

Using ontologies with case studies: an end-user perspective on OWL

J. Gary Polhill¹ and Gina Ziervogel²

¹Macaulay Institute, Aberdeen, AB14 0PR.

²Department of Environmental & Geographical Science, University of Cape town, Private Bag x3, Rondebosch 7701, South Africa

Email address of corresponding author: g.polhill@macaulay.ac.uk

Abstract. This document describes work undertaken to develop an OWL ontology of a case study of farmers in South Africa. This work was done with a view to seeing whether or not such an ontology would be of use in working with qualitative and/or quantitative case study data. For example, could the ontology be used to check whether the data were being interpreted consistently? We describe the process by which the ontology was developed and offer some reflections on tools and utilities that would make this easier and the strengths and weaknesses of OWL for representing concepts in a real-world sociological case study.

Introduction

The current trend for evidence-based policy suggests that social scientists advising policymakers in democratically accountable societies need to be able to demonstrate the evidence for the advice they give, and allow scrutiny of the processes by which the evidence is interpreted into such advice.

Whilst there are several methodologies for establishing the truth of hypotheses from evidence in the social sciences, this paper considers the potential role that ontologies might have to play in the toolbox for analysing such evidence, based on a real-world case study of community-garden farmers in a village in the Limpopo province of South Africa. The key advantage of ontologies is their foundation on logics such as description logics, allowing automated reasoning services to consistently apply definitions of key concepts. However, there are computer science issues pertaining to reasoning with ontologies that impose constraints on the expressivity with which concepts can be described. A key part of this exercise is therefore giving consideration to the limitations and prospects for using ontologies within the social sciences.

Scenario and source material

In 2002/03 there was a food-security crisis in South Africa (Mano et al., 2003). One of the issues identified was the availability and applicability of long-range climate forecasts to farmers growing food for home consumption. Farmers can use these climate forecasts to adopt strategies that mitigate against climate stress, such as planting drought-resistant cultivars. However, climate forecasts are aimed largely at the commercial farming sector, and farmers often operate in a multi-stressor environment where climate is just one issue among many others such as land access, political instability, market fluctuations, globalization and

HIV/AIDS. Farmers growing food for home consumption also tend to be women supporting their families while their husbands work in the city, and have been found to be less likely to be part of the climate information dissemination network (Archer, 2003). For some farmers, provision of the forecast needs to be conducted alongside advice on how to use it in their own circumstances.

A case study was conducted of farmers working in a community garden in a village in the Limpopo province of South Africa. The communal farming project was started in 1993 with the aim of supporting women in the production of vegetables for combating malnutrition among children (Archer, 2003). After the land was prepared, the first crops were grown for subsistence in 1996, but in later years vegetables have also been grown for selling. The project has had varying degrees of success depending on such things as whether the pump for the irrigation was working, or money was available for agricultural inputs. The case study involved structured interviews in 2001 (not used or referred to elsewhere in this document), followed by 51 interviews in 2003 asking similar questions. Nine in-depth interviews were then conducted in August 2003 involving farmers from the 51 farmer sample. The source data used in this paper consisted of a spreadsheet containing the data from the 51 interviews in 2003, and a document giving interview profiles of each of the nine-farmer in-depth interviews.

In developing OWL (Antoniou & van Harmelen, 2004) ontologies of this case study, we drew on the source data, and on three papers. One, by Emma Archer (2003), focused on issues in climate information dissemination worldwide, whilst the second (draft) paper (Archer & Easterling, n.d.) discussed the South African climate information dissemination system in particular, identifying areas in which it could be improved. The third paper (also a draft) (Ziervogel & Bharwani, n.d.) compared various strategies employed by farmers in the case study to adapt to multiple stressors (including climate) influencing their livelihoods, and thus had the greatest influence of the three papers on the ontology.

Ontology development methodology

Developing ontologies of case studies in social science is something that, to our knowledge, has not yet been tried. As a first attempt, the written material (the three papers) was used as a basis for deriving what might be termed a 'scenario ontology'. Such an ontology outlines the key aspects of the case study: what the major narrative agents are, and how they interact. As might be expected, the papers contained a rich vocabulary of different types of farmer. To validate the ontology, we attempted to enter some of the data from the spreadsheet into the ontology, with a view to using reasoning services to see if the farmers were classified according to the conceptualisation of the case study researcher (GZ).

