Recent developments in archiving social research

Louise Corti, University of Essex

Abstract

Recent developments in archiving have built on a fifty year foundation of sharing social survey data and are enabling the take-up of data curation practices on a wider scale. Advances in data archiving have been driven by the quest for comparable and harmonised data sources and mandates from sponsors of research to make data accessible - to provide both transparency and to maximise re-use value. In this paper I discuss four recent developments that are brining challenges for social science data archives: methods for archiving qualitative data; providing safe access to disclosive data; institutional data archiving initiatives; and dealing with the emergence of 'new' data types.

Keywords: data archiving; social science data; secondary analysis; data re-use, infrastructure

Introduction

Archiving social research data has a rich history dating back to the 1940s and its activities have been successfully professionalised over the years, through internationally collaborative efforts. Recent developments in data archiving methods have built on this foundation and are enabling the take-up of data curation practices on a wider scale. Developments have been largely driven by the need to locate and access comparative data sources, and also by new mandates from research funders across the world to open up research data to provide transparency, the possibility of verification of results and to maximised the potential of data. The 'open data' agenda is moving at a rapid pace (Cabinet Office, 2011).

This article provides a short review of some of the key issues I have encountered from my own experience in working in a national data archiving organisation, and being part of the international data archiving community (since 1995). In this paper I focus on four relatively recent developments in the field of data archiving and the challenges they bring to our practices: methods of archiving qualitative data; providing safe access to disclosive data; institutional data archiving initiatives; and archives needing to deal with new data types.

What is data archiving?

Data archiving is a method of conserving precious research resources and ensuring that their research potential is fully exploited. Social research data have often been collected at

significant expense and substantial expertise through national surveys or detailed qualitative approaches. If not preserved, these contributions may later exist only as analytic summaries in a small number of reports or publications, typically which do not exploit the whole data collection. For the social sciences, 'archiving' usually means keeping outputs from research projects together with sufficient contextualising information about the material for longer-term preservation to enable future re-use.

Social research data are collected and saved on a variety of media and, these days, are primarily digital. Within a very short space of time digital data are likely to become lost or obsolete as technology evolves.

There is an internationally established methodology for 'archiving' social research data and an established community of 'data professionals' who carry out the roles of actively acquiring, storing, and providing access to data for re-use in research, teaching and learning. In many developed countries there are national centres providing access to their country's national statistics and social surveys (IASSIST, 2011).

A potted history of social science data archiving

Prior to the 1960s survey archiving activity in Europe, some earlier initiatives from the US paved the way for the practices of data archiving. In 1945 Elmo Roper, one of the founders of survey research, gave his IBM punched cards from his 1930s opinion poll surveys to a university library in the US. Shortly afterwards George Gallup, inventor of the Gallup Poll, followed his lead and a dedicated unit was set up to hold the data. This led to the formation of the Roper Center, which opened up as an archive of international opinion polls in 1957 at the University of Connecticut (Scheuch 2006).

Digital data archiving initiatives date back to the early 1960s due to the foresight of a few key academics who were working on international efforts to facilitate access to social science data for cross-national and cross-cultural analysis. They persuaded the International Social Science Council (ISSC) to consider an infrastructure for this kind of empirical social research. In 1966, following a meeting, the ISSC agreed to constitute a standing Committee on Social Science Data Archives (CSSDA) to which UNESCO committed funding (Scheuch, 2003). This international collaboration which, very early on, built data inventories and retrieval systems, data archives management practices and training in secondary analysis evolved over the decades into the organisation which brings together social science data archives across the world today, the International Association of Social Science Information Systems and Technology (IASSIST) (O'Neill-Adams, 2006).

The Central Archive for Empirical Social Research at the University of Cologne, Germany (ZA now part of GESIS) was the first archive to be founded in Europe. The Inter-University Consortium for Political Research followed, setting up in 1962 in Ann Arbor, Michigan and in 1967 the SSRC Data Bank (now the UK Data Archive) was established by the Social

Sciences Research Council (SSRC) at the University of Essex in Colchester (UK Data Archive, 2007). Other research communities around the world followed these initiatives to preserve databases for future national and international research and teaching (Mochmann 2009).

