From 'Intruders' to 'Partners': The Evolution of the Relationship Between the Research Community and Sources of Official Administrative Data

Paul Jackson

1 Genesis

If the start of the modern period of research use of official microdata can be dated, then we might decide on 12 June 2003. On that day Julia Lane gave a keynote speech to the Conference of European Statisticians (CES) in Geneva,[1] describing the opportunities and the challenges of using confidential official microdata for research in a new way. Julia encouraged us all to recognise that the complexity of twenty-first century society requires statistical and other data-rich government institutions to work together with the research community in partnership. Julia's argument was that neither community would be able to meet the challenges on its own, but when working together their different strengths came to more than the sum of their parts. In 2003 a lot of work lay ahead if this partnership was to be possible, let alone successful.

2 Official Data

The focus here is on *official data*, also commonly referred to as 'administrative data', meaning the individual records of people and businesses that are obtained by public authorities in order for public services and administration to be carried out. These records are obtained under compulsion or provided in order to use public

[1] http://www.unece.org/fileadmin/DAM/stats/documents/ces/2003/crp.2.e.pdf

P. Jackson (✉)
Administrative Data Research Network, University of Essex, Essex, UK
e-mail: paul.jackson@essex.ac.uk

N. Crato, P. Paruolo (eds.), *Data-Driven Policy Impact Evaluation*,
https://doi.org/10.1007/978-3-319-78461-8_2

services. For Council of Europe member states, this can be seen as an interference with Article 8's right to private and family life—an interference that is justified when necessary in a democratic society and when carried out in accordance with the law.

If the partnership between the research community and the official data community is to be successful, these twin considerations have to be addressed—*lawfulness* and *necessity*. A lot of work has taken place since 2003 on the *lawfulness* of interference with privacy for research use of official data. Perhaps not enough has been done on the matter of *necessity*.

3 Research as a Lawful Interference with Private and Family Lives

The official data that third-party researchers have always turned to first are the data collected and collated by official statistics agencies. In the 1990s, the legal and policy frameworks of official statistics agencies prioritised the integrity of official data to build public and business confidence in published official figures. Guarantees for the confidentiality of official records were a very important part of these reforms. The Conference of European Statisticians drives the statistical work of the United Nations Economic Commission for Europe (UNECE). Its membership includes the European countries but also countries from North America, Australasia and Asia. The CES is a forum for agreeing, explaining and implementing transnational guidelines, standards and reviews of the production of official statistics. This role includes formal global assessments of national statistical systems. Taking their lead from the United Nations Statistical Commission, those global assessments required a benchmark and assisted national systems in achieving it. Thus, the CES had an important role to play in building the integrity of official statistics in its region and spent much of the decade designing and implementing the *Fundamental Principles of Official Statistics*.[2] Member countries had to address its new Principle 6:

> Individual data collected by statistical agencies for statistical compilation, whether they refer to natural or legal persons, are to be strictly confidential and used exclusively for statistical purposes.

Through the 1990s almost every European country modernised its legal framework for statistical use of official data. Implementation of the Fundamental Principles very often resulted in national legal and policy frameworks that considered research use of official microdata as a threat to Principle 6 and either ruled it out or established very severe controls. This sometimes put researchers into a class of undesirables labelled as 'intruders'. A lot of work went into designing policies and practices to deal with intruders, and sometimes an insufficient distinction was made

[2] http://unstats.un.org/unsd/dnss/gp/fundprinciples.aspx

between malicious intruders and those in research positions who had no desire to damage the integrity of official data. Nuanced messages are hard to communicate to a sceptical public, and it is much easier for public authorities to give unconditional assurances that nobody else will ever see your private information.

Through this same period, the volume and sensitivity of personal data held by government departments and statistics offices increased, as did the threats to data security. Digital government expanded the amount of information that could be maintained in active service, and to deliver more efficient and more joined-up services, these digital data were increasingly linked, often using a single personal identifier. This improved the power of the information but also put very large volumes of data at risk of unauthorised access and use. The information management and security controls for official data had to be substantially improved, and they were.

