Ontology Ontogeny: Understanding how an Ontology is created and developed

Hayley Mizen[1], Catherine Dolbear[1], and Glen Hart[1],

[1] Ordnance Survey, Research and Innovation, Romsey Road,
Southampton SO16 4GU United Kingdom
{Hayley.Mizen, Catherine.Dolbear, Glen.Hart } @
ordnancesurvey.co.uk

This paper describes the development of a systematic method for creating domain ontologies. We have chosen to explicitly recognise the differing needs of the human domain expert and the machine in our representation of ontologies in two forms: a conceptual and a logical ontology. The conceptual ontology is intended for human understanding and the logical ontology, expressed in description logics, is derived from the conceptual ontology and intended for machine processing. The main contribution of our work is the division of these two stages of ontology development, with emphasis placed on domain experts themselves creating the conceptual ontology, rather than relying on a software engineer to elicit knowledge about the domain. In particular, this paper concentrates on the creation of conceptual ontologies and analyses the success of our methodology when tested by domain experts.

1 Introduction

Ordnance Survey, the national mapping organisation for Great Britain, is investigating the potential benefits of introducing a Topographic Semantic Reference System to improve the integration of topographic and thematic data. The ultimate purpose is to enable machine understanding, which in turn provides the potential for data and service interoperability. An ontology is an important component of a semantic reference system, and we are therefore researching the nature of such ontologies and methods to create them. This paper describes our current work on developing a methodology to create domain ontologies. In part we have titled the paper "Ontology Ontogeny" to emphasis our interest in the development of ontologies, ontogeny being the development processes an animal undergoes from egg to adult; but in part we just thought it too good a conjunction of terms with similar roots to miss.

Section 2 provides background to the research, explaining our motivation and placing the research in context. We provide a brief review of other approaches to ontology construction in Section 3 and outline our own views on the structure of ontologies in Section 4. In Section 5 we describe our own methodology and in Section 6 provide an analysis of its success to date. Finally, Section 7 contains our closing observations and suggestions for future research directions.

2 Background

Ordnance Survey has the challenge of enabling third parties to integrate their data with the topographic data that it provides. In order for an organisation to complete its business related tasks it is frequently necessary for multiple data sources to be combined (integrated) and used together in a structured way. As there may be differences in semantics as well as in the structure of these datasets, the data must be adapted to fit the task, often with compromises being made. Currently, the cost of these integration and adaptation activities are a major barrier to the adoption and efficient exploitation of complex datasets. An important aspect of this integration process is the recognition of semantic differences between datasets. Often these differences are missed due to incomplete documentation, but more importantly mistakes occur because of misunderstanding due to assumptions made at the domain level. These mistakes may be costly: subtle differences in semantics may result in data being improperly integrated, which may not be noticed until after operational decisions are made.

We are investigating whether technologies currently applied to the development of the Semantic Web, particularly ontologies, may facilitate the capture of domain knowledge in such a way as to detect errors in data integration, or, due to the explicit nature of the semantics, prevent them occurring at all. Ultimately this technology could enable such integration and adaptation to occur "on the fly" – making the Semantic Web a reality. Given that this cannot be fully achieved in the near or medium term, our general approach is an incremental one. Manual processes will be systematically automated, eventually enabling some fully automated processes and services and others which are significantly automated, but still require some manual input. We are therefore initially placing an emphasis on ontologies being used as an aid to largely manual processes.

In order to increase the understanding and acceptance of the technology of Ontologies within Ordnance Survey, we have taken the notion of Semantic Reference Systems as proposed by Werner Kuhn [1] and broadened its definition. Whilst Kuhn describes such systems in terms of top level ontologies that provide grounding for other ontologies, we use it to also encompass what we term foundational domain ontologies. These are ontologies that are intended to establish de facto semantics for a particular topic area. In the case of Ordnance Survey, it would be to establish a Topographic Semantic Reference System. Kuhn rightly states that a Semantic Reference System is more than just an ontology: it must also support the transformations between domains. At this stage though, our research is limited to the development of the ontological component.

We see a Topographic Semantic Reference System as complementary to the existing Coordinate referencing system (The British National Grid) and the developing Feature Referencing System (OS MasterMap®) [2]. Its purpose will be to provide a common semantic definition of the principal topographic concepts applicable to this country, which will assist users of Ordnance Survey data to automatically conflate and adapt it with their own data. In order to build such a system however, we must first understand the necessary structure of the ontology and how it will be constructed.

