

Developing Ontologies to Support eSocial Science: The PolicyGrid Experience

A Chorley, F Hielkema, P Edwards

School of Natural & Computing Sciences

University of Aberdeen

{ a.h.chorley, f.hielkema, p.edwards }@abdn.ac.uk

Abstract

This paper describes the experience of the PolicyGrid project in developing ontologies for use in evidence-based policy assessment. We begin by outlining some examples of ontology use in other eScience applications, before discussing the requirements for an ontological framework to support provenance of resources in evidence-based policy research. We continue by discussing the various stages in our ontology development process, including the role played by social scientists. We conclude with a number of comments about our experience to date, and the wider implications for use of ontologies in eSocial Science.

Introduction

The Semantic Grid (De Roure, Jennings, & Shadbolt, 2005) is often described as an ‘extension of the current Grid in which information and services are given well-defined meaning, better enabling computers and people to work in cooperation’. Semantic Grids not only share data and compute resources, but also share and process metadata and knowledge. Several technologies play an important role in realising this vision, including: OWL (Web Ontology Language) and RDF (Resource Description Framework). Within the PolicyGrid¹ project we are exploring issues associated with the creation of a Semantic Grid to support social science research. While we have argued previously (Hielkema *et al.* 2007) that ontologies are only part of the answer for capturing metadata in eSocial Science, they do still play a vital role by providing a conceptual framework within which we can describe social science research activities. Using ontologies we are working towards the following:

- greater interoperability of social science datasets and tools;
- enhanced transparency of social science research outputs;
- increased potential for informed re-use of resources;
- improved facilities for resource and tool discovery.

Working with our social science collaborators, we are exploring these issues in the context of ‘evidence-based policy making’ (Bullock, Mountford, & Stanley, 2001). This came to the fore

¹ <http://www.policygrid.org/>

in the UK policy environment in response to a perception that government needed to improve the quality of its decision-making processes. Evidence can take many forms: research, analysis of stakeholder opinion, simulation modelling, public perceptions and beliefs, anecdotal evidence, cost/benefit analyses. In the remainder of this paper we discuss how ontologies can be used to provide support for evidence-based policy research, and our experience of ontology development. It is important to stress here that we did not set out to create a conceptual framework for all of social science, but rather, our aim was to investigate support for particular kinds of research activity. Later in the paper we reflect on what our experience tells us for eSocial Science in general.

A Role for Ontologies

Ontologies provide a formal specification of the concepts in a domain and the relationships between them (Gruber, 1993). Before we discuss the role for such a conceptual framework within evidence-based policy research, it is important to consider experience in other scientific domains. The Geodise project (Chen *et al.* 2003) demonstrated how ontologies could be used to enrich resource descriptions in an engineering design application, through semantic annotations. For example, log files produced by engineering design tools could be annotated to allow them to be indexed, queried and re-used. In another application area, SmartTea (Frey *et al.* 2004) employed an ontology to capture the *Processes* and *Materials* within synthetic organic chemistry experiments; this was used to describe laboratory plans and associated annotations via a tablet PC. Ontological support for social simulation experiments formed part of the FearlusG pilot project (Pignotti *et al.* 2005) in which resources including Perl scripts, parameter files and result files could be annotated with semantic metadata. In addition, an ontology was used to allow such resources to be situated within the wider context provided by a scientific argument (including hypotheses). In these example projects, ontologies are used to provide semantic enrichment to a range of digital resources and through these annotations to provide knowledge level services to end users.

In the *Green Book, Appraisal and Evaluation in Central Government* (HM Treasury, 2003) the UK Treasury recommends that as well as keeping a record of the resources and reports used in a policy assessment activity, a researcher should keep a record of what happened to enable the creation of the final report. This process documentation is often known as *provenance* and is an important aspect of record keeping. 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 (Chorley *et al.* 2007). Experience in other application areas would suggest that there is an obvious role for ontologies in supporting such a framework, by defining the resources and activities within a policy research exercise, and the relationships between them.

Metadata frameworks already play an important role in social science research. For example, the UK Data Archive², the largest collection of digital data in the social sciences and humanities in the UK, uses a schema based on the Data Documentation Initiative³ (DDI) and the Text Encoding Initiative⁴ (TEI). While the UKDA schema provides a mechanism for documenting social science resources, its reliance on XML means that tasks such as data integration are complex. It is for this reason (amongst others) that we have embraced OWL to develop our provenance framework.