The net result of the 1990s' changes in information volume and sensitivity, its management and its legal framework was to increase the gap between the settings applied to the data in government and the settings in place in the research community. Social and economic research in academia, which had most to gain from using the personal digital information held by public authorities, did not progress in a coherent and deliberate way with reforms that mirrored the changes the official data community were undergoing. The official data community was building transparent governance and high-profile public accountability. For example, in the United Kingdom (UK), the government published a consultation document on structural reform to build trust in official statistics in 1998,[3] followed by the *Framework for National Statistics* which introduced the first UK National Statistician role and a code of practice.[4] In the 1990s few, if any, European countries appointed a high-profile and publicly accountable champion of social and economic research with duties equivalent to those of the new National Statisticians, appointments which might have been made under a 'national research law'. Academia did not create and then bind itself to a national code of practice for the research use of official personal information or make a common undertaking to citizens about how confidentiality was to be respected. No regulator of social and economic research and researchers emerged. In the 1990s much of this did happen in the field of health and pharmaceutical research, but not in social and economic research.

When in 2003 Julia Lane asked us all to address these issues, the CES agreed to take on the challenge. Work began to enable the benefits of partnership with researchers without compromising Principle 6. New laws, policies and practices were needed. The CES commissioned a task force to produce principles and guidelines of good practice on managing statistical confidentiality and microdata

[3] https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/260823/report.pdf

[4] https://www.statisticsauthority.gov.uk/archive/about-the-authority/uk-statistical-system/history/key-historical-documents/framework-for-national-statistics.pdf

access. They were published in 2007,[5] in what has proved to be a very influential report. The report asserted that when certain conditions are met, the research use of confidential official data is not a breach of Principle 6. The task force's elaboration of these conditions, and the solutions that achieve them, has given us the conceptual building blocks that are elaborated in the legal and policy frameworks in place today across the UNECE region. Since 2003, almost every UNECE country has modified its legal and policy frameworks to implement the task force recommendations and the parallel policy initiatives. New tools and techniques have been developed to take advantage of the new policies and legislation. The main thrust of this work has been to enable conditional gateways through the non-disclosure laws and policies that apply to statistical and other government outputs derived from personal records.

It might be said that research access to official data has followed the same evolutionary change as access to music. At first, if you wanted to hear the music of your choice, you had to play it yourself, and if you wanted data for a particular purpose, you probably had to collect it yourself. Punched cards and then magnetic tape transformed the capture and storage, and reuse, of both music and data at about the same time. Lovers of music started building their own collections of recordings, and researchers started compiling their own copies of data. But then the histories of music and data separated temporarily; for a period the distribution of digital music threatened the property rights of performers and copyright holders, which was something controllers of personal data could never allow to happen. On the research community side, it took a while to accept that having no desire to damage official data's confidentiality is one thing, but inadvertently having that effect is quite another. The local compilation of confidential official data did not become anything more than a great record collection. Thankfully, there was never a 'Pirate Bay'[6] for personal official data. The solution music found was to replace downloads and copying with streaming and digital rights management. Distributing access to content with permissions, rather than distributing the content itself, is inherently safer for all concerned. The equivalent of music streaming in research use of official data is remote access or laboratory access through virtual desktops, and this is now the standard practice.

The UK has this year made primary legislation that provides for the reuse of personal information for research purposes.[7] Appropriately it is part of legislation for a digital economy. It is an example of what is now found in many UNECE countries—the substitution of the original legislative protections under which private personal information was first collected by a public authority with an equivalent but different set of legislative protections that are built around the framework of the 2007 CES Report. Chapter 5 of Part 5 of the Digital Economy Act 2017 provides that whatever statutory or other obligations pertain to the administrative records of

[5] https://www.unece.org/fileadmin/DAM/stats/publications/Managing.statistical.confidentiality.and.microdata.access.pdf

[6] https://en.wikipedia.org/wiki/The_Pirate_Bay

[7] http://www.legislation.gov.uk/ukpga/2017/30/part/5/chapter/5/enacted

personal information held by a public authority, they are not a legal barrier to the extraction, linking and disclosure of those records to a researcher. The barriers are not simply removed, of course; they are substituted by a set of conditions applying to the researcher, the research project, the parties who prepare the data for the research, the level of anonymisation the data must achieve and the working environment the research must take place in. The UK Statistics Authority is given the statutory duty to establish the criteria for these conditions and the function of accreditation against those criteria.

