, because of the steady stream of quantitative information about social and economic phenomena being published in the media. Various sections of the public, as well as non-governmental organisations, have also come to expect these numbers and the debates that accompany their announcement. Given this favourable context, social statistics should become a priority in the communications strategy of SSHRC. Their social statistics communications programme should be articulated around two streams of information, one coming out of research in academia, the other from the regular publications of Statistics Canada.

In the first place, university researchers produce a number of research findings that are of great interest to the public and to policy makers, and this will increase as Research and Training Groups come into being. The paradox here is that communications are best managed centrally, or at least from a few key locations where the information can easily be made available to all who could use it, while the research supported by SSHRC's programmes is produced in a multitude of research groups and centres, distributed in a number of institutions of higher learning. The challenge, then, is to gather and

organise the information and to prepare researchers for its effective transmission, while seldom having direct control over the way in which communication does take place.

The communications programme should first monitor the flow of research involving social statistics, using for this purpose connections to Research and Training Groups, to Senior Fellows in Social Statistics and, above all, to research forums. Indeed, the programme should be systematically represented on the Boards of the research forums. Most of these can be expected to have Web pages and there should be a central Web page of these Web pages. The media's attention should also be drawn to interesting and relevant research being produced or presented in these places; a bulletin could be published regularly, including for each item the name of reference persons that can be contacted about the findings. Moreover, the programme should have regular relationships with the communications services of universities, encouraging them to raise the profile of social statistics research. This could be achieved both by having universities signal such research to the programme, and by the latter providing support, when needed, to these university services. Finally, the programme could offer training workshops to researchers, for instance at the Congress of Social Sciences, on how to deal with the media.

In the second place, Statistics Canada regularly publishes a variety of social and economic statistics, thus bringing issues to public attention. When the profile of specific issues has thus been raised, SSHRC could take advantage of the moment to make the ideas of academic social scientists more visible to the public and the media. While the purposes and operations of the communications programme at SSHRC will remain quite distinct from those of the Communications Division at Statistics Canada, some co-operation would be helpful. For instance, given some advance notice about Statistics Canada's publication calendar, the communications programme could present the media with a current list of experts in any given area of research, and help identify potential contributors to debates on social statistics. Given the decentralisation of many aspects of social policy, attention could be paid to identifying experts coming from the various regions of Canada. The programme could itself organise presentations, and debates among experts, on issues where social statistics can be used to throw light on issues. The public for these events would often be the media, but it could also be public servants, politicians or various organisations. Co-operation and a division of labour with the Humanities and Social Sciences Federation would be useful in this regard.

Finally, the programme could, in co-operation with the research forums, organise lecture tours and workshops where research findings based on quantitative social science analysis would be presented to various groups, especially to potential new researchers in social statistics in universities. This activity could be undertaken in co-operation with Statistics Canada, which is already active in this field.

Co-ordination, costs and timetable

Co-ordinating the Canadian Social Statistics Research System

> *In order to co-ordinate the implementation of the programme in this report and to monitor progress, we recommend that SSHRC and Statistics Canada agree to a memorandum of understanding specifying their short- and long-term objectives under this initiative. The two organisations should also jointly appoint a co-ordinator for the overall Research System and an advisory committee of senior officials from academic and other organisations.*

The initiative in quantitative social research we have described involves an ensemble of interrelated programmes. A strong research environment involves a set of mutually reinforcing institutions, which is why we refer to an emergent "Canadian Social Statistics Research System". The ongoing co-ordination of various components could benefit from an explicit agreement on goals and the formal co-ordination of the programme elements. The actors filling the different roles in the system will need more systematic information about each other than informal networks can provide. There are also some overarching issues that do not fall within the mandate of any one of the components of the system, such as efforts to foster intellectual and methodological perspectives that cut across all research areas and activities. We recommend that this goal-setting and co-ordination take the form, first, of a memorandum of understanding between SSHRC and Statistics Canada, and, second, of a position for a co-ordinator of the initiative.

We recommend the appointment of a full-time co-ordinator at the director level. The co-ordinator should have extensive experience in research brokerage and management, as well as the intellectual breadth and research experience required to communicate effectively with a range of social scientists. Part of the co-ordinator's mandate should be to seek out and develop partnerships that will facilitate the overall research effort.

The co-ordinator could convene a liaison group, involving representatives from all the components of the system: the chairs and the administrative officers of the peer-review committees dealing with the Research and Training Group, fellowships and forums programmes; representatives of the Research and Training Groups and research forums themselves; co-ordinators of the summer workshops and Research Data Centres; a representative of the communications programme. The co-ordinator would report to the Chief Statistician and the President of the Social Sciences and Humanities Research Council.

To build and maintain strong linkages to the academic and policy research communities, it is recommended that a senior-level advisory committee be formed. Representatives on such a committee should include senior representatives from the academic community (*e.g.* Vice-presidents for research) and government (*e.g.* Assistant Deputy Ministers for policy), as well as international researchers or research managers from countries that have successfully implemented similar initiatives.

Costs and timetable

The proposed research initiative consists of a number of inter-related components. The Joint Working Group did an initial costing of the various components by considering the costs of similar programmes. We recognised that a much more extensive costing exercise would be required as the proposals are transformed into detailed activity work plans.

We also recognised that the full programme will take a number of years to reach full development, and the assumption has been made that the programme would be phased in over a five-year period beginning in fiscal year 1999-2000. The Working Group feels there is a pressing need to proceed quickly on a number of components, and that work could and should begin in this next fiscal year. The highest priority should be given to the start-up, at least on a pilot basis, of several Research Data Centres. It should also be possible to initiate the training programme by beginning with a few selected courses and to launch a few research forums by holding a number of research conferences on priority topic areas. Work should also begin at SSHRC on a communications strategy with respect to social statistics. It is recognised that major competitive programmes, in particular the Research and Training Groups and Senior Fellowships, will require a longer review process; they should consequently be started as soon as possible.

Annex A

SOME PRIORITY AREAS FOR RESEARCH IN SOCIAL STATISTICS

A wide range of policy issues and research areas could benefit from the strengthening of quantitative research which would put to use the available databases. The role of the Joint Working Group, however, is not to "pick winners" but to establish programmes with fair selection processes. The brief topic descriptions which follow are intended to suggest, but not restrict, the research that this initiative would stimulate.

Child development

In the past ten years, there has been an increased awareness that the quality of children's experiences during the formative years has long-term effects on their happiness and well-being, their future educational and occupational experiences, and their health status as adults. A research agenda on early childhood development in Canada could focus initially on the following questions:

- What is the prevalence of Canadian children who are vulnerable to unduly negative life experiences stemming from poverty, family violence, inadequate parenting, or racial and ethnic prejudice?

- To what extent is childhood vulnerability related to family structure, especially single vs. two-parent families, and socio-economic factors, such as family income and parents' education?

- In what ways do the answers to the first two questions depend on the age of the child and the cohort?

- What are the buffering mechanisms or protective factors associated with healthy child development?

As the policy community, including parents, teachers, administrators and government policy makers, attempts to design a new social policy for Canada, it must figure out ways to strengthen families and communities without dramatically increasing government expenditures. Among practitioners, there is a sense that clinical interventions for all children at-risk are too costly and do not adequately meet the needs of all of them. But are interventions targeted for particular groups a better alternative or would universal programmes likely have a stronger impact? If so, what types of programmes would be most effective? Research on child development could provide a means of monitoring our progress towards reducing childhood vulnerability and redressing inequalities.

To address these questions, the chief source of data is the National Longitudinal Survey of Children and Youth (NLSCY).

Youth in transition

The transition from high school to post-secondary education and from education into the labour market is problematic for many youth. While the research literature identifies the family background factors mainly associated with poor academic and occupational attainment, considerable research is necessary to understand the pathways to success and critical transitions for youth between the ages of 15 and 25. Some of the principal questions that might guide a programme of research include:

- Which groups are particularly prone to leaving school before graduation and have the most difficulty in making the transition to post-secondary education or to the labour market?

- What are the skills, attitudes and behaviour of youth who achieve successful transitions and those who do not?

- To what extent do graduation rates vary among schools within each province? What school-level factors contribute to successful graduation rates and to high academic achievement?

- Do programmes such as co-operative education, mentoring programmes, distance education and apprenticeships help students make these transitions?

At least five national and international data sets can be brought to bear on these questions: the National Longitudinal Survey of Children and Youth (NLSCY), the Programme of International Student Assessment, the planned Canadian Youth in Transition Survey, the International Literacy data, and the Education and Training Surveys.

Families in flux

Over the last three decades, family life has changed profoundly. The increasing proportion of women in the labour force (especially mothers with young children), as well as the growth of flexible and atypical employment have modified the gender division of labour, both within families and within society, and have led to a reorganisation of family time. The rise of separation and divorce, the decline of marriage and the increase in cohabitation have transformed the family trajectories of women, men and children. Researchers have just begun to understand the far-reaching implications of these changes. The questions to be answered include:

- What are the effects of the changing labour environment on the propensity of men and women to both start and maintain conjugal and family relations?

- How are the existing relations among family members (*e.g.* between conjugal partners, between parents and children, between children and grand-parents) modified and redefined once the family separates?

- What are the consequences of family disruption on fatherhood?

- How are resources shared between partners once a union is dissolved?

To address most of these questions, longitudinal data are required. Such data sets include the 1984 Family Survey, the 1990 and 1995 General Social Surveys on the Family, the National Longitudinal Survey on Children and Youth, and the Survey of Labour and Income Dynamics.

Growing old in Canada

In recent years, researchers, policy makers and programme managers have begun to direct attention to the process of ageing and to the status of being old in Canada. Major gaps in our knowledge about growing old in Canada include:

- Good descriptions of the uneven retirement process of Canadians from the age of 50. Canadian policies and research are largely based on the misconception that everyone in the labour force enters at about 20 and retires at age 65.

- Knowledge about the oldest old, aged 85 and over. Important questions concern the health of this group, care giving and receiving, living arrangements and income security of people with limited employment-based pensions (including CPP/RRQ).

- The characteristics of people who will likely enter and are currently in residential institutions. Knowledge about these groups is critical to health, housing and economic security planning and policies.

Numerous data sets exist to address these issues, including the National Population Health Survey, Survey of Labour and Income Dynamics, General Social Survey cycles on related topics, CARNET data, the Longitudinal Administrative Data files, the Canadian Study of Health and Ageing, the Survey of Ageing and Independence, the Health and Activity Limitation Survey, and the Residential Care Survey.

Education, skills and literacy

The development of skills and human capital, along with technological advancement (both "soft" and "hard") are seen as the primary forces driving productivity and hence the standard of living in modern economies. It is difficult to overstate the importance of education, training and skill development for most societies. While this topic has been the focus of major research efforts, rapid changes in the economy and society, and particularly changes in the role and significance of education and skills requires ongoing research. In a "knowledge-based" society, the following issues are currently on the policy agenda:

- The performance of Canadian students in a national and international context.

- The effect of changes in the education system on access to higher education.

- Lifelong learning and its implications.

- The role of literacy, *independent* of educational attainment, in labour market success and daily activities.

- The adequacy of training in Canadian firms.

- The link between human capital and technological change.

- Skills shortages and oversupply.

- The role of human capital in wealth development.

A number of existing data sources cover these issues, including Adult Education and Training Surveys, International Literacy Surveys, Academic Achievement Tests, Graduate Follow-up Surveys, the planned Workplace and Employee Survey, and traditional surveys such as the Labour Force Survey as well as censuses.

The distribution of wages and work

Developed economies today face central policy issues concerning changes in the distribution of wages and of work. In particular, many countries, including Canada, the United States and the United Kingdom, have seen a substantial increase in wage inequality in the last 20 years. Key research questions include:

- What has caused these trends: increased imports from low-wage countries, and the related phenomenon of outsourcing to those countries, or the introduction of new technologies that eliminate the jobs of less-skilled workers?