Entering the data directly into the scenario ontology, however, though possible, proved unwieldy, and it became clear that a better approach would be to build up the scenario ontology from the source data. The advantage of this is that the ontology development process follows the workflow of the case study researcher, from survey through to analysis. Using OWL's import facilities, a series of ontologies were developed, from the source data to a rudimentary scenario ontology (Figure 1).

Figure 1 shows the four stage process, which followed the workflow of the practitioner, involved in developing the final scenario ontology based on the source data. All ontologies were built using Protégé¹.

1. *Development of survey ontology.* This ontology used the column headings in the spreadsheet as a basis for creating an ontology. Classes were created to cover major groupings of columns (planting plans, households, income), which were linked back to the interview using object properties. Individual columns suggested datatype properties for the classes. Classes are also used to support qualitative answers to questions (i.e. textual responses not constrained to a predetermined set of options).
2. *Development of survey data ontology.* The survey ontology contains only the classes and properties; the survey data ontology then contains the data from the 51 farmer interviews of 2003: the entries in the spreadsheet. It is debatable whether the survey data ontology should be separate from the survey ontology, but there are good reasons to do this. Firstly, one can imagine using the same survey ontology for a different data sample (e.g. the 2001 sample or the 2003 sample in this case study); in this case it is desirable to keep the samples separate, which can only be achieved by separating the data from the survey ontology. Secondly, it is possible that one might find the need to create new classes or properties during the course of filling out the data. Having a separate survey data ontology documents the stage at which concepts are created, allowing explicit distinctions to be made between preconceptions as embodied in the survey ontology, and practicalities required for the survey data.

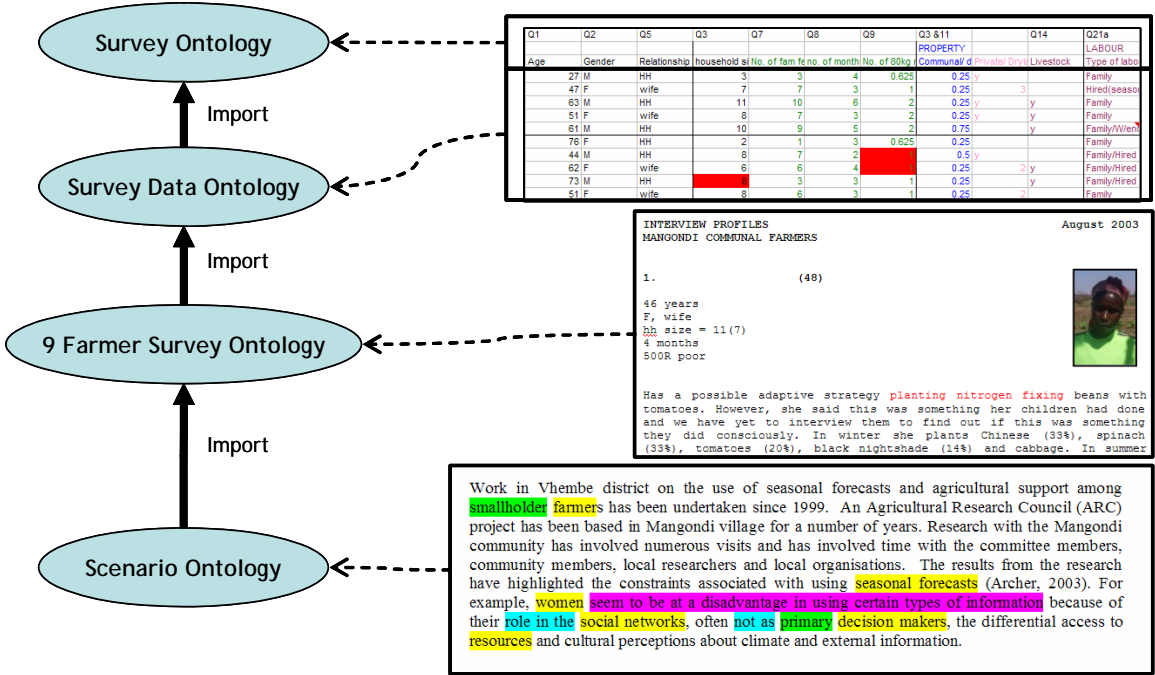


Figure 1. Relationships between ontologies and source material

3. *Nine-farmer survey ontology.* The 51 farmer survey was followed by in-depth interviews of nine of these farmers. This ontology captures extra information recorded by this survey. In this case, for example, the farmers were asked who approached them

¹ <http://protege.stanford.edu/>

for advice, and who they asked for advice, requiring extra properties to be added to the ontology to capture this information.