3 Previous Approaches

The creation of an ontology is usually viewed as a knowledge acquisition task, which, as defined by Kidd [3], involves eliciting, analysing and interpreting human expert knowledge, and transferring this knowledge into a suitable machine representation. Many other ontology methodologies are based around a similar structure, or contain similar design criteria, but all differ slightly and not one has become a formal or even de-facto standard. Uschold's methodology and Fernàndez-López and Gómez-Pérez's METHONTOLOGY are believed to be the most representative [4]. Both methodologies propose initial modelling phases that develop an implicit shared understanding and explicit informal human-readable glossaries before structuring the information in a logical ontology. Uschold and King first define their classes precisely and unambiguously using natural language which are structured as a semi-formal hierarchies before building a logical ontology [5]. METHONTOLOGY further develops a more systematic method for domain conceptualisation. It provides a set of tasks for assisting the ontology modeller in capturing and structuring the information required for a logical ontology using a series of tables, a "Data Dictionary", and a series of concept trees [6]. Some of these representations however, are clearly specific to their domain of Chemistry and would not be suitable for a geographic ontology. In other existing methodologies, the processes of knowledge capture and formal coding have been carried out at the same time (for example, [7] and [8]). However, we support the approach of Uschold and King [6] and Gómez-Pérez et al.[9], who advocate the use of separate stages in ontology development.

The most popular methodologies [6] and [9] promote the creation of concept trees and sub-groups of similar classes. These promote an early dependence on the structures of formal languages and encourage the ontology modellers to group classes under familiar headings that in some cases do not represent the true logic underlying the relationship. This is particularly true for sub-sumption relationships, for example in a topographic ontology, concepts may be unnecessarily divided under "natural" and "man-made" branches in a hierarchy. We believe an ontology should also be much more than a taxonomy, and in fact, we discourage the use of hierarchies altogether, as they decrease the potential for inference and reuse by creating dependency between concepts. Under the umbrella of risk management, outside the world of academia, we have found that not all domains have a clear classification structure and cannot always be divided into small bounded modules. We have yet to look further into overcoming difficulties found with ontology modularisation and scalability and have identified this as an area of future research. More detailed reviews and discussions of ontology methodologies can be found in [5] and [10].

Knowledge representation is procedural and people find it difficult to describe exactly how they carry out these procedures or tasks. As the expert becomes more competent in their activity, the more automatic their use of knowledge becomes, and the less accessible it is to the knowledge engineer [11]. Past approaches in the AI community as part of the development of expert systems have tended to view knowledge elicitation as a preliminary to the more serious business of encoding knowledge in a software language. Rather than placing emphasis on the importance of knowledge elicitation from a domain expert, our strategy is instead to provide the domain expert with a set of clear and systematic steps that enable them to author a first-stage or "conceptual" ontology themselves.

4 Our approach to ontology construction

While our methodology may be broadly applicable to the construction of any type of ontology, we are focusing on the development of domain ontologies in particular. A domain ontology is a formalisation of the knowledge in a subject area (domain) such as topography, ecology, biology etc, and differs from other types of ontology such as the task ontology (a formalisation of the knowledge necessary to solve a specific problem or task abstracted above the level of a specific situation or organisational context).

Each ontology can be thought of as a pair of two linked ontologies: a conceptual ontology and a logical ontology. The conceptual ontology is intended to be primarily for human consumption: it attempts to balance the need for maximal formality of the ontology whilst retaining clear human comprehension. It is a means for domain experts to capture domain knowledge, which encourages them to record and describe their ideas explicitly in a standard structure. It should be free from the constraints of the logical ontology, and should not be influenced by the structures or rules that description logics present. The logical ontology provides a machine interpretable representation, typically using a derivative of first order logic such as description logic and is produced by an ontology expert familiar with languages such as the W3C standard language for representing ontologies; OWL (Web Ontology Language). It is generated from the conceptual ontology and, as we have found, information will be lost during this translation due to the inability of description logics to represent the true complexity of a conceptual ontology[1]. We have considered the possibility of including an intermediate stage between the conceptual ontology and the OWL ontology, where information is transformed into a more expressive logic such as First Order Logic to achieve a more complete representation. The advantage of the SHOIN(D) logic on which OWL is based is however in the tractability of its reasoning. .We believe a split

[1] Information loss also occurs during the creation of a conceptual ontology but this is less easily measured.

between these two ontologies is important, given the difficulty most people have in comprehending description logics and their inability to fully express the full richness of a domain. We emphasise that the conceptual ontology should be constructed and verified by the domain expert themselves, rather than the ontology engineer, and cite this as an advantage of our two-stage methodology.

