Tools for Tracing Evidence in Social Science

A Chorley[1], P Edwards[1], A Preece[1] and J Farrington[2]

[1] Dept. of Computing Science, University of Aberdeen
[2] Dept. of Geography & Environment, University of Aberdeen
a.h.chorley@abdn.ac.uk

Abstract. Evidence-based policy assessment requires evidence from a variety of sources (quantitative and qualitative) to be gathered and then synthesised to form an evaluation of a policy's aims or outcomes. In this paper we argue that an appropriate *provenance* framework is an essential pre-requisite for any eSocial Science solution which aims to support such activities. Recent work applying provenance techniques to laboratory records in chemistry is reviewed, leading to a discussion of requirements for an equivalent infrastructure to support evidence bases in social science. Progress towards the development of such a provenance architecture is then described.

1. Introduction

Our work within the PolicyGrid[1] project is investigating how best to support social science researchers in their policy assessment activities through the use of Semantic Grid (De Roure, Jennings & Shadbolt, 2005) technologies. e-Science applications which utilise semantic technologies now exist in areas as diverse as life sciences, chemistry, and earth sciences. However, until recently there has been little work exploring the potential of these techniques within the social sciences, arts and humanities. The concept of 'evidence-based policy making' (Bullock, Mountford, & Stanley, 2001) came to the fore in the UK policy environment in response to a perception that government needed to improve the quality of its decision-making processes; it has been argued that in the past policy decisions were too often driven by inertia or by short-term political pressures. Evidence can take many forms: research, analysis of stakeholder opinion, simulation modelling, public perceptions and beliefs, anecdotal evidence, cost/benefit analyses; as well as a judgement of the quality of the methods used to gather and analyse the information.

In the *Green Book, Appraisal and Evaluation in Central Government* (HM Treasury, 2003) the UK Treasury recommends that in policy assessment:

> *"Reports should provide sufficient evidence to support their conclusions and recommendations. They should provide an easy audit trail for the reader to check calculations, supporting evidence and assumptions."*

In other words, as well as keeping a record of the resources and reports used in the assessment, a researcher should keep a record of what happened to enable the creation of the

[1] The PolicyGrid project is funded under the UK ESRC eSocial Science programme; award reference RES-149-25-1027. (http://www.policygrid.org/)

final report. This will document all the stages in the assessment and will include information on: what was done, how it was achieved, who did it, when it was done, and so on. This process documentation is often known as *provenance* (or lineage, pedigree, history) and is an important aspect of scientific record keeping across disciplines, including life sciences and chemistry. Groth *et al.* (2006) define the "*provenance of a piece of data as the process that led to that piece of data"*. With an appropriate provenance framework in place, pieces of evidence that form part of a policy assessment could then be traced back to their source, e.g. a published report, a process used to analyse a dataset.

The UK Government uses a range of evaluation methods to ensure that policies are as effective and efficient as possible. The *Green Book* (HM Treasury, 2003) presents the techniques and issues that should be considered when carrying out an economic appraisal or evaluation of a policy, project or programme. These activities form part of a broad policy cycle that is sometimes formalised in the acronym *ROAMEF* - Rationale, Objectives, Appraisal, Monitoring, Evaluation and Feedback, see Figure 1.

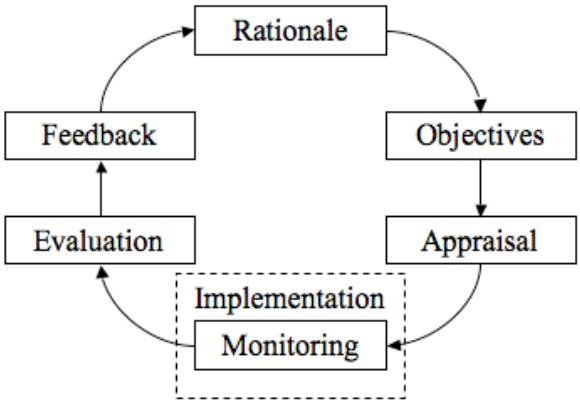

Figure 1: ROAMEF diagram. Reproduced from (HM Treasury, 2003)

