The Social Side of Science Data Sharing: Distilling Past Efforts

Peter Burnhill, Director, EDINA / Edinburgh University Data Library, UK
Robin Rice, Data Librarian, Edinburgh University Data Library, UK
Diane Geraci, Director, Science Libraries / Faculty Associate, ICPSR,
University of Michigan, USA

ABSTRACT
Our purpose is to review the suitability and generality of data curation practices and principles developed in the social, political and economic sciences for use in the life and physical sciences.

The secondary analysis of data, generated by third parties, in the social and economic sciences prompted the growth of data archives and data libraries, complete with international association of those who 'do data' (www.iassistdata.org). In large part this was because researchers could not command the authority and finances of government that were required to generate the data they needed. The fit was not perfect in several respects. However, comparable questions and arrangements arose for access to data from those academic research groups that were able to secure funds for their own data generation. What to make of this history of practice, in which there has been continuing, secondary access to primary data extending across a forty year period.

Drawing upon experience from working in a variety of data services in the UK and US, the authors have been critically examining how to apply this to 'science', from astronomy to zoology, and what can be learnt from the practices that have arisen in recent years in such fields as geomatics and bio-informatics.

Our assertion is that data sharing has been around longer, and is more common in the social sciences than most other disciplines; our purpose is to share our experience as data folk.

Underwritten by funding bodies on grounds of what would now be recognised as ‘open access’ principles, this data sharing was born of necessity, not collegiality or consensus. Academic researchers in the social sciences have had to rely upon secondary sources of data because they have generally lacked the authority and means to generate primary data. Their subject matter has context in both place and history, and the ability to examine change over time requires secondary analysis of datasets, often requiring geo-spatial referencing and the combination of data from different sources. There are both small-scale and large-scale surveys conducted by the academic sector but it has been cost-effective to look to re-analyse both those data and those produced by government agencies, such as the decennial population census, the annual large-scale surveys, and regular monitoring of economic activity. Those data have merit in their own right and also serve as context for purpose-specific social science enquiry.

Over the four decades or more since the 1960s, **an infrastructure of data archives and data libraries** has evolved to assist secondary analysis by the international social science community[1]. A cadre of data archivists and data librarians, drawn from a variety of professional background, has grown up around this infrastructure, meeting annually over a thirty year period and otherwise operating as what is recognizable as virtual, as well as formal organization: the **IASSIST**.[2]

These organizations have stated missions to ensure **continuing access to time-specific social science data**. Initially geared to meet the **peculiarities of the survey** dataset, the availability of small area statistics from population censuses, released in computer-readable form, spurred on **developments in geo-spatial data handling**.[3] An interesting side benefit of this infrastructure

was the value for government agencies: their data was analysed in novel ways, and they were also able to recover their own datasets through the *de facto* national data archives set up by academe. University data libraries (and some national services) then began identifying gaps in their collections and purchasing micro or macro-level data that filled an identified research need, such as data produced by NGOs or financial institutions. During the 1990s, the data services met with traditional libraries in the land of the digital library, as universities built means of access to digital collections. Assistance with the use of encoded numeric or geographic data has remained a largely specialist area.

This seems an apt time to review the experience of social sciences data users and curators, in the light of two factors: new paradigms of research in the life and physical sciences built around continuing access to data across networks; and triggers for the social sciences to engage in technological advance.

In our review of past **activity as data folk** we recall the attention given to enhancing the value of datasets through the data cleaning required for creating 'public use' versions, and through interaction with users who reported errors in the data or the documentation. The data archivists knew how to 'strip' identifiers to anonymize micro-data to ensure respondent confidentiality, if this had not already been done by the producer. This early attention to the requirements of confidentiality about respondent information and identity has developed into concerns about competing demands of benefit from statistical enquiry and the dis-benefit of unauthorised disclosure.[4] There was also need to provide means for re-calibration of derived variables, such as new forms of 'social class' categorization.

An early pre-occupation was on 'cataloguing of machine-readable data,' on finding aids, and on data citation.[5] Despite the considerable advances in descriptive metadata and in resource discovery across the Internet, proper **dataset citation** is rarely practiced using standard bibliographic references, notwithstanding the problematic nature of citation within datasets or dynamic databases.

One feature of social science computing in the late 1970s, and on, was the rise of the statistical package, such as SPSS. This also acted as documented database, with the effect of standardizing input and output formats and making long-term preservation of datasets simpler. It also assisted the shift of attention to the codebook, and other forms of data-level documentation, which would now be recognizable as 'semantic representation information' in OAIS terminology. This is now enhanced by metadata schemes such as the DDI (**Data Documentation Initiative**), documenting data elements and their relationships in micro (individual-level) and macro (aggregate) level social datasets in XML, so as to facilitate online query and interrogation. Such online systems facilitate quick and easy access, making the datasets more amenable for secondary use, particularly by teachers, students, and potentially researchers from other disciplines.

