

Data Provenance, Evidence-Based Policy Assessment, and e-Social Science

Lorna Philip^{1*}, Alison Chorley² John Farrington³ and Pete Edwards⁴

¹ & ³ Geography and Environment, ² & ⁴ Computing Science,
University of Aberdeen, Scotland, UK.

* corresponding author email l.philip@abdn.ac.uk

Abstract. This paper considers how the concept of data provenance, developed in *e-Science*, can be applied to *e-Social Science*, specifically Evidence Based Policy Assessment. It considers what Evidence Based Policy Assessment is and what types of information can be used in such research. Quantitative and qualitative data and analytical/ interpretative approaches are outlined and the implications of these for the development of a provenance architecture/ model for Evidence Based Policy Assessment are suggested. The schema for the implementation of an *e*-based provenance system, work undertaken as part of the wider PolicyGrid project, is outlined. The paper concludes by offering thoughts on the contributions that the application of data provenance concepts to Evidence Based Policy Assessment can make.

Introduction

This paper outlines and discusses the challenges posed and the issues raised when addressing the provenance of data used in evidence-based policy assessment (hereafter EBPA). It is based on work being undertaken by an inter-disciplinary team of social scientists and computer scientists as part of a larger project (PolicyGrid; www.policygrid.org) funded at The University of Aberdeen as a node in the ESRC's NCeSS *e-Social Science* project¹. As an introduction EBPA is defined and some thoughts about how provenance has been used in *e-Science*, and how this informs the development of a provenance architecture in an *e-Social Science* context, are outlined. Provenance issues associated with quantitative and qualitative data are considered before the paper turns to consider how provenance may be incorporated into systems designed to support EBPA.

Evidence Based Policy Assessment

Evidence-based policy and its assessment have become central themes in UK policy-making. The UK Government's Green Book (HM Treasury, 2003) sets out, as a best practice guide for central government departments and executive agencies, principles for the *ex ante* **appraisal** of policy, and for its *ex post* **evaluation** (both of which are embraced in our use of the term '**assessment**'). The Green Book concentrates on economic assessment, including the valuation of 'non-market policy impacts'. Key requirements include the provision that reports should provide *sufficient evidence* to support their conclusions and recommendations,

¹ ESRC Grant number RES-149-25-1027.

that there should be an easy *audit trail* to allow decision makers to understand the assumptions underlying conclusions and recommendations, and that there should be sufficient information to support any *later evaluation* (para 2.14, our emphasis). The Magenta Book (UK Cabinet Office Strategy Unit, 2003-5) recognises that policy evaluation can be *formative* (how, why and under what conditions does a policy intervention ‘work’, or ‘fail to work’?) or *summative* (what impact does a policy intervention have?), though these may overlap and interact (p.4). It also, crucially, recognises that a wide range of evidence, including qualitative data, can be useful in policy evaluation and does not seek to impose cost-benefit analysis as the only method of evaluation. ‘Evidence’ can be data and/ or conclusions from previous research; it may also be gathered to supplement previous evidence, to fill gaps in previous evidence, or to create new evidence. Questions that policy-makers should ask about evidence derived from systematic review of existing research include: was the evidence sifted and graded for quality?; were the inclusion and exclusion criteria explicit?; is the evidence easy to understand?; and has the strength of the evidence been assessed? The requirements for evidence outlined in the Green Book and the Magenta Book are essential parameters for inclusion in any *e-Social Science* tools being applied in EBPA, and pose significant challenges, notably with the use of qualitative data, as discussed below.

Provenance: From *e-Science* to *e-Social Science*?

All stages of an EBPA project, from research design through to the preparation of the final report can potentially be addressed through the concept of *provenance*². Groth et al. (2006, p2) define the provenance of a piece of data as “the process that led to that piece of data”. The computer based provenance architectures/ models reported in the *e-science* literature record provenance information about processes which are essentially quantitative methods. They have been applied in a variety of areas including: life sciences, e.g. myGrid (Stevens *et al*, 2003); chemistry, e.g. CombeChem (Taylor *et al*. 2006); and High Energy Physics (Branco & Moreau, 2006). A variety of architectures have been developed, some of which depend on the experimental process being modelled as a step-by-step workflow which the scientist executes in a linear manner, recording information at each step (Stevens *et al* 2003; Taylor *et al*. 2006). Buneman *et al* (2001) described how provenance information can be used to keep track of the evolving versions of large scientific databases. The provenance information can be used to roll back a current database to an earlier version, meaning that the earlier versions do not need to be archived. What lessons can be taken from the *e-Science* experience and what must be developed to support provenance for *e-Social Science*?