² <http://www.data-archive.ac.uk/>

³ <http://www.icpsr.umich.edu/DDI/>

⁴ <http://www.tei-c.org/>

We are also developing LIBER (Language Interface for Browsing and Editing RDF), an interface that allows users without previous experience of the Semantic Web to access and create metadata. The tool uses Natural Language Generation techniques to present metadata in a textual rather than a graphical format. We will rely on this tool to gather the metadata needed to populate the provenance framework from social science users. This tool is described in (Hielkema *et al.* 2007); Figure 1 shows a screenshot of a resource description created with LIBER.



Figure 1: Using LIBER to Describe a Questionnaire Resource.

Ontology Development

Several ontology development approaches have been described in the literature; one review of these was provided by Fernandez-Lopez and Gomez-Perez (2002). Many of the approaches share common characteristics, including: study of application context (usage scenarios), preliminary design of concept hierarchy, detailed ontology structuring and refinement. Noy and McGuinness (2001) highlight a number of high level “rules” of ontology development:

- There is no one correct way to model a domain – there are always viable alternatives.
- Ontology development is necessarily an iterative process.
- Concepts in the ontology should be close to objects (physical or logical) and relationships in the domain of interest.

Our initial development strategy called for a number of domain ontologies, in addition to those required to capture provenance information. However, we quickly discovered from our user requirement gathering workshops and related research that it would be an impossible task to model social science concepts in an ontology that even one research group would subscribe to, let alone an entire community. Concepts in social science are contested and mutable (Edwards *et al.* 2006). In our early discussions, one user even stated that a personal ontology would be acceptable to him. Such an approach seemed contrary to our aim to improve the

sharing of data. We therefore decided to avoid modelling domain concepts as much as possible. Instead we focused our ontology development effort on support for provenance networks.

These ontologies are then supplemented by folksonomies. A *folksonomy* (Guy & Tonkin, 2006; Gruber, 2005) is a social classification process where users can annotate their resources with keywords or tags, which are not restricted in any way. In some folksonomies, e.g. that used by the photo-sharing website Flickr, users can use other users' tags, so that a set of frequent tags emerges. Using a folksonomy, we can guide users when they are creating annotations or constructing queries. Furthermore, we can stimulate the emergence of community vocabularies, by presenting them with an overview of the tags popular with other users. Lightweight annotation (Goble *et al.* 2006) is the process of associating resources with tags, which can be derived from folksonomies or the names of ontology classes; experience shows that this form of annotation is straightforward for non-expert users.

From this point onwards, we focus our discussion on the ontologies required to deliver a provenance framework for evidence-based policy research. The requirements for these ontologies were as follows:

- Able to capture properties of a range of social science resources (e.g. papers, interview transcripts, datasets, etc.);
- Able to describe process information in order to capture the methodological context;
- Support for process inputs and outputs to facilitate the creation of an evidence network.

In the next section we describe our first attempt at building an ontology to satisfy these requirements.

Creating the Initial Ontology

For our first attempt, we analysed the descriptions (available online) of the various datasets held in the UK Data Archive, the curator of the largest collection of digital data in the social sciences and humanities in the UK. The descriptions were analysed to identify the concepts necessary to code a description of a social science resource. The result of this initial exercise was an initial OWL ontology which modelled objects such as 'document' and 'person', but avoided abstract domain concepts such as 'rural accessibility' and 'poverty'. Domain-specific information can instead be specified by providing free-text values (or *tags*) for properties such as *SamplingProcedures*. Figure 2 shows an extract of this ontology. More information about this initial ontology can be found in Chorley *et al.* (2007).

Developing the Project, Resource and Task Ontologies

Using this initial ontology we performed a pilot evaluation study of the LIBER user-interface (Hielkema *et al.* 2007). In the event, this was as much an evaluation of the underlying ontology. Its structure made it difficult for users to find out where and how to add certain kinds of information. Finding the correct options in the menus was very time-consuming. Subjects were frequently unsure whether they had found the correct option, as some options resembled others too closely. Although menu size can be reduced by creating more detailed classifications (e.g. by only allowing 'Interviews' to have interviewers, not 'Documents'), this tends to confuse subjects with little experience in ontologies or other classification schemes;

which are precisely the users we aim to support. Some subjects stated that they had difficulties deciding where to start, as the tool does not suggest which information should be added first. In general, the ontology simply did not seem to match the subject's model of resource descriptions closely enough to enable users to easily create descriptions that they were satisfied with. We needed an ontology that was both more comprehensive and contained less ambiguous or duplicate properties; that was structured to minimise the size of each menu, without confusing users with excessive use of hierarchies; and that enabled users to describe resources in a manner with which they were comfortable.