- Is the collapse of unskilled men's wages a result of the declining influence of unions? Do changes in the quality of our basic education system play a role? Does the relative supply of highly educated workers influence inequality? If more than one of these factors is at work, what is their relative importance and the pattern of their interplay?

- Does a country's institutional structure affect whether it has an unemployment problem (France and Germany) or a wage inequality problem (the United States), or both (Canada)? Does labour market policy respond to research findings, and if so, how?

- What will the role of social support systems be in the future?

Substantial research has been conducted on these topics, but many questions remain. New data sources, such as the Survey of Labour and Income Dynamics and the Workplace and Employee Survey, as well as more traditional sources such as the monthly Labour Force Survey, will shed new light on the role of skill-based technological changes and related questions.

Social and community supports

Over the past decade there has been an increasing recognition of the importance of unpaid work activities, including child-care, household work, care for the elderly and volunteer work. Of course, women continue to do the majority of unpaid work, despite the increase in their paid labour force participation. The federal government's Policy Research Initiative has identified changing time allocation over the life course and within each stage of the life course as underlying many of today's social policy challenges. Care for both children and seniors is also undergoing changes. Ageing of the Canadian population, coupled with shifts in the responsibility for care away from institutions and towards individuals and families, are major challenges. The combination of a declining age of retirement with increased life expectancy may also result in time imbalances at older ages. Questions are increasingly being raised about the erosion of community support or "social capital".

Data addressing these issues include: material in various years of the General Social Survey dealing with time use, and with social support and care giving; the National Survey of Volunteering and Giving; the national Censuses; the National Population Health Survey; the Canadian Study of Health and Ageing and the National Longitudinal Survey of Children and Youth.

Social impacts of science and technology on families/children and on well-being

Science and technology are dominating forces in this century, perhaps even *the* dominating forces. They are often argued to benefit quality of life and children's futures, and yet we know little about the longer-term social impacts of technological change. It is difficult to make broad statements about whether their effects are positive or, more generally, to describe how they work. Also, there are poorly understood distributional issues. Who benefits and who loses; for example, what is the long-term effect of the gap between children with school- and home-based access to computers and the Internet?

Little is known about the social impacts of science and technology on families and children, and on well-being. Current efforts by Statistics Canada to construct a coherent framework for the systematic development of statistical information for science and technology present opportunities for analytical exploitation of existing data and the production of new data vehicles and linkages.

A number of data sets exist, or will soon be available, to address these issues, such as the Innovation Surveys, the Workplace and Employee Survey, the Graduates Surveys, special surveys on Internet use, research capabilities and so on. The development of new data is also necessary to extend research areas.

Evolving workplace and technology use

The 1990s has seen the intersection of a number of technology-related phenomena that affect the workplace and workers, among which:

- Increased use of information technologies, including rise of the Internet and related communications technologies in almost all industries, accompanied by a concern about whether this has led to the desired productivity gains.

- Increased focus on the importance of innovation for firm survival and growth, and for productivity gains.

- The effect of technology on downsizing.

- Concern that technology may be leading to increasing polarisation in society.

- A focus on human resources issues such as training, pay practices, work schedules and new workplace practices implemented to achieve "high performance" workplaces.

Relatively little is known about the adoption and diffusion of technology and innovation in work organisations and its effects on the workplace and workers. While researchers have long been concerned with these issues, few large-scale data sources have existed to document the rate of implementation of technology and innovation, let alone its implications. This has led to the use of often-questionable proxies for technology use, or case studies.

More recently new bodies of data have evolved to assist in our understanding of issues in this area and to provide new research opportunities. Such surveys include technology surveys, innovation surveys, the Workplace and Employee survey and a survey of the determinants of firm growth.

Welfare, income and poverty

Research on welfare and material inequality addresses three main questions. First, what is the distribution among individuals, families, household units and "communities" of income, wealth and the necessities of shelter and food? Second, what are the personal consequences of this inequality on the quality of social life, generally and on specific issues such as health? The third question is what mechanisms reproduce, or alter, inequality over time, both over the lifetime of individuals and between generations? At the aggregate level, this question can be reframed in terms of the evolution of geographical and other forms of communities through time?

Studies of welfare, income and poverty are necessary to understand the effects of a very wide range of policies involving huge expenditures including: the redistributive effects of taxation, the efficacy of social welfare programmes, education and health programmes, and arrangements for the delivery of social services and health care.

Basic profiles of economic and social inequality have been available for several decades. But as policies and programmes change, new research is required to assess their impacts. The new longitudinal surveys provide a hitherto unavailable way to examine closely the process of change over time. The impact on the welfare of Canadians of changes such as the loss of a job or the dissolution of a marital union can be addressed. The short- and long-term impact of persisting conditions of severe deprivation on the well-being of children and young people, and on their physical and mental health can be assessed.

Data available to address these issues include the Survey of Consumer Finances, the Survey of Labour and Income Dynamics, the National Longitudinal Survey of Children and Youth, the National Population Health Survey, the Family Expenditures Surveys, and censuses.

Annex B

CURRICULUM FOR A SUMMER SCHOOL PROGRAMME

While it may take several years to reach a steady state, it is useful to think about the curriculum of a more mature summer programme, distributed over four weeks. For data analysis, the core of the programme should be two courses, each lasting two weeks. The first would provide a basic introduction to regression, designed for participants with relatively little background, with data analysis examples from cross-sectional surveys such as the General Social Survey. A second two-week course in linear models would extend regression techniques to censored, categorical, "count" and duration data. The two courses should combine lectures in the morning with mentored practice sessions in the afternoons. In this and other courses, there should be an emphasis, and specific instruction, on how to write about the results of data analysis.

In addition to the two core data analysis courses, on a revolving basis and according to the demand, each summer programme should include one or more intermediate-level courses on topics such as longitudinal analysis, hierarchical models, latent variable models and categorical data. The emphasis would be on building a corps of researchers equipped with modern data analysis techniques to conduct state-of-the-art research. It will often be appropriate to focus on a single survey, especially for longitudinal analysis, where considerable effort is required to gain familiarity with a data set. There should be a mixture of one- and two-week courses, depending on the topic. Some shorter workshops could address, on a rotating basis, more limited analysis topics such as weighting and estimation issues for complex samples, the analysis of pooled cross-sections, and robust statistical methods.

A third kind of methodological course, definitely worth trying, was proposed to the Working Group by two statisticians (one at Statistics Canada and the other university-based). The teaching groups would focus directly on a data set, chosen on the basis of teachers' and students' interests. Students would work as a team, led by a researcher and including a programmer, a survey methodologist, a specialist on the survey topic and a statistician. The course would revolve around using the chosen data to address a specific research problem, with design, statistical and substantive issues addressed as they arose.

Consideration should be given to staging a workshop, perhaps a week in length, on the philosophy, logic and strategy of data analysis. In recent years, there has been more interest in trying to make explicit, and consequently debatable, the thinking underlying the quantitative analysis of social data. For example, a considerable debate over the relative roles of description and causal interpretation of social phenomena was sparked by the publication of Stanley Lieberson's "Making It Count". Such a course might combine an introduction to the problem of what and how much particular data and data analysis can tell us about social processes, with a discussion of strategies for analysis, focusing particularly on complex survey data sets and on how to do analysis that does justice to a complex topic, but is still feasible.

In addition to the applied statistics courses, each summer session should include at least one course geared to developing analysis of a single survey, such as Survey of Labour and Income

Dynamics, the General Social Survey on a topic, the National Population Health Survey, or the National Longitudinal Survey of Children and Youth. These could be open to researchers prepared to make a commitment to conduct a piece of research on the relevant data set and should be led by two or more researchers working with the data set; they would also involve an explicit mentorship arrangement for junior researchers. More senior researchers would take the course mainly to familiarise themselves with a data set and might collaborate in mentoring. According to the topic, the data-familiarisation should be paired with a statistical workshop dealing with methods required to analyse the data. Where appropriate there might be follow-up activities, including a strategy for providing ongoing methodological advice, plans for reviews of manuscripts and, if the resources can be found, a later meeting to discuss the research results. Depending on the topic and the extent to which the course emphasised training, a course of this kind could last for one or for two weeks.

Each summer session should include at least one course, perhaps a short seminar, oriented around a substantive research topic. The idea would be to gather researchers to review work in an area and discuss research priorities. This should be coupled with a broad examination of the available data sources relevant to the topic. Hopefully, one would bring together a mix of junior and more senior researchers to stimulate research in the chosen area and the organisers should develop a strategy to encourage ongoing communication among the researchers.

On a revolving basis, the summer programme should offer courses on a variety of topics. An attractive idea would be a seminar on policy research with quantitative databases. Researchers from government departments, as well as policy organisations, could be brought in to describe their own work and current policy and research concerns in their areas. A short course on data concepts should also be considered. The idea would be to focus on the way that Statistics Canada, and survey researchers in general, conceptualise and measure the parameters of key social variables. Some obvious candidates include race and ethnicity, labour force participation, work experience and perceptions of the quality of the social environment.

Because there appears to be no Canadian graduate programme with advanced training in survey design, consideration should be given to including one in the summer programme. Training in data analysis should produce researchers who are aware of the limitations of their data, but is not a substitute for knowledge of the extensive literature on survey design. Ideally, a course would begin with a review of research on questionnaire design and strategies for pre-testing, evaluating and revising new surveys, then turn into a workshop where the participants would write, test and revise a new questionnaire. The second stage of the course would ideally be conducted in co-operation with Statistics Canada or a university-based survey research organisation.

Annex C

RESEARCH DATA CENTRES

To facilitate access to confidential micro-data for research purposes, researchers would become "deemed employees" of Statistics Canada. There are numerous legal and organisational issues that must be addressed under this scenario, and this appendix outlines one approach. One issue is physical proximity to the data. Access is difficult for many if the data are available only in Ottawa. Hence, it is proposed that this approach be developed within a framework of Research Data Centres.

What is a Research Data Centre?

A Research Data Centre would be a physical location that would have a secure environment capable of protecting confidential micro-data files, and would be an extension of Statistics Canada. It would have an affiliated research programme administered by SSHRC and Statistics Canada. Researchers sworn in under the Statistics Act would have access to confidential micro-data files maintained at the centres, thus allowing research to proceed. Researchers would be held accountable for the protection of confidentiality in exactly the same way as Statistics Canada employees are currently held accountable.

The major issues associated with developing a research data centre programme

The protection of confidentiality. This is of central importance, since confidentiality is the cornerstone of the statistical system. Any access to data by researchers who are not regular employees of Statistics Canada must be done in accordance with the Statistics Act.

Ensuring that the work and the researchers fall under the Statistics Act. In order to ensure confidentiality, the researchers must be sworn in under the Statistics Act, and be subject to the conditions of the Act in the same manner as a regular Statistics Canada employee. The Act lays out the conditions under which people can be given access to confidential data, and hence these conditions must be incorporated into any programme of the Research Data Centres.

Ensuring that the Centres will succeed in their objective. The Centres must be structured in such a way as to attract top-quality researchers and their students to ensure that the goals outlined earlier in the report are met.

Ensuring an appropriate organisational structure. There are issues regarding the manner in which such Data Centres are organised and managed and the composition of the approval and evaluation committee for the research proposals. It is proposed that SSHRC and Statistics Canada jointly administer the affiliated research programme, and that Statistics Canada manages the facility, as it is an extension of that agency. Operationally, the selection and vetting for the research

programme would be conducted by committees consisting of prominent researchers, in compliance with the requirements of the Statistics Act and confidentiality.

Ensuring appropriate physical safeguards for the data. To both meet confidentiality requirements and to openly demonstrate that they are being met, adequate physical security obviously must be in place.

These issues are addressed in the following outline.

