New Data for Understanding the Human Condition

International Perspectives

OECD Global Science Forum Report on
Data and Research Infrastructure for the Social Sciences

February 2013

OECD
BETTER POLICIES FOR BETTER LIVES

New Data for Understanding the Human Condition: International Perspectives

OECD Global Science Forum Report on *Data and Research Infrastructure for the Social Sciences*

Data-driven and evidence-based research is fundamental to understanding and responding effectively and efficiently to global challenges related to the health and wellbeing of populations around the world. Spurred by the rapid growth in new forms of data collected in conjunction with commercial transactions, internet searches, social networking, and the like, and by technological advances in the capacity to access and link existing survey, census, and administrative data sets, the potential payoff for international and multidisciplinary collaboration of scientific groups to address these challenges is increasing rapidly. The Global Science Forum established an expert group to review developments in international data availability, consider their suitability for comparative research, detail the challenges to be addressed, and make recommendations to respond to these new opportunities. This report presents the findings and recommendations of the expert group.

February 2013

Preface

Social science is changing, driven by the need to improve our understanding of the human condition. Many of the research issues we face will require social scientists to work in close collaboration with scientific investigators in other disciplines, notably the biomedical and natural sciences, and across national boundaries. In part these changes arise from the need to adopt a multidisciplinary approach in our search for the causal mechanisms underlying and potential spread of communicable diseases, migration and human responses to climate change. But they also derive from a more 'data driven' approach to scientific investigation. The advances we have made in terms of our ability to generate, capture and re-use information on all aspects of human behaviour places us in a sea of data that has the potential to inform and inspire innovative approaches to scientific investigation. Together with the radical improvements in global communication, these developments now make it imperative that we should take stock of our capability to undertake multidisciplinary research across national boundaries and identify existing and potential obstacles to scientific progress. This report, prepared by a group of leading social scientists, attempts to draw together relevant initiatives in the areas of data discovery, access and sharing, and makes a number of important recommendations to extend and enhance these efforts.

We would like to take this opportunity to thank the members of the Expert Group who gave freely of their time and expertise to assist with the preparation of this report. We have also benefited from discussions with and presentations from additional experts in the course of our work. We value all of the views and opinions they provided.

The UK Economic and Social Research Council funded the participation of the chair, vice chair and the secretariat for the Expert Group. These funds also enabled us to extend membership of the group to a number of experts from non-OECD countries. Stefan Michalowski and his staff at the OECD Global Science Forum helped us to navigate our way through the protocols of the Forum and gave valuable advice and guidance in the drafting of this report. Tom Swasey at the University of North Carolina was responsible for the design of the report. Special thanks go to Margaret Birch at the University of Warwick, who has assisted not only with the organisation of meetings, but generally kept us on our toes throughout the past two years and provided the title for our report.

Barbara Entwisle, Chair
Peter Elias, Vice Chair

THE UNIVERSITY *of* NORTH CAROLINA *at* CHAPEL HILL

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

THE UNIVERSITY OF WARWICK

ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT (OECD)

The OECD is a unique forum where governments work together to address the economic, social and environmental challenges of globalisation. The OECD is also at the forefront of efforts to understand and to help governments respond to new developments and concerns, such as corporate governance, the information economy and the challenges of an ageing population. The Organisation provides a setting where governments can compare policy experiences, seek answers to common problems, identify good practice and work to coordinate domestic and international policies.

The OECD member countries are: Australia, Austria, Belgium, Canada, Chile, the Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Israel, Italy, Japan, Korea, Luxembourg, Mexico, the Netherlands, New Zealand, Norway, Poland, Portugal, the Slovak Republic, Slovenia, Spain, Sweden, Switzerland, Turkey, the United Kingdom and the United States. The European Union takes part in the work of the OECD.

GLOBAL SCIENCE FORUM (GSF)

The GSF is a venue for consultations among senior science policy officials of the OECD member and observer countries on matters relating to fundamental scientific research. The Forum's activities produce findings and recommendations for actions by governments, international organisations, and the scientific community. The GSF's mandate was adopted by OECD science ministers in 1999, and has been extended until the end of 2014. The Forum serves its member delegations by exploring opportunities for new or enhanced international co-operation in selected scientific areas; by defining international frameworks for national or regional science policy decisions; and by addressing the scientific dimensions of issues of social concern.

The Global Science Forum meets twice each year. At these meetings, selected subsidiary activities are reviewed and approved, based on proposals from national governments. The activities may take the form of studies, working groups, task forces, and workshops. The normal duration of an activity is one or two years, and a public policy-level report is always issued. The Forum's reports are available at www.oecd.org/sti/gsf. The GSF staff is based at OECD headquarters in Paris, and can be contacted at gsforum.contact@oecd.org.

New Data for Understanding the Human Condition: International Perspectives

Contents

Executive Summary: challenges and recommendations

Data-driven and evidence-based research is fundamental to understanding and responding effectively and efficiently to global challenges related to the health and wellbeing of populations around the world. Spurred by the rapid growth in new forms of data collected in conjunction with commercial transactions, internet searches, social networking, and the like, and by technological advances in the capacity to access and link existing survey, census, and administrative data sets, the potential payoff for international and multidisciplinary collaboration of scientific groups to address these challenges is increasing rapidly. The Global Science Forum established an expert group[1] to review developments in international data availability, consider their suitability for comparative research, detail the challenges to be addressed, and make recommendations to respond to these new opportunities. This report presents the findings and recommendations of the expert group.

Challenge *1*

massive amounts of digital data are being generated at unprecedented scales and velocity, much of it from new sources such as the internet. The reliability, statistical validity and generalisability of new forms of data are not well understood. This means that the validity of research based on such data may be open to question.

[1] Details of the membership of the Expert Group and its Terms of Reference are included in the *Appendix*.

Finding: the expertise and knowledge required to exploit the scientific value of these data and to make them available for re-use is in many cases dispersed across countries and scientific disciplines. Opportunities to gain leverage and build on common international expertise are missed, and costs consequently incurred as a result of duplication.

> Recommendation 1: ***national research funding agencies*** *should collaborate internationally to provide resources for researchers to assess the research potential and to develop new methods to understand the opportunities and limitations offered by new forms of data to address important research areas.*

Challenge *2*

while many countries have vast amounts of more traditional forms of administrative, survey, and census data collected by and held by national statistical agencies and government departments, knowledge about the existence of such data as micro-data records is a precondition for the efficient and effective planning of international research.

Finding: many significant activities are underway across the world to make research data easier to find, but not all are documented to one or the other of the two international standards that now exist for data documentation and interchange. As a result, information about the existence of micro-data and their availability for re-use is often difficult to find.

> Recommendation 2: ***national statistical organisations and international organisations*** *should ensure that all data they collect and process are documented to agreed and common standards. Such documentation should be easily discoverable on their websites.*

Challenge *3*

new forms of personal data, such as social networking data, can provide insights into the human condition. However, the use of those data as research resources may pose risks to individuals' privacy, particularly in case of inadvertent disclosure of the identities of the individuals concerned. There is a need for greater transparency in the research use of new forms of data, maximising the gains in knowledge derived from such data while minimising the risks to individuals' privacy, seeking to retain public confidence in scientific research which makes use of new forms of data.

Finding: many national privacy laws authorise the re-use of personal data for historical, statistical or scientific purposes, provided appropriate safeguards are put in place. However, there is no internationally recognised framework code of conduct which deals specifically with the use of new forms of personal data for research.

Recommendation 3: ***research funding agencies and data protection authorities*** *should collaborate to develop an internationally recognised framework code of conduct covering the use for research of new forms of personal data, particularly those generated via network communication. This framework, built on best practice procedures for consent from data subjects, data sharing and re-use, anonymisation methods, etc., could be adapted as necessary for specific national circumstances.*

Challenge 4

barriers to access to social science data hinder national and cross-national collaboration which could exploit their research value. These barriers relate to a variety of obstacles (legal, cultural, language, proprietary rights of access) all of which have to be identified and removed if cross-national research is to be promoted.

Finding: a number of activities are underway across the world to develop and provide access to social science micro-data for comparative research purposes. These activities tend to be 'domain specific' (*i.e.* international studies of political behaviour, social attitudes and lifestyles, fertility and family formation). They are driven primarily by the interests of leading social scientists in these fields, less so by national statistical agencies.

Recommendation 4.1***: national statistical agencies*** *should establish mechanisms to improve access for comparative research by the research community to social science micro-data in their possession. These same agencies should collaborate internationally to further their efforts in this area, particularly with respect to the use of novel methods to facilitate secure access to such data where there is a risk of disclosure of identities.*

Recommendation 4.2: ***leading international agencies***
(e.g. World Bank Group, World Health Organisation, International Labour Office and Organisation for Economic Cooperation and Development) should collaborate in the formulation of a strategic approach towards the removal of obstacles to improved access to and sharing of micro-data in their possession and should provide a coordinated plan for the creation of data discovery and data management tools on their websites.

Challenge 5

the drive to address what is increasingly an interdisciplinary and international research agenda requires the use of existing capacity to its full potential.

Finding: data resources and capacity to analyse data exist separately in national official statistical agencies and the research community, both within and across countries.

> Recommendation 5: *mechanisms should be established which build upon and enhance further the efforts being made by producers of data (e.g. official data producing agencies, businesses, researchers) and the users of data (e.g. researchers, policy-makers) to share expertise, knowledge and resources, particularly in the areas of data access, linkage and integration and analysis.*

Challenge 6

comparative research in the social sciences, based upon data pertaining to different countries and regions, is an essential part of the research process and is set to become increasingly important. Without collaboration and sharing of experience between countries in the development of comparable data resources, the full benefits from comparative research will not be achieved.

Finding: there are few opportunities for data producers in different countries to share their knowledge and experience about data harmonisation across disciplinary domains and for various types of data.