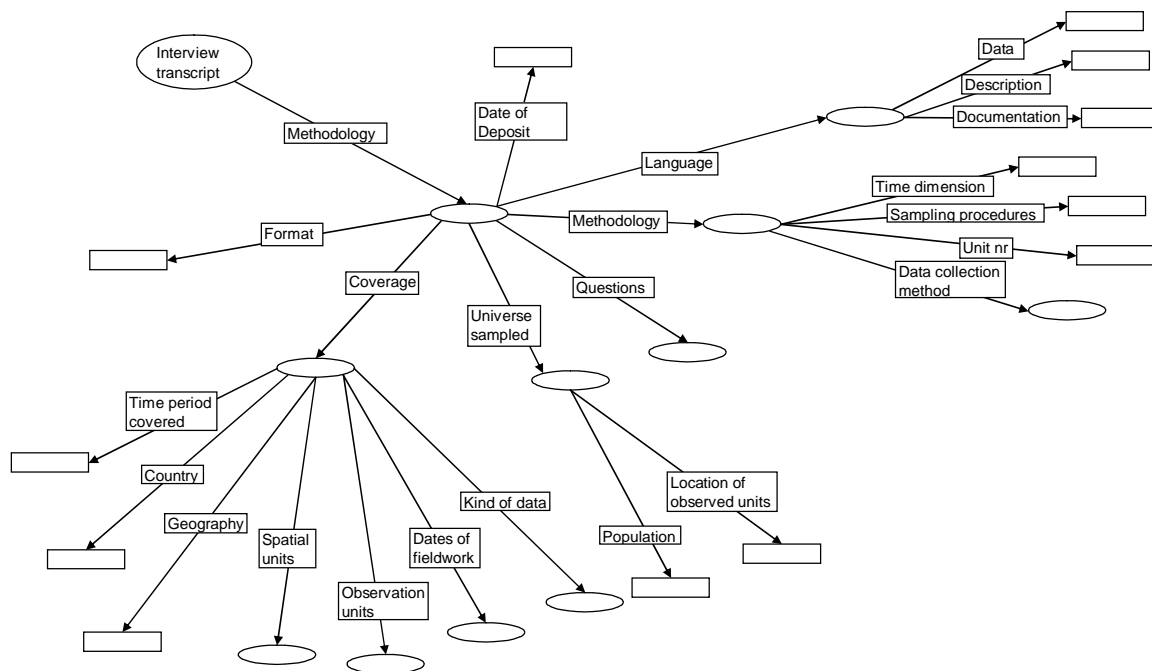


Figure 2: Excerpt from the Ontology Derived from the UKDA.

We thus set out to develop a revised set of ontologies that could be used by all tools and services that PolicyGrid is developing, while at the same time retaining compatibility with the UK Data Archive schema. In a series of interviews with social scientists, we listed the possible resource types that might be deposited, and the information that should be recorded about them. This led us to distinguish between resource types and the ‘tasks’ used to create them. For instance, a ‘questionnaire’ is used by one or more ‘questionnaire-surveys’, which can produce a ‘statistical-dataset’. As one resource can be used by more than one task, and much information can and should be recorded about each task (time, place, method, aims), these tasks should be separate entities with their own description.

We have constructed three different ontologies: *Utility*⁵, *Resource*⁶ and *Task*⁷. The *Utility* ontology is used to describe utility items such as projects and persons, while the *Resource* ontology describes resources, including information such as title, author, access rights and dates of creation and/or publication. The *Task* ontology is used to describe research tasks, e.g. focus groups, dissemination tasks or literature reviews, recording time, location, methodology,

⁵ <http://www.policygrid.org/utility.owl>

⁶ <http://www.policygrid.org/resource.owl>

⁷ <http://www.policygrid.org/task.owl>

etc. These ontologies are separate but compatible, so that the description of a resource and the process through which it was created use elements from all three. Figure 3 depicts parts of all three ontologies and how they interoperate. An interview transcript is a resource that is connected to a particular interview task, which could itself have produced other resources, e.g. a recording of the interview and factual notes about the interview. These tasks and resources all have properties such as ‘date of deposit’ or ‘date of interview’. They will also have properties describing roles, such as the interview transcript *was deposited* by a person and an interview task *was conducted* by a person.