The goals of the research data centre programme

The programme would have two goals:

- To promote quantitative research by academics, researchers in government agencies, research institutes or elsewhere in the public sector using Statistics Canada micro-data files, particularly household micro-data files.

- To improve the statistical programmes of Statistics Canada through feedback from researchers using the micro-data, and through the research and papers produced by the researchers.

Elements of the research data centre programme

The following are the essential elements of a research data centre programme:

- A secure, but user-friendly computer environment in which confidential micro-data could be stored would be required. Security would be at the standard maintained by Statistics Canada.

- A selection, approval and vetting process for the research based on confidential micro-data. A committee consisting of senior researchers (from academia and elsewhere) in a number of fields, as well as SSHRC and Statistics Canada officials would be created to administer the selection and vetting process. By and large, senior academics would determine which projects would be approved. There would be one peer-review panel for each major area or research, such as economics, sociology, health, education, and statistical methodology.[1] Academics or researchers at research institutes, government agencies or other research organisations could submit proposals.

- There would be at least one Statistics Canada employee on-site at the Data Centre, to manage the site, provide support and to oversee confidentiality issues.

- Only persons sworn in under the Statistics Act through the Research Data Centre Programme would have access to data at the Centres. Researchers would sign a contract that dealt with their obligations under the Act.[2]

- The Centres would have to be self-financing, with funding coming from the institutions running the Centre, granting councils such as SSHRC, or through the researchers.

Operating the Centres under the Statistics Act

Researchers with access to confidential data must be sworn in under the Statistics Act. Aside from regular Statistics Canada employees, the Statistics Act restricts access to confidential data to people who are "retained under contract to perform special services for the Minister (*i.e.* Statistics Canada)". These people become deemed employees of Statistics Canada. That agency retains researchers to conduct work when it does not itself have the resources to do all the work required. Statistics Canada clearly does not have sufficient resources to produce the analytical work needed to exploit the many new (and older) data sets. Such exploitation is needed to provide analytical insights of value in public policy development and debate, and in the promotion of basic research. Much of the capacity to take advantage of the rich data resources resides in the academic community and in other research organisations (in other federal agencies, think tanks, provincial governments, etc.).

In this environment, and under the current Act, in order to be sworn in under the Act the research done would be similar to that which Statistics Canada itself would normally conduct (if it had the resources). While this may sound restrictive, in fact Statistics Canada carries out a wide range of research; consequently, this is not likely to be an issue. In the selection process, the merits of the proposals would be paramount. This selection and review committee would oversee the review process after the completion of the project. Statistics Canada would obviously have some input to the process, since under the Statistics Act the work constitutes special services for the Minister.

Seeking approval from respondents to share the data with Data Centre users. There is an alternative means by which confidential data can be shared with users who are not covered by the Statistics Act. The Act includes a provision for the sharing of data with users. If the respondent's permission is sought at the time of data collection, confidential micro-data may be shared with selected users. Data sharing would be possible with SSHRC or an "Institute" created by that agency. Respondents would be informed about the nature of the institute (or SSHRC), and permission sought to share the data. Based on past experience, it is likely that most respondents would agree. Under such a scenario, Statistics Canada would then provide the raw micro-data to the Institute (or SSHRC), which would of course agree under a contract arrangement to maintain confidentiality.

This approach has the advantage of being straightforward. The selection and vetting process would then be entirely in the hands of the Institute; Statistics Canada would play little role other than setting up the conditions to ensure data confidentiality. The shortcoming of this approach is that it can only be applied to data collected in the future, and to some data sets. The data sets that have been created to date could not fall under such an arrangement. For that reason, this approach should be seriously considered for the future, but will not solve the current issues. Thus, there is the possibility of forming an "Institute" with which data sharing could take place. The propensity of respondents to share the confidential data with such an institute (or SSHRC) for research purposes could be tested. While this approach would have little benefit in the short run, in the longer run it is a potentially excellent solution to sharing selective data sets.

Handling the research paper upon completion of the project. Researchers may want to think in terms of two papers following completion of the project. A research report consisting of the quantitative analysis and an interpretation of the findings would be deposited with the Selection and Review Committee of the Data Research Centres. This product would become part of a research paper series,[3] and would be reviewed in a normal academic peer-review manner. The Committee (or its named designate) would run this process. The product would also undergo an "institutional" review by Statistics Canada. This is simply to determine if the work falls within the mandate of the agency for the purposes of the research paper series. Work conducted for the research paper series could be (and often should be) policy relevant but could not contain direct policy recommendations. Hence, the

institutional vetting by Statistics Canada is conducted primarily to ensure that there are no direct policy recommendations in the paper.[4] After the academic and institutional reviews, authors will make the necessary changes.

The researchers would, of course, be free to publish the research paper or an extended version of the research paper with policy comment and other additions they see fit, in any academic journal or any other venue. In short, beyond the initial quantitatively-oriented paper that is submitted to the research paper series, the researcher would be free to produce any other version of the paper and submit it for publication in any forum. In the event that the review of the paper by the Selection/Evaluation Committee and Statistics Canada leads to a decision not to publish the paper in their research paper series, the author will be able to submit it for publication elsewhere.[5]

Where the Centres might exist

The Centres may be affiliated with a university, research institute or research network. They could also be located at Statistics Canada regional offices, and for reasons outlined later, it is proposed that regional offices be employed initially. Obviously a physical location with a secure environment is needed. If located outside of Statistics Canada, competitions could be held to determine where such Centres might be placed. One would want to start with a very small number, as there would be a substantial resource impact on Statistics Canada through the servicing of such Centres. It seems likely that having such a Centre would allow the institute serving as a home base to develop a very strong empirical research capacity in potentially a number of disciplines. Very qualified empirically oriented researchers would be attracted to an institute with a Data Centre, allowing the university or institute to develop a strong programme.[6]

Starting small

This is an ambitious programme. A pilot approach may be appropriate. It is proposed that the data sets made available to the Centres initially include the new longitudinal surveys: the Survey of Labour and Income Dynamics (and the Labour Market Activity Survey, its predecessor), the National Longitudinal Survey of Children and Youth, the National Population Health Survey, the Workplace and Employee Survey, and related social statistics data sources required to conduct the research. It is here that the need for direct access to the micro-data is the greatest. It is also proposed that regional offices be the initial sites for the pilot project. This would allow the development of the Centres to proceed incrementally. Nonetheless, access to the confidential micro-data would be improved tremendously, and some fairness regarding the geographical location of the centres would be introduced, as regional offices are located across the country. If housed in Statistics Canada, the initial Centres may be less costly to run, as some of the infrastructure already exists. After the pilot project has been in place for some time (perhaps after two years), it is proposed that the programme be reviewed, and subject to this review, a competition for Research Data Centres at Universities or other research organisations be held.

The funding

Similar centres operated by the US Census Bureau and the US National Science Foundation have an annual budget of USD 250 000 per location. There are at least two alternative funding approaches. The first is similar to the way in which the Data Liberation Initiative (DLI) is financed. Under this scenario, the Centres would be entirely block funded. Such block funding would be provided by

SSHRC, the university or institute at which the Centre is located (after the pilot phase), and other organisations whose researchers use the Centre, such as government departments, research institutes, etc. Since it anticipated that primarily academic researchers would use the Centres, the majority of the block fund would be SSHRC-based. There would be no direct cost to the researchers using the Centres. The selection process run by the selection and review committee would regulate access to the centres. A variant of this approach would be to have much of the annual cost covered by block funding, and the remainder covered by fees paid be researchers using the Centre. University-based researchers whose projects were approved by the committee would receive a SSHRC grant (distributed through the review committee) to cover this cost. Researchers from other organisations would pay similar fees. It is proposed that one of these funding approaches be implemented.

Conclusion

This proposal urges the consideration of Research Data Centres as one solution to the issue of access to micro-data for researchers. Initially this would involve access to household survey data, particularly the new longitudinal household surveys. Business (establishment or company) surveys present unique confidentiality issues that may prevent them from being included in this endeavour, at least initially. The Centres could be supplemented by a remote access capability developed within Statistics Canada. The latter would be useful for researchers who are not geographically capable of using a Data Centre, or who for other reasons cannot or wish not to use the Data Centres. Smaller projects, for example, may be better dealt with through remote access.

Annex D

THE DEVELOPMENT OF A REMOTE ACCESS CAPABILITY AT STATISTICS CANADA

A remote access capability should provide researchers with tools allowing them to specify statistical procedures and to have these procedures applied to confidential data by Statistics Canada staff. The results would be screened by Statistics Canada for confidentiality and returned to the researchers.

Among the tools available to the researcher would be the public-use micro-data file (where available) and a permuted file which closely mirrors the confidential file in structure and detail, but with only a limited amount of real data. The researchers would specify their program or desired run as an SPSS/SAS job, using the permuted file, and then submit it to Statistics Canada via the Internet. The permuted files should contain sufficient detail to provide a basis for the testing and debugging of programs. Statistics Canada would not play a role in such debugging; program errors would be returned to the researcher. The main task of Statistics Canada employees would be to run the jobs against the master file and vet the output for confidentiality. Programmes that led to confidentiality problems would have to be modified by the researcher

The main advantages of remote access are its availability to researchers regardless of their physical location, and the simplification of the interface between them and Statistics Canada. Its disadvantages are its costs (when compared to the direct use of publicly available data files) and the possible delays incurred as the results are vetted. An informal working group has been set up within Statistics Canada to develop an implementation strategy that would minimise these disadvantages. For the moment, it goes as follows: the files would be produced by the subject area responsible for the survey, with the assistance of methodology staff skilled in the generation of such files. They would be distributed to researchers via the Data Liberation Initiative (DLI)/FTP site; DLI contacts at universities could help to facilitate access to the files as they do with the current DLI files. Statistics Canada would establish a central group, building on the present DLI team, to run the programs and screen the outputs for confidentiality. The results would be returned to the researchers using FTP.

The main challenges about this approach have to do with the range of permuted files that can be produced and made available, and with the effectiveness with which this can be done, with the effort required to screen the output for confidentiality, and with the turn-around time. Many of these questions will be answered only after a suitable test of the strategy. The latter has been implemented for the National Population Health Survey (NPHS) and it is anticipated that there will be four files available for testing in a remote access environment next year. This should provide researchers and Statistics Canada with enough information to decide how remote access can be used, in conjunction with other approaches, to improve access to data.

Annex E

A MODEL FOR RESEARCH FORUMS IN SOCIAL STATISTICS: THE CANADIAN EMPLOYMENT RESEARCH FORUM (CERF)

CERF is a non-profit corporation, whose primary goal is to increase productive interaction between researchers studying employment issues in the academic and government sectors, as well as policy makers themselves, and to increase both the volume and quality of policy-relevant research in this area. CERF is directed by a rotating volunteer board, at least one-third of the directors come from the academic community and from the government service. CERF receives core funding from Human Resources Development Canada, but it also raises considerable additional funds for its conferences from granting agencies, foundations and stake-holding government departments.

In the last eight years, CERF has organised a series of conferences (close to twenty, so far) on a wide variety of research topics. As a result, CERF has dramatically improved the access of policy makers to the technical expertise of academic researchers, at very low cost. CERF also sparked a dramatic increase in research using Canadian data. This was no accident: CERF conference organisers encouraged Canadian research, though they recognised the benefit of comparisons with the United States and other nations. Through contacts in government, CERF also facilitated access to confidential data. A list of conferences organised by CERF follows.