This ongoing international collaboration has led to the development of a distinct culture of data archiving, based on the concept of a 'study' as a unit of storage being akin to a book in a library. The survey's variables were akin to the entries in the book's index[1].

Thus the first social science data archives collected data of specific interest to quantitative researchers, such as opinion polls or election data, but as the trend for large-scale surveys grew in the late 1970s, archives began to acquire international comparative surveys, government surveys and censuses. Because of their large sample sizes and the richness of the information collected, these national and international surveys already used by governments for planning, policy and monitoring purposes represent major research resources for the social scientist. Examples of major on-going cross-national survey series include the barometers (Euro-, Latino-, Afro-, East Asia- and Arab-) and the International Social Survey Programme (ISSP) and the World Values Survey. These surveys constitute continuous comparative data collection that has been likened to large-scale equipment for the natural sciences, requiring significant and continuing investment (Mochmann, 2009). As such, the data archives which host the data together with the data collections agencies, are increasingly viewed as a large-scale facilities.

The UK Data Archive as a case study of archiving

By the 1990s the UK's own data collection had grown to thousands of datasets spanning a wide range of data sources relating to society, both historical and contemporary. Well-established academic and government surveys, longitudinal and cohort studies are deposited on a regular basis. In European countries similar studies are also available for research use, but access to survey data varies rather widely across Europe. The major cross-national European surveys are available for use in a timely manner via Germany's host archive (Cole et al., 2008).

In the 1990s in the UK research funders recognised the needs of both qualitative and historical researchers by supporting a qualitative data archive (Qualidata) and a history data service (HDS). As a result of Qualidata forming, a new culture of preserving and re-using qualitative emerged which in turn spawned a new literature on secondary analysis of qualitative data which has been making its way into key methods literature (some recent

[1] See an early scholarly article on the idea of a survey archive (Lucci, Rokkan and Meyetoff, 1957)

examples include Seale, 2011 and Moore, 2007). Qualidata as a pioneering archive today acts as consultant for other archives setting up qualitative strands to their data portfolio (for recent international progress see Corti, 2011). Since the mid-nineties in the UK and in the Netherlands, social historians have had access to collections spanning nineteenth and twentieth century statistics, census records, state finance data, demographic data, mortality data, electoral history and economic indicators (HDS, 2011; DANS, 2011).

What activities does 'data archiving' cover?

The roles of a data archive include the acquisition, preservation, documentation, processing, cataloguing, disseminating and supporting use of data. Formats preserved vary according to the type of data. Social research data are created in a wide variety of types and formats depending on the research method used. Survey data are generally stored as numeric codes, while transcribed text from recorded in-depth interviews and fieldnotes are typically stored as word-processed documents. These days, most data are usually collected in digital format and are analysed by computer software, either using a statistical package or computer-assisted qualitative data analysis software (CAQDAS).

Archiving digital data requires that data are preserved in formats that can be accessed by researchers, now and in the future. Data archives undertake various 'data processing' activities which include checking and validation, for example, by examining numeric data values and by ensuring data are anonymised to ensure that the risk of identifying individuals is minimal (unless permission is given to include identifying information, such as with an oral history testimony, where the author wishes to be attributed).

The next step is to adequately describe, or 'document' the raw data to enable informed use. Information is collated about the study, methods, questionnaires and data into a 'User guide'. A systematic catalogue or 'metadata' record is created for studies, providing an overview of the study, the size and content of the data files, availability and terms and conditions of access. Much work has been done on approving robust standards for data description. The international social science archiving community even pioneered its own data description or metadata standard, known as the Data Documentation Initiative (DDI) which is used across the world by archives providing access to national surveys (DDI, 2011). In this community, and increasingly in the field of population health studies, the DDI is seen as the de facto standard for archiving and providing systematic resource discovery and cataloguing for social science collections.[2]

[2] An example of a catalogue record for a national survey can be viewed from the Economic and Social Data Service (ESDS) catalogue (ESDS, 2011). Data processing methods are well-documented by the larger national data archives (UK Data Archive, 2010 and Interuniversity Consortium for Political and Social Research (ICPSR), 2009).