> Recommendation 6.1: ***national and international statistical agencies*** *should strengthen the efforts they are making to harmonise social and economic data at the international level, seeking to prioritise these activities in specific areas.*

> Recommendation 6.2: ***researchers*** *involved in the creation and maintenance of data resources designed specifically for comparative research and* ***those funding such research*** *should collaborate to provide mechanisms to foster an integrated approach to data design and harmonisation, access and sharing as stated in the OECD (2007) Council Recommendation concerning Access to Research Data from Public Funding.*

Challenge *7*

researchers have a responsibility to ensure that they have the skills and the resources necessary to ensure that the data they use for research are available for re-use and that plans are made to this effect before they engage in research.

Finding: only a small number of national funding agencies have requirements for researchers to make data management plans to accompany their applications for research funds.

> Recommendation 7: ***national funding agencies*** *should ensure that new research awards have accompanying data management plans and should assign resources for this purpose. They should cooperate at both national and international levels to share this information, publishing details about data and metadata to be created and plans for their preservation in formats that make this information readily accessible to the research community. The international community should agree on a standard semantic dictionary describing common elements of a data management plan to facilitate the discovery and use of the information contained in such plans.*

Challenge *8*

not all countries have invested in the skills, resources or infrastructure required to curate important datasets. This places such data at risk of loss, minimises the potential for their re-use and precludes their inclusion in an evolving global data ecosystem.

Finding: established social science data archives in a small number of countries have substantial expertise in archiving and curating datasets.

> Recommendation 8: ***social science research communities*** *in countries without institutional support for data curation or supporting infrastructure should conduct an assessment of their national needs and assets in this area that will contribute to national plans of action. Working with researchers in such countries, established social science data archives should assist them by developing an assessment instrument and providing expert advice in preparing plans.*

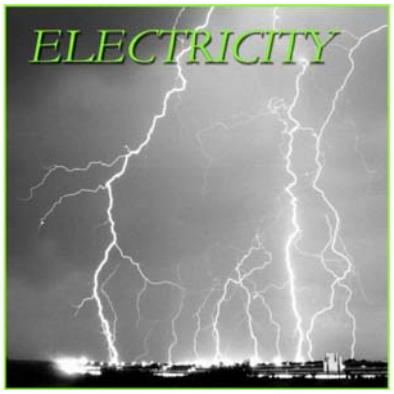
Challenge 9

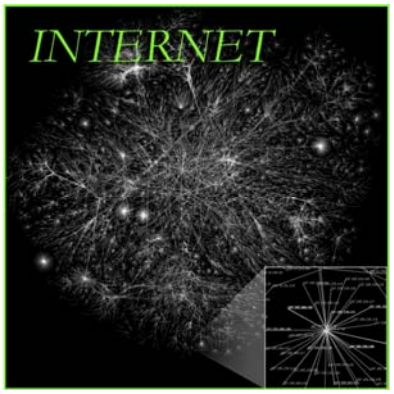
data sharing, including the creation of appropriate metadata to international standards is fundamental to the process of scientific enquiry. Researchers need incentives to ensure effective data sharing.

Finding: presently there are few incentives to encourage researchers to manage, maintain, archive and share data resulting from their research. Without clear incentives it is unlikely that the benefits of international collaboration will be fully realised.

> Recommendation 9.1: ***research funding agencies*** *should collaborate at the international level to ensure that a common system is adopted for referencing datasets in research publications. They should also ensure that the intellectual effort required for the creation and sharing of data is recognised in their evaluation of research activities.*
>
> Recommendation 9.2: ***publishers of research*** *should be encouraged to adopt guidelines for publications, stipulating that a common and internationally agreed referencing system for datasets is used within all scientific publications that have made use of data.*
>
> Recommendation 9.3: *the* ***employers of researchers*** *should recognise the intellectual efforts that have been made by researchers who generate significant data resources. This could be reflected via merit awards, promotions and other ways which acknowledge the professional contributions that have been made.*

Given the growing importance which now attaches to the value of social science data in its many forms, not just for research in the social sciences but for a wider and more multidisciplinary international research agenda, the recommendations made in this report should be pursued vigorously. The recommendations presented here form a coherent whole and will require global coordination to encourage and monitor their implementation. While specific agencies (research funding bodies, research groups, national statistical agencies and international statistical bodies) can point to actions they have initiated or are planning to undertake and which align with particular recommendations, this alone will not provide the impetus to foster and progress an international policy-relevant research agenda designed to improve understanding of the changing nature of the human condition.





1. Introduction

Three great advances in technology brought with them major changes in the structure of the global economy and societies. The harnessing of steam power distinguished the first; electricity generation and distribution the second; and global information connectivity the third. In common with these earlier technologies, the so-called 'web revolution' and its facilitating infrastructure has diffused rapidly throughout societies, brings changes which touch upon many aspects of human activity and comes with an associated surge in innovation leading to new products and services that were barely imaginable prior to technological adoption. This latest advance, now just twenty years from its inception, has important repercussions for the process of scientific enquiry, specifically upon the nature and volume of research data, methods of scientific investigation and the scope for international research collaboration.

The data issues addressed in this document are not simply of either academic or parochial interest. The recent turmoil in the Middle East was unanticipated and poorly understood at least partly because data on the new ways in which humans communicate were not well collected or analysed. Future capacity to predict and respond to financial crises will at least partially depend on the ability to collect, process and make sense rapidly of large quantities of financial information. The slow moving crisis of climate change requires integration of data on human beings and their interaction with their environment. The consequences of sudden disasters such as tsunamis and terrorist attacks are exacerbated by the lack of understanding of how humans process information to coordinate responses. Inadequate planning and lack of action about research data on the human condition have real and important economic and social consequences.

1.1 The challenges to be faced

The challenges are twofold. The first is to ensure that there is coordination of efforts being made in different parts of the world to develop access to all forms of research data and to capture the potential gains from research use of new forms of data. Much relevant expertise is vested in the public, private and research communities. Many significant activities are underway across the world to make research data easier to find, to improve access for research purposes and to harmonise different types of data in ways that make them more amenable for comparison. However, without efforts to coordinate these activities and to promote them in those areas where a lack of resources and skills hinders such developments, research on issues of global importance will be fettered.

The second derives from the fact that social science data are often derived from or relate to living persons or existing organisations, raising legal, ethical, confidentiality and privacy questions that can impede international research. In many countries these questions and associated issues about disclosure of identities hinder research access to micro-data. But some countries are now finding ways to overcome such access problems. The challenge is to ensure that technology is used in ways that encourage the spread of best practice access arrangements, while ensuring that legal safeguards are adhered to and that the risk of inadvertent disclosure of identities is minimised.

This report documents these core challenges for the identification, collection, management, processing and sharing of data on the human condition[2]. It provides a set of recommendations for leveraging the international and interdisciplinary research agenda and facilitating cross-national collaboration. Some of these recommendations propose to build upon existing activities in ways which will enhance global collaboration in social scientific research activities. Others identify gaps in areas which will have to be addressed, but require a longer time frame, concerted action by various bodies and more resources. For this reason the Expert Group will submit an action plan to the Global Science Forum, identifying the 'quick wins' that can be gained from coordination of activities over the next two years, while setting out the mechanisms for achieving the longer term recommendations.

[2] The scientific disciplines that cover research on the human condition include economics, sociology, law, human geography, demography, political science, social psychology, anthropology, city and regional planning, public policy, public health and the environmental sciences.

1.2 Plan of the report

The plan of the report is as follows. Section two defines some terms and describes this 'science-driven' approach to the international social science agenda. Given the sheer volume and variety of data that have research potential, how should priorities for improvements in data discovery, access and sharing be determined? The section outlines the need for an approach to this issue that is driven not by the availability of data but by the science that relevant data can inform and the value of their information content.

Section three forms the core of this report. It presents information about some of the important developments that are facilitating an international approach to social scientific research, highlighting where the Expert Group would like to see further efforts made to address the challenges posed by the increasing availability of data with research potential and recommend actions to grasp the opportunities these challenges present. These are directed at specific bodies or agencies that have the remit and the resources to take what are identified as the necessary actions.

This report makes recommendations to international organisations, national statistical offices, funding agencies, researchers and other research bodies to address progress in these areas. None of these recommendations will be easy to implement. There are no 'low hanging fruits' to be picked. For this reason, section four concludes by giving consideration to the ways in which these recommendations should be taken forward.

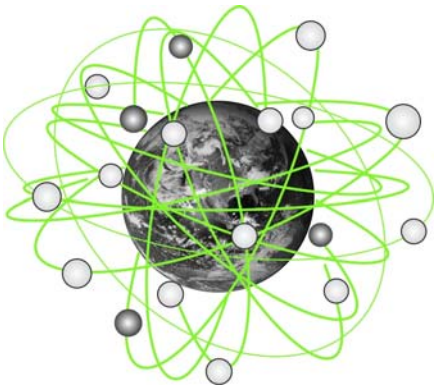
1.3 The Expert Group and its Terms of Reference

Details of the membership of the Expert Group and its Terms of Reference are included in *Appendix One*. The first of these terms of reference was

> *'to review and advise on the major data series which ideally would be available for research from every country, or initially at least from OECD countries, either through survey or administrative data collection, which can be collected or combined to standardised formats and deposited, codified and made accessible for truly international comparative purposes'.*

The Expert Group agreed at an early stage that it was infeasible to prepare such a list, simply because the preconditions for its fulfilment had not been met. While there are certain areas where considerable work has been undertaken to harmonise data, fundamental problems relating to their discoverability, management, access and comparability remain. Accordingly, the group focused attention on the identification of these problems and made recommendations for their solution. The second and third terms of reference required the Expert Group to consider the research potential of new forms of data, developments in technology and methods for the realisation of this potential and the ethical issues associated with this. Separate subsections are devoted to these issues in section three.

2. The advantages of a science driven approach to the international social science data agenda



2.1 Background

In formulating the recommendations made in this document, a number of studies and reports have been especially influential. These include: *Harnessing the Power of Digital Data for Science and Society* (2009); *Riding the Wave: How Europe can gain from the rising tide of scientific data* (2010); *Big Data: The next frontier for innovation, competition, and productivity* (2011); and *Science as an Open Enterprise* (2012). These reports have a wider remit than the present focus. They consider the need for international collaboration for the technical development of the infrastructure that underlies the web revolution, new ways to assess the research value of data, the training of the next generation of data scientists, management of the so-called 'data deluge', and the transparency of scientific research. The focus of this report is on the need to build on and leverage existing expertise about access to and use of data to inform the expanding international research agenda on the human condition.