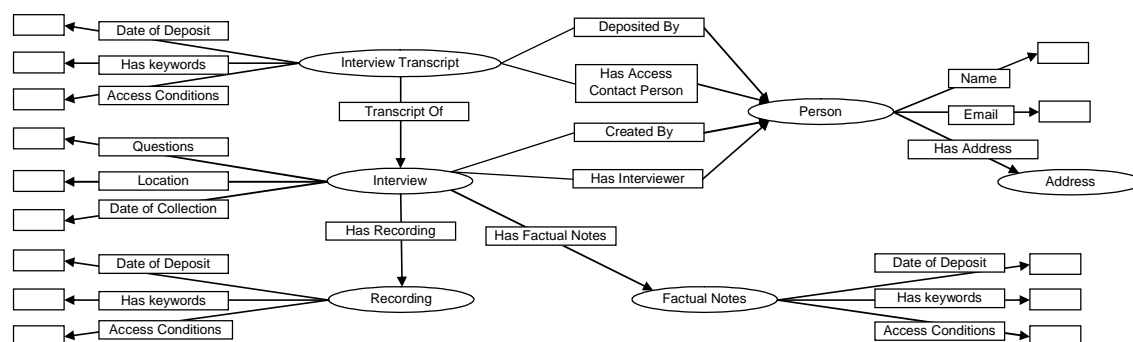


Figure 3: Using the *Resource*, *Task* and *Utility* Ontologies to Describe an Interview Transcript.

Assessing the Ontologies Through Case Studies

To determine if the ontologies were capable of capturing all of the provenance metadata required we then held a series of case study discussions with social scientists in which they were asked to describe one of their previous projects, from start to finish. They summarised each of the tasks that had been performed and the resources that were used by, or created as a result of these tasks. Such projects are highly complex with many interconnected resources and tasks. Figure 4 presents part of one of these project descriptions with the tasks depicted as boxes (labelled with the task details) and the resources depicted as circles (labelled with the resource type). The diagram summarises resources used or produced by the tasks. Once a number of these exemplars had been acquired they were used to assess the ontologies by ensuring that in each case the concepts, properties and relations were sufficient to allow the provenance record to be represented.

Each case study discussion consisted of interview sessions with social scientists in which they were asked to describe their projects and to represent the relationships between resources and tasks on a whiteboard. The output of each discussion was an interview transcript, plus a diagrammatic representation of the project using the notation shown in Figure 4. Publications and other forms of written output produced by the case study were also analysed to ensure that all aspects of the project were represented. The diagrammatic representation of the project was then shown to the social scientists for any necessary feedback and refinement.

The project depicted in Figure 4 and described in Philip and Macmillan (2005) was to explore the use of a new technique called *CV Market Stall* in investigating attitudes towards the control of wild animal species in Scotland. It is designed to allow participants in a Contingent Valuation study more time to decide what they would be willing to pay for certain wildlife projects and also to discuss the issues with other participants. Contingent Valuation studies normally involve face-to-face interviews where the participant is given information on a particular issue and are asked to make a decision in a very short time.

CV Market Stall has the participant complete a background questionnaire, then organises the discussion of the issues in a focus group. After the discussion, the participants are asked how much they would be willing to pay to have a particular wildlife project implemented, and just as importantly, *why* they have decided this. They are then sent away with a diary which they complete for a week, which is then followed-up by a telephone interview that ascertains whether their opinions and attitudes have changed after the week of reflection. This technique allows for group discussions and for detailed reflection about issues which leads to participants making reasoned decisions. It also records detailed qualitative information about the participants' decisions.

These tasks and their associated resources can be captured using our ontologies as shown in Figure 4. First, a list of participants had to be compiled in order to invite them to the focus groups. This list was then used in the questionnaire survey and the interview tasks. For the questionnaire survey, the questionnaire first had to be designed and was then used in the survey to produce completed surveys. Similarly, the telephone interview produced interviewer notes, which were transcribed and then analysed and coded.