Provenance applied to *e-Social Science* would provide information about how data were created, but must also provide information about the context of the data. Such context could include, for example, an account of the characteristics of secondary data and why it was used; an account of data collection methods, including who collected the data, from whom, when and where; an account of the analytical/ interpretative process including who was involved and any assumptions that were made about the data; and a record of what conclusions were drawn, how data were written up and what dissemination has taken place. In EBPA, context is particularly important because of the need to evaluate the quality (and thus the reliability) of data, the robustness of analysis, the generalisability and the validity of findings/ conclusions. Context also helps in data re-use particularly when those who created the data

² A variety of terms are used in the computer science literature to describe the origins, analysis, transformations, assumptions and conclusions drawn from data, including *lineage*, *pedigree*, and *genealogy*. In this paper, and the wider PolicyGrid project, we are using the term *provenance*.

and/or conducted the initial analysis are not involved. Parallels may be drawn between some research activities in the natural and social sciences, notably the application of statistical and modelling techniques to large quantitative data sets. However, qualitative research is arguably as important to the social science community as quantitative research, and the policy community has become increasingly receptive to qualitative data in the UK. A provenance architecture/ model for EBPA in the social sciences should thus be capable of handling quantitative, quantitative and mixed methods approaches.

Quantitative and Qualitative Data for Evidence Based Policy Assessment

A consideration of how social scientists typically 'use' data in an EBPA situation must start with recognising the qualities of quantitative and qualitative data/ information. They have their own characteristics, similarities, differences, strengths and weaknesses, imply different ways of looking at the social world, different stances being adopted by the researcher, the use of a different language and are associated with different research processes.

Quantitative approaches use numbers and categories to describe phenomena. Quantitative data-collection and analysis approaches follow well-established conventions which resonate with *e*-Science ideas of provenance. The quantitative researcher is assumed to be objective and detached from the subjects of study. Descriptive, analytical, single and multi-variate statistics and modelling techniques are standardised. The use of confidence levels ensures that established conventions for identifying significant findings are followed. Reports of quantitative research include figures containing 'raw' or aggregations of raw data - evidence of how data were analysed/ conclusions drawn. If the raw data were made available to other researchers, the analysis could, in principle, be replicated. Indeed the re-use of secondary data has a long history and it has been noted that such use is generally considered to be unproblematic (MIQDAS online guide; Fielding, 2004). However, conclusions drawn from the analysis of quantitative data may not necessarily be replicated: what one researcher selects as significant findings may not be the same as someone else working with the same data set.

Qualitative research is non-numeric, using words and images that can reflect variable understandings and experiences of the social world. Qualitative samples tend to be illustrative, not representative, and generalising from the findings is not normally the objective and, for philosophical reasons, is often discouraged (although the concept of *moderatum* generalisation as espoused by Payne & Williams (2005) is useful when considering the use of qualitative data for EBPA). The researcher is immersed in the research process – subjectivity, reflexivity and personal reflection are actively encouraged. Data collection and in particular data analysis/ interpretation is an iterative process. Well established approaches to the interpretation of qualitative texts, such as the development of thematic codes, content analysis and semiotic analysis, rely upon the unique interpretation and decisions made by the researcher: their application cannot be standardised. To critics of qualitative research this subjectivity implies a less robust process, but we adopt the view that it is simply one of a number of ways in which research may be approached.

Calls for the usefulness of a 'mixed methods' approach have been articulated in a number of social sciences disciplines including, for example, human geography (Philip, 1998; McKendrick 1999) and management science (Mingers & Brocklesby, 1997; Mingers, 2006).

While some EBPA will draw upon only quantitative or only qualitative materials, a ‘mixed-method’ approach is increasingly common in this type of research (e.g. Philip *et al* 2003; Farrington *et al.* 2004). Provenance architecture/ models in EBPA must therefore be capable of supporting qualitative and quantitative data, and a mixed methods approach.