The social sciences play an important role not only in terms of how they contribute to a better understanding of the human condition, particularly the nature and consequences of the decisions and actions undertaken by individuals, or in social and political groups, but also as a bridge between knowledge about human behaviour and the design and implementation of public policies focused on a wide range of issues. Increasingly, the social sciences partner with other disciplines to address far-reaching questions about the state of the world economy, factors affecting the health and wellbeing of the world's population, and the human dimensions of

global change. Social scientific investigation, combining the analysis of appropriate forms of data with methods for their analysis, contributes to scientific discovery in many disciplines.

Box 1 describes the essential nature of scientific inquiry and defines the role of data in relation to information and knowledge creation. Basic to this process are the concepts of transparency and replicability. The accessibility of data, its organisation and management, its quality and its preservation for re-use are essential components of scientific progress. Section three of this report expands upon these issues within an international context.

Box 1

Scientific inquiry, data, information and knowledge

Open inquiry has been the essential engine of scientific progress. Published communication of scientific theories and of the experimental and observational data on which they are based has permitted others to scrutinise them, to test the replicability of experiments and observations, to support, reject or refine theories, and to re-use data to create further understanding. It has permitted error to be identified and new hypotheses to be created.

Scientific knowledge is built up incrementally. Data from human or machine observation are numbers, characters or images that refer to an attribute of a phenomenon. They become information when they are interpreted and combined together in ways that have the potential to reveal patterns in the phenomenon. This becomes knowledge when the information leads logically to an understanding of the phenomenon. Data are the bedrock on which scientific knowledge is built. Its accessibility to scrutiny by others than the originators is essential to the progress of science.

But disclosure of data has little value in itself. It must be communicated effectively, which means that it must be accessible, so that it can be readily located. It must be intelligible to those who wish to scrutinise it. It must be assessable so that judgements can be made about its reliability and the competence of those who created it. And it must be usable so that it can be re-used, which requires the provision of appropriate metadata (data about the data).

Reproduced (in edited form) from 'Science as an Open Enterprise' The Royal Society 2012.

2.2 Defining types of data on the human condition

Traditionally, research data in the social sciences have been specifically designed for that purpose. For example, social survey data have been used to describe and monitor people's ideas and behaviours since the middle of the last century. Population censuses provide almost total coverage of national populations and include information of interest to demographers, economists, sociologists, planners and businesses. Increasingly, however, forms of social science data not specifically designed for research purposes are emerging as important alternatives and additions to more standard sources. Various types of administrative data, while not new, have become newly accessible in the form of electronic records, while entirely new forms of social science data have emerged as a consequence of the internet revolution. Box 2 gives examples of the wide variety of data that are now becoming more available or are potentially available for research purposes. Six categories of data are identified. These are:

Category A: Data stemming from the transactions of government, for example, tax and social security systems.

Category B: Data describing official registration or licensing requirements.

Category C: Commercial transactions made by individuals and organisations.

Category D: Internet data, deriving from search and social networking activities.

Category E: Tracking data, monitoring the movement of individuals or physical objects subject to movement by humans.

Category F: Image data, particularly aerial and satellite images but including land-based video images.

What all of these categories have in common is the fact that the data concerned are in digital formats, making them potentially more discoverable, accessible and useful for research than has hitherto been the case. Names such as 'administrative data', 'transactions data' and 'new forms of data' are used in this report. Categories A, B and C cover the first two groups. Categories D, E and F constitute what are termed 'new forms of data'. Further details about these categories and examples are shown in Box 2.

2.3 International research in the social sciences and research data

As the scope, detail, and availability of social science data expand, the degree to which they can inform research and policy also increases, especially when joined with data from other disciplines. Four broad areas are identified where there are growing international research challenges which will drive the need for a wide variety of types of research data.

Box 2

Forms of data with new research potential

Broad category of data	Detailed categories	Examples
Category A: Government transactions	Individual tax records	Income tax; tax credits
	Corporate tax records	Corporation tax; sales; tax, value added tax
	Property tax records	Tax on sales of property; tax on value of property
	Social security payments	State pensions; hardship payments; unemployment benefits; child benefits
	Import/export records	Border control records; import/export licensing records
Category B: Government and other registration records	Housing and land use registers	Registers of ownership
	Educational registers	School inspections; pupil results
	Criminal justice registers	Police records; court records
	Social security registers	Registers of eligible persons
	Electoral registers	Voter registration records
	Employment registers	Employer census records: registers of persons joining/leaving employment
	Population registers	Births; marriages; civil unions; deaths; immigration/emigration records; census records
	Health system registers	Personal medical records; hospital records
	Vehicle/driver registers	Driver licence registers; vehicle licence registers
	Membership registers	Political parties; charities; clubs
Category C: Commercial transactions	Store cards	Supermarket loyalty cards
	Customer accounts	Utilities; financial institutions; mobile phone usage
	Other customer records	Product purchases; service agreements
Category D: Internet usage	Search terms	Google©; Bing©; Yahoo© search activity
	Website interactions	Visit statistics; user generated content
	Downloads	Music; films; TV
	Social networks	Facebook©; Twitter©; LinkedIn©
	Blogs; news sites	Reddit
Category E: Tracking data	CCTV images	Security/safety camera recordings
	Traffic sensors	Vehicle tracking records; vehicle movement records
	Mobile phone locations: GPS data	
Category F: Satellite and aerial imagery	Visible light spectrum	Google Earth©
	Night-time visible radiation	Landsat
	Infrared; radar mapping	

- Population dynamics and societal change – migration, ageing populations, population growth, welfare/wellbeing.
- Public health risks – spread of communicable diseases, lifestyle factors and non-communicable diseases.
- Economic growth, innovation, research and development activity – science, education, workforce of the future, global trade and financial stability.
- Social and environmental vulnerability and resilience– environmental change, dynamics of poverty and related policy evaluations.

Across the globe there is an interest in the factors driving population change. In Europe and many countries in Asia, migration is the major determinant of population size and characteristics. Migration is best understood in relation to conditions and policies of both origin and host countries as well as the networks that link them. Ageing processes and the changing age structure of populations are also of great interest. A full understanding of these processes requires information about the social and economic situation of elderly populations combined with health data and information about the built environment, the availability of various services, *etc.* Knowledge about the welfare and wellbeing of human populations is critically bound up with the investigation of population dynamics. Comparisons between countries provide important information about the ways that these processes and their consequences might be addressed.

The second area requires cooperation between biomedical and public health scientists and social scientists. Long-term longitudinal studies of individuals and families provide much-needed data for scientific investigation in areas relating to non-communicable diseases. When linked to information about the places where people live and work, it is possible to trace exposures to poverty, as well as harmful and/or beneficial environments, as people move through their lives. In the study of transmission mechanisms for communicable diseases, social science provides the framework for modelling social interaction as well as long- and short-distance population movements critical to their spread.

In the third of these areas, economics and business studies are key disciplines that underlie the research interests. Until recently many economists relied on aggregate indicators of economic, scientific and financial performance. Now there is a growing demand for micro-data – for science and technology, the demand is for up-to-date longitudinal information about who is engaged in research and development, how they are funded, and what their activities are generating. For economic growth, there is increasing realization of the importance of building datasets that capture the structure of innovative organisations, as well as detailed information about the scientists, innovators and entrepreneurs that populate those organisations. Unlike aggregated data, micro-data on individuals or organisations retain all the original variation that is critical for modelling human behaviour. Where these exist the high demand for access indicates their research value.

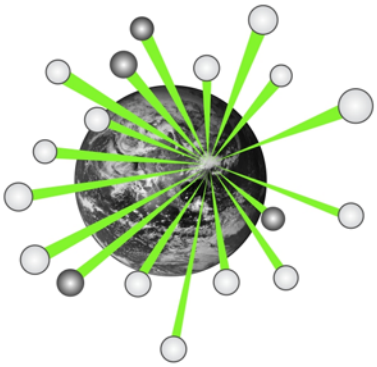
The fourth area combines the interests of sociologists, economists, environmental scientists and legal expertise in policy evaluation. Leading research in this area takes advantage of large household-based longitudinal studies and long-term longitudinal studies of different population age groups, combining these with spatially-referenced information mapping various types of vulnerabilities. Linking such information to administrative data on incomes, taxation and social welfare benefits provides the detail required for micro simulation modelling of social and economic interventions.

An international research agenda is, in many of these areas, a fundamental factor driving the need to identify, gain access to and analyse data from more than one country. But there are other reasons over and above the transnational nature of the phenomena under investigation which favour an international approach. Combining datasets across countries can make it possible to study rare groups, occurrences or combinations of characteristics that would not be possible with data from one country alone. Also, the emergence of multinational organisations weakens the relevance of national boundaries for the production of data. This is especially the case for data that are generated via or related to global communications networks.

In all of these areas, some basic data requirements can be discerned.

- Data are most useful for research purposes where they are available as micro-data, providing information pertaining to individuals and/or organisations.
- Repeat observations on the same unit of observation are required in order to study processes of change.
- Data need to be up-to-date. In a changing world, and with the new capacity to collaborate across nations and in multidisciplinary teams, there is an imperative to study phenomena as they emerge.
- Data must be comparable across cultures, languages and environments. The concepts used to implement and communicate data at the international level should derive from universally recognised methods and standards.
- Social science data infrastructure significantly extends the capacity to accumulate knowledge from existing sources, and extend its relevance, but also it will increase the potential impact on the knowledge base from well organised new sources through identifying the most relevant areas for harmonisation.

Individual country decisions about widening access to micro-data are likely to involve small incremental changes from long standing practice unless the experience base of countries with a variety of practices is drawn on. The growth in cross country comparative studies that require sharing of micro-data in well trusted contexts should increase the interest and leverage of researchers in all countries to have access to comparable microdata.