Data should be safely stored, in trusted digital repositories[3] and be available under terms and conditions that are appropriate for a particular collection, and which meet any ethical and legal obligations. Users typically want data in a user-friendly format, such as a common statistical or word-processing package. Larger data archives provide web-based delivery of data via instant download facilities in a secure and managed environment, where, for example, an SPSS file can downloaded, after user authentication and authorisation. Online guides, web pages, support and training courses offered by the archives typically provide users with in-depth help in finding the right datasets and using them to their full potential.

In the last ten years we have seen a move towards online data exploration tools, where users can search, browse and tabulate data via their web browser. Examples are the survey data exploration software tools, Nesstar (Nesstar, 2010) and Survey Documentation and Analysis (SDA, 2011).

Finally, ensuring long-term accessibility relies on technical procedures for data storage, preservation, security and access. The larger social science data archives have dedicated preservation policies that aim to meet international information technology and security standards, e.g. UK Data Archive (2011) and ICPSR (2011).

Recent developments: the last five years 2006 – 2011

More recently data archiving has expanded to serve specialist data types and user communities, and are starting to work in the broader data environment. I provide some examples of these developments through four use cases:

New approaches to archiving qualitative data

In the UK, qualitative researchers are routinely depositing data for sharing, are requesting access to other's data sources, and are contributing to the body of literature devoted to debate surrounding methods of sharing and re-use of data. In 2005, in the UK the Economic and Social Data Service (ESRC) supported an innovative funding scheme known as Qualitative Archiving and Data Sharing Scheme (QUADS), that aimed to develop and promote innovative approaches to the archiving, sharing, re-use and secondary analysis of qualitative research and data beyond the 'Qualidata' model it had set up ten years previously (Corti, 2011).

From the funded projects, four common areas of interest arose: metadata standards for qualitative data; defining and capturing data context; the challenges of audio-visual archiving; consent, confidentiality and intellectual property rights issues. While the last two

3 See the European Framework and Audit and Certification for Digital Repositories http://trusteddigitalrepository.eu/Site/Trusted%20Digital%20Repository.html

have largely been addressed through diminishing data storage costs and more sophisticated control of access to data, and detailed guidance on legal and ethical issues[4], the first two are still actively being researched.

Metadata standards for formally describing qualitative data, focusing on how to define, identify and relate data from complex qualitative data collections, are being developed by the DDI Alliance Working Group on Qualitative Data (DDI Alliance, 2011a).

The debate on *capturing context* is an area that is perhaps most controversial and worth pursuing in more detail here. How to represent the context of qualitative research typically arises in any mention of re-using qualitative data collected by someone other than the original researcher. The loss of the essential contextual experience of 'being there' and the lack of being able to engage in reflexive interpretation may then be viewed by some critics of archiving to be insurmountable challenges to returning to already collected qualitative data.[5] The QUADS scheme made some progress in this area and documented some practical methods of recording context, dependent on the nature of the research situation and setting. Attributes of contextual description derive from the context of interaction at the interview level to cultural context at the macro level.[6]

Finally, the challenges of archiving qualitative longitudinal data have been taken up in the 1990s in the US by Harvard's Henry A. Murray Research Archive (James & Sorensen, 2000) and from 2007 onwards by the UK Timescapes project. Timescapes was funded as a feasibility study to explore of the possibility of undertaking a large scale qualitative longitudinal study and to invest in and scale up from existing qualitative data resources (Timescapes, 2011). The project built a dedicated archive, providing useful and relevant data for further integrative and comparative analyses by the teams involved and others, to advance knowledge and theory and to guide policy and action. Close working of the archive's team and the researchers from the very start enabled a common understanding of some of the barriers to sharing potentially sensitive data more widely.

---

[4] Examples of practical and regularly updated guidance include the UK Data Archive's detailed content on Managing and Sharing Data (UK Data Archive, 2011a) and the Finnish Social Science Data Archive's advice on Informing Research Participants (FSD, 2008).

[5] One of the first artiche to challenge the UK's Qualidata archiving model was contributed by Mauthner et al. (1998). Since then replies and further debate have ensued. See Irwin and Winterton (2011) for an overview.

[6] The Scheme produced an edited volume of seven interesting contributions on defining and capturing context for qualitative data (Corti, 2006).

The approaches and solutions employed by Timescapes have certainly advanced our thinking about how to gain consent to share data for longitudinal studies, how to organise and describe data over time, and how to control access to a 'living' archive.