4. *Scenario ontology.* The scenario ontology was then built importing the nine-farmer survey ontology as a foundation, and adding material based on the texts written about the case study in the three articles mentioned above. The process for developing this ontology is closely related to the methodology used for the original scenario ontology.

The original scenario ontology was developed from the source texts using a seven-stage manual process outlined below. Automatic ontology learning from source texts is an on-going research area, which Gómez-Pérez & Manzano-Macho (2005) conclude has yet to establish a detailed methodology (p. 207), and none of the tools available are able to evaluate their accuracy (p. 208). That said, the process could no doubt have been facilitated by such tools. As it is, the ability to search for words and automatically highlight them (standard functionality provided by word processing software) proved useful (figure 1).

1. *Assembling source material.*
2. *Deciding high-level classes.* The source material was searched for the key concepts and central themes that would suggest the classes and relationships that would be needed in the ontology. In this case, these were such things as farmers, resources, and social networks. In figure 1, they are highlighted in yellow.
3. *Detailed analysis.* Again, using the source material, instances of the high-level classes were highlighted, and linguistic modifiers (e.g. adjectives) of them analysed to suggest subclasses, and datatype or object properties (highlighted in green in figure 1). Several subclasses of farmer were identified by this process, including `WomenFarmers`, and `SmallholderFarmers`. From the verbs in the sentences using these terms, it was possible to discover differences among the various classes and areas of commonality, e.g. that farmers choose the crops they will plant. One issue that became apparent in this stage of the process was the difference between ontological aspects of the scenario and research findings. For example, in gender analysis it might be that one starts with the assumption that women are disadvantaged (i.e. this fact is part of the ontology). Equally, however, one might prefer to define what it means to be disadvantaged and subsequently find that members of this class are predominantly women (magenta in figure 1). Confusing the two would be undesirable as it could lead to seeking evidence in the data, for or against a tautology.
4. *Developing support classes.* Support classes, for the purposes of this paper, are classes that are not defined as part of the scenario, but are assumed as common knowledge, such as social networks. Ideally, such classes would be imported from other ontologies, something OWL is specifically designed to do. However, in practice, searching for ontologies to import (e.g. using engines such as Swoogle²) proved too laborious: the tools are not yet mature enough to enable users to quickly sift through the results to determine which are appropriate.
5. *Class definition.* The distinction between primitive and defined classes in OWL is significant when it comes to the application of reasoning services, since only defined classes can be directly inferred to have members.

² <http://swoogle.umbc.edu/>

6. *Testing.* The ontology can be tested using some built-in features of Protégé, as well as checking for concept consistency, and checking the classification of test individuals using the reasoner. Data from the questionnaires can also be entered into the ontology, to check for the absence of concepts that applied to the gathering of evidence, but did not appear in the case study texts.
7. *Refinement.* The ontology is refined by iterating the process to bring in extra source material, respond to the results of the tests, or import existing ontologies for support classes.

Of these seven stages, when developing the final scenario ontology, searching for key terms and providing detail on them (stages 2 and 3), to such a point that they can be defined (stage 5), is of interest in developing an ontology that can be used to verify/support statements in the source texts.

OWL and ontologies in social science

Beyond the question of whether ontologies are something that could ever be of use in the practice of social science is the question of the adequacy of OWL for the task. OWL is rapidly emerging as the de facto standard ontology language. Its integration with the web, through an XML based syntax, makes it an integral part of the efforts to create the semantic web (Berners-Lee et al., 2001), and the associated facility to import ontologies has been used here. OWL is being used in particular in biomedical domains.