3. International data collaboration: the key challenges

3.1 Understanding the research potential of new forms of data

> ***Challenge 1:*** *massive amounts of digital data are being generated at unprecedented scales and velocity, much of it from new sources such as the internet. The reliability, statistical validity and generalisability of new forms of data are not well understood. This means that the validity of research based on such data may be open to question.*

Traditional forms of data in the social and economic sciences rely on survey techniques associated with representative samples of populations (of households, business organisations, *etc.*) or census methods. New forms of data arise from the growth of electronic communication, from internet usage, the growth of web-based services and from the expansion of the digital economy. Such data have potential research value and will become ever more pervasive.

New forms of data are likely to become increasingly important for international and comparative research as response rates associated with traditional and relatively expensive survey-based data collection methods continue to decline. Section two has identified six broad categories of data which have research potential but are not specifically designed with their use for research purposes in mind. Some are not 'new' in the sense that they may have been in existence as electronic records (or, at least, in some kind of machine-readable format) for more than 60 years.

Administrative data are included in this definition. While some countries are well advanced in terms of the ways in which they use such data for research, many countries have yet to make such data readily available as research resources. The personal nature of such microdata and legal restrictions on their re-use can hamper collaborative research efforts.

While these examples serve to indicate the wide range of data types with research potential, a number of important points need to be made:

- Data from different sources (surveys and censuses, administrative and internet-produced data) can enhance one another. Their complementarity still has to be explored, on the one hand to study the degree to and ways in which specific groups in the population use online services or participate in actions which generate administrative data, and on the other hand for the methodological purpose of efficient and optimal collection of representative and detailed information about social processes and trends.
- The internet crosses national boundaries; many transactions and communications do the same, and the large companies collecting these data are often multinational concerns. Legal regulations concerning property rights, rights of use, privacy, and confidentiality are, however, nationally based. This leads to confusion for those collecting, archiving and using these data.
- These legal issues have parallels and extensions in ethical considerations. The novel nature of online communications and transactions is leading to new and as yet hardly charted ethical questions that are to some extent related to legal issues but not totally covered by them.
- Formats of digital data are much more complicated and diverse than the traditional format of a data set consisting of a number of attributes of various units in some population. Communications and transactions involve two or more units, and they can be represented by networks or more complicated structures. This leads to the necessity of using non-traditional methods for data management and data analysis, an area in which much methodological and skills development is needed.
- Hardly any experience exists on the combined collection and analysis of survey/census, administrative and new forms of data such as those derived from internet activities, and cross-national or internationally comparative dimensions further compound this lack of knowledge. There is a need to stimulate further international collaborative research efforts in this domain, focusing on research questions that demonstrate the social importance of developments connected with online communications and transactions, elaborating the possibilities and charting the difficulties.

New forms of data bring considerable opportunities to have more frequent reporting, serial recording and rich geographic identification. The very much greater volume and velocity of information will need to be complemented by gains in methodology to assess and analyse these data and by new information management practices, including visualisation methods. It is

expected that there will be a much faster rate of change in access to public and commercial records in the future than in previous decades. Plans to explore and capitalise on this potential will benefit from cross-national co-ordination and collaboration.

Why is this important and what needs to be done?

Traditional forms of data collection are often expensive and response rates are increasingly a challenge. New forms of data may complement these methods, holding out the promise, yet to be established, of cheaper and faster ways of collecting data for research. Funding agencies, research institutes and interested companies should take initiatives enabling pilot projects (Box 3, for an example of this approach). This is already being done in a few countries, e.g. the Digging into Data call (Box 4); the archiving of Twitter by the US Library of Congress; and the concept of 'place-based' forms of data collection, but should be extended to focus specifically upon the research value of such data and the development of methods to make use of them.

Box 3

Linking Surveys to the World: Administrative Data, the Web, and Beyond

This project, recently funded within a joint research programme administered by the US National Science Foundation and the US Census Bureau (the NSF Census Research Network) will develop and evaluate methodologies that use the vast variety of data generated by households and businesses in the course of their ordinary activities. The project will examine administrative records created by businesses, individuals, and governments, streams of data from social media sites on the World Wide Web, and detailed geospatial data. The project will analyse these multiple sources of data and relate them to data collected on surveys. It aims to improve survey measurements of economic and demographic data and potentially supplement or replace surveys with statistics based on administrative, Web-based, and geospatial data. Applications of these approaches include the following: using linked survey-administrative data to assess attrition, selective non-response, and measurement error in surveys; using Web-based social media to measure job loss, job creation, small business creation, and informal economic activity; using administrative geo-spatial data to enhance small-area estimates; and training in the use and creation of linked survey-administrative datasets.
Edited from Project Abstract shown at:
www.nsf.gov/awardsearch/showAward.do?AwardNumber=1131500&WT.z_pims_id=503587

Box 4

Digging into Data

The Digging into Data Challenge is an international competition sponsored by major research funders from Canada, the Netherlands, the United Kingdom, and the United States. The competition funds interdisciplinary teams of scholars and scientists who are studying how "big data" changes the research landscape for the humanities and social sciences. Now that massive databases of materials are used by scholars in the humanities and social sciences -- ranging from digitised books, newspapers and music to transactional data like web searches, sensor data or cell phone records -- what new, computationally-based research methods might be applied? As the world becomes increasingly digital, new techniques will be needed to search, analyse, and understand these everyday materials. Digging into Data challenges the research community to help create the new research infrastructure for 21st century scholarship.

An example project funded under this competition is:

Mining Microdata: Economic Opportunity and Spatial Mobility in Britain, Canada and the United States, 1850-1911

This project will make use of novel data-mining technology to exploit one of the largest population databases in the world, a vast collection of harmonised 19th and early 20th century census microdata from Britain, Canada, and the United States originally digitised for genealogical research. The goal is to shed light on the impact of economic opportunity and spatial mobility on social structure in Europe and North America.

Recommendation 1: ***national research funding agencies*** *should collaborate internationally to provide resources for researchers to assess the research potential and to develop new methods to understand the opportunities and limitations offered by new forms of data to address important research areas.*

3.2 Discovering data

> ***Challenge 2:*** *while many countries have vast amounts of more traditional forms of administrative, survey, and census data collected by and held by national statistical agencies and government departments, knowledge about the existence of such data as micro-data records is a precondition for the efficient and effective planning of international research.*

Data discovery describes the processes that are required from data producers and those responsible for the curation of data, to make information pertaining to specific datasets widely available. Data discovery does not necessarily imply that such data are readily accessible, but places a researcher in a more knowledgeable position regarding the potential availability of data.

It is tempting to assume that the power of internet search engines and the sheer scale of the information that is searchable in this way make it easy to locate data that have relevance for a specific research purpose. While internet search engines can return enormous amounts of information, the provenance of this information is difficult to ascertain and there is no guarantee of its completeness. Where data are found in this manner, how were these data created, and by whom? What do they measure and can they be trusted? More importantly, though, some data which could be highly relevant for a specific research purpose may not be discoverable via web search methods because no information about them has been placed on an accessible website. Limitations on access and inadequate archiving increase the likelihood that data will not be re-used simply because other researchers are unaware of their existence. This can lead to situations where scarce research sources will be used to reproduce what is already known somewhere, and also limit the potential leverage of new sources in increasing the value of the existing knowledge base.

In examining the issue of data discovery, it is useful to characterise the way data originate. Four broad categories of producers of traditional forms of social and economic research data can be identified.

- National statistical agencies and bureaus, mandated by their national governments to collect information primarily for policy purposes.
- Researchers, either solely or more usually working in national or international teams, who collect data for general or specific research purposes.
- Research bodies funded specifically to create data from groups of countries for comparative research purposes.
- International organisations funding common data collection efforts in specific countries for policy monitoring purposes.

National statistical agencies publish much of their statistical output on their websites, but these tend to be aggregated data relating to specific groups or classes of individuals or organisations and organised in tabular form or as accounts. Discovering the nature of the underlying micro-data and its accessibility is often difficult unless there exists some form of agreement and/or protocol whereby such micro-data are archived and the information about them and conditions of access are published.

Researchers, funded to collect data via public research funding agencies and often working in national or international teams, are more likely to promote the discoverability of data they have created. Problems may exist if research groups are unwilling to alert others to the fact that they have created specific research data, an issue that research funding agencies can help avoid by ensuring that the researchers working on research that they fund are obliged to share data with others.

Research groups which work specifically to create comparative data are most likely to encourage further use of their data and to promote data sharing. Nonetheless, it is surprising to find that knowledge about such data resources is often confined to 'disciplinary-based' groups of researchers.

The international organisations have, until fairly recently, been more concerned with efforts to collect data for their specific policy needs than to prepare such data for sharing and re-use. This situation is now changing. Strong efforts are being made by organisations such as the World Bank, the International Labour Office and the World Health Organisation to make available for re-use data for which they have responsibility. Some have taken steps to ensure that micro-data they have previously collected are placed in the public domain (see Box 5). But how should they do this, and how might such data producers generally act to make their data discoverable?

Box 5

The World Bank and the Open Data Initiative

The International Household Survey Network (www.ihsn.org)), and the World Bank through its Open Data initiative, work closely with the international research community at universities, leading international data archives and other international organizations to develop and promote standards, software and guidelines for best practice in microdata curation and dissemination.

The World Bank, in partnership with international and regional organisations, is providing financial and technical support to the implementation of these standards (in particular the Data Documentation Initiative – www.ddialliance.org) and good practices by national statistics agencies in over 60 developing countries. This has resulted in large improvements in the way in which survey data are curated, preserved and disseminated in (and by) participating countries. The World Bank and IHSN have developed an open source web application which the World Bank (http://microdata.worldbank.org) and countries are using to establish microdata libraries and to make their survey data more discoverable and increasingly accessible to the research community.

Metadata (information about the way data are generated) play an important role in making data discoverable. By developing agreed procedures whereby data can be described, placing such information within a standardised web mark-up language (e.g. html, xml), data about data can be transferred between hosting websites, and can be interpreted and understood by machines. This transforms data discovery from an error-prone process of luck (or misfortune) to a systematic process similar in nature to the cataloguing of physical books in a library but with the added richness of detailed content descriptions.