To explore the issues surrounding eScience support for evidence-based policy assessment (hereafter, *EBPA*), we are using a particular case study. APAT (Accessibility Policy Appraisal Tool) (Farrington *et al,* 2004) is a specialised policy assessment methodology that was designed to examine and evaluate the accessibility impact of policies, using a mixed-method approach. It aims to improve understanding by participants of the accessibility implications of a policy through reflection and analysis and also generates and evaluates alternative policy options. A researcher conducting a policy assessment exercise will employ some methodology to evaluate the policy's impact (or possible impact) on the community. They may send out questionnaires to members of the public in certain areas of the country, or organise town meetings and focus groups to assess public opinion. They may interview policy makers to gather information about the impact of the policy on the community or other policies. They may perform a cost-benefit analysis in order to assess the fiscal impact of the policy. Such an approach is termed 'mixed method' - as the researcher uses a variety of methods and tools, both qualitative and quantitative, to evaluate the policy. Quantitative techniques use data obtained from questionnaires and surveys and can be analysed statistically to generate numerical evidence. Qualitative methods use data obtained from interviews, town meetings and focus groups and are usually subject to textual analysis against some conceptual 'coding' framework.

Philip *et al* (2007) consider some of the issues relating to qualitative, quantitative and mixed method approaches and how they impact upon a social scientist's view of provenance. In

summary, they conclude that the challenges faced are as follows: to track the evidence-gathering and evidence-analysis/conclusion process for qualitative and quantitative research in the social sciences, to consider the issue of data re-use for EBPA, and to be sensitive to epistemological concerns which express an uneasiness about reusing qualitative data in particular.

2. Related Work

In recent years, there have been a number of attempts to develop provenance solutions as part of the UK eScience programme. Perhaps most notable amongst these are the work of the CombeChem (Taylor *et al*, 2006), myGrid (Stevens *et al*, 2003; Goble *et al*, 2006) and PASOA (Groth *et al*, 2006) projects. myGrid and PASOA are concerned with managing provenance in the context of computational activities implemented as Web or Grid services; metadata is generated when services are invoked (by a person or by another service), while they are executing and when they terminate and return results.

In this paper we are particularly concerned with the approach taken by the CombeChem project, as we feel that the eSocial Science community can learn from the experience of this earlier project; CombeChem models human-centred activities in the chemistry laboratory (*in vitro* experiments) as well as computational activities (*in silico* experiments). Before we discuss the approach taken by CombeChem, it is appropriate to provide some context. In the past, individual researchers published their results in paper based proceedings and journals. Now with the increased use of the Internet these papers are available online in repositories or on the author's Web site. If another chemist wishes to reuse the result from a paper, they may be able to access the relevant data via a repository - but the details of what exactly was done to produce that data may be unclear. CombeChem seeks to change this by providing links from the publication back to the recording of the original experiment in a lab book. Hence by clicking on the result in the paper, the reader will immediately be taken to the data and be shown a record of the experimental process. This is the idea of publication@source (Hughes *et al*, 2004).

Development of the CombeChem infrastructure was guided by the following design principles:

- The approach should be grounded in established operational practice;
- It should capture associations (using metadata) between various entities;
- Capture of metadata should be as automated as possible;
- Flexible information re-use should be supported;
- Management of data and metadata should be given equal consideration.

To perform an experiment, the chemist first has to design the experiment and specify the material to be used. To facilitate this, CombeChem developed a planner which the chemist uses to describe step-by-step every activity they plan to do in the lab. As the experiment is performed, the chemist uses a tablet PC to record annotations associated with each step; these serve as provenance information. Figure 2 shows (in diagrammatic form) part of a CombeChem experiment plan (taken from Taylor *et al*, 2006). Such experiments are recorded using an ontology[2], the scope of which encompasses experimental activities such as materials planning, planning of procedural steps, and experiment recording. Information recorded using this ontology captures the human-centred activities inherent in much of experimental

[2] Ontologies for computer scientists are data models representing a set of concepts within a domain (e.g. `Document` and `Author`) and the relationships between those concepts (e.g. `Document` *has an* `Author`).

chemistry. The two central concepts within this ontology are `Materials` and `Processes`. A `Material` is either a chemical entity or a dataset, while a `Process` is either an *in silico* process, *in vitro* process or a hybrid process; specialist sub-types of `Process` exist to allow experiment plans to capture details such as "Reflux", "React" and so on. The experimental metadata produced are stored in a persistent Jena RDF store[3] backed by the MySQL database platform.