Providing secure access to disclosive microdata

In 2008, in the US, the National Opinions Research Centre (NORC) set up the first 'Data Enclave' to provide a confidential, protected virtual environment within which authorized researchers could access sensitive microdata (Lane et al., 2008). Following this lead, in the UK a new Secure Data Service was launched in 2011 providing UK researchers with secure remote access to business, economic and social survey data deemed too detailed, sensitive or confidential to be made available under the standard licences operated by regular data services (SDS, 2011).

Both services allow researchers to analyse the data remotely from their institution with access to familiar statistical software and office tools, such as Stata, SPSS, and Microsoft Office. Limited business and social survey data are available with detailed geographic information, such as postcode-level variables or grid-referenced versions. The provision of access to these restricted data is through a 'membership' model, where 'approved' members receive the training, support and advice necessary to manage and analyse the data in order to maximise research outputs while protecting respondents' privacy. Based on trust, a two-way commitment is upheld through a User Agreement and a shared Code of Practice.

Other countries are taking up this model of remote access to restricted social and economic data rather than relying on their existing procedures which require researchers to visit secure sites in person. In Europe a funded European Union project, Data without Boundaries (DwB), is exploring supporting equal access to official microdata for the European Research Area, making use of these secure systems (DwB, 2011).

Institutional archiving initiatives

Since the mid 2000s, funders of social research have recognised the value of keeping research data for future use and have set up data sharing policies requiring researchers to document and share data from their own projects. Through these initiatives we have seen Universities begin to take more responsibility for supporting their research staff and research 'assets'. This can be seen in the rise in institutional repositories (IR) set up primarily to host materials such as journal articles, theses and dissertations now attempting to handle research data. These initiatives can provide wide-scale visibility for an institution's scholarly research (Laakso, 2011; RIN, 2010).

At the time of writing, in 2011, IRs are in a state of experimentation, testing procedures and tools for ingesting and curating their data assets. Across the disciplines, data are so vastly heterogeneous in nature that any common solutions for archiving are challenging - perhaps the only common denominator being a top level description of a defined collection. In time we would expect some of the more mature IRs to hold and provide access to social science data. However, given that it has taken 45 years and millions of pounds to establish dedicated social science data infrastructures and agreed practices, it will be interesting to track how IRs can gain the expertise to fully support contextualised collections in every research domain.

One might envisage a new landscape where traditional social science data archives and IRs live together, with the former providing access to nationally-acclaimed data 'jewels' (e.g. cohort studies) and providing expert guidance and capacity building to IRs who hold locally created research data collections. Social science data archives already provide training on managing and sharing data aimed at researchers, from gaining consent, data formats, copyright, anonymisation, methodology and so on.[7] A final rather useful output to come out of the open access agenda and IR movement has been the adoption of Digital Object Identifiers (DOIs) for making data permanently citable. This methodology gives greater visibility to digital outputs and should make publishing their own data more attractive to researchers (Datacite, 2011).

New data types

Data archives are used to dealing with 'predictable' data from traditional research methods. New forms of data collections are in use, such as internet surveys and blog capture, but most of the national archives have yet to deal with these on a large scale (see England and Bacchini, this issue). In the UK, the ESDS ingested a collection which made use of teenagers' chat room contributions, which were presented as a multi-person conversation transcript. While chat room members' names may at first seem anonymous, of course they often are often not, and could be identified by some investigative work on the Internet. As the literature on how to use new data sources grows (for example, online research methods as in Fielding et al., 2008), so the archives will need to consider how best to describe these data for re-use. Describing the methodological attributes of a research 'study' is more challenging when much of the context relating to data gathered informally via the Internet, is unknown.

A further new type of data for data archives is that which forms an integral part of a publication. More recently, experimentation has been done on richer kinds of academic outputs, known as the 'enhanced publication'. This interactive publication draws on viewing actual data excerpts as the reader proceeds through the written commentary. Work on how

[7] Much cited and re-used guidance on, and training resources for, managing and sharing research data have been published by the UK Data Archive (Van den Eynden et al., 2011; Corti et al., 2011).

to achieve these links, technically, has been carried out in the Netherlands (SURF Foundation, 2009). Archives will need to join hands with publishers to explore how best they can enable continuous access to the original data source via the research article.