A vital part of the research process stems from publication of research findings. This can stimulate the interests of others to attempt to replicate research results, to extend and improve the research methods and to seek alternative types of data from different sources to determine the robustness of research findings. By adopting a standard approach to the way in which the data sources are referenced in publications, this, in turn, can help others to locate data for their research. This issue is addressed in section 3.8 which recommends ways in which incentives for data sharing can be improved.

Why is this important and what needs to be done?

Knowledge about the existence of micro-data with research potential is an essential part of the process of planning and undertaking research. Such knowledge can lead to efficiency gains in the research process. If data that already exist can be located, significant costs in time and money can be saved through their re-use. Data discovery can also lead to research discovery as paths are traced into earlier uses of existing data. The use of agreed metadata mark-up protocols vastly enhances the prospect for data discovery. What is needed is widespread agreement among data producers to adopt such standards and to place the resulting information in locations where it can be 'harvested' by the major data archives. Agreements between researchers and those who publish their research findings, to cite the sources of data they have used and to make these sources accessible to others, not only enhances the 'discoverability' of data but improves the research process.

Recommendation 2: ***national statistical organisations and international organisations*** *should ensure that all data they collect and process are documented to agreed and common standards. Such documentation should be easily discoverable on their websites.*

3.3 Legal and ethical issues arising from research use of new forms of data

Challenge 3: *New forms of personal data, such as social networking data, can provide insights into the human condition. However, the use of those data as research resources may pose risks to individuals' privacy, particularly in case of inadvertent disclosure of the identities of the individuals concerned. There is a need for greater transparency in the research use of new forms of data, maximising the gains in knowledge derived from such data while minimising the risks to individuals' privacy, seeking to retain public confidence in scientific research which makes use of new forms of data.*

Whilst developments in science and technology often run ahead of the ability to think through their ethical implications, the rate of change is now accelerating. Over the next few years the volumes of data used in research will increase and better, more powerful, computational methods for processing these data will be developed. Through these developments more detailed studies of the behaviours and interactions between people and how they work and collaborate will become feasible.

This view of a data-rich environment in which the complexities of human behaviour can be explored poses both legal and ethical issues. For example, the integration of multiple sources of data may increase the potential to identify individuals either directly or through deduction, based on combinations of characteristics including their association with others (see Box 6).

Social scientists have developed thoughtful approaches to the protection of more traditional forms of social science data, which can be leveraged to guide use of newer forms of data also. These range from data releases that reduce the level of detail provided about individuals, to restricted use contracts, or to a requirement that use should only occur in and be carefully monitored by special secure data facilities. At the same time, advances in computer science suggest that there may be novel ways in which researchers can benefit from the utility of new forms of data without endangering the privacy of the individuals concerned.

Sheer computational power applied to voluminous data sources will not answer the questions about the rights and wrongs of particular forms of research and the associated methods used. For example, studies that draw on data from online systems such as Facebook and Twitter are generating specific concerns. Whilst it is recognised that these online systems provide unprecedented access to data about what people think, what they do and when, there are issues relating to ethics that are presently muddled. The key issue here is that the ethics of online studies are currently confused and the consequences of harvesting this material are as yet unknown.

Box 6

OECD Guidelines on Protection of Privacy and Personal Data

The OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data (1980) define personal data as "any information relating to an identified or identifiable individual". Many of the types of data on the human condition outlined above (section 2.2) will be considered personal data according to this definition. The re-use of those data for research or other purposes may be subject to restrictions contained in national privacy laws.

Many national privacy laws authorise the re-use of personal data for historical, statistical or scientific purposes, provided appropriate safeguards are put in place. Traditionally, such safeguards have involved the "anonymisation" or "de-identification" of data, with a view of concealing the identities of the individuals concerned. Over the past decade, however, it has become clear that not all anonymisation and de-identification techniques are equally robust. As a result, the use of these techniques to eliminate privacy risks is increasingly questioned (OECD 2011).

Why is this important and what needs to be done?

The 'web explosion' and the associated growth of new forms of data with research potential positions social scientists as the key group qualified to ensure that such data are used in ways which are seen as both beneficial and non-intrusive into personal freedoms. It is imperative that new and internationally recognised codes of conduct be established to prevent the potential abuse of personal information and to ensure that public support for research which takes advantage of new forms of data is not eroded as a result.

At a strategic level, there is a need for a fundamental reassessment by national statistical agencies and by researchers of their views on ethics and privacy in light of the changes wrought by the Internet. The current model is largely focused on limiting the identifiability of the individuals concerned, which may prove difficult to sustain over time. Furthermore, most research is of minimal substantive concern to people, suggesting the need for greater flexibility in allowing new forms of data to be re-used for research purposes. The future model of ethical data use might focus more on transparency around how data is being used, rather than preventing specific types of use.

Recommendation 3: ***research funding agencies and data protection authorities*** *should collaborate to develop an internationally recognized framework code of conduct covering the use of new forms of personal data, particularly those generated via network communication. This framework, built on best practice procedures for consent from data subjects, data sharing and re-use, anonymisation methods, etc., could be adapted as necessary for specific national circumstances.*

3.4 Issues in access to social science data

Challenge 4: *barriers to access to social science data hinder both national and cross-national collaboration which could exploit their research value. These barriers relate to a variety of obstacles (legal, cultural, language, proprietary rights of access) all of which have to be identified and removed if cross-national research is to be promoted.*

A lack of access to research data clearly inhibits data sharing. Data may be made discoverable through common referencing systems in research publications, or through the existence of data management plans (see section 3.7.1), but if access for re-use is not granted or is limited in some way, research plans based around such data will either fail or will become less efficient.

This section considers the main obstacles to access and examines how these are being avoided or circumvented in different countries. Again, the focus of attention is on access at the international rather than the national level.

Unlike many of the data required for research in the physical sciences, information required by social scientists often relates to people or organisations that are potentially identifiable in micro datasets. This gives rise to problems of access relating to the nature of the consent(s) that data subjects may have given for use of their personal information; the need to protect and keep secure data that could reveal identities; and the problems relating to the sharing of such data across different legal jurisdictions. Further access issues derive from questions of ownership, particularly where data have commercial value, intellectual property rights and the desire of some research groups to retain, beyond a reasonable period, proprietary rights of access to data they have created.

Access problems and issues at the international level can be resolved in a number of ways. First, it may be the case that the desire to create high quality comparable data causes research groups and their funders to collaborate in the creation of internationally shared data resources. Examples of these are given in section 3.6, and include national census data collections (see Box 7), longitudinal studies of ageing, demographic health surveys and international social

survey programmes. Second, international funding agencies and other international organisations may work to set up major programmes in which the government agencies and national statistical agencies cooperate to share microdata resources. An example of this approach is shown in Box 8. A third approach is where national funding agencies, data producers and social scientists have set up national bodies to help resolve national problems of access to data, then these bodies link in a cross-national collaboration to share their solutions.

Box 7

IPUMS: Integrated Public Use Microdata Series

IPUMS-International (www.ipums.org/international) disseminates extracts of integrated census samples to researchers and students world-wide, free-of-charge. Official statistical agencies of 99 countries encompassing 87% of the world's population have embraced a uniform set of protocols to facilitate access from the IPUMS website.

Currently (January 2012), 185 samples are accessible totalling 400 million person records for the period 1960-2010. Each year, 20-25 new samples are integrated into the database with recent censuses having the highest priority. Integration not only facilitates comparative research but it also adds value by means of "pointer" and other variables unique to IPUMS.

Users get data by means of a series of point-and-click menus to select countries, censuses, variables, and even sub-populations. The resulting extract is then downloaded for analysis using SPSS, SAS, STATA or other statistical software. Extensive custom-tailored integrated metadata are also downloadable—in text, image and DDI form.

The IPUMS collaboratory is led by the University of Minnesota Population Center in cooperation with leading academic and statistical institutions. Sustained funding is provided by the National Science Foundation and National Institutes of Health (USA).

Box 8

Partnership in Statistics for Development in the 21st Century

*The **Partnership in Statistics for Development in the 21st Century (PARIS21)** is a global partnership of national, regional, and international statisticians, analysts, policy-makers, development professionals and other users of statistics. It is supported by a Secretariat hosted within the OECD's Development Co-operation Directorate.*

PARIS21's goal is to develop a culture of Management for Development Results.

PARIS21 pursues this goal primarily by encouraging and assisting low-income and lower middle income countries to design, implement, and monitor a National Strategy for the Development of Statistics (NSDS).

The Partnership's activities also include facilitating the co-ordination of stakeholders to better address an evolving agenda, advocating for increased involvement of national stakeholders in statistical development and enhancing the status of statistics in major international initiatives, and stimulating increased demand for and better use of data.

The latter work stream involves the co-management (with the World Bank and other partners) of the International Household Survey Network (IHSN) and the implementation of the Accelerated Data Program (ADP).

The ADP helps national statistical offices in about 70 developing countries improve data access, with a focus on properly documented household survey micro-data. Outputs of the ADP include the publishing of 25 web-based national survey catalogues to facilitate access by researchers to those datasets.

For further details see: http://www.paris21.org/

Some countries (*e.g.* UK, Germany) have established institutions through which representatives of the research community, official statisticians and research funders meet to highlight developments of mutual interest and to outline potential problems associated with access to official statistics for research (see Box 9). Where such bodies exist, they could act to coordinate their agenda and encourage the exchange of knowledge at the cross-national level. The EC Global Research Data Infrastructure 2020 project is an example of an international initiative to develop mechanisms of data exchange across a variety of organisations. Data without Boundaries is another example (see Box 10).

Whatever mechanisms are employed, the essential point is that these communities should be brought closer at both national and international levels to develop and implement standards for data and metadata and to share methodological insights, particularly with regard to the use of administrative data for research and associated issues of data sharing, access and linkage. Where new data standards are required, researchers, standards agencies and official statisticians should cooperate to develop standards that maximise their value both for comparative research and for official statistical requirements.