OWL ontologies consist of a set of concepts (also referred to as classes) arranged in a hierarchy using a relation `rdfs:subClassOf` in which **X** is a subclass of **Y** iff an instance of **X** is necessarily an instance of **Y**. Classes are not assumed to be disjoint, unless they are asserted to be so using the `owl:disjointWith` relation. There is also a set of properties (sometimes referred to as roles), which allow classes to be associated with each other and with primitive datatypes (such as numbers or strings). Properties have a domain and a range, each of which is a set of classes that could be said to be the subject and the object of the property respectively. Properties with primitive datatypes as range are referred to as datatype properties, whilst those with classes as range are referred to as object properties. Properties can be functional (among other characteristics), which means that anything having that property can only have one value for it. Properties can also be arranged in a hierarchy using the `rdfs:subPropertyOf` relation, in which *A* is a subproperty of *B* iff for all *x*, *y* such that *x A y* it is necessarily true that *x B y*. (For example, if *x* is-the-mother-of *y*, it is necessarily true that *x* is-the-parent-of *y*.) OWL classes may have restrictions put on the properties that they are the domain of. These restrictions define necessary conditions for individuals to be members of a class. If such restrictions are also sufficient conditions for instance membership, then the class is said to be defined. An OWL ontology is effectively a set of axioms (or assertions) that may be used by an inference engine to construct proofs (e.g. that an individual *x* must be a member of class **Y**).

The limitations of OWL are closely tied with the theoretical limitations of computers. Reasoning in OWL-DL is worst-case NEXPTIME, which would traditionally be regarded as intractable, though fortunately typical-case reasoning is feasible. OWL-DL is currently the most expressive species of OWL with decidable reasoning. The loss of the ability to use reasoning services would detract considerably from the potential utility of ontology development in the social sciences in establishing consistent use of terminology. However, OWL-DL has fairly limited expressivity, and whilst it may be adequate for the domain of

biomedicine, it may not be possible to create a sufficiently expressive ontology language (with decidable reasoning algorithms) to do justice to the complexity of the social sciences.

This section explores some of the issues in defining terms in this case study. Although OWL-DL has proved sufficient for defining some of the terms, there are particular issues with OWL that mean that other terms are beyond its expressive powers (even that of OWL-Full). Though the scenario ontology uncovered more than ten subclasses of farmer of potential interest, it will suffice to demonstrate both the potential and the limitations of OWL by exploring just a few: farmers who are wives to the head of the household; wealthy/poor farmers; and smallholder farmers.

What we are interested in is the ability to define the concepts based on the evidence supplied by the source data, and on case study knowledge of the practitioner. (The separation of the survey data ontology from the scenario ontology allows explicit representation of which is which.) Farmers who are wives to the head of the household are an interesting group of people to identify, because they are to an extent subservient to the head of the household, and thus not wholly autonomous when it comes to making decisions. The questionnaire explicitly asks what relationship each interviewee has to the head of the household. Creating a defined class in OWL to automatically identify such farmers is thus a relatively trivial matter. In the questionnaire ontology, a class `Relationship` was created, and individuals belonging to that class for each relationship that might appear (wife, husband, self, daughter, etc.). One of the properties of the `Person` class is the `hasRelationshipToHouseholdHead` property, with range the `Relationship` class. `Farmer` is a subclass of `Person`. When entering data about each farmer, the appropriate relationship individual is then selected from the available list. A defined class then simply has the conditions that the individual is a `Farmer`, and that they have the wife individual as the entry in the `hasRelationshipToHouseholdHead` property. Figure 2 shows the OWL code generated by Protégé from the creation of the class `HHWifeFarmers`.

```
<owl:Class rdf:about="#HHWifeFarmers">
  <owl:equivalentClass>
    <owl:Class>
      <owl:intersectionOf rdf:parseType="Collection">
        <rdf:Description rdf:about="VillageSurvey.owl#Farmer"/>
        <owl:Restriction>
          <owl:onProperty rdf:resource="VillageSurvey.owl#hasRelationshipToHouseholdHead"/>
          <owl:hasValue rdf:resource="VillageSurvey.owl#wife"/>
        </owl:Restriction>
      </owl:intersectionOf>
    </owl:Class>
  </owl:equivalentClass>
</owl:Class>
```

Figure 2. OWL code for the `HHWifeFarmers` class in XML syntax.

Defining wealthy and poor farmers is rather more difficult. Although the ontology captures income data (from farming, pensions and employment), OWL cannot process numeric entries in datatype properties. It is thus impossible to define a class using the income data as a basis. However, one thing added by the nine-farmer data was a wealth category with values: ‘poor’, ‘average’ and ‘wealthy’ (the term ‘better-off’ was used in the ontology). Creating a defined class of poor farmers from these nine farmers is thus a trivial matter of inspecting the value of the `hasWealthClass` datatype property of the farmers and selecting those with value `poor`.³

³ There is a potential design pattern in linking properties with numeric datatypes and their categorizations (e.g. age in years, and whether the individual is ‘old’ or ‘young’), and using the categorization as a basis for defining class

It might be possible to generalize from these data to the rest of the 51 farmer survey using the income data of the nine farmers and the category they have been given as a basis, but in doing so, it is clear that there is considerable overlap between the categories; figure 3 shows in particular the overlap in categories between poor and average farmers.