Conclusion

The fertile hotbed of collaboration amongst quantitative social scientists over the last fifty years has created a bedrock of research data that future social scientists can exploit. In the last decade, qualitative data, more disclosive data and new types of data, including linked data sources have joined this enterprise, bringing with them a host of challenges for the archives community. I have described some of these challenges here – capturing sufficient context for data, providing secure access to data and enabling meaningful linking of data and outputs. The other challenge I alluded to was the expansion of the social science archives community to include research collections sitting in universities' own in-house data repositories.

A new era is upon us – the data deluge - as it has been termed, being driven by the open data agenda. 'Open data' means more data and the need for greater infrastructure and connectivity.

The data landscape is changing rapidly and data archivists will look to e-infrastructure solutions and tools to help us search across the many hundreds of siloes of data, visualise and interact with them. Resource discovery is a huge challenge for the data archivist, with the primary goal being to provide intuitive and simple pathways to data and to make connections with other resources. Fortunately, the data archiving community, through key collaborations – the Council of European Social Science Data Archives (CESSDA), the International Association for Social Science Information Services and Technology (IASSIST) and the DDI Alliance - are spearheading metadata and technical initiatives that will contribute to this challenge. The secure data service model discussed provide a very good example of how data archives are exploiting e-infrastructure to provide facilities to empower researchers.

Ongoing collaboration will enable us to align and harmonise our research and development to achieve joint goals. The CESSDA has been nominated as one of a series of new formalised research infrastructures - the CESSDA European Research Infrastructure Consortium (ERIC) - forming as this is being written. This formal entity will have a central hub, a common data portal and shared theasuri providing easier access to data across the European countries.

References

Cabinet Office (2011). Making Open Data Real: A Public Consultation. The UK Cabinet Office. http://www.cabinetoffice.gov.uk/resource-library/making-open-data-real-public-

consultation.

Cole, K., Wathan, J., and Corti, L. (2008). The provision of access to quantitative data for secondary analysis. In *Sage Handbook of Online Research Methods*, pp. 365-384. Edited by N. Fielding, R. Lee and G. Blank. London: Sage Publications.

Corti, L. (2011). The European landscape of qualitative social research archives: methodological and practical issues. *Forum: Qualitative Social Research*, **12** (3). http://www.qualitative-research.net/index.php/fqs/article/view/1746.

Corti, L., Van den Eynden, Bishop, L. and Morgan-Brett , B.(2011). *Managing and Sharing Data: Training Resources*, UK Data Archive.http://www.data-archive.ac.uk/create-manage/training-resources.

Corti, L. (ed.) (2006). Making qualitative data more re-usable: issues of context and representation. *Methodological Innovations Online*, **1** (2). http://erdt.plymouth.ac.uk/mionline/public_html/viewarticle.php?id=33&layout=html.

DANS (2011). *Dutch historical data collection*. Data Archiving and Networked Services (DANS0, The Hague. http://www.dans.knaw.nl/en/content/humanities-%E2%80%93-historical-collection.

DDI Alliance (2011). *What is DDI?,* Data Documentation Initiative Alliance. http://www.ddialliance.org/what.

DDI Alliance (2011a). *Qualitative Data Exchange Working Group*. Data Documentation Initiative Alliance. http://www.ddialliance.org/alliance/working-groups#qdewg.

Datacite (2011). *Why cite data?*, Datacite. http://datacite.org/whycitedata.

DwB (2011). *Data without Boundaries – DwB.* Data Without Boundaries. http://www.dwbproject.org/

ESDS (2011). *Economic and Social Data Service catalogue record for a UK survey*. University of Essex. http://www.esds.ac.uk/findingData/snDescription.asp?sn=6732&key=health+survey.

Fielding, N., Lee, R. and Blank, G. (2008). *The Handbook of Online Research Methods*. London: Sage.

History Data Service (HDS) (2011). *About the History Data Service Collection.* University of Essex. http://hds.essex.ac.uk/history/data/introduction.asp.

International Association for Social Science Information Services and Technology (IASSIST) (2011). *About IASSIST.* IASSIST. http://www.iassistdata.org/about/index.html.

ICPSR (2009). *Guide to Social Science Data Preparation and Archiving.* University of Michigan. http://www.icpsr.umich.edu/files/ICPSR/access/dataprep.pdf.