By bridging across national communities of official statisticians and researchers, there is much to gain from a shared understanding of risks, opportunities and constraints that widening data access brings. Without this many national statistical agencies will remain inhibited from acting to widen access. The research community and the OECD need to make this a high priority and take full advantage of the regular international conferences among official statisticians.

Why is this important and what needs to be done?

Good access to relevant data is an essential component of the research process. Where access is hindered or prevented through issues relating to the ownership of data, mechanisms must be found to resolve these issues and make data accessible for research purposes. Where access issues relate to concerns about the identification of individuals and/or organisations, the sharing between countries of best-practice in secure access methods should be encouraged.

Recommendation 4.1: ***national statistical agencies*** *should establish mechanisms to improve access for comparative research by the research community to social science microdata in their possession. These same agencies should collaborate internationally to further their efforts in this area, particularly with respect to the use of novel methods to facilitate secure access to such data where there is a risk of disclosure of identities.*

Box 9

National Data Forums for Social Science Data

The idea underlying the creation of a forum for the development of social science data and related data resources is that of improved coordination and communication between social scientists, data producers (national statistical agencies, government departments, large private sector businesses and sources undertaking academic direction), and those involved in the curation of data.

Such forums are established in different ways. The UK Data Forum is essentially a voluntary grouping of these organisations, whereas the German Data Forum has a constitution and structure established via legal instruments. Despite these differences, the strategic approaches they have adopted for the development of resources and their work to resolve obstacles to national research programmes have been notable.

See:
http://www.esrc.ac.uk/funding-and-guidance/tools-and-resources/research-resources/data-services/NDS/UK-data-forum.aspx
www.ratswd.org

Box 10

Data without Boundaries

The Data without Boundaries (www.dwbproject.org) project is designed to support the establishment of a new European data infrastructure (CESSDA-ERIC). It will promote equal and easy access to official micro-data for the European Research area, within a structured framework where responsibilities and liability are equally shared. Europe needs a comprehensive and easy-to-access research data infrastructure to be able to continuously produce cutting-edge research and reliable policy evaluations. Among the project's aims are:

- *Improving access to official statistical micro-data – by building a remote access network between existing research data centres;*
- *Improving resource discovery for official statistics – by ensuring that common standards for metadata are incorporated within official datasets made available for research re-use.*

Recommendation 4.2: ***leading international agencies*** *(e.g. World Bank Group, World Health Organisation, International Labour Office and Organisation for Economic Cooperation and Development) should collaborate in the formulation of a strategic approach towards the removal of obstacles to improved access to and sharing of micro-data in their possession and should provide a coordinated plan for the creation of data discovery and data management tools on their websites.*

3.5 Data linkage and data integration

Challenge 5: *the drive to address what is increasingly an interdisciplinary and international research agenda requires the use of existing capacity to its full potential.*

There is a vast array of data which is collected by government departments (and private sector organisations) which could be used to inform research and provide policy guidance. Via linkage methods and integration techniques, such data are used to great advantage in some countries, notably in Scandinavia and more recently in the Netherlands and in Scotland (see Box 11). This reduces the cost of official statistics and can produce powerful new research databases, but other countries have been slow to maximise the advantage of resources which are paid for from the public purse.

What are the specific obstacles to data linkage?

There are various reasons why linking data across different government departments might be difficult. These include legal and cultural barriers, public concerns about the use of personal data, skills deficits, and technical barriers.

- Legal and cultural barriers: depending upon the perceived sensitivity of the data and/or the legal framework governing data sharing arrangements some departmental 'gatekeepers' regulate access conditions tightly.

- Public concerns: to date there has been relatively little public engagement to explain the potential of administrative data linkage, and the methods that are used to protect individual confidentiality when such linkages are made.

- Skills barriers: even though techniques for record linkage are now well developed, and are used by numerous organisations regularly, the capacity with which to carry out successful linkages may be in short supply.

- Technical barriers: while various models for secure data access exist in some countries, the expertise, hardware and software to implement secure access is unevenly distributed among countries.

The opportunities presented by making better use of data linkage are considerable. For this to happen there is a need to galvanise support within national statistical organisations, to clarify the legal position and to establish a credible research network with the expertise to undertake this work. If this can be achieved it could greatly enhance the research potential of routine data, and would have significant benefits for research and public policy.

Box 11

The Scottish Longitudinal Study

The Scottish Longitudinal Study is a large-scale linkage study which has been created by using data available from current Scottish and administrative data sources. These include Census data, birth, marriages and deaths records, cancer registrations and hospital admissions. The study began with 1991 Census data as its foundation, selected to cover over 5% of the Scottish population. Linked records are available for analysis as anonymised individual-level data. A range of measures has been taken to ensure confidentiality whilst facilitating detailed research. Data are held in a secure room and accessed only by authorised researchers. Outputs are screened to ensure that no inadvertent disclosure of identities is possible.

Further details about these data and access procedures are available at

www.lscs.ac.uk/sls/access.htm

Why is this important and what needs to be done?

Significant progress in data linkage and integration between datasets is being made in some countries and regions. An example from the European Union is shown in Box 12. The OECD Statistics Committee has created in 2011 an *Expert Group for International Collaboration on Microdata Access*. The new Group will propose solutions for facilitating cross-border access to microdata; this implies establishing agreed procedures for efficient cross-border access to microdata, including improving the availability and accessibility of metadata on microdata.

Such progress can benefit from close cooperation between national statistical agencies, (acting as guardians of datasets which have the potential for linkage) and the research community. This is an area where there is a need to bridge between these groups, enabling national statistical institutions to benefit from the skills and expertise available within the academic sector, and for more detailed knowledge about data sources and methods housed within statistical offices to be gained by researchers.

Recommendation 5: *mechanisms should be established which build upon and enhance further the efforts made by the producers of data (e.g. official data producing agencies, businesses, researchers) and the users of data (e.g. researchers, policy-makers) to share expertise, knowledge and resources, particularly in the areas of data access, linkage and integration and analysis.*

Box 12

Eurostat survey data harmonisation programme

Eurostat and the Member States of the European Union have embarked upon an ambitious reform and modernisation of the ESS (European Statistical System). Details can be found in the vision document proposing new production methods for EU statistics and related implementation measures.

Amongst the 10 principles presented in the 'joint strategy' the first emphasises the provision of statistical information to fulfill the needs of multiple users in decision making, including researchers, policy makers and EU citizens. The seventh principle states that this modernisation should be based on standardisation and integration of production processes using generic and exchangeable methodologies and ICT tools.

Priority actions for the modernisation process identified by Eurostat are:

- *modernisation of European household/individual micro-level surveys;*
- *promoting new data sources (administrative, non ESS surveys, internet);*
- *promoting greater use of data linking/matching to increase the potential of existing data;*
- *working across domains, or vertically where appropriate, to reengineer the production and dissemination process, e.g. through the definition of a common validation environment.*

The vision document can be viewed at:
http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2009:0404:FIN:EN:PDF

and the Joint Strategy is presented at:
http://epp.eurostat.ec.europa.eu/portal/page/portal/ver-1/about_eurostat/documents/ESSC20100506ENvisionfinal.pdf

3.6 The comparability of research data on the human condition

Challenge 6: *comparative research in the social sciences, based upon data pertaining to different countries and regions, is an essential part of the research process and is set to become increasingly important. Without collaboration and sharing of experience between countries in the development of comparable data resources, the full benefits from comparative research will not be achieved.*

Comparative research in the social sciences, based upon data pertaining to different countries and regions, is an essential part of the research process and is set to become increasingly important for a number of reasons.

- The advantages gained from studying social and economic phenomena in varied social, cultural and legal environments.
- The need to address research issues which transcend national boundaries.
- The global nature of the data available for research.

Comparative research requires comparable data. This is often difficult to achieve even for research resources which are designed specifically for comparative research (see Box 13). Where data arise from different agencies (national statistical agencies, national research groups *etc.*) it is essential that data definitions and classifications adhere to international standards wherever these exist. In some research areas standards may still need to be developed further. Equally important is the availability and quality of associated metadata, especially for more complex datasets such as longitudinal data.

Box 13

The European Social Survey (ESS)

The ESS research infrastructure was established in 2001. It aims to chart stability and change in the social structure, conditions and attitudes of Europeans and to interpret how the social, political and moral fabric is changing. It also aims to achieve and spread higher standards of rigour in cross-national social science research, including for example, in sampling methods, data collection, the reduction of bias and the reliability and validity of questions. Its data archive currently allows access to more than 200,000 completed interviews with individuals in over 30 European countries.

In order to make its data, methods and protocols readily available to as wide an audience as possible, the ESS offers swift and fully open access arrangements to academics, professional users, policy makers and the public. This open access is a hallmark of the ESS's philosophy. In addition, the on-line training resource, 'EduNet' and the online data analysis tool – NESSTAR - seek to promote and support exploration of this rich dataset.

The ESS website is the portal through which researchers access both documentation and data (www.europeansocialsurvey.org). The range of data, metadata and paradata available within the ESS archive enables analysts to assess possible methodological effects on data. Users have immediate access to all resources following a short registration.

Why is this important and what needs to be done?

There is an increasing range of social science data resources which are designed specifically for comparative research. These are generated by research groups, by the OECD and, in Europe, through the work of agencies such as Eurostat. It is essential that the experience that these groups and bodies have gained in the harmonisation of concepts, definitions and classifications is shared. For new forms of data, problems of comparability may be minimised where they are generated via global mechanisms, or may be particularly acute in, for example, the comparability of administrative data from different countries.

Without the knowledge and confidence that data from different domains are truly comparable, the value of research evidence based on them is diminished.

