Information is Power: Overcoming Obstacles to Data Sharing

Professor Denise Lievesley

Head of School of Social Science and Public Policy,

King's College London

and President,

International Statistical Institute

Sharing data – two important publications

- Fienberg S., Martin and Straf (1985) ‘Sharing research data’ National Academy Press
- Arzberger P., Schroeder, Beaulieu, Bowker, Casey, Laaksonen, Moorman, Uhlir, Wouters (2004) ‘Promoting Access to Public Research Data for Scientific, Economic, and Social Development’ *Data Science Journal*

- Statistical datasets
- Data produced for other purposes (often administrative or management)
- Research data

Sharing Statistical data

Aim – to encourage the widest possible **informed** use of data consistent with the responsibilities with respect to confidentiality etc

Collect once, use many times

Benefits to data providers of sharing data with the research community

- Development of knowledge
- Encourage greater exploitation of data
- Contribute to sound policy decisions
- Foster multiple perspectives on data
- Facilitate comparative research
- Create knowledgeable data community
- Provide feedback on data and improve data quality
- Improving teaching and ensuring relevance to official statistics

Reduction of response burden

- Compliance costs important especially in small countries and in surveys of elites, businesses, institutions
- Fresh data collection takes time and resources
- Secondary data analysis can take place in resource –constrained environment

- There is growing awareness that failure to exploit the full potential of official data has costs for society and many official agencies now espouse the aim of ensuring that data are used as extensively as possible.
- For purposes of public accountability it is important that official data are made available.
- Often these are data which the research community could NOT collect themselves

Sharing administrative data

- Unrivalled & untapped level of detail
- Survey data has limitations
- Administrative data may have full coverage
- and better temporality
- Reduces respondent burden
- Has potential cost benefits
- Opportunities for data linkage with other sources
- Local ownership and involvement

Sharing research data

"Publicly funded research data are a public good, produced in the public interest. As such they should remain in the public realm. Availability should be restricted only by legitimate considerations of national security restrictions; protection of confidentiality and privacy; intellectual property rights; and time-limited exclusive use by principal investigators."

Scientific paradigm

- The ISI declaration on professional ethics states that “A principle of all scientific work is that it should be open to scrutiny, assessment and possible validation by fellow scientists.”
- One of the fundamental principles of scientific scholarship is that research findings together with the underlying data should be available for others to confirm, refute, clarify or extend the findings.
- Promote deliberate replication, avoid ignorant duplication

"In recent years, the debate on e-science has tended to focus on the "open access" to the digital *output* of scientific research, namely, the results of research published by researchers as the articles in the scientific journals. This focus on publications often overshadows the issues of access to the *input* of research - the research data, the raw material at the heart of the scientific process and the object of significant annual public investments. In terms of access, availability of research data generally poses more serious problems than access to publications." Arzberger et al (2004)

Incentives in academic system

- In 1985 the report of the US committee of national statistics pointed out that '*A scientist is recognised and rewarded through the scientific community and its institutions. Researchers will have greater incentives to share data if the community and its institutions foster the idea that the practice advances science and is part of what is recognised as necessary and proper scientific behaviour".*
- Competition, performance targets, etc
- Role of the Research Assessment Exercise

Barriers to data sharing

- Confidentiality and sensitivity of data
- Legal restrictions
- Promises made to respondents
- Concerns about misuse of data
- Ensuring equity of access
- Need for revenue generation
- Ambiguities over data ownership
- Concerns about data quality

Responsibilities of data users

- acknowledge and give credit
- respect conditions of access
- provide feedback on use
- ensure the quality of their analysis
- avoid bringing the data providers into disrepute

Value and role of data intermediaries

Importance of establishing policies on data access, sharing and preservation

- of official agencies
- funding bodies
- universities
- professional societies

Example policy

(UK Economic and Social Research Council)

restricts new data collection,

encourages secondary analysis,

requires deposit of new data and derived data in UK data archive,

sets standards for documentation,

provides resources for data access and preservation,

supports training of users,

builds data commons.

Access – one size doesn't fit all

- **Needs of users/usages differ**
  - especially in relation to their sophistication and the need for individual level data
- **Diverse data sets especially in relation to sensitivity of content and possibility of disclosure**
- **Integrated, longitudinal and spatially disaggregated data pose particular challenges**
- **So do administrative data**
  - Good practices exist for survey data but not for admin. data
- **And cross-national data**
  - European social survey

Preservation

Having collected data at some cost to society, it behoves us to manage them well.

Alongside dissemination, this entails data preservation.

Due to poor data management, human error as well as technical change and inadequate use of technology, many data sets are no longer readable.

Thus all that remains of this important legacy are the, often quite superficial, reports that were produced at the time.

To this extent an important part of our heritage is lost and we will be severely limited in our analysis of change.

Metadata

It is necessary not only to preserve data but also to create and preserve metadata and contextual information. This is essential to ensure that the interpretation of the data will be informed.

The documentation should include

- data collection instruments and forms
- instruction manuals
- definitions and concepts
- descriptions of scope and coverage and other aspects of quality
- codebooks
- basic tables
- records of validation checks

Case study– building the secondary uses services

National Health Service in England individual patient care records

- The collection of data which records every interaction with the health service from conception to autopsy

Two committees

One on potential usage

- Conducting audits of clinical practice;
- Surveillance of infectious diseases
- Management of the health system
- Monitor equity of access and provision;
- Evidence-based health policy
- Providing better information to the general public
- Improving the quality and safety of care

Second committee on governance

- Hierarchy of data access consistent with ensuring lowest risk of patient identification
- Need to know
- Role of honest brokers and safe havens
- Development of ‘virtual’ safe havens

Information governance of Secondary Uses Service

- aggregate data widely available
- default anonymised
- - or pseudonymised
- if identifiers needed consent should be obtained
- full justification in terms of benefits to be made for exceptions
- exceptions assessed by transparent, equitable, replicable and open process involving patients representatives
- requirement for safety and security of information (ie accountability)

Concluding remarks

We create a diverse range of datasets, many of which are unique, rich in information content and incapable of replication.

Sharing allows scientists to extend the value of these datasets through new, high quality, ethical research and exploitation. It also reduces unnecessary duplication of data collection.

Building preservation and documentation systematically into routine data management is part of good practice: it strengthens quality, enables replication and audit, and provides a sound basis for data sharing.