Global integration of new forms of data: problems and possibilities

Peter Elias Peter.Elias@warwick.ac.uk

Key words: administrative data; social media data; transactions data; Global Science Forum; OECD

Introduction

New forms of data, arising from the operation of administrative systems, the use of internet search engines, from social media, telecommunications platforms, customer databases and various sensing and imaging devices, are viewed as providing a 'new frontier' for scientific research in the social and economic sciences and for research at the boundaries between these disciplines and medical and environmental sciences. Interest in the use of such data for scientific research has been developing rapidly over the last few years. Researchers and statisticians in many countries are exploring the potential research value of these digital resources. The time seemed right, therefore, to draw together expertise from a number of countries to explore the range of problems that are faced by those wishing to develop this potential in order to gain new insights into human behaviour.

In 2010 and through its Science Ministry[1], the UK Economic and Social Research Council put a proposal to the OECD Global Science Forum[2] to examine the potential value of new forms of data for research in the social and economic sciences. Having adopted this proposal, countries were invited to nominate experts to form an Expert Group. The report prepared by the group was approved by the Global Science Forum and published in March 2013[3].

The group worked by defining the scope of 'new forms of data', then identifying a series of challenges that needed to be addressed in order to pursue an international research agenda which could benefit from the exploitation of new forms of data.

What are the new forms of data?

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. Figure 1 gives examples of the wide variety of data that are now becoming more available or are potentially available for research purposes.

1 The Department for Business, Innovation and Skills.

2 See http://www.oecd.org/sti/sci-tech/oecdglobalscienceforum.htm

3 The full report, including Terms of Reference and membership, can be downloaded at http://www.oecd.org/sti/sci-tech/new-data-for-understanding-the-human-condition.htm

Figure 1: Forms of Data with 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	

What all of these categories of data have in common is that the data concerned are in digital formats and are generally preserved for a period of time, making them more discoverable, accessible and useful for research than has hitherto been the case. Their most important common feature is that they are not specifically designed for research purposes, but have potential value as research resources.

The challenges identified and recommendations made

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.

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 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: there are risks relating to the inadvertent disclosure of the identities of individuals and organisations arising from the use of some new forms of data as research resources, e.g. social networking data. There is a need for greater transparency in the research use of new forms of data, balancing the gains in knowledge derived from such data with the risks of disclosure, seeking to retain public confidence in scientific research which makes use of new forms of data.

Finding: there is no framework code of conduct covering the use of new forms of personal data for research.

> ***Recommendation 3:*** ***research funding agencies*** *should collaborate to develop a 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.*

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 by the research community to social science micro-data in their possession. These same agencies should collaborate internationally to share their experience, 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.*

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 agenc****ies 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.*

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.

Next Steps

Following publication of the Expert Group's report, a number of national research funding agencies[4] considered these recommendations and agreed to take forward those which would make a significant contribution to both national and cross-national use of new forms of data for social and economic research. In the area of social media data, the ESRC is in the process of developing a call for proposals to establish a Centre

[4] These are the UK Economic and Social Research Council, the German Deutsche Forschungsgemeinschaft, the Netherlands Organisation for Scientific Research, the US National Sciences Foundation and the French Agence National de Reserche.

for International Social Media Research. For improved access to customer databases, a specification is being developed for infrastructure that will provide improved access to private sector data in a secure environment[5].

While these activities represent significant steps towards meeting some of the recommendations made by the Expert Group, much remains to be done. The chair and vice chair of the Expert Group have proposed a two year programme of work, consisting of desk research, expert group meetings and a workshop, which would take forward the development of issues relating to the ethical use of new forms of data for research. The focus of this work will be on the ethical use of new forms of data for research purposes. At present there are no clear guidelines for researchers on a number of issues. For example, how might such data be reused for research purposes, how can any misuse of such data (*e.g.* by inadvertent or deliberate disclosure of identities) can be minimised. Other issues include whether or not consent for reuse for research purposes is required and, if so, from whom and how durable is such consent? For varying data types there will be different answers to these questions, and experience in countries will vary.

[5] This is currently termed 'Business Datasafe' and will commissioned in 2013/14.