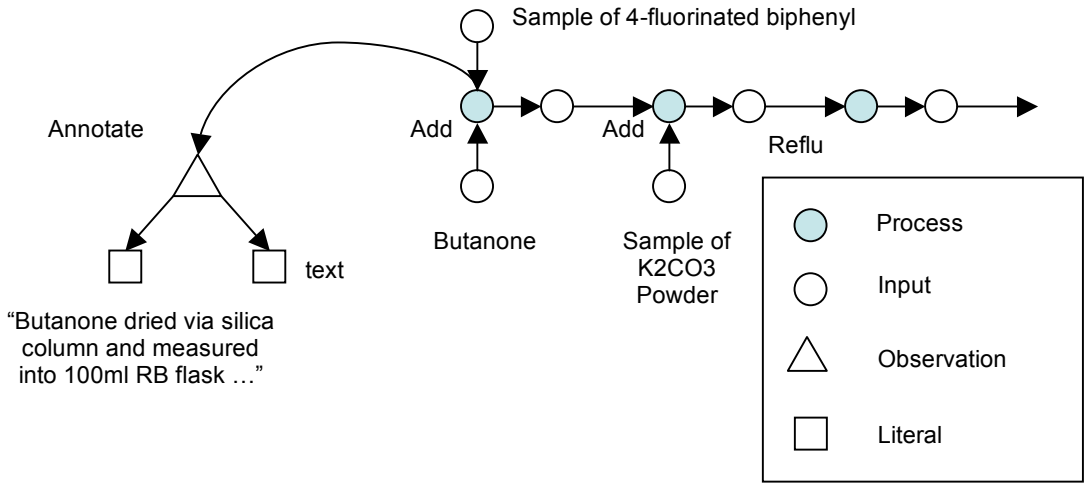

Figure 2: A CombeChem process-product spine illustrating a process annotation (after Taylor *et al*, 2006).

3. Requirements for a Provenance Architecture to Support EBPA

All stages of an EBPA project, from research design through to the preparation of the final report can potentially be supported through an appropriate provenance framework. In this section we will return to our earlier remarks on evidence-based policy assessment and the experiences of the CombeChem project to outline a set of requirements for an architecture to support management of provenance. Before we do, it is necessary to review the scope of provenance: What precisely is it that we are trying to capture? How can the provenance information be used? Goble (2002) presents the "7 W's of Provenance": *Who*, *What*, *Where*, *Why*, *When*, *Which* & (W)*How*. *Who* - describes who deposited the resource, who the author was, if it is a qualitative interview then it describes the interviewer and interviewee. *Where* - this could be where the author works (their biographical information), or for an interview, where the interview took place. *When* - when the resource was created, deposited or the last time it was modified, or if for an interview, when the interview took place. *What* - what was done to create the resource. *How* - this is closely related to the *What* provenance but describes how in this particular instance the resource was created. *Why* - why was this method used in the first place or why were certain things done in a specific way. *Which* – might describe which method was selected from a set of possible approaches. Goble also describes some of the uses of provenance, which include as a means to estimate data quality

[3] Jena is an open-source Java framework for constructing ontology driven applications (http://jena.sourceforge.net/). Resource Description Framework (RDF) is a metadata language used to make *predicate(subject,object)* statements about resources.

and reliability, to allow replication of data derivation, to establish ownership of data, and as a context for data interpretation.