The UK's Digital Economy Act and its equivalents in other countries are now providing the basis for the lawful interference with private and family lives for the purpose of research. The substitution of one set of protections with another requires the research community to build and maintain the facilities and behaviours necessary to meet the criteria. The Seventh Framework Programme (FP7) project 'Data without Boundaries'[8] explored how the social science data archive community can work in partnership with producers of official data to provide extra capacity and capability for research data access in a form that can achieve accreditation against any reasonable criteria. Data archives, and equivalent social science infrastructures such as the UK Data Service[9] are able to provide accredited research facilities to host official data and to host the approved researchers as they use them. Data archives have capacity and capability advantages that are unlikely to be found in public administration departments. They have the ability to create excellent metadata and data documentation, and they can retain anonymised data for reuse by other projects, saving the multiple extraction costs. Importantly, the providers of many of these services have expertise in linking and anonymisation to create powerful, but very low-risk, research datasets. They have established remote access and laboratory facilities that distribute access, not data. In many countries now, the research community through its funding councils have established world-class services for the preparation and use of research datasets, representing an excellent route for the safe exploitation of data as a key national economic resource. The importance of this service is recognised in the European Strategy Forum on Research Infrastructures (ESFRI)[10] Road Map, with the Consortium of European Social Science Data Archives (CESSDA) recognised as one of its success stories.[11] CESSDA is now a European Research Infrastructure Consortium, with a legal personality and an ability to enter legally binding contracts and receive grants in its own name. CESSDA-ERIC[12] is an example of how the research community is now establishing working environments for the provision of access to data for researchers in accordance with common and accreditable standards in public trust, information security, researcher training, metadata and data documentation and cataloguing.

[8] http://www.legislation.gov.uk/ukpga/2017/30/part/5/chapter/5/enacted

[9] https://www.ukdataservice.ac.uk/

[10] http://ec.europa.eu/research/infrastructures/index_en.cfm?pg=esfri

[11] http://www.esfri.eu/esfri_roadmap2016/roadmap-2016.php

[12] https://www.cessda.eu/

CESSDA-ERIC and its members are excellent partners to public authorities who wish to improve lives through the better use of their data, in accordance with the law.

The effect of regulatory legislation can be positive or negative. Those countries subject to the new General Data Protection Regulation[13] may still be undecided as to its effect on research use of official data. Article 6 does not contain a lawfulness provision that unambiguously refers to research in academic and other non-government research institutions, but it does provide for processing in the public interest to be determined lawful. Article 5 provides a positive definition of pseudonymisation which allows for data that have been subject to identity encryption to be used for research purposes without those data being personal data, as long as the encryption key remains beyond the use of the researcher. Article 89 in particular ensures proper recognition of scientific research, suitable technical and organisational measures for protection of any personal data in the information used and an obligation to use data minimisation techniques such as pseudonymisation where possible, to ensure that personal data are used only where necessary. The important concept to retain is that the General Data Protection Regulation is there not to stop the use of personal data for research but to provide a regulatory framework for that processing as an economic imperative in a democratic country pursuing economic well-being.

4 Research as a Necessary Interference with Private and Family Lives

Merely because something is lawful does not make it something that should happen, and the Article 8 right to a private and family life makes this clear. If there is to be interference with privacy, it needs to be both lawful *and* necessary 'in the interests of national security, public safety or the economic well-being of the country, for the prevention of disorder or crime, for the protection of health or morals, or for the protection of the rights and freedoms of others'.

Economic research providing evidence that promotes economic well-being can claim a clear provision in Article 8; but only some aspects of social research are similarly provided for. We may regret its absence, but 'the better understanding of society' is not overtly provided for as a just cause for interfering in personal privacy. It would seem sensible to align with the economists and argue the necessity of social research as a factor in the economic well-being of the country.