- Founding CERF Conference. Kingston, Ontario, 31 May 1991.
- Policy Research in Training, Unemployment and Income Support, and Immigration. Aylmer, Quebec, 6-7 March 1992.
- Labour Markets in the Last Two Recessions: A Comparative Perspective (with Statistics Canada). Ottawa, Ontario, 5 March 1993.
- Income Support Programmes and Policies (with Health and Welfare Canada). Ottawa, Ontario, 24 September 1993.
- Immigration and the Labour Market. Hull, Quebec, 11 March 1994.
- Labour Markets and Income Support: Focus on British Columbia (with B.C. Ministries of Social Services and Skills Training). Vancouver, B.C., 25 March 1994.
- Workshops in conjunction with the Learned Societies Annual Conference. Calgary, Alberta, 13 June 1994.
- Youth Labour Market Adjustment (with BC Ministry of Education and UBC), Vancouver, B.C., 25 June 1994.

- International Conference on the Evaluation of Unemployment Insurance. Ottawa, Ontario, 14-15 October 1994.

- Policy Responses to Displaced Workers (with the Université du Québec, Montréal and Ekos Research Associates). Montreal, Quebec, 2-3 December 1994.

- Retooling the Employed Workforce: Focus on New Brunswick (with HRDC, NB Region and UNB). Fredericton, N.B., 31 March-1 April 1995.

- Environmental Policies and Labour Markets (with HRDC Innovations). Ottawa, Ontario, 15-16 September 1995.

- The Canada - US Unemployment Rate Gap (with Centre for the Study of Living Standards). Ottawa, Ontario, 9-10 February 1996.

- Changes in Working Time in Canada and the United States (with Upjohn Institute, Statistics Canada, HRDC). Ottawa, Ontario, 13-15 June 1996.

- Social Experiments, Evaluation and Employment and Social Policy. 1996.

- Immigration, the Labour Market, and the Economy. Richmond, BC, 17-18 October 1997.

- Participation Rates and Employment Ratios. Ottawa, Ontario, 17 December 1997.

- Labour Market Transitions and Income Dynamics (co-sponsored by Statistics Canada and HRDC). In conjunction with the CEA meetings in Ottawa, Ontario, May 28-29, 1998.

- Self Employment Conference (with the Canadian International Labour Network and the OECD). Burlington, Ontario, 24-26 September 1998.

More information on CERF is available at its Web site: http://cerf.mcmaster.ca

Annex F

MEMBERSHIP OF THE JOINT WORKING GROUP

Paul Bernard, Chair
Département de sociologie
Université de Montréal

Betty Havens
Community Health Sciences
University of Manitoba

Peter Kuhn
Department of Economics
McMaster University

Céline Le Bourdais
INRS – Urbanisation

Douglas A. Norris, Director
Housing, Family and Social Statistics Division
Statistics Canada

Michael Ornstein
Institute for Social Research
York University

Garnett Picot, Director
Business and Labour Market Analysis
Statistics Canada

J. Douglas Willms
Atlantic Centre for Policy Research in Education
University of New Brunswick

Martin Wilk, Former Chief Statistician
Statistics Canada

Assisted by Hélène Régnier, Policy Analyst, SSHRC

NOTES

1. Preference might be given to empirical research that has policy relevance. To allow the system to be responsive to rapidly emerging policy questions, a two-tier approval process should be considered. One tier would be designed to render quick decisions for short-term feasibility studies with only minimal funding, and another directed at longer-term projects with more substantial resource implications.

2. In order to promote the training of a new generation of researchers with expertise in the use of these data sets, and in order to facilitate the research itself, this should include research assistants, especially graduate students, employed by the academic researcher or under his/her supervision in dissertation research.

3. The research paper series developed for the publication of the initial quantitative analysis would likely be best maintained in one central location. Managing the research paper series centrally ensures that there is some consistency in the way the papers are handled. There may be resource considerations for Statistics Canada and the Committee. Vetting the research papers to ensure that the work is properly conducted may at times be time consuming. Such vetting will be largely done through the academic refereeing process. At times, however, work may be required to validate the use of the data. It is important that the person responsible for maintaining the series understands research and the research world, and has the resources necessary to prevent a backlog of papers.

4. Direct policy recommendations or comments refer to direct evaluation, criticism or advocacy of existing or proposed government programmes. This does not exclude research on policy relevant topics, which is in fact encouraged. These restrictions are placed on Statistics Canada output (and the joint research paper series) to protect its neutrality and objectivity.

5. Normally the copyright on work that has been done under contract as a special service for the Minister (as required by the Act) would be vested with Statistics Canada. However, in the contract struck before the work begins, it is proposed that Statistics Canada agree to vest the copyright with the researcher. Statistics Canada would retain the right to vet all publications stemming from the project for confidentiality and data reliability. In the same contract Statistics Canada would retain the right to reproduce the paper if it chose to do so.

6. Certainly the NBER (National Bureau of Economic Research) in the United States, one of the first institutes to have a Census Bureau Data Centre, has developed a very strong empirically based research programme, presumably in part due to the superior data access available to researchers affiliated with the NBER.

Chapter 8

NEW HORIZONS FOR THE SOCIAL SCIENCES: GEOGRAPHIC INFORMATION SYSTEMS

by

Michael F. Goodchild*
National Center for Geographic Information and Analysis, and
Department of Geography, University of California, Santa Barbara

Introduction

Geographic information systems (GIS) are not exactly new to the social sciences – the United Kingdom's Economic and Social Research Council funded a network of GIS-based laboratories at UK universities in the late 1980s and early 1990s, and the US National Science Foundation's Directorate for Social, Behavioral, and Economic Sciences funded the National Centre for Geographic Information and Analysis to promote GIS-based research from 1988 to 1996 (both GIS organisations continue to exist). But there are nevertheless good reasons to include GIS in a discussion of the future of the social sciences. The use of GIS has spread very widely among the sciences, and it is now an accepted tool among all of the disciplines that deal with the surface of the Earth and its human population. Moreover, the concept of GIS has evolved substantially, and I propose in this chapter to take a deliberately broad view of the term's meaning. GIS also claims to be an *integrating* technology, spanning disciplines and blurring the distinctions between them, both important prerequisites for any broadly useful research infrastructure. Finally, the use of GIS has prompted interest in a number of fundamental issues that are collectively identified as geographic information science.

The chapter is organised as follows. The next section explores the nature and history of GIS, and the contemporary meaning of the term. It includes what I hope is an honest assessment of the technology's strengths and weaknesses. The third section includes a personal selection of the key concepts and principles of GIS. This is followed by a brief review of geographic information science. The final major section discusses the concept of Digital Earth, and its possible value as a motivating force. The chapter closes with three final points.

---

* National Center for Geographic Information and Analysis, and Department of Geography, University of California, Santa Barbara, CA 93106-4060, United States.
Tel: +1 805 893 8049. Fax: +1 805 893 3146. E-mail: good@ncgia.ucsb.edu.
A shorter version of this chapter is published as Michael F. Goodchild, "New Horizons for the Social Sciences: Geographic Information Systems", *Canadian Journal of Policy Research/Revue canadienne de recherche sur les politiques*, ISUMA, Vol. 1, No. 1, Spring 2000, pp. 158-161.

The nature of GIS

Overview

It is very appropriate that this conference is being held in Ottawa, since the city has the strongest claim to be the home of GIS. In the early 1960s, the federal and provincial governments had funded the Canada Land Inventory, a massive effort to assess the current and potential uses of Canadian land in a belt extending north from the US border well beyond the areas of widespread settlement. The objectives of the project required detailed analysis to determine the areas in use or available for such activities as forestry, agriculture or recreation. Maps of different themes were to be overlaid to determine correlations and conflicts, but in manual form both area measurement and overlay are highly labour-intensive and crude operations. Roger Tomlinson was able to show that computerisation was cost-effective, even in the environment of the mid-1960s, with its primitive and expensive computers and no tools for meeting the special requirements of handling map data. The Canada Geographic Information System (CGIS) was born out of a simple analysis of the relative costs of processing geographic data by hand and by computer.

The subsequent history of GIS has been described in detail (Foresman, 1998). Much important work in the late 1960s and 1970s was conducted at the Harvard Laboratory for Computer Graphics and Spatial Analysis, under the direction of Howard Fisher, William Warntz and Brian Berry. Many important roots lie in landscape architecture and in the computer-assisted design (CAD) systems developed at Cambridge and elsewhere. But the beginnings of the modern era of GIS, with its widely available commercial software products, dates to the early 1980s and the dramatic reductions in hardware costs that began then and continue today.

Today GIS is a major computer application with uses that range from the management of natural resources by government agencies and corporations, to the operations of utility companies, to support for scientific research and education. The software market is currently dominated by Environmental Systems Research Institute (ESRI), of Redlands, CA, and particularly by its ARC/INFO, ArcView, and SDE products. ESRI's annual sales are in the region of USD 250 million, its users number in the hundreds of thousands, and close to 10 000 people attended its most recent user conference in July 1999. Other significant GIS vendors include Autodesk, MapInfo, Smallworld and Intergraph.

More broadly, GIS is part of a complex of geographic information technologies that includes remote sensing, the Global Positioning System, and geographic information services offered on the World Wide Web (WWW). In a 1993 study, the US Office of Management and Budget estimated that federal expenditures on digital geographic information amounted to USD 4.5 billion annually, and figures of USD 10 billion to USD 20 billion for annual global expenditures seem reasonable. The term "GIS" is used increasingly to encompass all of these, and phrases such as "doing GIS", "GIS data", the "GIS community" suggest a willingness to see "GIS" as a shorthand for anything that is both digital and geographic in nature. Longley *et al.*, (1999) provide a recent review of all aspects of GIS.

Definition

GIS is defined most generally as technology for processing a specific class of information - geographic information. *Processing* is understood to encompass creation, acquisition, storage, editing, transformation, analysis, visualisation, sharing and any other functions amenable to execution in a digital domain. *Geographic information* is readily defined as information linking locations on the Earth's surface with specific properties, such as name, feature, population, elevation, temperature. More generally and precisely, it consists of atoms of information or *tuples* of the form (location, time,

property. To be communicable, a scientist would argue that all three components must be well defined, using terms that are known to both sender and receiver of information. In the case of location, this argument clearly favours general standards such as latitude and longitude over more problematic specifications of location such as place names. But there are strong arguments for including information in GIS that is poorly defined, vague or subjective, because of the importance of these forms to human communication, and there has been much interest recently within the research community in the problems of handling vague geographic information.

This definition of geographic information is deceptively simple. Unfortunately, the geographic world is continuous and infinitely complex, and there are therefore an infinite number of locations in space and time to be described. In practice, geographic information must somehow approximate, generalise or simplify the world so that it can be described in a finite number of tuples. There are an unlimited number of ways of doing this, in other words an unlimited number of ways of mapping the real geographic world into the contents of a GIS database, which requires an alphabet comprised of only two symbols, 0 and 1. Many such mappings or *representations* have been implemented in various disciplines and areas of application, and many are implemented in the standard GIS software products as *data models*.

Data models fall into two broad but imperfectly defined categories – *raster* and *vector*. In a raster representation, the world is divided into an array of cells of fixed size (note that some distortion is implied, since the curved surface of the Earth cannot be covered by uniformly sized, non-overlapping square cells). All properties of the surface are expressed as uniform properties of the cells, and all sub-cell information is lost. Moreover, rasters are not convenient for capturing geometric structures larger than the cell, since it is generally difficult to link cells together. In a vector representation, properties are associated with geometric points, lines or areas, and the locations and shapes of these objects are defined by co-ordinates. Areas are approximated as polygonal figures by connecting points with straight lines, and curved lines are similarly approximated as *polylines*. Vector representations readily accommodate variable spatial resolution, links between objects and complex geometric structures, and are strongly favoured in applications of GIS to social phenomena.

Large and comprehensive software environments such as GIS are possible because of strong economies of scale in software production. Once a basic framework has been built, by implementing a limited number of basic data models with associated tools for creating, editing, visualising and sharing data, additional functions can be added very easily and cheaply. This principle is clearly evident in spreadsheets, word processors, statistical packages and GIS, all of which are defined by basic data models.