The CombeChem provenance design principles presented earlier provide us with a starting point for our own requirements for EBPA provenance. Although those principles were developed in a chemistry scenario, they are sufficiently generic that they apply equally well to an eSocial Science application. However, there are a number of additional constraints that we must accommodate in the EPBA context:

- The approach should capture *data-oriented* as well as *process-oriented* provenance (Simmhan, Plate & Gannon, 2005) – as we are interested in resources and the method by which they are generated/revised/analysed;
- Different methodological perspectives must be supported, e.g. although a *survey* is in itself a common social science research tool – its execution may reflect differing underlying methodological approaches, depending upon the perspective of the individual researcher, group or community;
- Evidence assertions derived from data and/or analysis should be made explicit.

Figure 3 presents an example of an EBPA process taken from the APAT case study. We have extended the diagrammatic notation used in Taylor *et al*, 2006 to include additional symbols for detailed methodological information, and evidence statements. It should be noted however that while CombeChem users are asked to specify an experimental plan at the outset, in our framework this is not the case; instead, provenance assertions are used to specify process relationships between resources.

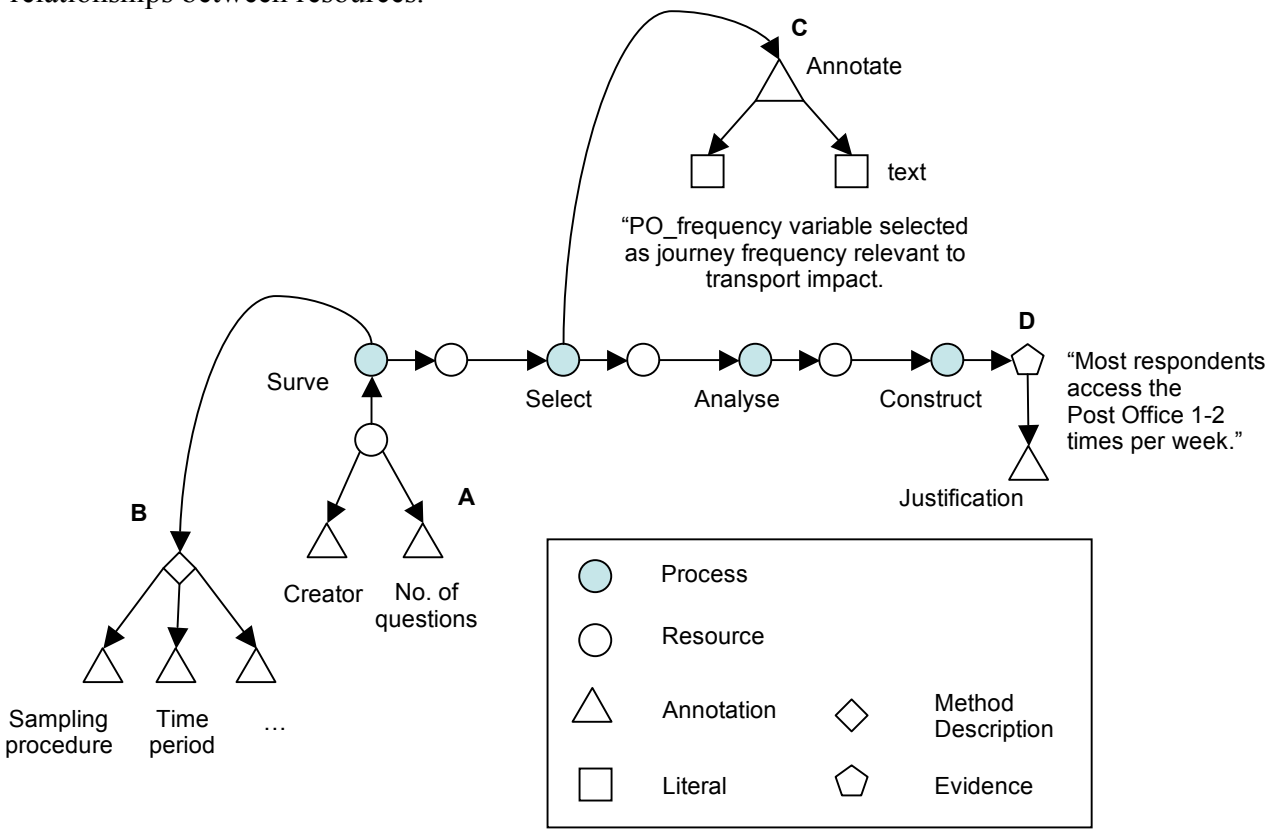

Figure 3: An example evidence path taken from the APAT case study, illustrating different forms of metadata: *resource (A)*, *methodology (B)*, *annotation (C)*, *evidence (D)*.