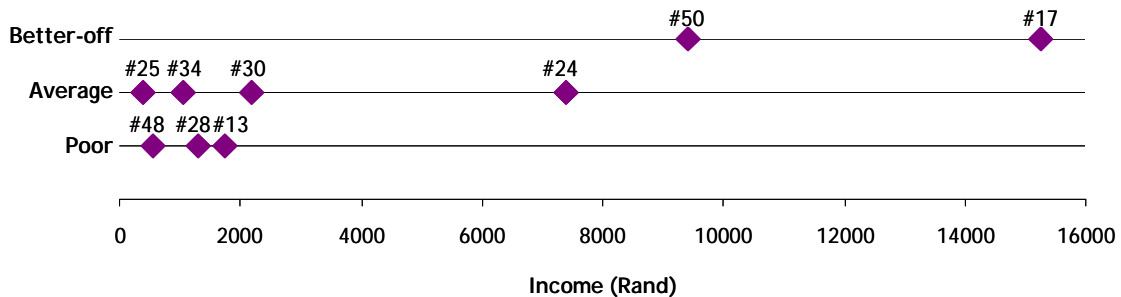


Figure 3. Wealth category and income for the nine-farmer survey.

Clearly income is not all that determines wealth, and other factors are at play. For example, if a farmer belongs to a wealthy household, but does not herself earn very much money, her personal income may be low, but she may be regarded as wealthy nevertheless. The 51 farmer survey data as recorded in the spreadsheet does not capture which farmers belong to the same household, but if it did, there might be scope for using this information in a class definition. A second issue is the source of the income. Many of the farmers have permanent (off-farm) employment and/or pensions of one form or another. These are more secure sources of income than income from farming, which is much more uncertain, depending on variable factors as climate and market demand. Thus an individual whose income comes solely from farming may be regarded as poor even though in a particular year they happened to have made a good income.

Thus, we might define a poor farmer as one who does not belong to a household with a person asserted to be wealthy in it, and whose income comes solely from farming. However, perhaps even this is insufficient, as surely the size of the holding is of significance. The concept of the smallholder farmer is thus potentially relevant. Defining such a class using an ontology would be of particular interest in comparing case studies: the word ‘smallholder’ clearly evokes a question of spatial scale of land holding, but in different geographical situations, the size of holding required for class membership will be different. Here it is a holding of roughly 1 Ha. The weakness of OWL in its ability to process numeric datatype properties now prevents any hope of defining this class, however. Short of creating a class that defines a smallholding as the union of all values in the range required for class membership (not very practical), we must instead rely on assertions to indicate membership.

The question, then, for quantitative aspects of analyzing case studies using OWL, is whether numeric conditions on datatype properties can be added to OWL for defining classes without creating an undecidable logic. For more complex classes that OWL is too weak to define, there may still be a utility in asserting concept membership given the availability of software to mine the ontology for commonly held properties of such classes as opposed to their disjoints. Such an analysis might enable the creation of more precise definitions, and the search for members whose asserted membership is dubious.

membership. However, it might be simpler from the point of view of the ontology just to create disjoint classes (‘OldPerson’ and ‘YoungPerson’) and assert membership.

For qualitative aspects of case studies, ontologies allow explicit representation of where classifications have been made on interview responses, or where responses are regarded as equivalent. For example, one of the questions asked of the 51 farmers was how they responded to the climate forecast. The practice in the survey data ontology used was to select individuals from a class of strategies for responding to climate. However, it would have been better to record the full text of each individual's response uniquely, and then in the scenario ontology, organize the responses into a series of classes using assertions. This would allow explicit representation of how the researcher had classified individuals' responses, giving other researchers the opportunity to debate them.

Tools and technologies

Ontology development is quite a time-consuming process. There are various technologies involved, and the usability of these technologies is critical in determining the level of uptake of any methodology using ontologies. The three key technologies currently are OWL itself; applications for creating ontologies (of which Protégé is probably the most popular); and reasoning software.