Provenance and e-Social Science

Tracking evidence-conclusion chains in any type of social science research potentially raises philosophical, ethical and legal issues. Ethical/ moral drivers are very important in the social sciences. Not only must those working in the policy field conduct their research in an ethical manner, they must be able to produce, through transparent processes, a robust system for the analysis of data which forms evidence of policy ‘success/ failure’. Legal issues relating to research materials, for example, provisions under data protection, copyright, intellectual property rights and freedom of information legislation (*c.f.* Bishop, 2005; Parry & Mauthner, 2004; 2005) also apply. For some qualitative researchers each data collection event is unique and unreplicable; the narrative content is singular and not capable of analysis by anyone other than the researcher who collected the data. Others view data as being capable of analysis by others, and also subscribe to *extension through generalisation*. Can, or should both positions be captured by a provenance architecture/ model? Further, epistemologically informed positions surrounding the secondary use of qualitative data (Mauthner *et al* 1998; Fielding, 2004) can affect attitudes towards and the practicalities of archiving data and allowing other researchers access to source materials, regardless of whether those researchers be from the policy (government), academic or consultancy communities. Notwithstanding these concerns, in the context of UK government practice there is a case for developing a provenance architecture/ model to support the range of data used in social science research.

The challenge, then, is to track the evidence-gathering and evidence-analysis/ conclusion process for qualitative and quantitative research in the social sciences, consider the issue of data re-use for EBPA *and* be sensitive to epistemological concerns which express an uneasiness about reusing qualitative data in particular. The PolicyGrid project is addressing this challenge through its work on ontologies³ which need to be robust and capable of capturing not only information about resource (e.g. data) content, but also its *context*.

The work in PolicyGrid has produced a schema for an *e*-based provenance system (Chorley *et al*, 2007). The architecture defines provenance as a description of how a resource came about; in other words, it captures information about the activity (or set of activities) performed to create the resource. At its simplest this provenance information represents the association between two or more resources. For example, a quantitative data set would be associated with the questionnaire that elicited the information. The context for the association would also be recorded (e.g. the date a questionnaire was administered, the response rate, etc.). This provenance information can then be used to answer queries such as “how was this evidence derived?”, providing evidence that would allow a third part to ascertain the robustness, or *truthfulness*, of the data collection process. Information about analysis/ interpretative processes could also be captured because the provenance architecture/ model allows the researcher to record what they do. For example, when working with interview transcripts a description of the process by which thematic codes were developed

³ Ontologies for computing scientists are data models representing a set of concepts within a domain (e.g. Document and Author) and the relationships between those concepts (eg. Document *has an* Author) and should not be confused with the term ‘ontology’ which, for the social scientist, refers to what can be accepted as facts, and what the world must be like for knowledge to be possible, as reflected in theoretical positions such as empiricism, realism etc.

could be recorded, akin to the use of memos or other analytic notes in qualitative research. If those interview transcripts were re-used and the second researcher drew additional, or different, conclusions from the text, a second layer of analytical information could be added.

The provenance information is collected by the user (researcher) who describes the resource using a metadata-elicitation tool, based on Natural Language Generation techniques, developed by PolicyGrid (Hielkema *et al* 2007). A generic ontology which is not specific to any domain will be used for the general resource description and supplementary domain specific ontologies will be used to allow the user to describe specific methodologies used to create the resource. There will be domain specific ontologies based on quantitative methods, qualitative methods and others. At present the researcher/ user supplies all of the provenance information, however, in future, if provenance-aware software is used to create the resource then some of the provenance information will be automatically created for the user and presented in the description, thus easing their workload. The design of the provenance model and its use in the tool will not constrain the user in any way because they will not be forced to supply information they do not have or do not want to share. Some items in the description are mandatory (e.g. the name and author of the resource) but other items are optional. Of course a resource that is described in as much detail as possible is the optimal outcome but if some fields are optional the need for anonymity, compliance with data protection etc. is accommodated and use (hopefully) promoted.

Conclusions

This paper has demonstrated that, drawing upon the experience of *e*-Science, there is potential to apply the concept of provenance to an *e*-Social Science context. A provenance architecture/ model capable of supporting both quantitative and qualitative data and analytical/ interpretative methods is being developed which will allow users to record whatever they do: it is thus capable of accommodating the different styles of quantitative and qualitative research (at least in the EBPA context), may be capable of furthering the application of a mixed methods approach in the social sciences because *contextual* information demonstrates how useful combining different types of data can be, and it facilitates the re-use of data. Encouraging social scientists to work with the 'provenance concept' has the potential to support methodological rigour and overall good research practice because it will support and encourage researchers to be reflective and reflexive at all stages of their research. However, challenges remain. How receptive will qualitative researchers be to formally recording the analytical/ interpretative process? How can the trade-off between developing generic and widely applicable provenance architecture/ models on the one hand or developing a range of provenance architecture/ models tailored to specific needs or specific types of research on the other be reconciled? The next stage for the PolicyGrid team and collaborators is therefore to test the provenance model that has been developed, a process which no doubt will answer some questions and raise more.