But herein lies one of the fundamental weaknesses of the GIS idea – there are simply too many possible geographic data models. In order to accommodate the needs of new applications, software vendors have repeatedly extended the basic data models of their products. One of the most persistent problems is associated with time, since early GIS were built largely to accommodate the static data of maps, and their data models have been extended with varying success to deal with temporal change (Langran, 1992). The problems of dealing with data on networks, necessary in many transport applications, has led to the emergence of products specifically targeted to this niche. Today, a vendor such as ESRI offers a suite of products rather than a single, comprehensive GIS. Each product is designed for a particular class of applications, or for a user community with a particular level of sophistication. The products are able to share data, and many of the concepts on which they are based are common. But with the present trend towards unbundling of software in favour of modular code for specific applications, it seems likely that the days of monolithic GIS are numbered. Instead, we are likely to see much smaller software components that can be mixed to service particular applications, held together through common specifications and standards. Efforts are under way through the Open GIS Consortium (www.opengis.org) to standardise across the entire vendor community, but whether

this will be successful or whether standardisation will be achieved only across the products of each vendor, remains to be seen.

Key concepts

It is not at all clear that a technology designed to process geographic information is of significant value to the social sciences, or that it has potential as research infrastructure. In this section I examine six key concepts and briefly explore their value to the research enterprise.

Integration

One of the commonest ways of introducing GIS – the basis of the cover design on many textbooks – is the *layer-cake*, a representation of the way a GIS database integrates many properties, variables and themes through common geographic location. To Tomlinson and CGIS, this was the computational equivalent of overlaying a number of maps portraying different uses of the same area. If a GIS database links properties to locations, it can clearly also link properties to properties through common location. In the literature of GIS, the technology is often presented as the *only* basis for integrating the departmentalised operations of an organisation, and the only way of achieving an integrated perspective. For example, the US Geological Survey is organised according to four distinct themes: geology, water, biology and topography. Yet for many of its users, this organisational structure impedes rather than facilitates, since it makes it difficult to determine all that the USGS knows about a specific place. US socio-economic data are similarly partitioned among different surveys, agencies and production systems.

While this argument is most obviously made about data, it can also be made about process. Demographic and economic processes, for example, interact at common locations. By creating representations that are spatially and temporally explicit, GIS databases permit coupled and integrated modelling of multiple processes that would otherwise be studied only separately and within the domains of different disciplines.

Spatial analysis

Of necessity, much socio-economic information is collected in *cross-section*, and the construction of *longitudinal* series is beset by problems of continuity, budgets and changing technology. Spatial analysis, or spatial data analysis, comprises a set of techniques and tools designed to analyse data in spatial context. A GIS database captures not only links between properties at the same place, but also such spatial concepts as proximity, containment, overlap, adjacency and connectedness. Visualisation in spatial context (commonly, in the form of a map) is an obvious and powerful way of detecting pattern, anomaly, outlier and even causation. Of course, the forms found in cross-sectional data can never confirm cause, since the same forms can always be created by many processes. Nevertheless, spatial data can be powerful sources of new insights and hypotheses, and powerful bases for confirmatory tests.

Spatial analysis is often best portrayed as a collaboration between mind and machine, combining the power of the eye and brain to detect pattern and scan complex visual displays quickly, with the machine's power to compare layers, apply statistical tests and perform transformations. Spatial analysis has undergone substantial change in recent years, as more and more of the power of the desktop computer has been allocated to achieving ease of use through user interfaces that are visual

and intuitive. In the early days of computing the machine was expert, and the user *submitted* tasks to it. But today's designs make it much easier to support collaboration, and the concept of spatial analysis has broadened accordingly, to encompass everything from visual examination and exploration of mapped data to complex confirmation of spatial statistical models.

Spatially explicit theory and modelling

A model or theory is *spatially explicit* if it is not invariant under relocation; in other words, if changing the locations of the objects that participate in the theory changes the theory's predictions. For example, spatial interaction models are widely used to predict choices made by consumers among shopping destinations (and a variety of other forms of interaction over space as well, including telephone traffic, migration and commuting). Distance appears explicitly in the model, and transformation of space, for example by construction of a new transport link, changes the model's predictions.

Whether space can ever *explain*, or whether it must always be a surrogate for something else (in the case of the spatial interaction model, for the disincentives of travel time or transport cost, for example) is a moot point. Environmental determinism, or the hypothesis that location determines aspects of human behaviour, is now largely discredited within the discipline of geography. In other disciplines, notably economics and ecology, there is much current interest in explicit theorising about space – in its simplest form, through the partitioning of populations into sub-populations whose spatial separation is modelled as a source of imperfect communication. Space is also explicit in many forms of micro-simulation, in which intelligent agents representing individual actors are allowed to move and interact according to well-defined rules.

Place-based analysis

The previous argument is taken a little further in the current interest in *place-based* or *local* techniques of analysis. Earlier debates in geography, notably in the 1950s, had pitted the champions of a *nomothetic* approach, whose aim was the discovery of principles that applied uniformly everywhere (*general* geography, in the sense of Varenius) against the champions of *idiographic* geography, whose focus was the description of the unique properties of places (the *special* geography of Varenius). Nomothetic geography was held to be scientific while idiographic geography was *merely* descriptive.

In the past decade something of a middle ground has emerged between these two positions, aided and abetted by GIS. In this new approach the parameters of models are allowed to vary spatially, and their variation is interpreted and used as the basis for insight and further analysis. For example, suppose that some model $p = f(z)$ is hypothesised to apply to human societies. Given the extreme variability of humanity, it seems unreasonable to believe in a single confirmation of the model conducted in a single city – but on the other hand, an experimental design that samples all of humanity's variability is clearly impossible. Instead, place-based analysis focuses on how the parameters of the model vary from place to place, and draws insights and conclusions from those variations. It thus deals explicitly with the problem of *spatial heterogeneity*, or the notion that no geographic area, however large, can be representative of humanity or the Earth's surface unless it encompasses the entire Earth – geography has unlimited variance up to the scale of the Earth.

The set of techniques designed to support place-based analysis includes adaptive spatial filtering, geographic brushing, geographically weighted regression and local statistics. Fotheringham (1997) provides an excellent recent review.

Knowledge and policy

GIS is widely used both inside the academic research community and also in government agencies, corporations and NGOs. Its applications thus span the distinction between pure and applied, or curiosity-based and problem-driven research, and it provides a clear bridge between them, echoing the arguments of Laudan (1996) that no effective demarcation exists today between science and problem-solving. General knowledge of the ways human societies operate must be combined with data on local conditions to make effective policy, and this is perfectly captured in the ability of a GIS to combine local detail (the contents of the database) with general principles (the algorithms, procedures and data models). GIS is used to simulate the operations of processes under local conditions, and to examine the impacts of general principles in explicit spatial context.

One of the compelling attractions of GIS to a government regulatory agency appears to lie in its procedural nature, which to a scientist might seem overly simplistic and naïve. For example, it is easy to write into law that no industry should locate within 1 km of a residential area, and easy to implement regulation based on GIS analysis, by computing 1 km buffer zones around industries or around residential areas. Scientifically this is naïve, since we have no reason to assume that the effects of industrial pollution are the same upwind as downwind, for example; but the simple procedure stands up well to court challenge, since it can be applied uniformly and diligently. It raises the interesting question of whether effective policy can ever be based on good but complex and often equivocal science.

Place-based search

Narrowly defined, geographic information provides a representation of spatial variation of phenomena over the Earth's surface. It includes maps and images, which provide *exhaustive* representations of the entire surface within their limits, and sampled data that provide information only about a selection of places. Recently, however, there has been much interest in a third class of information that is not strictly geographic but nevertheless can be geographically referenced – that has some form of geographic *footprint*. For example, a tourist guide to Paris has such a footprint and, while it may contain maps, it also contains many other forms of information, some strictly geographic and some not.

Interest in such geographically referenced information arises because of the potential for using geographic location as a basis for search over large and possibly distributed collections of information. For example, geographic location is one of two primary organisational keys for the Electronic Cultural Atlas Initiative (www.ecai.org), an international effort to make primary data in the humanities accessible over the WWW (the other key is time). The University of Southern California is building a major digital archive of its collection of historic photographs of Los Angeles, using the same principle.

A *geolibrary* is defined as a library whose primary search mechanism is geographic. Location has not fared well as a basis of search and organisation in the traditional library for largely technical reasons, but there are no technical reasons that prevent a digital library being organised to respond to the query "what have you got about *there*?" . A recent report of the US National Research Council (NRC, 1999) elaborates on the concept and describes many current prototype implementations.

Geographic information science

GIS has developed as complicated and sophisticated technology for support of science and policy making, but it has done so largely in the absence of a coherent body of theory or language. In this it stands in sharp contrast to the statistical packages, which developed to support an existing and widely used set of techniques underpinned by well-defined theory. If the statistical packages are implementations of statistical theory, then where is the theory that GIS implements?

One consequence of this lack of pre-existing theory is the diversity of languages and standards that have emerged from a largely unco-ordinated GIS software industry. GIS products appear to their users as highly intuitive and pragmatic, rather than as implementations of some universally accepted set of principles, which perhaps explains their popularity. But it means that the GIS community is deeply divided into distinct *information communities*, each with their own set of norms, standards and terms. There are very high costs associated with moving data from one product to another or retraining staff.

Geographic information science seeks to develop the science behind the systems, and to address the fundamental issues raised by GIS. Its focus is well described by the research agenda of the University Consortium for Geographic Information Science, an organisation of major US research universities that now includes some 60 members (www.ucgis.org). The agenda was developed by consensus at the Annual Assembly of UCGIS in Columbus in 1996 (UCGIS, 1996), and contains ten topics:

- *Extensions to representations*, or research to elaborate the set of data models that form the basis of GIS, notably to include time, the third spatial dimension and level of detail.
- *Scale*, or research on the characterisation of level of detail, transformations that aggregate or disaggregate, and the role of scale in modelling process.
- *Uncertainty*, or research on the characterisation of data quality, its impacts on the results of modelling and analysis, and its visualisation and communication.
- *Cognition*, or research on the ways humans understand, reason about and work with geographic information.
- *Spatial analysis*, and the development of new techniques and tools for analysis of spatial data.
- *Distributed and mobile computing*, and the opportunities offered by new technology for new uses of GIS in the field and distributed over electronic networks.
- *Interoperability*, or research on the problems caused by lack of standard protocols and specifications, and the development of new theory-based terminology.
- *Acquisition and integration*, or research on new sources of geographic information and their integration with existing sources.
- *Spatial information infrastructure*, or policy-oriented research on the production, dissemination and use of geographic information.
- *GIS and society*, or research on the impacts of GIS on society, and the context provided by society for GIS.

Digital Earth

In a speech written for presentation at the opening of the California Science Center in Los Angeles in January 1998, US Vice President Al Gore proposed "a multi-resolution, three-dimensional representation of the planet, into which we can embed vast quantities of geo-referenced data" (www.opengis.org/info/pubaffairs/ALGORE.htm). In the speech, Digital Earth is an immersive environment through which a user, particularly a child, could explore the planet, its environment and its human societies. It might be available at museums or libraries, and a more modest version might be available through standard WWW browsers running on a simple personal computer.

Digital Earth is interesting for several reasons, and the concept has attracted widespread interest (the first International Symposium on Digital Earth will be held in Beijing in December 1999). First, it has some of the properties of a *moonshot*, or a vision that can motivate a wide range of research and development activities in many disciplines. It challenges our state of knowledge about the planet, not only in terms of raw data, but also in terms of data access and the ability to communicate data through visualisation. How, for example, would one portray state of human health or quality of life to a child? Moreover, it challenges our understanding of process in the invitation to model, simulate and predict, since the concept should not be limited to static portrayal.