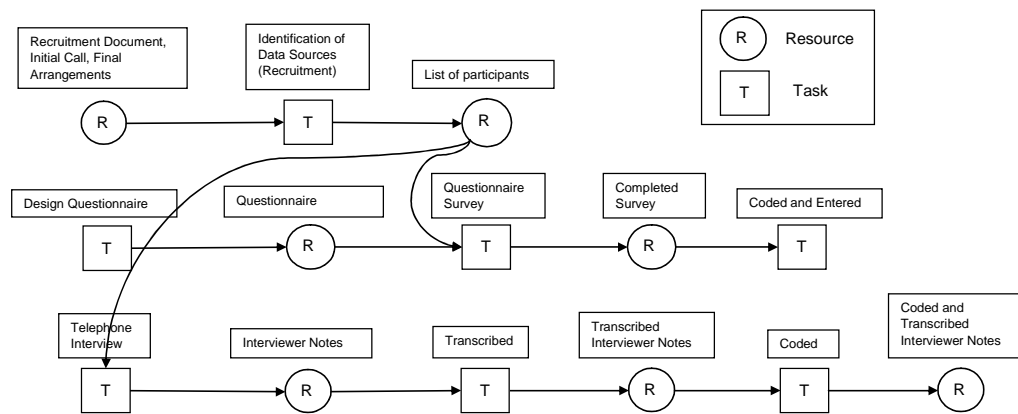


Figure 4: Provenance Extract from the CV Market Stall Case Study.

We conducted several of these case studies and in each one, we were able to demonstrate that our ontologies could capture the metadata required to document the provenance record. Metadata about all of these tasks and resources could be recorded and used to provide an audit trail for someone (e.g. funding body or government department) to examine the project.

Adapting Ontologies for Usability

Although we aim to hide the ontologies from view in our software tools, their structure influences the usability of these tools. This is especially true in LIBER, our natural language metadata creation interface. LIBER maps property and class names to text with only minor changes, i.e. mapping capitals and spaces to underscores. For instance, the property names are presented as options for adding additional information in a menu (e.g. the menu in Figure 1). This means they have to correspond to a natural language phrase that clearly presents their purpose. The tool generates a natural language text from metadata based on this ontology. Its class names are frequently inserted in the text as noun phrases (e.g. 'the person' or 'an interview transcript'). This means that certain class names, such as 'statistical data', cannot be used as 'a statistical data' is not an acceptable noun phrase.

LIBER assumes that the user will deposit or search for a certain resource. Its first act is to show the hierarchy of resource types to the user, and ask him to select one. As the class hierarchy is thus shown directly to the user, we have to avoid 'spurious' classes which do not correspond to any entity in the user's domain model. In ontology building, such classes can be introduced to group properties and improve symmetry; but in LIBER, such classes would only confuse the user.

Finally, LIBER does not provide an overview of the ontology, as its target users do not have any experience with ontologies and therefore would be unlikely to understand such an overview. However, this means that it can be difficult for users to find out how to add certain information. The PolicyGrid ontologies described above force users to describe their methodology using a *Task* object. If users cannot find out how to create a *Task*, they cannot state anything about the research that led to their resource. The ontology has to be carefully structured so that it is clear to the user where and how all information can be added. For instance, if the user deposits an interview transcript, there should be a clear option to describe the interview(s) it is based on (e.g. 'interviews transcribed').

Discussion

Our ontology development process has been a multi-faceted activity, involving: examination of existing representations and standards; face-to-face interviews with end users; assessment of prototype ontologies versus exemplar use cases; usability evaluation; iterative refinement. Balancing the sometimes conflicting demands of end users, software tool developers, usability constraints, and elegant knowledge representation continues to be a difficult challenge. In order to develop ontologies, it is essential to build an ontology development community. That community development depends on the provision of effective support for facilitating collaboration as an explicit part of system design (Ure *et al.* 2007). In PolicyGrid we have tried to involve our user community in our ontology development throughout the various stages. We hope that this has not only ensured that the ontologies reflect the user requirements, but has also established a sense of ownership amongst the end users.

It is important to compare our approach with that of the Data Documentation Initiative. DDI provides a powerful mechanism to allow the results of social science research to be archived; in contrast, our framework has been designed to support ongoing project activities in an 'organic' fashion, with new resources and task information being added throughout an investigation. The end result of this organic process is an actual research record, rather than the formal, archival, 'after the fact' presentation that DDI archivists would produce. We argue that the former is vital in the context of evidence-based policy research, to ensure that there is a high degree of transparency and accountability. While DDI (version 3.0) provides support for grouping of studies that are related along one or several dimensions (time, geography, etc.) our approach would allow such relationships to be deduced rather than explicitly coded. Despite these comments, we do acknowledge the significance of the DDI schema as a standard in social science and have incorporated many of its features into our ontologies.