In turn, it is likely that the public authority most able to make informed national policy to promote economic well-being is also the public authority that holds a great deal of the official data needed to generate the evidence for that policy-making. It follows that the necessity test is met best when a research project is complementary

[13] http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679&from=EN

to the analysis of the department. Interference with the privacy of the department's clients may be necessary if the outcome is evidence that fills a knowledge gap in the department or provides elaboration or greater context and insight to the evidence the department already has. The more distant the research product is from the civic functions performed by the public authority, the more difficult it is to argue the interference with privacy is *necessary*. The research community has the capacity and capability to make its work very necessary indeed. Interdisciplinary and longitudinal linking by experts in the research community creates datasets that are potentially more powerful for analysis than the data held by a department acting alone. Researcher training and a career building the skills and knowledge needed to comprehend the messages to be found in the data are to be found across the research communities of all UNECE countries. The challenge to the research community is to ensure that its work is clearly relevant to the economic well-being of the country, and the challenge to the official data community is to recognise that contribution and to value it when it is presented as part of a request for access to personal information.

Both the research community and the official data community want the same thing: better information, for better decisions, for better lives. However, for a long time there have been differences in opinion as to how best to go about it. An official in government lives in fear of data loss, because of the direct responsibility they have to the citizen who gave them their private information and nobody else. A researcher lives in fear of missing a funding window and publication deadlines for prestigious journals. These pressures need not be conflicting, but in practice they too often are. Having the main purpose, the economic well-being of the country in common means that both communities should work hard to bring their perspectives closer together.

From 2011 to 2013, the OECD Expert Group on International Collaboration on Microdata[14] examined the non-legislative barriers to better use of official data. It concluded that a common *language,* a build-up of *trust,* a transparent understanding of *costs versus benefits* and making the provision and use of data for research a *business as usual* activity are every bit as important as the *lawfulness* of data access. The group emphasised that mature partnerships are the best way to address these issues.

The group produced a glossary of terms, in a chapter of its report called 'Speaking the same language'. It is impossible to make a partnership agreement on access to data if the parties to it have different understandings of fundamental terminology. Whole infrastructures can be undermined, even lost, if we are incapable of describing to the public in a coherent manner whether a person can, or cannot, be identified in a research dataset. Initiatives such as the Anonymisation Code of Practice, issued by the UK's Information Commissioner, will only help improve this situation. Our responsibility is to make sure we use these excellent glossaries and guides.

[14] http://www.oecd.org/std/microdata.htm

The OECD group also examined *trust*. Those who wish to use confidential official data have in the past expected the data owner to have faith in them rather than trust in them. Trust is faith with evidence. When a public authority uses their own staff to produce policy evidence, it has reason to believe that this job can be done without risk to privacy or to the integrity of the data. The staff are subject to departmental employment rules; they have been selected at recruitment and trained through their careers; they use official and supported information technology, and they have line management to monitor their conduct. The data are familiar, and any metadata are readily available. The outcomes are known before they are published. The staff know the context of their work and the intention of the policies their evidence is designed to support. Of course, this may also be true of the staff of a research organisation; the question is whether or not there is evidence of equivalence. Unless compelling evidence of equivalence is provided, the data owner cannot have equivalent trust in staff that are not in their employment, trust in information and communications technology and metadata systems that are not under their control or trust in the timing and impact of the publication of evidence. The onus is on the research team to provide that compelling evidence. Independent accreditation may help.

The OECD Expert Group also examined *information as an economic resource*. It is important not to forget the reality of costs and measurable benefits. There is an assumption that the huge budgets of large policy departments and statistics offices can always provide the relatively small resources needed to support extractions of data for research purposes. That may be true, but it is the predictability and notice period of these requests that is important. Budgets, typically, are allocated at the start of the financial year and then spent on the allocated task only. In many departments, in-year flexibility between allocations may be limited. If the part of a department that uses an administrative data source is not allocated a budget for extracting data, building metadata and documentation, attending meetings of the research project team, checking the suitability of the project team and their research environment, checking lawfulness and necessity, etc., then it is not likely to be able to allocate resources for ad hoc project requests. It is incumbent on the research community to co-ordinate its requests for access to data, to compromise on the detail of the data to be extracted from administrative systems and where possible to bundle together a number of projects that can be satisfied with the data that administrative departments are willing and able to provide. This approach offers the most benefit for the least cost, but it does require an infrastructure or another form of strategic leadership to co-ordinate and find compromise on behalf of the community as a whole.

