MANCHESTER
1824

The University
of Manchester

Institute for
Fiscal Studies

Data Sharing and Data Management in the Social Sciences

James Banks

University of Manchester and Institute for Fiscal Studies

ERC Brussels, 23 February 2015

ELSA English
Longitudinal
Study of Ageing

Overview

Conventions and issues in the social sciences with regard to:

- Data archiving and access
- Funding of data activities
- Rewards to, and benefits of, data sharing

with reference to conclusions, recommendations and ongoing activities of ESRC/MRC/Wellcome/CRUK **Expert Advisory Group on Data Access (EAGDA)**:

- Incentives for data sharing
- Confidentiality
- Data access protocols and systems

http://www.wellcome.ac.uk/About-us/Policy/Spotlight-issues/Data-sharing/EAGDA/index.htm

Background

My perspectives on this issue are those of someone who is:

- Data user: Applied Microeconomist, although also working with researchers in other disciplines as well (predominantly social science and epidemiology). Mainly using UK microdata but also US and EU
- Data generator: Co-Principal Investigator of English Longitudinal Study of Ageing (ELSA); Chair of Scientific Advisory Board of Understanding Society
- Advisor to funders: EAGDA member and Trustee of Nuffield Foundation
- Positively inclined towards data-generation and data-sharing

General data sharing issues in the social-sciences

Types of data

Social-science data pertains to human subjects, but is diverse:

- Large scale quantitative survey data
  - Cross-section or longitudinal samples
  - perhaps including biomarker and anthropometric data
- Geographical and geocoded data
- Government administrative records
- Aggregate political and socioeconomic indicators
- Derived datasets combining some or all of the above
- Qualitative data (focus group transcripts etc.)
- Commercial data (e.g. supermarket or market research data)
- ... and much more

Benefits to data sharing

Strong belief within the social-science research community, and from funders of social science research, in the social benefits and externalities to data sharing. Based on:

- Replication and robustness analysis are necessary for scientific quality
- Reducing duplication of data-collection can generate cost efficiencies
- Public data sharing and archiving protocols preserve data better for future generations of researchers
- New lines of research, unanticipated by study PIs, can be opened up

Creates an expectation that any publicly-funded research should be expected to provide such benefits

Costs and risks inherent in data sharing

There is also a recognition of potential costs of the data-sharing and the need for these to be acknowledged

- Data-out costs and archiving efforts can be substantive
    - including ongoing support and maintenance activities after initial data-collection and distribution
- Some perceive a reputational risk to researchers
    - Through discovery of errors or bad design
    - Research opportunities missed if done by others
    - Insufficient reputational reward for data-generators
- Many acknowledge possible reputational risk to study and to potential abuse of respondent consent. With longitudinal studies these could threaten long-term viability of study

But these concerns are perhaps less, on average, in social-sciences than some other fields e.g. health

My own view

- Designing your own data gives you a comparative advantage even if others can have the data too
- You need to exploit this by making sure you prioritise research on the data without ignoring data curation and distribution costs. This usually requires supplementary grants
- If we really itemized all the data costs funders would not be able to afford it — considerable cross-subsidisation (within projects and across projects) is required to make things work
- Informally, the profession does value data generators highly
- The main problem is with the (increasing) use of metrics, either at the institution or the funder level
- UK REF2014 allowed datasets as outputs but (anecdotally) not well-used. Perhaps partly because documenting the impact and usage of your data is really hard

Tiers of access and types of data

Best practice recognises a combination of:

- Public-release anonymised datasets for bonafide research
    - Often through repositories if available (e.g. UK Data Archive)
- Low-risk ‘restricted release’ data available under special license, typically from study via Data Access Committee
    - e.g. geographical data, some administrative linked data, biosocial data, text items
- High-risk disclosive data accessible in restricted environments
    - Secure enclave on site of data-owner
    - Third-party secure setting
    - Secure remote server access, e.g UKDS Secure Lab, ONS VML

But not always consensus over what types of data fall in to what arrangement. Different studies, funders, and countries all take different stances.

Consent and data ownership

For survey data, respecting the nature of the consent that the respondent has given is the prime consideration.

- Who did they consent to? (Funder, PI, host institution?)
- For what purposes?
- And by whom, and in what situations?

Consent can still be very general- many respondents want their data to be used as much as possible, subject to them not being identifiable. (People who do not collect data often don't understand this)

Going forward, for those collecting primary datasets, getting the wording of the consent right is becoming increasingly important

Specific issues and answers to common questions

Specific questions and answers: 1

What data need to be shared?