To date we have a number of applications which use the *Resource*, *Task* and *Utility* ontologies. These include the LIBER tool (discussed earlier) which facilitates metadata annotation, browsing and querying; a prototype virtual research environment (ourSpaces) which integrates support for collaboration between social science researchers with access to resources (and associated provenance information); a desktop qualitative analysis tool (Squanto) which interoperates with our resource repository and provenance annotations to identify relevant resources (based upon coding activities) in real time.

As stated earlier, our aim has been to develop ontologies to support provenance in evidence-based policy research. The resource types and methodological information coded within our framework should, we believe, be applicable to a much wider range of social science research activity. However, at this stage we have insufficient evidence in order to be able to identify any boundary conditions regarding its applicability; investigating this issue further forms part of our research agenda.

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). We thank Dr Lorna Philip, Dr Colin Hunter, Professor John Farrington and Louise Reid for providing case studies and valuable input into the ontology development process.

References

- Bullock, H., Mountford, J., and Stanley, R. 2001. Better Policy-Making. *Centre for Management and Policy Studies Technical Report*, Cabinet Office.
- Chen, L., Shadbolt, N.R., Tao, F., Puleston, C., Goble, C. and Cox, S.J. 2003. Exploiting Semantics for e-Science on the Semantic Grid. *Web Intelligence (WI2003) Workshop on Knowledge Grid and Grid Intelligence*, 13-16 October 2003, Halifax, Canada.
- Chorley, A., Edwards, P., Preece, A. and Farrington, J. 2007. Tools for Tracing Evidence in Social Science. *Proceedings of the Third International Conference on eSocial Science*, Ann Arbor, Michigan.
- 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., Aldridge, J. and Clarke, K. 2006. A Tree Full of Leaves: Description Logic and Data Documentation. *Proceedings of the Second International Conference on e-Social Science*, Manchester, UK.
- Fernandez-Lopez, M. and Gomez-Perez, A. 2002. Overview and Analysis of Methodologies for Building Ontologies. *The Knowledge Engineering Review*. 17(2): 129-156.
- Frey, J.G., Hughes, G.V., Mills, H.R., schraefel, m.c., Smith, G.M. and DeRoure, D. 2004. Less is More: Lightweight Ontologies and User Interfaces for Smart Labs. *Proceedings of the UK eScience All Hands Meeting 2004*, Nottingham, UK.
- Goble, C., Corcho, O., Alper, P. and De Roure, D. (2006): 'e-Science and the Semantic Web: A Symbiotic Relationship', *Proceedings of Discovery Science 2006 LNAI 4265*, Barcelona, Spain, pp. 1-12.
- Gruber, T. 1993. A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition*, 5(2), 199-220.
- Gruber, T. (2005): 'Ontology of Folksonomy: A Mash-up of Apples and Oranges', <http://tomgruber.org/writing/ontology-of-folksonomy.htm>, 2005.
- Guy, M. and Tonkin, E. (2006): 'Folksonomies: Tidying up Tags?', *D-Lib Magazine*, vol. 12, no. 1, 2006.
- Hielkema F., Edwards P., Mellish, C. and Farrington, J. 2007. A Flexible Interface to Community-Driven Metadata. *Proceedings of the Third International Conference on eSocial Science*, Ann Arbor, Michigan.

- HM Treasury, 2003. The Green Book: A Guide to Appraisal and Evaluation, London, HM Treasury.
- Noy, N.F. and McGuinness, D.L. 2001. Ontology Development 101: A Guide to Creating Your First Ontology. *Stanford Medical Infomatics*, Stanford, CA.
- Philip, L.J. and Macmillan, D.C. 2005. Exploring Values, Context and Perceptions in Contingent Valuation Studies: The CV Market Stall Technique and Willingness to Pay for Wildlife Conservation. *Journal of Environmental Planning and Management* 48(2) 257-274.
- Pignotti, E., Edwards, P., Preece, A.D., Polhill J.G. and Gotts, N.M. 2005. Semantic Support for Computational Land-Use Modelling. *Proceedings of the Fifth IEEE International Symposium on Cluster Computing and Grid (CCGrid 2005)*, IEEE Press.
- Ure, J., Procter, R., and Lin, Y. 2007. A Socio-Technical Perspective on Ontology Development in HealthGrids. *Proceedings of the UK eScience All Hands Meeting 2007*, Nottingham, UK.