Recommendation 6.1: ***national and international statistical agencies*** *should strengthen the efforts they are making to harmonise social and economic data at the international level, seeking to prioritise these activities in specific areas.*

Recommendation 6.2: ***researchers*** *involved in the creation and maintenance of data resources designed specifically for comparative research and* ***those funding such research*** *should collaborate to provide mechanisms to foster an integrated approach to data design and harmonisation, access and sharing as stated in the OECD (2007) Council Recommendation concerning Access to Research Data from Public Funding .*

3.7 Research data management plans and data curation

3.7.1 Research data management plans as policy

Challenge 7: *: researchers have a responsibility to ensure that they have the skills and the resources necessary to ensure that the data they use for research are available for re-use and that plans are made to this effect before they engage in research.*

Research in the social sciences, particularly that which involves international collaboration, is becoming increasingly complex. Different types of existing data may be combined and new data collected by various methods such as surveys, from administrative sources or from observation. While research teams are often challenged to develop and analyse relevant data, the process of managing this information, including plans for its subsequent re-use, is equally challenging and is increasingly viewed by research funders as an essential part of research planning.

Recognising this need, many research funders are now requiring that applications for grant funding should be accompanied by a data management plan. Such a plan describes the nature of the intended research outputs (data, software and other materials produced in the course of the project), strategies for sharing these with others, access arrangements associated with data sharing and the long-term preservation and organisation of data and other research products. Examples of these requirements are shown in Box 14.

Box 14

Data management plans

A number of research funding agencies now require research proposals to include a data management plan. The purpose of these plans is to ensure that researchers have given consideration to the nature of the data that will be created and/or used in their proposed research, covering issues such as:

- *a description of the types of data to be produced in the course of the proposed research;*
- *the standards to be used for data and metadata formats;*
- *policies for access to and sharing of data, with appropriate provisions for privacy, security, intellectual property;*
- *policies for re-use of data;*
- *plans for archiving data.*

For further information about how such plans are specified in the USA and the UK, see:
http://www.nsf.gov/eng/general/dmp.jsp
http://www.esrc.ac.uk/about-esrc/information/data-policy.aspx

Why is this important and what needs to be done?

Data management planning is an effective way of ensuring that research data are discoverable, accessible and curated in ways which conform to evolving standards of data management. Such plans should also promote collaboration amongst researchers and those providing data curation services early in the life of a research project. Working closer together should result in data and metadata that are better prepared for ingest into long-term data management infrastructure. This will contribute to greater efficiencies in the deposit of data as well as lessening the amount of retrospective clean up needed in preparing the data for preservation. As yet, however, there is no general agreement between countries regarding the requirement to produce such plans, their content and the ways in which the plans themselves are made available across the scientific community. There is a clear need to promote and coordinate these activities in ways which will provide much-needed information about current and future research activity.

Recommendation 7: ***national funding agencies*** *should ensure that new research awards have accompanying data management plans and should assign resources for this purpose. They should cooperate at both national and international levels to share this information, publishing details about data and metadata to be created and plans for their preservation in formats that make this information readily accessible to the research community. The international community should agree on a standard semantic dictionary describing common elements of a data management plan to facilitate the discovery and use of the information contained in such plans.*

3.7.2 Data curation, infrastructure and policy

Challenge 8: *not all countries have invested in the skills, resources or infrastructure required to curate important datasets. This places such data at risk of loss, minimises the potential for their re-use and precludes their inclusion in an evolving global data ecosystem.*

Data curation embodies those activities that manage data, making them readily discoverable, documented and accessible for re-use, protected against technological obsolescence and preserved indefinitely. These particular activities, while part of the overall management of data across the complete research lifecycle, are often beyond the scope and timeframe of original research projects. As such, the data from these projects require additional services and infrastructure to ensure their long-term access and re-use. A number of developed countries, having previously invested in data services and infrastructure, are now exploring ways of forming a wider research data ecosystem. This larger environment for data will be composed of a variety of organisational models – including the modern data archive – that support data curation functions and perform data stewardship roles, spanning domains of scientific enquiry within and across national borders. A vision of a global scientific data ecosystem in which data are readily exchanged and combined to tackle new scientific challenges has become an international priority amongst these leading data nations. Vital to this vision are data curation services and research data management infrastructure.

In many developed countries, the social sciences are well served by their data archives which are actively engaged in becoming globally connected. Possibly the best example of a close international network of connected social science data archives is CESSDA – the Council for

European Social Science Data Archives, now preparing to become recognised as a legal European institution[3] (see Box 15).

While part of the scientific community is pushing forward to develop a global data ecosystem, many countries are left on the side-lines because of an absence of national research data management infrastructure or the failure to coordinate their activities in this area. For example, some countries have a multiplicity of institutional structures that provide for the curation for specific datasets. Others rely upon *ad hoc* activities undertaken by social scientists to ensure that data they have collected or for which they have guardianship are preserved. Often it is both a lack of resources and an absence of coordination across scientific domains that lead to a plethora of approaches for technical management. Without clear and well established mechanisms for data curation, research data are 'at risk' in terms of their long-term preservation.

Box 15

Council of European Social Science Data Archives

The Council of European Social Science Data Archives, CESSDA, is the result of more than 35 years of cooperation on social science data archiving in Europe. Central elements in this cooperation are cross-border data dissemination, development and sharing of standards and tools, knowledge exchange and education. Presently CESSDA has 20 member institutions spread over all of Europe. CESSDA makes itself available to the public through the CESSDA data catalogue http://www.cessda.org/accessing/catalogue/index.html but most prominently through the services offered by the member institutions. Presently preparations are being made to establish CESSDA as a European Research Infrastructure Consortium (ERIC). This will greatly enhance the CESSDA services available to the research community and will help make the access to European social science and humanities data even more smooth and seamless. The CESSDA ERIC will have Norway and Germany as important partners.

There is an urgent need to take stock of the situation in many countries with respect to the curation of scientific datasets. In some instances national statistical agencies may be positioned to provide the lead, given their need to preserve official statistical sources. In other cases it may be appropriate for international agencies to provide the resources and skills required for the safekeeping of data that they have helped create. Specific domains in some countries have infrastructure in place which, if shared, could leverage the development of data curation services. An awareness and willingness to cooperate across domains is a hurdle to overcome. In all cases it is important that the social science research community, both researchers and their institutions, should take the initiative to ensure that best practice in the curation of valuable data is

[3] This is known as a European Research Infrastructure Consortium (ERIC), a legal instrument that is designed to facilitate the joint establishment and operation of research infrastructures of European interest.

encouraged, resourced and implemented in ways that help develop their capacity to engage with scientific communities in other countries.

Why is this important and what needs to be done?

Data curation, including ingest functions, documentation, management, sharing and long-term preservation of microdata, and the infrastructures underlying these activities are vital to scientific progress. They facilitate the re-use of data, either to replicate and verify research findings, or to extend analysis in new and interesting directions. Not all countries have established plans to develop the expertise or infrastructure to curate data in ways that promote their re-use and preserve them against obsolescence. While such plans should come from the social science communities within countries, established social science data archives can provide assistance and direction in the development of national plans. This would be a bottom-up initiative with some top-down guidance.

Recommendation 8: ***social science research communities*** *in countries without institutional support for data curation or supporting infrastructure should conduct an assessment of their national needs and assets in this area that will contribute to national plans of action. Working with researchers in such countries, established social science data archives should assist them by developing an assessment instrument and providing expert advice in preparing plans.*

3.8 Improving incentives for international sharing of research data

---

Challenge 9: *data sharing, including the creation of appropriate metadata to international standards is fundamental to the process of scientific enquiry. Researchers need incentives to ensure effective data sharing.*

Research is increasingly a competitive venture, and there is much benefit to the public deriving from competition in terms of the speed with which important research is undertaken and the ways in which the best scientific expertise forms into groups with a strong international reputation. But competition can inhibit data sharing, which in turn may stifle the ability of other scientific groups to replicate and extend important research findings. There are a number of important recommendations made in this report which, if enacted, will make data more discoverable, transform international access and improve their comparability. However, equally important is the need to improve the incentives that the research community has to engage in good data management practices and encourage further the culture of data sharing that typifies the social sciences.

Why is this important and what needs to be done?

Scientific progress in all disciplines thrives on the sharing of research knowledge and the need to replicate, test and extend research findings. This, in turn requires scientists to share their data. Where this can be done safely and securely it should be done. For this to be encouraged, appropriate incentives should be put in place. Although social science data are the focus of this report, the incentives identified here are quite general and apply to research across all disciplines.

Recommendation 9.1: ***research funding agencies*** *should collaborate at the international level to ensure that a common system is adopted for referencing datasets in research publications. They should also ensure that the intellectual effort required for the creation and sharing of data is recognised in their evaluation of research activities.*

Recommendation 9.2: ***publishers of research*** *should be encouraged to adopt guidelines for publications, stipulating that a common and internationally agreed referencing system for datasets is used within all scientific publications that have made use of data.*

Recommendation 9.3: *the* ***employers of researchers*** *should recognise the intellectual efforts that have been made by researchers who generate significant data resources. This could be reflected via merit awards, promotions and other ways which acknowledge the professional contributions that have been made.*

4. Conclusions

Data-driven and evidence-based research is fundamental to understanding and responding efficiently and effectively to global challenges related to the health and wellbeing of populations around the world. As requested by the Global Science Forum, the Expert Group on Data and Research Infrastructure for the Social Sciences has reviewed developments in international data availability, considered their potential and suitability for comparative research, detailed the challenges to be addressed, and has made recommendations to respond to these new opportunities.

These recommendations address the activities of national funding agencies, national and international statistical organisations and researchers as well as other bodies with a direct interest or stake in some particular relevant issue. They consider the potential of new data, data discovery, ethical issues, access, data linkage and integration, comparability, data management and curation, and data sharing. Progress in some of these areas is already underway, but is somewhat of a patchwork. There is little global coordination and the potential benefits of a more strategic approach are not yet realised.

Coordination is at the very heart of some of the recommendations put forward in this report. The comparability of research data from different sources and countries is an area where coordination is fundamental for the implementation of agreed standards and for validation of outcomes. Examples are included in this report which indicate the progress being made in separate areas (*e.g.* population data, household income surveys, social surveys), but much more could be achieved through the sharing of expertise and experience across different data domains.