References

- Branco M and Moreau L (2006): 'Enabling provenance on large scale e-Science applications', in *Proceedings of the International Provenance and Annotation Workshop (IPAW'06)*, volume 4145 of *Lecture Notes in Computer Science*, pp55-63, Chicago, Illinois.

- Buneman P, Khanna S, and Tan WC (2001): 'Why and Where: A Characterization of Data Provenance', Proceedings of the 8th International Conference on Database Theory, pp.316-330, January 04-06, 2001.
- Bishop L (2005): 'Protecting respondents and enabling data sharing: reply to Parry and Mauthner' *Sociology*, vol. 39, pp. 333-336.
- Chorley, A., Edwards, P., Preece, P. & Farrington, J. (2007): 'Tools for Tracing Evidence in Social Science' *Proceedings of Third International Conference on eSocial Science*.
- Farrington, J, Shaw, J, Leedal, M, Maclean, M, Halden, D, Richardson, T, and Bristow, G. (2004): *Settlement, Services and Access: The Development of Policies to Promote Accessibility in Rural Areas in Great Britain*, H.M. Treasury, The Countryside Agency, Scottish Executive, Welsh Assembly Government.
- Fielding N (2004): 'Getting the most from archived qualitative data: epistemological, practical and professional obstacles', *International Journal of Social Research Methodology*, vol. 7, no. 1, pp. 97-104.
- Hielkema, F., Mellish, C., and Edwards, P (2007): 'Using WYSIWYM to Create an Open-ended Interface for the Semantic Grid' in S. Busemann, ed.: *Proceedings of the 11th European Workshop on Natural Language Generation*, pp. 69-72.
- HM Treasury, (2003): *The Green Book: A Guide to Appraisal and Evaluation*, London, HM Treasury.
- McKendrick J (1999) 'Multi-method research: an introduction to its application in population geography', *The Professional Geographer*, vol. 51, no. 1, pp. 40-49.
- Mauthner NS, Parry O and Backett-Milburn K (1998): 'The data are out there, or are they? Implications for archiving and revisiting qualitative data', *Sociology*, vol. 32, no. 4, pp. 733-745.
- Mingers J (2006): 'A critique of statistical modeling in management science from a critical realist perspective: its role within multimethodology' *Journal of the Operational Research Society*, vol. 57, pp. 202-219.
- Mingers J and Brocklesby J (1997): 'Multimethodology: Towards a Framework for Mixing Methodologies', *Omega, International Journal of Management Science*, vol. 25, no. 5, pp. 489-509.
- Parry O and Mauthner NS (2004): 'Whose data are they anyway? Practical, legal and ethical issues in archiving qualitative research data' *Sociology*, vol. 38, no. 1, pp. 139-152.
- Parry O and Mauthner NS (2005): 'Back to basics: who reuses qualitative data and why?', *Sociology*, vol. 39, no. 2, pp. 337-342.
- Payne G and Williams W (2005): 'Generalization in Qualitative Research' *Sociology*, vol. 39, no. 2, pp. 295-314.
- Philip LJ (1998): 'Combining quantitative and qualitative approaches to social research in human geography – an impossible mixture?' *Environment and Planning A*, vol. 30, no. pp. 261-276.
- Philip LJ, Gilbert A, Mauthner N and Phimister E (2003): *Scoping Study of Older People in Rural Scotland* Scottish Executive Central Research Unit, Edinburgh (118pp).
- MIQDAS online guide, available at <http://www.cf.ac.uk/socsi/hyper/QUADS/index-html> accessed 9th August 2007.
- Stevens, R.; Robinson, A.; and Goble, C. (2003): 'my-Grid: Personalised Bioinformatics on the Information Grid'. *Bioinformatics*, vol. 19, no. 1, pp. 302–304.
- Taylor, K.; Essex, J. W.; Frey, J. G.; Mills, H. R.; Hughes, G.; and Zaluska, E. J. (2006): 'The Semantic Grid and Chemistry: Experiences with CombeChem' *Journal of Web Semantics*, vol. 4, no. 2, pp. 84–101.
- UK Cabinet Office Strategy Unit (2003-5): *The Magenta Book*, Government Chief Social Researcher's Office, London.