OWL is not as easy to get into as it might appear. The first hurdle is understanding the differences between the three main species of OWL: OWL-Lite, OWL-DL and OWL-Full, which in the order stated allow increasing expressivity at the cost of tractability of reasoning. The major issue, however, is understanding the subtleties of each assertion made. For example, asserting that the range of a particular property is a particular class does not constrain individuals asserted or inferred in the range to be known already to belong to the class stipulated, but instead allows the reasoner to infer that the individuals belong to the class. Equally potentially unexpected is the assertion that a property is functional, which allows the reasoner to infer that two or more individuals in the range must in fact be the same individual, rather than imposing a constraint. Without a thorough understanding of these subtleties, it is very easy to create an inconsistent ontology. OWL's usability is further constrained by such things as the inability to put spaces in the RDF ID field, making classes, properties and individuals look more technological.

Protégé is an application still very much in its infancy. The facilities it provides for creating an ontology are comprehensive, but its user interface is probably still not suitable for those without some experience in computer science. It is prone to crash or hang unexpectedly, resulting in loss of data for those who do not save their work regularly. Visualisation tools are also still rudimentary. However, as the popularity of OWL continues to increase, so a larger user-base will demand better ontology creation and visualisation tools. Protégé also provides a facility for entering individuals into the ontology, which was used here to enter the 51 farmer data into the survey data ontology. This facility is not well-suited to the needs of the practitioner when entering data; the ability to use a more familiar (and cleaner) interface such as a (preferably automatically generated) web-based entry form or even a spreadsheet would be much better.

Reasoning software can tell you that an ontology is inconsistent, infer the satisfiability and equivalence of classes, and infer membership of classes. It is the last of these facilities that will be used here to establish the consistent use of terminology. However, thus far, there are no reasoners that can integrate with Protégé and explain how a particular conclusion is reached. In the case of inconsistency, this is vital in assisting with debugging the ontology (Kalyanpur et al., 2005). The reasoner used for this work was RACER version 1.7.24, a

precursor to the now commercially available RacerPro⁴ that is not downloadable any more. Commercial licences pose a financial obstacle to adoption of technology in the community and do not usually allow inspection and modification of source code. Manchester University's FaCT++ reasoner⁵ uses ordering heuristics (Tsarkov & Horrocks, 2005) that make it quicker than RACER 1.7.24, and released under a GNU General Public Licence, making it much better suited to academic work. The University of Maryland's SWOOP⁶ ontology editor (Kalyanpur et al., 2006) is integrated with its Pellet⁷ reasoner (Sirin et al., n.d.) product, however, and claims to expose the workflow of the reasoner for purposes of debugging ontologies. It has not, at the time of writing, been tried by the authors, though both are released under open-source licences.

A final but significant issue is that all but the smallest ontologies are difficult to explore and understand without a user guide of some sort. Although provisions are made for comments, there are no documentation standards for ontologies, nor tools to help with documentation. In ontologies that are supposed to be based on evidence, recording the provenance of particular assertions could be useful. Ontologies are supposed to make assumptions explicit, but until it is easy for those not initiated to the world of ontologies to navigate, explore and understand an ontology, this claim does not hold. Embryonic facilities do exist, such as the tool-tips facility in Protégé to display somewhat unwieldy natural language translations of OWL phrases, and visualisation tools such as OWLViz and OntoViz. Tools are needed that allow the construction of a narrative tour of an ontology to facilitate familiarisation with it and criticism of it, for users who are experts in the field that the ontology pertains to, but not experts in ontologies themselves.

Reasoning with case study data

To demonstrate how the reasoning services associated with the ontology could be used, we consider here four statements made in the source texts for the scenario ontologies:

Statement 1. “a farmer who is the head of the household prefers radio as a medium of [climate] dissemination.” (Archer, 2003, p. 1529)

Statement 2. “Farmers who are wives to heads of the household ... prefer that seasonal forecasts be provided through the extension officer.” (ibid.)

Statement 3. “women are ... more concerned with having crops for home consumption” (Ziervogel & Bharwani, n.d., p. 17)

Statement 4. “wealthier farmers pursue more market-driven strategies” (ibid.)