Second, Digital Earth is interesting because of its implications for the organisation of information. The prevailing metaphor of user interface design is the office or desktop, with its filing cabinets and clipboards. Many prototype digital libraries employ the library metaphor, with its stacks and card catalogues. But Digital Earth suggests a much more powerful and compelling metaphor for the organisation of geographic information, by portraying its existence on a rendering of the surface of the Earth. The idea can be seen in limited form in many current products and services, including Microsoft's Encarta Atlas.

Finally, Digital Earth is a fascinating example of a *mirror world* (Gelernter, 1991). Just as a map, it captures a particular state of understanding of the planet's surface, and the data and information available to its builders. But since it cannot be a complete representation, it is interesting in what it leaves out and in how it reveals the agendas of its builders.

Closing comments

I would like to make three brief points in conclusion.

First, GIS seen narrowly is an important and growing application of computing technology. It includes software, today largely developed and marketed by the private sector; data, increasingly available in large quantities through the medium of the WWW; and tools for analysis and modelling that focus on the spatial aspects of data, and increasingly on the temporal aspects. As such, GIS is of increasing importance to those social sciences that deal in one way or another with activities and phenomena that distribute themselves over the surface of the Earth, and with understanding the processes that lie behind them.

Second, GIS seen broadly raises a number of challenging and fundamental issues that range from human spatial cognition to the modelling of complex spatial processes. Collectively, they motivate a multidisciplinary effort to advance what can be termed geographic information science, and many of these issues intersect and engage the social sciences.

Finally, GIS seen broadly is intimately related to the concept of Digital Earth, or the development of an accessible, unified emulation of the surface of the planet and the processes that affect it, both human and physical. As a vision, Digital Earth may or may not be achievable, depending on the assumptions one is willing to make about future technologies, the availability of information, and our ability to characterise and understand process. But as a moonshot it is an idea that can motivate a broad spectrum of activities across many disciplines.

REFERENCES

Foresman, T.W. (ed.) (1998), *The History of GIS: Perspectives from the Pioneers*, Prentice Hall PTR, Upper Saddle River, NJ.

Fotheringham, A.S. (1997), "Trends in Quantitative Methods. 1: Stressing the Local", *Progress in Human Geography* 21 (1), pp. 88–96.

Gelernter, D. (1991), *Mirror Worlds: Or the Day Software Puts the University in a Shoebox. How it will Happen and What it will Mean*, Oxford University Press, New York.

Langran, G. (1992), *Time in Geographic Information Systems*, Taylor and Francis, London.

Laudan, L. (1996), *Beyond Positivism and Relativism: Theory, Method, and Evidence*, Westview Press, Boulder, CO.

Longley, P.A., M.F. Goodchild, D.J. Maguire and D.W. Rhind (eds.) (1999), *Geographical Information Systems: Principles, Techniques, Management and Applications*, Wiley, New York.

National Research Council (1999), *Distributed Geolibraries: Spatial Information Resources*, National Academies Press, Washington, DC (also at www.nap.edu).

University Consortium for Geographic Information Science (1996), "Research Priorities for Geographic Information Science", *Cartography and Geographic Information Science* 23 (3), pp. 115–127.

Chapter 9

NEW HORIZONS FOR QUALITATIVE DATA

by

Paul Thompson*
Director, Qualidata, University of Essex, Colchester

Introduction

It may be helpful for me to start by explaining my own experience. I am primarily a researcher, and for nearly 30 years I have been using life-history or oral-history interviews for researching social change in a series of different areas ranging from marriage and stepfamilies and from to fishing communities to international financiers. My book, *The Voice of the Past* (1978, 1988, 1999), is internationally the best-known work on this research approach. In 1987, I set up the National Life Story Collection, an independent charity based in the British Library National Sound Archive, which houses the principal national centre for the collection of recorded life-history material, and was its Director until 1997. Through my work, I had become aware of the lack of any archiving policy for important earlier qualitative research projects and, after our pressing this issue, the Economic and Social Research Council eventually decided to fund the establishment of Qualidata, the Qualitative Data Archival Resource Centre. From its launch in 1994, I have been its Director, and Louise Corti its Manager. Qualidata's task is to seek out and rescue valuable earlier as well as current qualitative research material, including especially life-story interviews and ethnographic fieldwork notes and, where appropriate, to arrange for it to be archived. I want to focus particularly on its work, since Qualidata is the only national centre of its kind in the world, and I believe that similar initiatives in other countries would greatly enhance the international potential of qualitative research.

Sharing qualitative data

From its origins in the 18th century, the progress of social science has been essentially cumulative. Knowledge has been built up incrementally, resting on the foundations of earlier findings, and interpretation has always depended upon comparisons: with other social groups, other contexts, other cultures, other times.

* Professor Paul Thompson is Director of Qualidata, University of Essex, Colchester, Essex CO4 3SQ, England. Tel: 01206-873058.

Comparison can only be effective when the data is sufficient to allow convincing re-evaluations. Fortunately, many social scientists grasped this relatively early. For example, the original returns of the British population census were kept as public records and have proved an invaluable basis for re-analysis in recent years. And, when Beatrice and Sydney Webb had completed their pioneering study of British trade unionism, they archived their notes on their interviews carried out throughout the country in the newly-founded London School of Economics, where they remain the principal source of information on late 19th century trade unionism.

It was in this spirit that in Britain the Economic and Social Research Council's Data Archive was set up in 1967 in order to retain the most significant machine-readable data from the research which it funds. Thus, crucial survey data can be re-analysed by other researchers, so that the money spent on research becomes not only an immediate outlay, but a resource for other researchers in the future.

There was, however, a significant gap in this policy. Although advances in word processing now mean that most research of any kind is machine readable, until recently most machine-readable data was statistical, based on surveys. Qualitative (non-statistical) research was paper-based, and it presented different problems for archivists in terms for example of storage and of confidentiality. Hence, the Data Archive archived scarcely any of the qualitative material collected by research projects funded by ESRC.

There was no intellectual reason for this. Qualitative and quantitative research are equally based on comparison. Classic British re-studies include not only Seebohm Rowntree's three surveys of poverty in York, and Llewellyn Smiths's repeat of Charles Booth's in London, but also the two successive community studies of Banbury. In North America, similar key examples include the three successive studies of Middletown, or Glen Elder's re-analysis of the fate of Californian children of the depression years,[1] or, to take an anthropological instance, the controversial re-study and reinterpretation by Oscar Lewis of Redfield's Tepotzlan in Mexico.

Nor can it be convincingly argued that the re-use of qualitative research material is less likely than that of quantitative data. This was brought home to me by the experience of my first oral history/life story research, the "Family Life and Work Experience before 1918" project which was carried out in 1970-73 specifically for my book *The Edwardians* (1975, 1992, etc.), a social history of Britain between 1900 and 1918. I drew especially heavily on the interviews for the chapters on childhood and family life. In terms of writing from the research team, the project also resulted in Thea Thompson's *Edwardian Childhoods* (1981), a chapter in my *I Don't Feel Old* (1990), and many of the insights in *The Voice of the Past*. But, from the start, we recognised that the interviews could be invaluable to others.

As part of the analysis they were coded and this statistical material was given to the Data Archive. It has been rarely used, perhaps half a dozen times in 20 years, and I am unaware of any publication resulting from such users. We also offered the Data Archive the transcripts of our interviews, but these were refused. We therefore decided to set up an elementary archive ourselves.

The transcripts of the 444 interviews – all on paper, given the date – were organised in two versions: *i)* the interview as given; and *ii)* cut-up versions re-sorted according to the 20 main themes of the interview guide. Once the project was ended, due to the generosity of the Department of Sociology we were able to maintain the transcripts in a special room and – without officially advertising ourselves as an archive – allow access to *bona fide* researchers who have approached us in a continuing stream for over 20 years. The collection has proved a very valuable inspiration to many of our own students and it has also resulted in some notable publications by visiting scholars. These include, for example, Standish Meacham, *A Life Apart: The English Working Class* (1980);

Charles More, *Skill and the English Working Class* (1980) and Michael Childs, *Labour's Apprentices* (1992) on work; John Gillis, *For Better, For Worse* (1985) on marriage; and David Crouch and Colin Ward on *The Allotment* (1988), as well as articles on class by Patrick Joyce, on social mobility by David Vincent, on education by Jonathan Rose, on religion by Hugh McLeod, on stepfamilies by Natasha Burchardt, and on women and the family by Ellen Ross. Sometimes rereading the material made us aware of issues which we could have too easily pursued at the time, for example on stepfamily relationships and, above all, made us regret that, just because the interviews had been carried out for a book which was to end in 1918, we obtained scarcely any information on the last 50 years of the lives of those we interviewed, and indeed too often cut them off from telling us about what had since happened to them. Nevertheless, as time went on, it became increasingly clear that this representative set of interviews, now unrepeatable, constitutes a priceless historical resource for the future. Already, judged purely in terms of the number of publications, the original output of the research team has increased more than five-fold.

This experience led me to question what had happened to the data from other important projects using qualitative methods. Eventually, in 1991, the Economic and Social Research Council asked us to carry out a small pilot study of what was happening to qualitative data from projects which it had funded. This revealed that 90% of qualitative research material was either already lost or at risk, mostly in researchers' homes or offices. Even some the material which researchers reported as "archived" turned out to be deposited in so-called archives which had none of the basic requirements of an archive, such as physical security, public access, reasonable catalogues or with recorded material and listening facilities. In the worst instance of all, generations of distinguished anthropologists had been depositing their lifetime's work with the Royal Anthropological Institute in London, which had no catalogue, no archivist, no access and not even physical security; the collections appeared to be slowly mouldering into oblivion.

We calculated that it would have cost at least GBP 20 million to re-create a resource on the scale of that at risk. For the older material, moreover, the risk was acute, and the need for action especially urgent. This was subsequently borne out by the fate of the research material on child-rearing which John and Elizabeth Newson had been collecting in Nottingham since the early 1960s, consisting of over 3 000 high-quality in-depth interviews with parents and children. It need hardly be said that parenting is and will remain a major issue for both research and public debate, and the Newsons were the outstanding pioneers of social science work in the field. Only their earlier series of interviews were fully exploited in their own writing, and the possibilities of drawing on the unused material, and also of carrying out further follow-up studies, would have been immense. Tragically on their retirement, just weeks before Qualidata opened, the Newsons decided that their lifetime's research data should be destroyed.

As a result of our pilot study report, we were invited by ESRC to develop our proposed solution, and the outcome was the founding in October 1994 of Qualidata. Qualidata has a double purpose. The first has been a salvage operation – to rescue the most significant material created by research from previous years. The second has been to work with ESRC and the ESRC Data Archive to ensure that for current and future projects the unnecessary waste of the past does not continue. All qualitative researchers are obliged, as part of their research applications, to consider the potential archival value of the data they create and, where appropriate, to co-operate in securing its deposit in an appropriate archive. We are concerned not only with life-story and oral-history projects, but also with other contemporary sociological studies, with anthropology, linguistics, education, geography and, indeed, with the whole range of social science disciplines.

Qualidata is not an archive itself: it is an action unit. We locate and evaluate research data, we catalogue it, we organise its transfer to suitable archives and we publicise its existence to researchers.

We consult with ESRC and other funding bodies on qualitative aspects of data-set policy. We provide advice to researchers on the implications of archiving for research, both through organised workshops and through individual consultations.

The first question we had to solve was where we could best deposit the material which we traced. We have visited and selected a list of potential repositories which we could recommend as providing good facilities, such as the British Library and its National Sound Archive, the Modern Records Collection at Warwick University, the Institute of Criminology at Cambridge, the Institute of Education in London, and the Mass Observation Archive at Sussex University, and we have worked out with them suitable forms of agreement for the conditions of deposit. Rather to our surprise, however, we found that no library was willing to take large collections on social policy or social change. In the event, we have been able to set up a special new archive for these areas at Essex itself, with initial support from the Joseph Rowntree Foundation.