ICPSR (2011).Interuniversity Consortium for Political and Social Research *Digital Preservation Policy Framework.* University of Michigan. http://www.icpsr.umich.edu/icpsrweb/ICPSR/curation/preservation/policies/dpp-framework.jsp.

Irwin, S. and Winterton, M. (2011). Debates in qualitative secondary analysis: critical reflections. *Timescapes Working Paper 4.* University of Leeds. http://www.timescapes.leeds.ac.uk/events-dissemination/publications.php.

James, J. and Sorensen, A. (2000). Archiving longitudinal data for future research. Why qualitative data add to a study's usefulness. *Forum: Qualitative Social Research*, **1**(3). http://www.qualitative-research.net/fqs-texte/3-00/3-00jamessorensen-e.htm

Laakso. M., Welling, P., Bukvova, H., Nyman, L. and Björk, B-C. (2011). *The Development of Open Access Journal Publishing from 1993 to 2009*. PLoS ONE, **6**(6): e20961. doi:10.1371/journal.pone.0020961.

Lana, J., Heus, P. and Mulcahy, T. (2008). Data Access in a Cyber World: Making Use of Cyberinfrastructure. *Transactions on Data Privacy,* **1**(1). http://www.tdp.cat/issues/tdp.a002a08.pdf.

Mauthner, N.S., Parry, O., and Backett-Milburn (1998). The data are out there, or are they? Implications for archiving and re-using qualitative data, *Sociology*, **32** (4) 7233-745.

Mochmann, E. (2008). Improving the evidence base for international comparative research. *International Social Science Journal*, **59** (193-194) 489-506

Moore, N. (2007). (Re)Using Qualitative Data. *Sociological Research Online* **12**(3). doi:10.5153/sro.1496http://www.socresonline.org.uk/12/3/1.html

Nesstar (2010). *About Nesstar.* Norwegian Social Science Data Services. http://www.nesstar.com/about/about.html.

O-Neill Adams, M. (2006). The origins and early years of IASSIST. *IASSIST Quarterly*. **Fall**, 15-13. http://www.iassistdata.org/downloads/iqvol303adams.pdf.

Research Information Network (RIN) (2010). *An introduction to Open Access.* Research Information Network. http://www.rin.ac.uk/system/files/attachments/open_access_booklet_screen_0.pdf

Scheuch, E.K. (2003). History and visions for the development of data services for the social sciences. *International Social Science Journal*, **53** (4) 384-399.

Seale, C. (2011). Secondary analysis of qualitative data. In Silverman, D. (ed.) *Qualitative Research* (3rd edition). London: Sage.

Secure Data Service (SDS) (2011). *About the Secure Data Service*. University of Essex http://securedata.data-archive.ac.uk/.

SURF Foundation (2009). *Enhanced Publications: Linking Publications and Research Data in Digital Repositories.* Amsterdam: Amsterdam University Press. http://dare.uva.nl/document/150723.

Survey Documentation and Analysis(2011). *SDA Features.* University of California, Berkeley. http://sda.berkeley.edu/index.htm.

Timescapes (2011). *Data Archive.* University of Leeds. http://www.timescapes.leeds.ac.uk/data-archive/.

UK Data Archive (2007). *Across the Decades: 40 years of Data Archiving.* University of Essex. http://www.data-archive.ac.uk/media/54761/ukda-40thanniversary.pdf

UK Data Archive (2010). *Quantitative Data Processing Procedures*. University of Essex. http://www.data-archive.ac.uk/media/54770/ukda081-ds-quantitativedataprocessingprocedures.pdf

UK Data Archive (2011). *Preservation Policy*. University of Essex. http://www.data-archive.ac.uk/curate/preservation-policy

UK Data Archive (2011a). *Create and Manage Data: Consent and Ethics*. University of Essex. http://www.data-archive.ac.uk/create-manage/consent-ethics.

Finnish Social Science Data Archive (2008). *Informing Research Participants*. http://www.fsd.uta.fi/english/informing_guidelines/index.html.

Van den Eynden, V., Corti, L., Woollard, M., and Bishop, L. (2011). *Managing and Sharing Data: Best practice guidance for researchers.* University of Essex. http://www.data-archive.ac.uk/media/2894/managingsharing.pdf.