Both statements 1 and 2 are supported by evidence given in Archer (2003, fig. 1), who shows 11 of 19 heads of household preferring the radio, and 17 of 27 wives, children or mothers of the head of the household preferring an extension agent. To test this using the ontology, we created subclasses of farmer for household heads (`hasRelationshipToHouseholdHead: self`) and wives of the household head (figure 2). We also created subclasses of farmer based on their answer to the question, “What would be your chosen method of receiving [a

4 <http://www.racer-systems.com/>

5 <http://owl.man.ac.uk/factplusplus/>

6 <http://www.mindswap.org/2004/SWOOP/>

7 <http://www.mindswap.org/2003/pellet/>

forecast]?” This includes options to select ‘radio’ or ‘extension agent’, which are encoded in the ontology as individuals of the class `ClimateInformationNetworkNode`. There are then four defined subclasses to create to cover each possible intersection between household head / wife and preference for radio / extension agent. Using reasoning software, membership of these classes was as shown in Table I.

Preference	Household Head	Wife to HH	Men	Women
Radio	13	14	9	22
Extension officer	8	8	4	13

Table I. Results using the reasoner of numbers of individuals in each subclass

These data give support to Statement 1, but disagree with Statement 2. Wives to the household head have a clear trend to prefer the radio as opposed to the extension officer, albeit that there is insufficient evidence in the data to confirm this with a binomial test with the null hypothesis that they prefer an extension officer (p-value 0.14, B(22, 0.5)).

Archer concludes that “a focus on characteristics that may marginalize a key user subgroup ... would and should reinforce and inform operational efforts aimed at improving forecast applications for underserved user groups” (p. 1530). A farmer that prefers the extension officer is pretty much as likely to be a household head as a wife to the household head on the basis of the above data. There is thus no evidence in these data that wives of household heads are significantly more marginalised or underserved than household heads as regards preference for climate information dissemination. However, the sample does contain an unusually high proportion of female-headed households. Table I also contains data by gender rather than relationship to household head. Looking at preferences for an extension officer, 17 of 48 people (35%) prefer the extension officer and 35 of all 48 people (73%) are women, compared with 13 women of the 17 who prefer the extension officer (76%). The percentage of women is slightly larger in the extension officer preferring subpopulation than in the population as a whole, which is a trend towards what needs to be shown to state that women are underserved if extension officers are less available than the radio. However, there is far from sufficient evidence to reject the null hypothesis that those who prefer the extension officer are more likely to be men using a binomial test (p-value 0.49, B(17, 0.271)).

Turning to Statement 3, there is no direct evidence in the original data of whether crops are specifically grown for home consumption. Some interviewees do say that they do this, but there is nothing to suggest that those who have not mentioned it do not. The survey data contains only details of which crops were planted and when (recorded in the `Plantings` class in the ontology), but certain crops are more likely to be grown for home consumption than for the market (e.g. maize, and to a lesser extent, black nightshade and Chinese lettuce). On this basis, certain crops are asserted to be `HomeConsumptionCrops`, and it can thus be inferred that certain `Plantings` are `HomeConsumptionPlantings` (if `HomeConsumptionCrops` appear in the previous or current year of the `Planting`). A `HomeConsumptionFarmer` is then a farmer with a `HomeConsumptionPlanting` as the `hasCropsPlanted` property value. Using the reasoner, we then determined that 2 of 12 men plant for home consumption as opposed to 20 of 39 women. Statement 3 is therefore confirmed by the evidence insofar as the assumptions about the crops hold. There is not quite sufficient evidence, however, to reject at the 0.05 level of significance the null hypothesis that men are more concerned with planting for home consumption using a binomial test (p-value 0.08, B(22, 0.235)).

Statement 4 suffers from the issues discussed earlier about determining the wealth of farmers. Putting that to one side, there is then the ontological issue of defining those who pursue more market-driven strategies. The interview data have several relevant questions that could be drawn on to evaluate this. Firstly, farmers are asked directly whether or not they market their crops. The farmers asked belong to a group of farmers who own plots on a communal garden, and a condition of plot ownership is that some of the produce grown is sold to local schools. The answer to that question is thus perhaps of less relevance in determining how market driven the farmers are. Another question asks where the farmers sell their produce, to which some have answered that they sell in neighbouring villages, in Thohoyandou (a town near the case study site), and in Johannesburg. With the assumption that more market driven farmers are more likely to sell their produce away from their immediate locality, we can define a class `OutsideVillageMarketingFarmer` and inspect which of those are `BetterOffFarmers` or `PoorFarmers`.