One of the most difficult subjects to discuss is the management of media and public comment on important social and economic research findings. One of the issues addressed in the *Fundamental Principles* and most codes of practice for official statistics is the manner and timing of release of key statistics. Trust in government and its statistics has been low in the recent past. When trust in official statistics is low, it is essential for producers of official statistics to concentrate on their reputation for integrity and trust and therefore to keep the publication agenda on a topic such as crime or poverty, utterly predictable and independent of any

other narrative. Pre-announced and independent publication of statistics is essential to prove separation of statistical results from the political narrative, whether the narrative is constructed by government or by single interest groups. The inclusion of published official statistics and other data in research publications does not harm trust in official figures; in fact it probably enhances trust. However, the use of unpublished data from official sources raises a number of public confidence issues. How was it decided which projects would get access to unpublished data and (especially) which would not? Was there interference with the data, or the project, by the provider of the official data? Does the department wish to interfere with the timing and manner of the release of the results for political reasons? On the other hand, does the research project approach a social or economic issue from a particular campaigning position or perspective, selecting for its inquiry an analysis of all the harms—but none of the benefits—of a policy in action? Do the research results tell the markets (or the researcher) anything about the direction of travel of an economic indicator before the official figures are released? It can be difficult to respect the importance of independent research and academic freedom and at the same time maintain public confidence in the timely, predictable, pre-announced and independent release of similar information through official figures. Add a media eager to find information that supports its position on social and economic impacts from political decisions, and it is only a brave and confident national statistician or other administrative data owner who enables the production of research results without knowing how and when those results will be released. It may be beneficial to the two communities to introduce a third. If a partnership includes an 'evidence intermediary', being an organisation established to discover, collate and reveal evidence for key areas of policy, then the partnership has its own independent arbiter of what evidence is needed when, and why.

5 A Future in Partnerships

For this latter reason especially and for all the other reasons elaborated here, the research community and the owners of administrative data should seek to enter into partnership agreements. Such agreements would establish the expectations, the contributions, the desired shared outcomes and the details of common actions necessary to get the best evidence from administrative data, thereby enabling the best decisions for the best lives of citizens.

Within partnership agreements we should find, among other things, agreement on:

- The necessity of the research enabled by the partnership to economic well-being
- The lawfulness of the research within the conditions established in legislation
- The accreditation procedures for researchers, projects and their working environment
- The language that will be shared to describe the use of the data

- The costs, and the benefits, the partners expect to experience
- The routine for decision-making through the period of the partnership
- The schedule of work and the production of findings and policy evidence
- The evidence intermediaries that can help identify the most important shared areas of research interest
- How trust will be built and maintained throughout the relationship

The creation of partnership agreements would deliver the spirit of Julia Lane's 2003 keynote address and take us into a new period of genuine collaboration, for the mutual benefit of data owners, data users and the citizens whose lives are affected positively by the intelligent use of data.

Paul Jackson joined the UK Office for National Statistics in 1998. He developed the protocol in the National Statistics Code of Practice for confidentiality and data sharing and established ONS's Microdata Release Panel and Approved Researcher conditions to improve access to ONS confidential microdata. Paul has worked on the European Statistical Law including the regulation for research access to Eurostat's microdata and chaired the OECD Expert Group on international collaboration on microdata. Paul was on the steering committee of the FP7 project 'Data Without Boundaries' and was the first managing director of the Consortium of European Social Science Data Archives (CESSDA). Paul is now the strategic data negotiator for the UK's Administrative Data Research Network, tasked with improving the flow of data from government departments to researchers in a new research council-funded infrastructure.