4. A Provenance Solution for EBPA

The provenance architecture defines provenance as a description of how a resource was created, modified, used and re-used; it records provenance metadata according to the 7 W's described earlier. It should be noted here that the provenance metadata layer operates on top of a layer of resource metadata (labelled A in Figure 3), which captures properties of the resource such as author, date of creation and so on (see Figure 4). We categorise provenance according to the following types:

- Type 1 (*Methodology*)
  To characterise the process by which a social science resource was created we require a mechanism to support the capture of provenance metadata from both human centred processes and computational (*in silico*) processes. Metadata frameworks are needed to describe several, very different social science methodologies including quantitative, qualitative and social simulation modelling. As mentioned above, each of these frameworks must support differing methodological perspectives, as it would not be appropriate for us to force researchers to conform to just one (standard) view of qualitative analysis, for example.

- Type 2 (*Annotations*)
  The experience of the CombeChem project was that chemists used the electronic lab notebook to annotate experimental activities with text and diagrams; we aim to provide the same support for social science researchers. It is our view that such lightweight annotations are an excellent way of capturing *Why* provenance.

- Type 3 (*Evidence*)
  We have chosen to create a special category of provenance metadata to support derivation of evidence and its subsequent use. This captures information about the construction of evidence from other data and resources, as well as descriptions of the use of evidence assertions in policy argumentation.

When a researcher is describing a resource, for instance an interview transcript, they will use a resource ontology to describe *resource* information. They will then use one of the specific methodology ontologies (in this case the qualitative methodology ontology) to describe the *provenance*. Information might include: details about the interviewer and interviewee, when the interview took place, where the interview took place, etc. The provenance architecture must also record the associations between resources, for example, an interview transcript will be associated with the set of questions that were asked at the interview. The same set of questions may well be associated with several interview transcripts. In a similar way a quantitative dataset will be associated with its survey questionnaire.

Our provenance architecture currently uses several ontologies derived, in part, after study of the UK Social Science Data Archive[4] schema (which is itself based on the Data Documentation Initiative[5]). In addition to a social science resource ontology, we have three methodology ontologies modeling quantitative methods, qualitative methods and social simulation methods. Figure 4 shows part of the resource ontology describing an interview transcript, while Figure 5 highlights the interview component of the qualitative method ontology.