At the same time we have evolved a model pathway along which any data set was intended to travel, from being traced, evaluated and selected as a priority for archiving, through cataloguing to deposit in a chosen archive. For this, we developed forms of agreement for depositors and produced a set of documents for the guidance of researchers which set out a strategic practical path through the key issues, such as confidentiality and copyright, which must be confronted during the process of deposit. They have required a great deal of discussion and we regard them as an important step forward in themselves.

The material which we have deposited includes my all own research interviews: we felt it right that I should be the first guinea-pig for testing out our transfer procedure (and indeed it proved a humiliating process, revealing the extent to which I could not remember what I had done or why on earlier research projects). Other major deposits have included, for example, the lifetimes' research material of Peter Townsend, Britain's leading researcher on poverty since the 1950s; Ray Pahl's research on urban and family sociology since the 1960s; letters for Annette Lawson's book on adultery; qualitative material from classic British studies such as *The Affluent Worker*; and the sustained research studies of George Brown on the social origins of depression since the 1960s.

It is important to emphasise that many of these studies are of much more than historical value. In some cases, they may provide key benchmarks for repeat studies or attempts to measure social change. Others may be of value precisely because they offer research data which cannot be repeated. This is perhaps most obvious in the case of anthropological studies of unique societies which have since changed fundamentally. But equally, in the case of Townsend's data it seems clear that the sheer scale and depth of his studies of poverty and old age are unlikely to be repeated in the foreseeable future, so that they must continue to provide an exceptionally rich quarry of experience to guide contemporary social policy experts who cannot afford to replicate such studies.

One of our most basic challenges has been to get a measure of how much material of comparable significance may survive from earlier research projects. Establishing this is a very substantial task in itself, and has involved setting up three different surveys which together take us back to 1945. We believe that the information which we are gathering will be of considerable intrinsic value for researchers, as well as a foundation for our own activities.

Once we have established the existence of qualitative material and the willingness of the researcher to consider deposit, we seek the advice of specialists in the relevant discipline in deciding which projects should be evaluated. In prioritising, we take into account both intellectual and practical criteria. In intellectual terms, we ask, is this research recognised as of high quality, influential in its field, or representing the working life of a significant researcher; and does it have a high value of

potential for reanalysis or comparative use? Practically, we consider whether it is at high risk of destruction; can be made freely available for research use, with copyright and confidentiality not too restrictive for reasonable access; and whether it is legible or audible and in reasonable physical condition, and sufficiently documented for informed re-use.

Ethical issues

However, the mention of confidentiality and copyright leads to the thorniest problem an archiving retrieval programme must face.

We have been well aware from the start that both for researchers and for archivists confidentiality and copyright must be basic issues in considering the archiving of qualitative material. The two are different, but intertwined. The ethical importance of confidentiality has always been important. Whether or not confidentiality has been explicitly promised, all social science researchers have a responsibility not to expose their informants to potential injury, whether through publicising confessions of illegal activity, or to libel suits, or simply to risk of scandal and ridicule. For this reason, some research material, such as studies of criminal behaviour or paramilitary groups, may be intrinsically impossible to archive. Others might be accepted, with a timed delay for research access or with access restricted to *bona fide* researchers only. Some material can be anonymised. However, the degree of anonymisation which is possible in survey work through the disaggregation of individual cases is intrinsically undesirable for qualitative material, for it destroys the meaning of the data; in many contexts, a partial anonymisation, even through the changing of all names, may not render the informant absolutely unidentifiable. The balance in terms of cost, degree of anonymity and integrity of the data resulting from alternative possible approaches have to be weighed up for each set of material.

Copyright, on the other hand, has become a more serious problem for British researchers since the 1988 Copyrights, Designs and Patents Act. Until then, an interviewee's willingness to be interviewed was held to imply an oral contract to be interviewed and for the researcher to hold the copyright of the interview. The position has now been reversed in European and in British national law. These changes were not made with the needs of social researchers in mind, and copyright law is certainly one issue on which a watchful eye by an international social science research committee could be important for the future. Under the new British law, the interviewee is assumed not to have transferred copyright unless there is a written contract. Although this will be possible in some research projects, in many contexts social researchers are likely to feel that attempting to obtain copyright transfers will jeopardise the success of the research. Nor is it at all clear that a total transfer of copyright from the informant, for example in the case of a life-story interview, is ethically justified. Fortunately, it is still held under British law that a willingness to be interviewed implies a licence for the researcher to use the material. If the request for the interview is framed in a sufficiently broad way, this licence may also extend to the use of other social researchers. The wording of the introductory section of interviews is therefore of particular importance if qualitative data is to be available to other researchers.

As for research already carried out, in more recent cases it may be possible to write to informants seeking written copyright agreements, but these will very rarely cover all those interviewed. More usually, in our view, we have to weigh the copyright difficulty against the potential value of the particular material, and the knowledge that legal suits against researchers are extremely unlikely since in most cases their publication value is trivial. It is important to remember that archivists have long been accustomed to accepting and allowing the use of written material whose copyright was held by (often unknown) others, but this has rarely caused serious difficulties.

Thus, in our view, while in theory copyright may seem the more difficult issue, in practice it is more likely to prove to be confidentiality.

New databases

To fully realise the aims of Qualidata, we would need not only to succeed in archiving significant qualitative material, but also to influence the culture of this whole school of social science research. In the past, partly because of the absence of a policy for the archiving, and also because research material was usually paper-based and in some cases (such as anthropologists' field-notes) hand-written and difficult to understand, qualitative researchers rarely thought of archiving, or (except historians) of the re-analysis of earlier data. Now, with most transcripts machine readable, the possibility is opening up for a crucial shift in attitudes among social science researchers and students towards a new culture of secondary analysis, based on the assumption that the creation of in-depth interview material should be for the benefit not only of the individual investigator, but of the research community as a whole.

In that spirit we should, I believe, look more closely at the practice of ethnographic researchers in Scandinavia. In Stockholm, the Nordic Museum Archive, with a staff of 250, provides a national service for museums encompassing libraries, photographs, exhibitions and objects. It has a separate "Memory" section, led by Stefan Bohman, with a staff of ten. The archive has been collecting material resulting from the researches of academic ethnographers for almost a hundred years: substantial in quantity, well kept and indexed and regularly used in an attractive reading room. I was told that such research material was useful only when the original words of informants were recorded; when they had merely been summarised, they were too much reshaped by the academic preoccupations of the time to be of much research interest to contemporary ethnographers.

Still more strikingly, the archive has been organising regular autobiographical competitions since 1945, and since 1928 it has been collecting special thematic essays from a panel of 400 correspondents throughout Sweden. The themes have gradually shifted from earlier historical preoccupations to encompass all aspects of contemporary everyday life, including even computing. The archive also has a notable collection of diaries. All the basic thematic and autobiographical data is well indexed and this has now been computerised, so that very full information on the collection in Stockholm can be available at centres throughout the country. In short, the Nordic Archive provides an impressive example of what can be achieved when research money is put into the creation of general research resources as well as specific projects.

The National Life Story Collection, an independent charity which I launched at the British Library National Sound Archive at the end of the 1980s, was also intended to collect material, not for a single book, but as a wider resource for other researchers, writers and broadcasters. This time, too, the plan was to carry out full life-story interviews, rather than just interview people about their earlier lives. We envisaged the creation of a national autobiography in sound, which would encompass, on the one hand, any man or woman notable enough to deserve an obituary notice in the daily press, and on the other, a continuing cross-section of ordinary men and women throughout Britain. If we could have enthused a generous billionaire, we might have succeeded in this but, unfortunately, none of those whom we approached found the overall project sufficiently compelling. So we were thrown back on fund-raising for separate projects. We have managed to do some useful things: our projects include lives of sculptors and painters (with the Tate Gallery), Jewish Holocaust survivors in Britain, the financial elite of the City of London, and workers at all levels of British Steel. But the national cross-section we envisaged proved insufficiently attractive for private funding.

We were, however, successful in raising the sponsorship needed to carry out the first British national competition for written and oral autobiography, for which we received nearly a thousand

entries. This has proved the catalyst for a much more ambitious national BBC local radio project for the Millennium, now approaching completion. We participated in drawing up background information and interview guidelines for the thematic life-story interviews, and over 6 000 have been recorded during the last year on minidisk. They have been systematically summarised, although unfortunately none of the interviews have been transcribed in full. These recordings, along with coded summaries provided by the interviewees, will shortly be deposited in the National Sound Archive, and will be available for public use next year. Since this material has been collected from every part of the country, and because it is digitally recorded so that the sound as well as the text can be thematically recovered and edited, I believe that this project has created a remarkable potential resource for social researchers, and also for students. It may require fine-tuning of the summaries and indexing and also some transcription to make it more usable, but it deserves to be recognised as a major new resource.[2]

I had hoped that private sponsorship would enable us to sustain a more continuous and broader recording programme than is possible through research funding. In practice, that has not proved the case. In Britain, private sponsorship remains narrow and rather unimaginative in its aims. From our experience, it remains, despite the cuts in public funding, much more possible to gain substantial financial support for research resources from research foundations and councils.

Interaction

For social scientists using qualitative material, it would be a substantial step forward if there was even a modest policy of life-story recording with an evolving sample of different age groups, etc., following a basic core framework, but extended to address in greater depth themes which changed to represent new research interests. The Swedish example of written and recorded work of this type indicates that it could be used as a resource throughout the educational system, and not just by researchers.

For researchers, a particular attraction would be that informants might be chosen on the basis of existing national survey samples. Experience has shown that in-depth interviewing of informants in longitudinal studies increases rather than decreases their likelihood to respond positively to the next request for a survey interview. This opens the possibility of an illuminating interaction between the quantitative survey evidence and the qualitative life-story material. It needs to be recognised, however, that such an interaction will only prove valuable if the survey material is accessible in a suitable way: that is to say, firstly, that new hypotheses generated by the life-story interviews can be readily tested without a big time input, and secondly, that the evidence can be looked at case by case, and that scrambling for anonymity has not made this impossible.

A second type of interaction can also be envisaged, as new technologies advance, between different types of presentation: visual, aural and textual.

New technologies

The implications of the new technologies for researchers are much more profound than this. In my view, qualitative social scientists face a double technical challenge in the next few years and the outcome will decide whether a lively qualitative research movement survives into the 21st century. The first is video. At the moment, audio recording remains justifiable because it is less disturbing and very much cheaper than video. But, when high-quality video with equally good sound, simultaneously operable by an interviewer, becomes almost as cheap as audio recording, the arguments against it will vanish. I have already seen researchers using a camera-sized handheld Japanese video which can be

tucked onto one's shoulder, making it possible to do completely informal outdoor or indoor video interviews, and over the next three years the price of such equipment is likely to fall as fast as its quality rises. Life-story interviews using only sound may become archaic survivals of a past technical era.

Multimedia raises more dramatic possibilities for research, archiving and publishing – but also, for the moment, more doubts. In principal it is already possible to organise an archive so that you could read a transcript and then at will switch over and hear the sound of the same passage and see the expressions of the speaker – this represents an enormous advance in the accuracy and potential interpretative insights of researchers using interviews. But the cost of doing this for a large collection remains prohibitive (and there are also technical problems in the amount of sound and video which can be included). In practice, multimedia has proved commercially viable principally with publications for children (who take more quickly to new media and are being used as the lead-in for a future adult market) and for reference works such as dictionaries and encyclopaedias.