Conceptualising a domain before processing it in a logical ontology can play a more significant role that simply collating information to be modelled. When separated from the formalisms of logical modelling, the structure can be used by domain experts themselves to record their knowledge and interpretations of their domain. In some instances, the domain expert may not have any existing complete documentation of their domain, in which case these stages of conceptualisation and knowledge capture are a useful mechanism for exposing domain information. While ontology experts' modelling techniques tend to pre-empt the knowledge structure imposed by description logics and ontology languages such as OWL, we assume that the domain experts are unfamiliar with ontologies and their rigorous structures. Instead of communicating the methodology using jargon familiar only to ontology engineers, we use common terms that can be easily understood by our target audience. For example, instead of using terms like "classes", "properties", and "attributes" we use the words "concepts", "relationships" and "characteristics". Our methodology is presented using a systematic structure, similar to the task-based structure used by Gómez-Pérez et al.[9], but is additionally supported by illustrations, examples, and written guidelines. A systematic task list promotes the use of a standard ontology structure and ensures the ontologies are produced consistently, which maximises the potential for interoperability between different ontologies.

5 Method for constructing a conceptual ontology

Our approach is to provide domain experts with a comprehensive and systematic set of criteria and guidelines to assist them through the entire conceptual ontology life-cycle. The methodology is still being developed, and we describe the basic skeleton of tasks for building a domain conceptual ontology only, supported by examples from the flood risk management ontology. The methodology comprises four main tasks: deciding on the requirements and content of the ontology; populating a knowledge glossary and constructing a set of triples (relationships between concepts); evaluating the ontologies; and finally, documentation of the conceptual ontology.

Stage 1- Preparatory

Task 1: Identifying the requirements

At the very onset of modelling the domain knowledge, the domain expert formulates a set of requirements for the ontology. This will provide the modeller (the domain expert) with a clear focus for ontology content and scope. It can be used throughout the ontology life-cycle as an evaluation tool. The criteria for identifying the requirements is similar to that identified by both Uschold and King [6] and Grüninger and Fox [7]. Primarily, the modeller records their definition of an ontology, their purpose for building it (which determines which type of ontology they produce), the scope of the intended ontology (based on the purpose), and a set of competency questions. We advise that the scope should be contained and restricted in size, so that ontologies produced are manageable and consistent. If the scope is large (e.g. the domain of topography) then the modeller may wish to sub-divide the domain into further domain ontologies (hydrology, urban areas, etc.), and integrate the modules together when they are all complete. The competency questions will differ depending on which type of ontology is being built. For domain ontologies, the competency questions are formulated so that they can be used to check at each stage of ontology construction whether the correct relationships have been created between the concepts, and whether the relationships created sufficiently describe the domain. To define competency questions, some pre-conceptions about which concepts are core to describing the domain are required. Generic examples include*, "Does the ontology sufficiently describe the domain to a level of granularity suitable for the purpose? Do all concepts have at least one link to another concept?".* Examples specific to a hydrology domain ontology within the topographic field would be: *"Have I sufficiently described the essence of being a"River" in terms of its relationships to its characteristics and links to other concepts? Have I made the distinctions clear in the relationships describing "River" and "Stream?"* .

Task 2: Collecting the data

Here, we acquire the input knowledge base needed to construct the conceptual model, based on the purpose, scope and competency questions. When appropriate, the modeller should reuse other ontologies that also suit the purpose of the ontology they are building. We are currently developing our research for reusing single concepts and sets of concepts and relationships from other conceptual ontologies, and the reuse of full conceptual ontologies.

The modeller should identify any documentation that captures the knowledge they wish to be in the ontology. The information must be suited to the purpose, be within scope, and be true to their representation of the domain in question. Where documentation is not available or sufficient, the ontology will be built using the domain expert's knowledge of the domain. Either manually or through using semi-automated data mining programmes, the modeller should extract the semi-structured sentences that contain information required to be in the ontology. These should contain important descriptor terms such as "and", "or", "sometimes", and "not"; terms that describe probability: "must", "likely", "might", "maybe", "sometimes"; and terms that describe possibility, including "usually" and "typically". It should then be verified that these sentences are complete within themselves, and complete in terms of recording all necessary information required. The aim is to reduce ambiguity by restructuring sentences, but ensure information is not lost. The sentences are then validated against the goals or purpose. It is well understood [12] that the linguistic and logical meanings of " and" and "or" are different. By recording these semi-structured sentences, our methodology provides the logical ontology modeller with a documentation trail so that he or she can check back to understand exactly which of the two possibilities the domain expert meant.