---

[4] http://www.data-archive.ac.uk/

[5] http://www.icpsr.umich.edu/DDI/

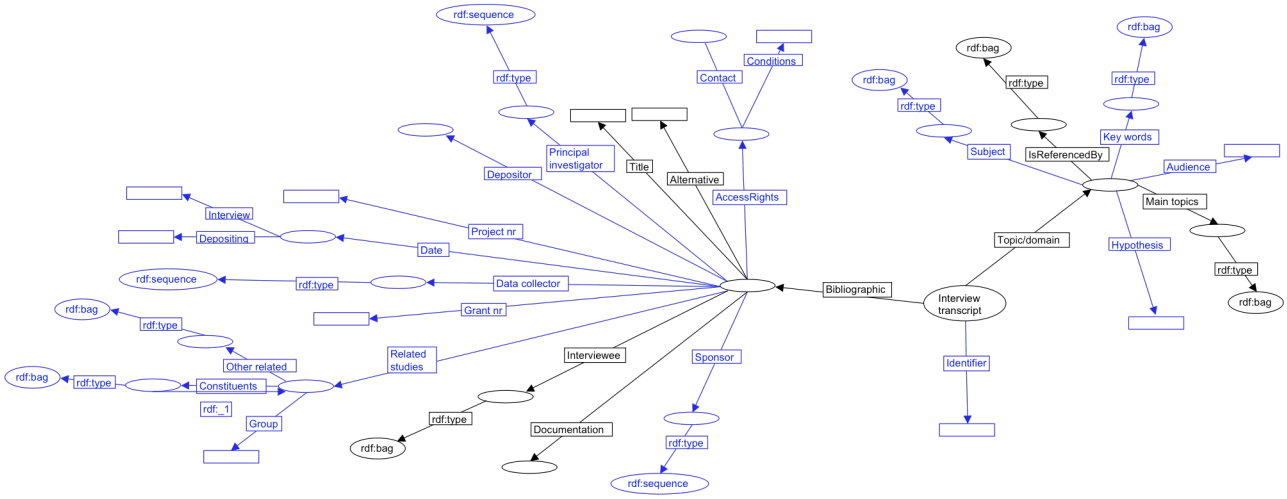

Figure 4: Part of the resource ontology describing an interview transcript, derived from the UK Data Archive.

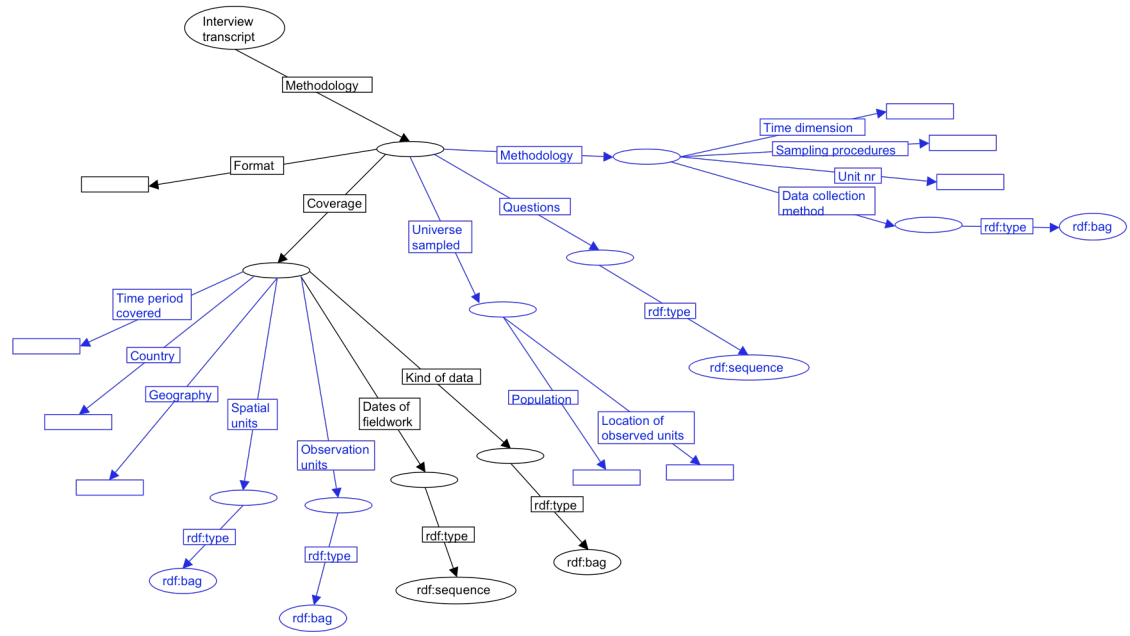

Figure 5: Part of the qualitative method ontology describing an interview transcript, derived from the UK Data Archive.

One of the most significant challenges to be faced with a provenance architecture such as this is how to acquire the metadata. Currently, in our implementation users are required to supply all the metadata by hand. In the future, however, if provenance-aware software were used to create and analyse social science resources then some of the provenance information could be automatically created, significantly easing user workload. Figure 6 shows a screenshot from a prototype Grid-enabled qualitative analysis tool (*Squanto*) to which we are adding just such a provenance component. *Squanto* (Edwards *et al*, 2007) supports qualitative analysis of interview transcripts and other textual resources, either via free text or structured coding frameworks; to add provenance support to the tool we are investigating the use of a third form

of coding – to allow resources to be annotated with methodological information. As it is unlikely that tools such as *Squanto* will ever capture all provenance information, others within the PolicyGrid project are exploring ways of creating metadata using software based on Natural Language Generation techniques (Hielkema *et al,* 2007).