In other cases, explicit coordination of activities would yield efficiency gains, potentially very important given the current economic climate. For example, the exploration of the research potential of new data can progress more rapidly when funding agencies collaborate in reviews

and awards as well as sharing what is learned. 'Digging into Data' is a good example of interagency cooperation to explore the research value of new forms of data for understanding the human condition. But this was very much a 'bottom-up' approach driven by researcher curiosity. Some countries are now looking hard at the potential value of new forms of data to supplement data collected by more traditional methods. Also, the scope to integrate different types of data is relatively unexplored. Commercial organisations hold large amounts of data within their customer databases, but their willingness to allow such data to be used for publically-funded research is, as yet, largely untested.

As another example, many national and international bodies are grappling with privacy concerns in the integration and re-use of micro data. Currently, solutions developed by one group are being borrowed by others in a 'bottom-up' approach to the problem, but much remains to be done in this area. The Expert Group identified data linkage as one of the key mechanisms through which many important research issues can be addressed, allowing for the extension of existing data in ways that are efficient and potentially informative. However, data linkage increases the risk of revealing the identity of the persons whose data have been linked. This issue is being pursued in various ways in different countries. There is a need for an audit of practice in this area. What work is underway to facilitate various forms of data linkage in a secure manner and how might this information be shared between researchers and national statistical agencies within and between countries? This is an area where a coordinated approach built around common definitions and shared practice could really make a difference.

Data management plans are a further example. As funding agencies increasingly require the submission of such plans as part of the proposal process, a coordinated and perhaps centralised source of approaches and good practices could represent a substantial savings of time and effort while simultaneously improving knowledge about data and metadata prior at an early stage in their development.

In still other cases, a coordinated approach may yield a much greater impact. This is perhaps most clear with respect to incentives for data sharing. If funding agencies, publishers, and employers combined together to insist on, recognise, and reward the creation and sharing of research data, it is likely that the desired behavioural change (improved data sharing) would be effected rather quickly.

The recommendations of the Expert Group are directed to multiple national and international bodies and groups of researchers within and across countries. They cover a wide variety of topics under the general heading of data and research infrastructure. Progress is being made on some of these topics, as illustrated in the boxed examples given in the report. Individually and collectively, these activities serve as a platform on which to build for the future. However, there is no global coordination of activities, guided by the need to inform the development of a more collaborative and productive international research agenda on the changing nature of the human condition. The need for global coordination of activities designed to address the challenges identified in this report is paramount.

Appendix

The OECD Global Science Forum gave its approval in April 2010 to a proposal from the United Kingdom, to undertake a global initiative on data and research infrastructure for the social sciences. The work involved in this initiative was taken forward via the establishment of an Expert Group consisting of up to twenty members, the majority nominated by OECD member countries with additional members from non-OECD countries as deemed appropriate by the OECD. Full membership of the Expert Group is shown below. The secretariat for the Group was provided by the University of Warwick, England organised through Professor Peter Elias, acting on behalf of the UK Economic and Social Research Council. A series of five meetings and workshops involving the full membership of the Expert Group was held between November 2010 and November 2011.

Terms of reference for the Expert Group

- To review and advise on the major data series which ideally would be available for research from every country, or initially at least from OECD countries, either through surveys or administrative data collection, which can be collected or combined to standardised formats, and deposited, codified and made accessible for truly international comparative purposes.
- To review the potential availability for research over the next decade or longer of new forms of data generated by cyber and related activity
- To review new developments in technology and methodology for access to and digging into data in all its new forms, including disclosure analysis around confidential data, and the ethical issues involved in this

The programme of work reported here grew out of an initiative discussed by the International Data Forum at its foundation meeting in Beijing in June 2007, aimed at exploring the mechanisms through which data needs for future cross-national collaborative research on social scientific issues can be identified and prioritised, and how efforts by national research funding agencies and statistical authorities to make data more widely available for research purposes might be co-ordinated.

Membership of the Expert Group

Chair United States of America	Barbara Entwisle	Vice Chancellor for Research University of North Carolina, Chapel Hill
Vice Chair United Kingdom	Peter Elias	Institute for Employment Research University of Warwick
Australia	Garth Bode *(to Sept 2011*	Social Statistics Division Australian Bureau of Statistics
	Susan Linacre *(Nov 2011 – Jan 2012)*	Deputy Statistician Australian Bureau of Statistics
	Gemma van Halderen *(from Feb 2012)*	Australian Bureau of Statistics
Brazil	José Eduardo Cassiolato *(from June 2011)*	Economics Institute Federal University of Rio de Janeiro
Belgium	Patrick Deboosere	Department of Social Research, Vrije Universiteit Brussel (VUB)
	Françoise Thys-Clément	Centre de l'Economie de la Connaissance
Canada	Chuck Humphrey	University of Alberta and representing the Social Sciences and Humanities Research Council of Canada
Denmark	Ole Gregersen	Danish National Center for Social Research
	Niels Ploug	Statistics Denmark
European Commission	Michel Glaude *(to Nov 30, 2010)*	Quality, Methodology and Information Systems, DG ESTAT.B
	Pascal Jacques *(from Dec 1, 2010)*	Head of Sector "Research in Statistics" LISO - Local Informatics Security Officer Eurostat
	Dimitri Corpakis *(to Dec 31, 2010)*	Horizontal Aspects and Coordination DG RTD-L.1
	Philippe Keraudren *(from Jan 01, 2011)*	Science, Economy and Society DG RTD-L2
	Maria Theofilatou	Research Infrastructures, DG RTD-B.3
Finland	Sami Borg	Finnish Social Science Data Archive FSD
France	Roxane Silberman	Ministère de la recherche
Germany	Gert G. Wagner	RatSWD
	York Sure	GESIS-Leibniz-Institute for Social Sciences / University of Koblenz-Landau
	Heinz-Herbert Noll *(from Jan 2011)*	GESIS - Leibniz Institute for the Social Sciences 'Social Indicators Research Centre' (ZSi)
India	Dr S Durai Raju	Central Statistical Organisation, Ministry of Statistics and Programme Implementation, Government of India New Delhi
Netherlands	Peer Scheepers	Radboud University Nijmegen
	Tom Snijders	University of Oxford / University of Groningen
New Zealand	Len Cook	Institute of Policy Studies
	David Thorns	School of Social and Political Science
Norway	Bjørn Henrichsen	Norwegian Social Science Data Services
South Africa	'Maseka Lesaoana *(from April 2011)*	School of Mathematical & Computer Sciences University of Limpopo
United Kingdom	Vanessa Cuthill	ESRC Methods and Infrastructure Group
United States of America	Myron P. Gutmann	National Science Foundation
	Julia Lane	National Science Foundation
OECD	Stefan Michalowski	OECD GSF Secretariat
	Mika Shozaki	OECD Global Science Forum
(Administration)	Margaret Birch	Institute for Employment Research University of Warwick

Invited participants

Dave De Roure – Oxford e-Research Centre, Oxford University, United Kingdom.

Marina Jirotka – Oxford e-Research Centre, Oxford University, United Kingdom.

John Hobcraft – ESRC Strategic Advisor for Data Resources, University of York, United Kingdom.

Karen Dennison – UK Data Archive, University of Essex, United Kingdom.

Gaëlle Pierre – World Bank Group, Washington DC, United States of America.

List of websites and other references

Linked Data Berners-Lee, Tim, (2006). http://www.w3.org/DesignIssues/LinkedData

Riding the Wave: How Europe can gain from the rising tide of scientific data (2010). http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf

Harnessing the Power of Digital Data for Science and Society (2009). http://www.nitrd.gov/About/Harnessing_Power_Web.pdf

Big data: The next frontier for innovation, competition, and productivity (2011). http://www.mckinsey.com/Insights/MGI/Research/Technology_and_Innovation/Big_data_The_next_frontier_for_innovation

Science as an Open Enterprise (2012). http://royalsociety.org/policy/projects/science-public-enterprise/

Global Research Data Infrastructures: the GRDI 2020 vision (2012) http://www.grdi2020.eu/Pages/SelectedDocument.aspx?id_documento=6bdc07fb-b21d-4b90-81d4-d909fdb96b87

OECD (1980), "Guidelines on the Protection of Privacy and Transborder Flows of Personal Data", available at http://www.oecd.org/internet/interneteconomy/oecdguidelinesontheprotectionofprivacyandtransborderflowsofpersonaldata.htm

OECD (2007) "Council Recommendation concerning Access to Research Data from Public Funding", available at: http://www.oecd.org/science/scienceandtechnologypolicy/38500813.pdf

OECD (2011), "The Evolving Privacy Landscape: 30 Years After the OECD Privacy Guidelines", OECD Digital Economy Papers, No. 176, OECD Publishing, http://dx.doi.org/10.1787/5kgf09z90c31-en

Box 3: Linking Surveys to the World: Administrative Data, the Web, and Beyond
www.nsf.gov/awardsearch/showAward.do?AwardNumber=1131500&WT.z_pims_id=503587

Box 5: The World Bank and the Open Data Initiative
www.ddialliance.org
http://microdata.worldbank.org

Box 7: IPUMS: Integrated Public Use Microdata Series
www.ipums.org/international

Box 8: Partnership in Statistics for Development in the 21st Century (PARIS21)
http://www.paris21.org/

Box 9: National Data Forums for Social Science Data
http://www.esrc.ac.uk/funding-and-guidance/tools-and-resources/research-resources/data-services/NDS/UK-data-forum.aspx
www.ratswd.org

Box 10: Data without Boundaries
www.dwbproject.org

Box 11: The Scottish Longitudinal Study
www.lscs.ac.uk/sls/access.htm

Box 12: Eurostat survey data harmonisation programme
http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2009:0404:FIN:EN:PDF
http://epp.eurostat.ec.europa.eu/portal/page/portal/ver-1/about_eurostat/documents/ESSC20100506ENvisionfinal.pdf

Box 13: The European Social Survey
http://www.europeansocialsurvey.org/

Box 14: Data management plans
http://www.nsf.gov/eng/general/dmp.jsp
http://www.esrc.ac.uk/about-esrc/information/data-policy.aspx

Box 15: Council of European Social Science Data Archives
http://www.cessda.org/accessing/catalogue/index.html