So far, I know of no European social science or oral-history project which has brought out its material in multimedia form. I have seen a few North American examples. But, to date, the most sophisticated to have been produced are the series of multimedia CD-ROMs from projects carried out by Karen Worcman and her team at the Museu da Pessoa in São Paulo, Brazil. These are either commercially or city-sponsored projects on topics such as a football club, a trade union or a firm. Whatever the theme, they combine interview texts with shorter extracts of audio and video, along with maps, genealogical trees, old film and other documents. What is novel, and in these CDs beautifully designed, is the way in which all this information is interconnected so that you can switch easily between one type and another. The elegance of the design and the colours, and the touches of wit such as the faces which ask for a click with a smile or the winking eyes, are all there to help you to enjoy moving around. Multimedia is clearly a new art form.

We have already had to learn how different an interview is on audio tape from a typed transcript with words only. The typed form can never convey more than a hint of the tones, accents and emotions of the spoken word, and the irregular pauses of speech necessarily disappear behind the logical sequence of grammatical punctuation. The testimony will vary yet again if it is to be published in printed form: editing will render it smoother and more condensed, cutting out repetitions and filler-words, helping the reader to focus on what the editor thinks matters most. Sooner or later, it will become clear that multimedia requires its own kind of reshaping of the recorded life story; and also that a certain type of teller is the best for the medium.

But I am still puzzled by another question: I do not yet understand what the Museu da Pessoa's CD-ROMs are meant to do for their audience. With the São Paulo football club project, for example, you can stand in front of the monitor, press your finger on the screen, and opt for the 1930s, then for notable goals, see the goal scored on film and hear the uproar of the crowd, and then switch to the footballer who scored, and discover about his life and memories. In theory, you could use this multimedia design to build up a real historical resource on the club's history, with dozens of testimonies from retired players, managers, patrons and generations of fans. But, in practice, it seems to me that the fascination is in flicking from one sort of information to another, rather than in exploring anything in depth. Ultimately, it is another form of reference work – and a dictionary, however fascinating, can never be a novel. Hence, while multimedia can store life stories, it is designed for a form of use which is fundamentally inimical to any sustained narrative or authorial argument. In other words, it depends on what its users make of it. Karen Worcman argues that – especially in a country like Brazil, where there is little popular interest in history – it can hook in people who would never be attracted to more conventional forms of publication. We shall see. But

certainly multimedia is a new form of publication. And social researchers are going to be forced to come to terms with it.

The future is all the more challenging, and also more difficult to predict, because of the simultaneous explosion of not only multimedia but also the Internet. Clearly the Internet is now well established as one of the principal means of communication between researchers worldwide. It also provides a new way of sustaining close friendships across the globe. In principle, this could include the exchange of major qualitative research databases. And there is no technical reason why it should not become a principle multimedia form. However, whether this comes about is likely to depend partly on solving the difficulties of recovering the production costs of multimedia packages. There are likely to be crucial battles between the computing and publishing giants over copyright, markets and profits, in which researchers of any kind will be merely marginal. For the same reason, we may not see the full flowering of the interactive possibilities of the Internet, through which users might modify the multimedia packages they receive, inserting material of their own and reshaping them, just as the original sources have already been reshaped to become part of the package. This may be too democratic a possibility for publishers and for academics too. But of one point there can be no doubt: if researchers are to remain communicators, and if their work is to have any substantial influence, we must move with the technology of communication.

International potential

As a qualitative social science researcher, I strongly support the formation of an international committee on large-scale infrastructural needs. In part this could have a defensive value, helping to ensure that future changes in the international law of copyright do not make qualitative research more difficult or even practically impossible.

More positively, with the further development of multimedia systems and also of computerised translation, there is a tremendous potential for the international exchange of data sets. A British researcher on mining communities, for example, through international research agreements might in the future have access to data enabling comparisons with mining communities in Canada, Latin America, China or South Africa.

The combination of such exchanges with the development of systematic qualitative research archiving programmes in other countries – and currently, the only other unit we have found comparable to our own is the Murray Research Center at Harvard, which is a resource centre for research on women – would, in my view, lead to a new era in qualitative social research. Through becoming, at the same time, more cumulative and more comparative, it could shake off much of its current fragmentation and realise again its full potential as a second pillar of social science research.

NOTES

1. Elder's work is based on a longitudinal study of Oakland children which was carried out from the 1920s to the 1990s, combining quantitative and qualitative information with exceptional richness. This study was based at the Institute for Human Development at the University of California at Berkeley, and has been the basis of many other important studies of the family and marriage, including notable work by Erik Erikson, Tamara Hareven and Arlene Skolnick. I doubt if there is any group of people in the world born before 1960 who have been so richly and continuously documented. Nevertheless, this major resource has now been closed to researchers, its archivist has been sacked and its future survival seems very uncertain.

2. It may be worth adding that we also hope to organise the permanent archivist and opening to research use of the material created by a BBC television project, *Video Nation*, which over the last five years has trained a careful cross-section of 250 men and women from the general public to keep daily diaries and record videos over a period of several months each. This collection of some 5 000 videos with accompanying diaries is very well indexed. It is not open to non-BBC researchers and no plans have been made for its preservation when the *Video Nation* series ends. It is, however, a striking example of how public funding through the media can help to create a first-rate research resource.

Social Sciences for a Digital World: Building Infrastructure for the Future

6-8 October1999, Ottawa, Canada

LIST OF PARTICIPANTS

Co-chairs:

Mark Renaud
President, SSHRC

Bennett Bertenthal
Assistant Director, Directorate for Social, Behavioral and Economic Sciences, NSF

Austria	**Müller, Karl** Department Head Institute for Advanced Studies, Dept. of Political Science and Sociology
Belgium	**van Langenhove, Luk** Deputy Secretary General Ministry of Science Policy
	Luwel, Marc Flemish Ministry for Science Policy
	Van Doninck, Bogdan Director of Agora Federal Office for Scientific, Technological and Cultural Affairs
Canada	**Bernard, Paul** Professor, Département de sociologie Université de Montréal
	Clements, Patricia Professor, Department of English University of Alberta
	Felligi, Ivan Chief Statistician Statistics Canada
	Hollett, Alton Director Newfoundland Statistical Agency
	McDonald, Lynn Director Centre for Applied Social Research University of Toronto

Canada *(cont'd)*

Norris, Douglas
Director
Housing, Family and Social Statistics
Statistics Canada

Orstein, Michael
Director
Institute for Social Research

Péruse, Denise
Analyste
Fonds pour la formation de chercheurs et l'aide à la recherche (FCAR)

Ritchie, Pierre
Executive Director
International Union of Psychological Sciences

Strangway, David
President
Canada Foundation for Innovation

Rowe, Penelope
Executive Director Community Services
Council of Newfoundland

Sheridan, Michael
Assistant Chief Statistician
Social, Institutional and Labour Statistics
Statistics Canada

Watkins, Wendy
Chief Data Librarian
Carleton University

Picot, Garnett
Director
Business, Labour and Market Analysis
Statistics Canada

Gaffield, Chad
Director
Institute for Canadian Studies
University of Ottawa

Mark, Tim
Executive Director
Canadian Association of Research Libraries

Moffat, Marshall
Director
Knowledge Infrastructure Division
Industry Canada

de la Mothe, John
Professor
Faculty of Administration
University of Ottawa

Dufour, Paul
Senior Policy Analyst
Secretary of State for Science and Technology

Canada *(cont'd)*	**Clement, Wallace** Director Institute for Political Economy Professor, Sociology and Anthropology Carleton University
	Weir, Lesley Assistant Chief Librarian University of Ottawa
	Crabbé, Philippe Professeur Départment de économique Université d'Ottawa
	Landriault, France Directrice Politiques, planification et collaboration internationale Conseil de recherches en sciences humaines du Canada
	Croux, Denis Directeur Programmes stratégiques et initiatives conjointes Conseil de recherches en sciences humaines du Canada
	Calvert, Ian Director Information Systems Social Sciences and Humanities Research Council of Canada
	Ellis, Ned Director General Programs, Policy and Planning Social Sciences and Humanities Research Council of Canada
	Wiggin, Pamela Director of Communications Social Sciences and Humanities Research Council of Canada
	Lamoureux, Michel Vice-President External Relations Canada Foundation for Innovation
Denmark	**Pederson, Lene Wul** Researcher Danish Data Archives
Finland	**Borg, Sami** Director Finnish Social Science Data Archive
	Hörkkö, Sirkka-Leena Senior Advisor Ministry of Education, University Division
France	**Silberman, Roxanne** IRESCO/CNRS

Germany	**Mochmann, Ekkehard** President International Federation of Data Organisers for the Social Sciences
	Wagner, Gert Director, German Socio-Economic Panel Study (GSOEP) European University Viadrina at Frankfurt (Oder)
	Hotopp, Ulrike Bundesministerium für Bildung und Forschung (BMBF)
Japan	**Negishi, Masamitsu** Professor & Director Developmental Research Division, Research and Development Department National Center for Science Information Systems
	Takeishi, Akira Associate Professor Institute of Innovation Research Hitotsubashi University
Luxembourg	**Schaber, Gaston** Director, CEPS/INSTEAD
Netherlands	**Dekker, Ron** Director, Research Council for Social and Behavioural Research Scientific Statistical Agency
	Hooimeijer, Pieter Scientific Director Netherlands Graduate School of Housing and Urban Research (NETHUR)
	Soete, Luc Director Maastricht Economic Research Institute on Innovation and Technology
Norway	**Henrichsen, Bjorn** Director Norwegian Social Science Data Services (NSD)
	Lande, Trygve Science Advisor Culture and Society Division Research Council of Norway
Portugal	**Bonfim, José** Advisor Institute for Scientific and Technological Co-operation
	Caraca, Joao Director, Science Department Calouste Gulbenkian Foundation

Spain	**Barrio, Jose Felix** Advisor for Education and Science Spanish Embassy, Ottawa, Canada
Sweden	**Vogel, Joahim** Professor Statistics Sweden and Department of Sociology, Umea University
	Ohngren, Bo Deputy Secretary General Swedish Council for Research in the Social Sciences and Humanities
Switzerland	**Farago, Peter** Programme Director, NSF Research Programme “Switzerland: the Future” Swiss National Scientific Research Fund
	Joye, Dominique Director SIDOS
United Kingdom	**Thompson, Paul** Director Qualidata Project, Essex University
United States	**Bainbridge, William** Science Advisor, Social, Behavioral and Economic Sciences National Science Foundation
	Goodchild, Michael Professor and Chair National Centre for Geographic Information and Analysis Department of History
	Lane, Julia Professor of Economics Department of Economics, American University
	MacWhinney, Brian Professor Department of Psychology Carnegie Mellon University
	Newlon, Dan Director, Economics Program National Science Foundation
	Rockwell, Richard Executive Director, ICPSR Institute for Social Research
	Ruggles, Steven Professor

European Commission	**Kourtessis, Artemios** DG XII Programme
	Mitsos, Achilleas Director DG XII/F,SDME 3/59
Israel	**Kamen, Charles** Director Israel Central Bureau of Statistics Department of Social and Welfare Statistics

International Organisations

European Science Foundation	**Jowell, Roger** Professor, National Centre for Social Research
	Smith, John Head of Unit – Social Sciences
International Social Science Council	**Kosinski, Leszek** Secretary General
UNESCO	**Lievesley, Denise** Director – Institute for Statistics UNESCO
	De Guchteneire, Paul Co-ordinator Social Sciences Division, MOST Programme, UNESCO

OECD Secretariat

Aubert, Jean-Éric	Principal Administrator Directorate for Science, Technology and Industry
Bayar, Viviane	Principal Administrator Directorate for Science, Technology and Industry

SSHRC Secretariat

Lauziere, Marcel	Special Advisor to the President
Moorman, David	Policy Analyst

OECD PUBLICATIONS, 2, rue André-Pascal, 75775 PARIS CEDEX 16
PRINTED IN FRANCE
(93 2000 03 1 P) ISBN 92-64-17678-0 – No. 51265 2000