The design of the provenance model should not constrain the user and should not force them to supply information they do not have or do not want to share. Some items in a description may be mandatory (e.g. the name and author of a resource) but other items can and should be optional. Of course a resource that is described in as much detail as possible is the preferred outcome, but if some fields are optional the need for anonymity, compliance with data protection, etc. is accommodated and use (hopefully) promoted.

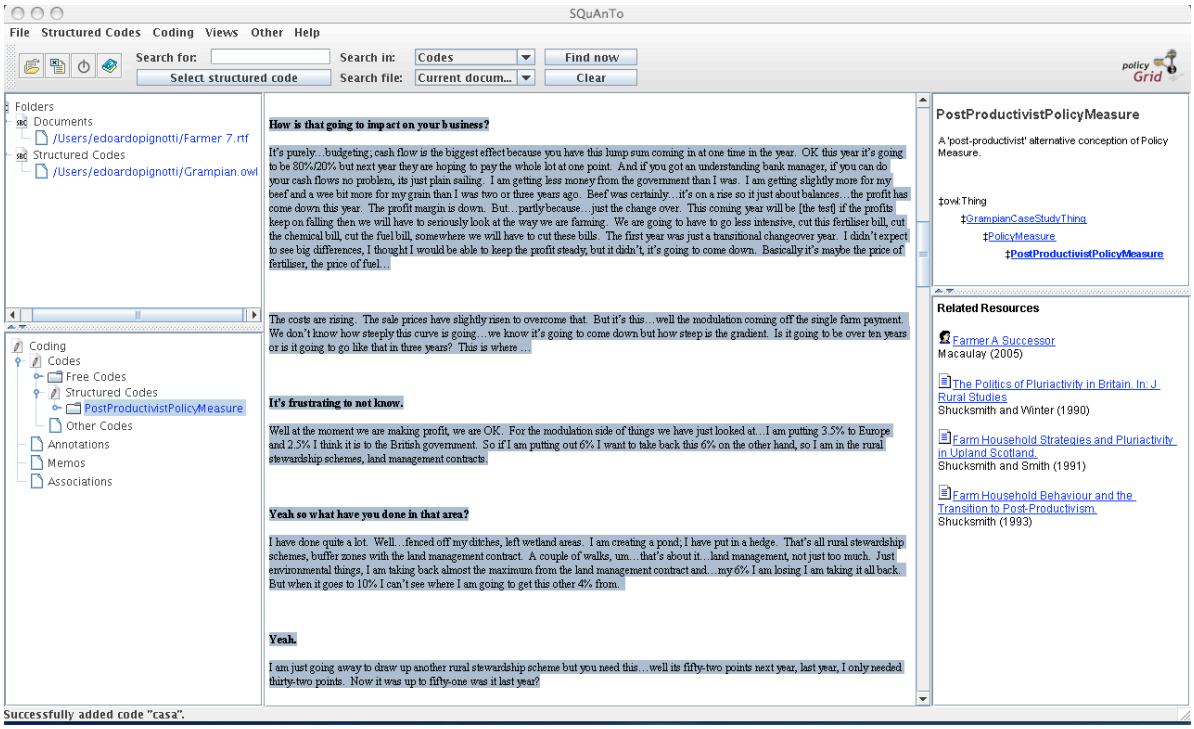

Figure 6: A Grid-enabled qualitative analysis (coding) tool.

5. Discussion & Future Work

We have highlighted the need for a mechanism to record the process of creating resources which allows research methods to be scrutinised. This provenance metadata can be used to answer queries such as "how was this evidence derived?" - providing details that would allow a third party to ascertain the robustness, or *truthfulness* of a data collection and analysis process. The network of resources inter-linked by provenance information can also be used to highlight where resources, and hence evidence, are missing, and guide the researcher in providing them or explaining why they are absent. Provenance can also be queried to show what worked (or not) in a policy assessment lifecycle and can provide feedback to a new policy assessment activity.