Stage 2: Populating a knowledge glossary

The first step in capturing and structuring the domain knowledge is to populate a knowledge glossary. Comparisons can be drawn with the “Data Dictionary” and the “Tables of attributes” proposed by Gomez-Perez et al. [9], but the glossary is more suitable for an audience less familiar with “classes” and “attributes”. We have used common natural language for the glossary headings and provide guidelines to assist the domain experts in identifying the correct information. Table 1 provides an example of two concepts from the flood risk domain ontology populated in a knowledge glossary.

Table 1. Knowledge Glossary

Term	**Synonym term**	**Natural language text definition**	**Linguistic term**	**Conceptual ontology term**	**Core / Sec.**	**Core concepts chars**	**Value and units**	**Rules, constraints and assumptions**
Flood risk map	Flood map	A map classifying risk into risk levels applicable to different areas.	Noun	Concept	Core	Has scale Shows risk level	Scale: 1:25000 to 1:100000	Scale is for regional maps
Is an input of		A relationship term to describe the link between two concepts, where one is used in the creation of the other.	Verb	Relationship	Core			Has inverse relationship (has input)

The information required for the glossary is extracted from the semi-structured sentences and enhanced by the domain expert. The modeller is encouraged to record the linguistic definition of a term (e.g. noun, verb) as an intermediate step to identifying which terms are concepts in the ontology and which are relationship terms or characteristics (attributes). The nouns are more likely to be concepts and verbs are most likely to be relationship terms. Defining the terms and recording these is a useful means for the domain expert to clarify their definition and interpretation of the term and its use within the ontology. The definitions will also be used in later stages of the methodology to identify relationships to other terms. The “core concepts” which are key to describing the domain are distinguished from the “secondary concepts” which either describe aspects of the core concepts or have differentiating relationships with them. This is useful for later stages of modelling. Secondary concepts are not members of the domain under consideration, but are necessary to enable concepts in the domain to be related to other domains. For example in the case of hydrology a core concept *“River”* could define a relationship to a secondary concept *“Field”* that would rightly belong to a different domain. Core concepts are vital to the ontology and are presumed to have the most relations to other concepts. They should be described within the ontology not only by their relations to other concepts, but also by their relation to their attributes (e.g. has size, has location), or as we term them in the conceptual ontology methodology, “characteristics”. The domain expert is encouraged to identify these using the semi-structured sentences and their own knowledge, and will use this information to explicitly describe the core concepts by their wholes and parts in the conceptual ontology. Characteristics of secondary concepts are not required in the conceptual ontology. The domain expert uses the glossary to record any assumptions, rules or restrictions governing the use of the definition, the characteristics or values within the ontology to reduce the assumptions made when creating the network of relationships between concepts and to avoid information loss at this early stage in development.

We appreciate that not all the knowledge required for the ontology will be captured from the semi-structured sentences and domain expert's knowledge, and that the glossary will undoubtedly be added to when the ontology is developed further. However, when the modeller is content with the information they have captured, the glossary should be validated against the purpose and scope set in the requirements stage. We are currently developing more efficient techniques than populating a table for composing the glossary and more formally testing the content of the glossary against the semi-structured sentences.

Stage 3: Creating a semantic network of triples

The next stage is to use the information captured in the knowledge glossary to construct a concept network that describes the domain in question. A concept network visualises an ontology as nodes (concepts) and links (relationships between concepts). This is much more than Gomez-Perez's "Concept Classification Trees" [9] which organise domain concepts in taxonomies. Our approach limits the use of hierarchical relationships that can encourage the creation false groupings of concepts or unnecessary divisions between groups of concepts (e.g. the division of "natural" and "man-made" concepts in a traditional topographic object classification), although these are not completely prohibited. Instead, we argue that richer inference can be achieved if the concepts are defined within themselves and through a range of relationships to other concepts (i.e. concept-to-concept relations and concept-characteristic-relations), so the shape and form of a semantic net is more comparable to a lattice than a hierarchy.