Using the assertions of wealth category provided by the nine-farmer data, none of the better-off or poor farmers are found using the reasoner to belong to `OutsideVillageMarketingFarmers`. Assuming, however, that anyone with an income lower than the poorest asserted averagely well-off farmer in figure 3 is poor, and that anyone with an income greater than the poorest asserted wealthy farmer is better-off, we can expand the membership of `PoorFarmers` and `BetterOffFarmers` with a further series of assertions based on wealth. With this expanded membership, only one better-off farmer belongs to `OutsideVillageMarketingFarmer`, but no `PoorFarmers`. Even without using the expanded membership, the means, medians, and upper and lower quartiles of the distribution of income of the `OutsideVillageMarketingFarmers` are all greater than those of the population of interviewed farmers as a whole. Thus there is some support in the evidence for Statement 4.

It is worth briefly reflecting on the question of whether all this analysis really needs to be done with OWL and reasoning software. The original data came in a spreadsheet, and clearly social scientists are used to working with spreadsheets to store and analyse their data. With the appropriate application of formulae in the cells, it would not have been difficult to conduct the same analysis as that done above. Arguably, the definitions of concepts would be explicit as expressed in the formulae, and whether one prefers such a formula to the visualisation provided by Protégé may well be a matter of taste. However, formulae in a spreadsheet are not explicitly a set of assumptions, whereas ontologies are designed to define terminology. It may be, however, that providing appropriate facilities within spreadsheet software would be a better way to encourage adoption of ontologies.

Conclusion

This paper has shown how ontologies of a real social science case study can be created and used to establish empirical support for statements made in written material. We have also discussed the limitations of OWL and associated tools for creating such ontologies. Whilst we have been able to show the potential of ontologies in a few cases for evaluating the evidence for statements in source texts, it is clear that there are sufficient limitations in the technology as it stands to prevent widespread adoption of these techniques. These limitations can only be overcome through continued application of ontologies to real-world case studies in partnership between computer scientists and social scientists. The goal, of creating transparent analyses of source data with consistent application of terminology, is surely one worth pursuing. If so, it is imperative that the social sciences engage with the development of relevant technologies, tools and services to ensure their needs are met as a matter of core priority rather than as an afterthought.

Acknowledgments

This work was funded by the Scottish Executive Environment and Rural Affairs Department, and by the EU Framework Programme 6 NEST Pathfinder Initiative on Tackling Complexity in Science.

References

- Antoniou, G., & van Harmelen, F. (2004) Web Ontology Language: OWL. In Staab, S., & Studer, R. (eds.) *Handbook on Ontologies*. Berlin: Springer, pp. 67-92.
- Archer, E. R. M. (2003) Identifying underserved end-user groups in the provision of climate information. *Bulletin of the American Meteorological Society* November 2003, pp. 1525-1532.
- Archer & Easterling (n.d.) Mitigating climate stress: Strengthening the 'end-to-end' climate information system in South Africa. *Draft paper to be submitted to Environment*.
- Berners-Lee, T., Hendler, J. & Lassila, O. (2001) The semantic web. *Scientific American*
- Gómez-Pérez, A. & Manzano-Macho, D. (2005) An overview of methods and tools for ontology learning from texts. *The Knowledge Engineering Review* 19 (3), pp. 187-212.
- Kalyanpur, A., Parsia, B., Sirin, E., & Hendler, J. (2005) Debugging unsatisfiable classes in OWL ontologies. *Journal of Web Semantics* 3 (4), 268-293.
- Kalyanpur, A., Parsia, B., Sirin, E., Cuenca-Grau, B. & Hendler, J. (2006) SWOOP – A web ontology editing browser. *Journal of Web Semantics* 4 (2).
- Mano, R., Isaacson, B., & Dardel, P. (2003) Identifying policy determinants of food security response and recovery in the SADC Region: The case of the 2002 food emergency. *FANRPAN Policy Paper, keynote paper prepared for the FANRPAN Regional Dialogue on Agricultural Recovery, Food Security and Trade Policies in Southern Africa, Gabarone, Botswana, 26-27 March 2003*.
- Sirin, E., Parsia, B., Cuenca Grau, B., Kalyanpur, A. & Katz, Y. (n.d.) Pellet: A practical OWL-DL reasoner. *Submitted to Journal of Web Semantics*.
- Tsarkov, D., & Horrocks, I. (2005) Ordering heuristics for description logic reasoning. *Proceedings of the International Joint Conference on Artificial Intelligence IJCAI-05*, pp. 609-614. <http://ijcai.org/papers/1352.pdf>
- Ziervogel, G., & Bharwani, S. (n.d.) Adapting to variability: Pumpkins, pumps, poverty and the role of climate. *Draft paper dated February 2004*.