None of the existing provenance approaches in eScience fully support our requirements; some only model computational processes, while CombeChem (which does have many useful features) requires the experiment to be completely defined before it is executed. However, we are taking aspects of these existing architectures and building upon them to produce our provenance architecture.

For future work we want to extend the domain specific ontologies to cover more methodologies and then evaluate the architecture with social scientists to ensure that it records all of the provenance metadata they would wish to capture, and provides a way of utilising that metadata that is useful to support esocial science. A visualisation tool is also to be implemented to enable a researcher to view a graph of all of the resource associations, which could be integrated with natural language generation tools to allow detailed descriptions of the resources and the processes to be viewed.

One of the most significant challenges we face is non-technical. While disciplines such as Chemistry have a very long tradition of using lab books, meticulously recording every step they perform, the same practice is not so well embedded in social science. This means that it will be a challenge to persuade social scientists to record their data and methods of working in the detail they require for reuse. It is also difficult to obtain concrete requirements because the social scientists themselves do not know precisely what it is they want to record or even query using the provenance metadata. This means that the design of the architecture has to be flexible enough to develop as our work progresses.

Acknowledgments

The work described in this paper is funded by the UK Economic Research Council as part of the PolicyGrid Node of the National Centre for eSocial Science (award reference: RES-149-25-1027). The ideas contained here have been influenced by several of our collaborators, including Lorna Philip, Gary Polhill, and Nick Gotts.

References

Bullock, H., Mountford, J., and Stanley, R. 2001. Better Policy-Making. *Centre for Management and Policy Studies Technical Report*, Cabinet Office.

De Roure, D., Jennings, N., and Shadbolt, N. 2005. The Semantic Grid: Past, Present, and Future. *Proceedings of the IEEE*. 93(3):669-681.

Edwards, P., Chorley, A., Hielkema, F., Pignotti, E., Preece, A.D., Mellish, C., and Farrington, J. 2007. Using the Grid to Support Evidence-Based Policy Assessment in Social Science. In S. Cox, ed.: *Proceedings of the UK eScience All Hands Meeting*, 345-352.

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.

Goble, C. 2002. Positions Statement: Musings on Provenance, Workflow and (Semantic Web) Annotations for Bioinformatics. In *Workshop on Data Derivation and Provenance*, Chicago.

Goble, C. Corcho, O., Alper, P., and De Roure, D. 2006. E-Science and the Semantic Web: A Symbiotic Relationship. *Discovery Science*. Lecture Notes in Artificial Intelligence (LNAI) 4265, pp1-12.

Groth, P., Jiang, S., Miles, S., Munroe, S., Tan, V., Tsasakou, S., and Moreau, L. 2006. An Architecture for Provenance Systems. ECS, University of Southampton.

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*, 69-72.

HM Treasury, 2003. The Green Book: A Guide to Appraisal and Evaluation, London, HM Treasury.

Hughes, G., Mills, H., Roure, D. D., Frey, J. G., Moreau, L., schraefel, m., Smith, G. and Zaluska, E. 2004. The Semantic Smart Laboratory: A System for Supporting the Chemical eScientist. *Org. Biomol. Chem., 2,* 1-10.

Miles, S., Groth, P., Branco, M., and Moreau, L. 2007. The Requirements of Using Provenance in e-Science Experiments. *Journal of Grid Computing*. 5(1):1–25.

Philip, L., Chorley, A., Farrington, J. & Edwards, P. 2007. Data Provenance, Evidence-Based Policy Assessment, and e-Social Science. In *Proceedings of Third International eSocial Science Conference*.

Simmhan, Y.L., Plate, B., and Gannon, D. 2005. A Survey of Data Provenance in eScience. *ACM SIGMOD Record* 34(3): 31-36.

Stevens, R., Robinson, A., and Goble, C. 2003. my-Grid: Personalised Bioinformatics on the Information Grid. *Bioinformatics* 19(1):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* 4(2):84-101.