- Survey data: Yes, with multi-tiered arrangements
- Qualitative data: Yes, but some disagree, and text is potentially highly disclosive
- 'Derived' datasets: Perhaps, even if not created by data-generators

What are the publication requirements?

- Mixed, but within economics at least it is now common for empirical papers to **require** a data annex with:
    - Full data for replication if author is allowed to distribute it
    - Links to the public-release files and instructions or code to facilitate replication if author cannot distribute data. How to obtain data if it is more restricted.
    - Viewed as a necessary evil, and sometimes difficult to fulfill given restrictions placed on users in their access conditions

Specific questions and answers: 2

How common are repositories?

- Common, not universal, nor universally used where they exist
- UK (one centrally funded ESRC archive) somewhat unusual in this respect. Considered a gold standard

What are requirements for data-generators?

- Public funders in social sciences typically require release of data within a short time after completion of the project, e.g. ESRC: within 3 months of end of award
- Even without a formal requirement it is unusual for data-generators withhold data to allow self-publication. The standard practice is to release it whenever it is ready. Sometimes before ("Dirty Data")
- PI's are expected to itemise data costs in budget justification

Issues: 1

Substantial resource costs of good practice

- In practice it is difficult to get sufficient resources in face of cost pressures
  - Acknowledging the true costs would make many applications 'uncompetitive'
  - Most funders make post-award cuts and this is often the first thing to go
  - Often a fungible budget item (across other grants and or with institutional contributions) and hard to manage post-award
  - It is not just about data-processing and distribution. Much ongoing work needs to be put into documation, data discovery tools, monitoring of usage and subsequent publications...
- And what happens after the end-of-award? Many activities need to continue at least for a while if not longer

Issues: 2

Concern with career incentives and structures not as great as within e.g. health, but still exists

- Adequate scientific recognition for PIs and study researchers
- Career structures and recognition for Data Managers

Not only are data inputs not valued enough, but other inputs are becoming increasingly measured so performance along those other domains is becoming important

Issues: 3

Citation and referencing

- **Digital Object Identifiers** are increasingly used, but still not universal
  - Even if they exist, users rarely use them
  - Without them measuring and documenting data-usage is very hard
- Study or cohort profiles as published papers can serve as a citable link to datasets
  - Only limited journal and publication opportunities for such papers

Issues: 4

Administrative records

- Use of government administrative records has become acknowledged as a source of potential value
  - Scandinavian register databases
  - UK Administrative Data Research Network and Administrative Data Task Force
  - Survey-Admin linked databases (when respondent has given consent for linkage - c.75% do)
- Disclosiveness and privacy concerns are strong. Secure environments are the norm.
- Researchers tolerate this, with some concerns:
  - How are access conditions determined- government officials do not tend to cede control to Data Access Committees and may not understand nuances of the issues
  - To date, harms and violations have not tended to come from academic researchers

Issues: 5

Longitudinal studies create particular problems

- A huge ongoing workload designing and collecting new waves of data can crowd out resources for working on past waves
- An ongoing relationship with respondent and need to ensure continuing participation can create risk-averse PI's with regard to reputational risks of study
- Longitudinal datasets are an order of magnitude more complicated to construct, document and deposit
- Consent may be withdrawn in the future and past waves of data need to be adjusted and redistributed

Issues: 6

International or cross-country projects can create particular problems

- Different national laws and protocols for data sharing and release are often inconsistent
- Institutional differences across countries means a lot of work has to be done on data to make them useable
- International users can raise difficult issues, e.g. monitoring and enforcement of adherence to usage agreement and set of possible sanctions

But international comparative research can bring particularly high returns. Some funders now very committed to

- International harmonisation and distribution
- International data-discovery and metadata tools

Issues: 7

Derived data: A very grey area

- Suppose a researcher supplements public-use data with extensive analysis of other public information to create new 'derived' variables. Should these new data be made available?
- Repositories and data-generators will often allow linkage to original files so the opportunity is there.
- But many third-party users may be protective of their intellectual property and past investments of time and effort. In addition, funders requirements for 'secondary data' projects are not so clear
- Typically this depends on the attitudes and goodwill of the individual researchers

Conclusions

Even in an area which has a good reputation with regard to data sharing there are still many difficult issues where agreement between funders, data-owners and researchers is needed, e.g.

- Funders need to adequately resource data sharing activities
- More work to be done to get agreement on use and release of disclosive data, particularly internationally
- Repository use should be encouraged- brings cost efficiencies, quality standards and aids discovery
- Ongoing and everchanging issues of appropriate respondent consent in the light of the recent European Directive

Despite more recognition from funders than in other areas, I would say that data sharing in social sciences is still predominantly driven by convention, coupled with the goodwill of specific researchers. Still much more that needs to be done.