The costs of preparing datasets for access were high, and remain so. Data archives and services therefore engage in **appraisal** activities, if only to prioritise effort required to support the various aspects of data curation: long-term storage; creation of the public use version; mark-up in XML; 'publication' for online browsing and access. Datasets may also be rejected for long-term preservation if there are insufficient metadata to establish context and relevance.[6] The onus should be on the data producers to create metadata, as a planned activity during the data

collection stage rather than post hoc. This begs questions on what **incentives to create metadata** data producers might recognize in order to "reduce costs of archiving, accelerate release of data, and improve its quality."[7]

This **evolution in the social sciences** stands comparison with the more recent and dramatic rise of 'collaboratories', grid computing, distributed databases and the explosion of data volumes in the physical and life sciences. The drivers for the changes in methods and practice of science are well-known, articulated as 'data deluge' arising from high-throughput experiments, supercomputer simulations, sensor networks, and satellite surveys. Example is given of particle physicists: the Large Hadron Collider under construction at CERN to generate several petabytes of data per year and involve collaborations of over 1000 physicists from over 100 institutions in Europe, America, and Asia. The focus for infrastructure is on middleware that "will enable physicists to set up appropriate data sharing/replication/management services and to facilitate decentralized computational simulations and analysis."[8]

Not all in the physical and life sciences is 'big science', but the trend towards institutionalised data sharing appears to be irreversible. As Mike Lesk asserts, the scientific paradigm itself has now shifted: from the 'old style' (hypothesis, design experiment, run experiment, analyze results, evaluate hypothesis) to the 'new style' (hypothesis, look up data to test it, evaluate hypothesis). This is cost effective in both time and money. Lesk points to molecular biology as the first to shift to the new style, astronomy next, and predicts that many other fields will follow.[9] This new style of 'doing science' has led to calls for better practices in data citation, as well as improvements in database structures for recording annotation and provenance information that remains linked to the cited data as it is re-used in new studies.[10]

Technological advances in computing power also have effect for the social sciences but not as simple data deluge. There are more voluminous data of potential interest to social scientists being generated by the **automation of everyday transactions**: retail shopping data from 'loyalty cards', patient medical records, outdoor CCTV monitoring, and business records of private firms serve as examples. **Socio-legal barriers to use** these new data sources are high however, and access is not generally assured: some barriers are ethical, others financial. Many social scientists continue to work alone or in small groups, collecting their own small-scale data through interviews, surveys, or direct observation captured on digital media, while adding value through accessing the kind of large-scale 'benchmark' datasets we have discussed here.

Advances in computing have prompted the **development of new methods** in the social sciences including modeling through data simulation, mixture models (combining data from different levels of effect and agency), data and text mining (signaling use of quantitative and qualitative method), extracting value from geo-spatial referencing, and means to make better evidential use of visual and sound material.

The importance of data sharing and data curation is now recognized across the arts and sciences, not least through the establishment of the Digital Curation Centre in the UK, and a full programme of activity in the USA under the Library of Congress' National Digital Information Infrastructure and Preservation Program (NDIPP). Policies are emerging for institutional and centralized repositories that will establish the direction of data curation for the future.[11] We are

now in it together, with opportunity for mutual learning, geared we hope to assist data folk in universities and other research institutions support this shift in research activity.

NOTES

[1] ICPSR (Inter-University Consortium for Political and Social Research), based at the University of Michigan, was established in 1962, funded by institutional memberships. The UK Data Archive at the University of Essex, was established in 1967, funded by the Economic and Social Research Council (ESRC). About this time, data libraries were set-up in universities in the USA and Canada, Edinburgh University Data Library among the first to be set-up in the UK, in 1983/4. Recently, Statistics Canada has been active in supporting the training of data support specialists based in universities across Canada to support academic use of national datasets and local data libraries.

[2] IASSIST–the International Association for Social Science Information Service and Technology–is an individual membership organisation which has been reinventing itself since 1974. See http://www.iassistdata.org/membership/about.html.

[3] Craglia, Max, et al (2005). 'Building Bridges between Social Science, Grid, and Geospatial Communities: a Reflection on Practice.' *First International Conference on e-Social Science.* Manchester, UK: 22-24 June, 2005. [Available: http://www.ncess.ac.uk/events/conference/programme/.]

[4] Dale, Angela (2003). 'Research Access to Microdata: an Attempt to Provide a Context.' *IASSIST Quarterly* (27:4) 16-19. [Available: http://iassistdata.org/publications/iq/iq27/iqvol274dale.pdf.]

[5] Drolet, Gaëtan (2005). 'Citing Statistics and Data : Where Are We Today?' *IASSIST/IFDO 2005 Conference.* Edinburgh, UK: May, 2005. [Available: http://www.iassistdata.org/conferences/2005/presentations/h1drolet.ppt.]

[6] ERPANET / CODATA Seminar Report (2003). 'The Selection, Appraisal and Retention of Digital Scientific Data.' Biblioteca Nacional, Lisbon: December 15-17, 2003 (18-20). [Available: http://www.erpanet.org/events/2003/lisbon/LisbonReportFinal.pdf.]

[7] That is the topic of a project funded under the digital preservation research programme of the US Library of Congress and the National Science Foundation, see http://www.si.umich.edu/research/project.htm?ResearchID=73.

[8] Hey, Tony and Anne E. Trefethen. (2005). 'Cyberinfrastructure for e-Science.' *Science* (308): 817-21. American Association for the Advancement of Science: 6 May 2005. [Available: http://www.sciencemag.org/cgi/content/short/308/5723/817.]

[9] Lesk, Mike (2004). 'Online Data and Scientific Progress: Content in Cyberinfrastructure' *Presentation given as part of the UK Digital Curation Centre's Visitor Programme*. Edinburgh: 24 September, 2004. [Available: http://www.dcc.ac.uk/docs/bl-sep04a.ppt.]

[10] Buneman, Peter, Bose, Rajendra and Denise Ecklund (2005). 'Annotation in Scientific Data: a Scoping Report (Draft).' Edinburgh: 1 May, 2005. [Available: http://www.dcc.ac.uk/docs/scoping.pdf.]

[11] For one recent example see the Draft Report of the National Science Board (2005). *Long-Lived Digital Data Collections: EnablingResearch and Education in the 21st Century.* Washington, DC: May 26, 2005. [Available: http://www.nsf.gov/nsb/documents/2005/LLDDC_report.pdf.]