We have adapted Gruber's five design criteria to reflect our own interpretations [13]. These criteria should be used throughout the ontology life-cycle to enforce consistency and coherence. The modified criteria are:

1. Clarity: Definitions should be expressed unambiguously to ensure the intended meanings are comprehensible. They should represent the modellers interpretation of their domain.
2. Coherence: Relationships should be consistent with definitions.
3. Extendibility: It should be possible to add new terms without the revision of existing definitions accepting the addition of new relationships.
4. Minimal encoding bias: The choice of terms should not be made purely for convenience or implementation.
5. Minimal ontological commitment: Secondary concepts should be described using the weakest model only. These do not need to be described in terms of their characteristics. Gruber suggests that all terms should be defined using the weakest model,

thus making as few claims as possible. But although this maximises reusability, if ontologies are to be integrated through techniques such as semantic similarity, identification of matches between concepts will be essential. Core concepts should therefore be described additionally by their wholes and parts through relations to their characteristics although these should be both necessary and sufficient for the purpose and scope.

We specify a number of rules for creating a concept network to enforce consistency of the ontologies, including the following:
a. The modeller should work bottom-up, building the ontology with the most specific concepts which can then be generalised when necessary (identifying super-ordinates), to prevent groups of concepts being grouped under hierarchies or false semantics. Membership of a concept to another should be created instead by inference.
b. Multiple inheritance should only be created when the concept can inherit all of the characteristics of both super-ordinate concepts.
c. We advise only creating hierarchies when necessary for describing the domain, where the sub-ordinate inherits all the characteristics of its super-ordinate plus other characteristics, or when the ontology needs to move between different levels of granularity. The modeller should consider whether an alternative relationship can be used instead.
d. If new concept or relationship terms (i.e. those that are not already in the glossary) are needed when building the concept network they should be validated against the scope, goal or purpose, and added to the glossary before adding them to the conceptual ontology; this will ensure the term is used consistently with its definition.
e. If information can not be captured in the concept network, it should be recorded as semi-structured sentences or as an example for the logical ontology modeller who will attempt to include this information in the logical ontology.
f. If concepts or small groups of concepts are found to have no links into the rest of the concept network, the modeller should review their inclusion in the semantic net. If their inclusion is not suited to the scope or description of the domain they should be disregarded.

The domain expert should choose which method of representation both suits their ontology and their personal preference. To date we have used two methods of visualising the concept network: using network diagrams for graphically displaying links between concepts (Figure 1 illustrates an example from the flood risk management ontology), and creating a list of “conceptual ontology triples” where the concepts and relationships are recorded as subject-predicate-object. Both can be difficult to manage if the scope of the ontology is large, and the former does not facilitate the capture of “restrictions, assumptions and constraints”. Cyclicity and repeated triples are also difficult to manage in a list of triples. Similarly with the glossary, we are developing more sophisticated tools for capturing the triples using a user-friendly interface.

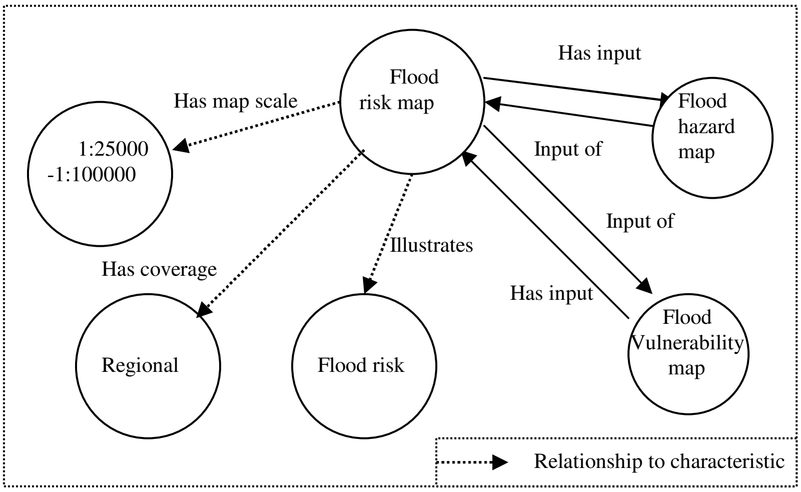

Figure 1. Concept network for concept "Flood Risk Map"

The domain expert should use the information captured in the knowledge glossary, plus their own knowledge to complete a concept network by completing the following tasks systematically:

Task 1: Create the links between the core concepts and their characteristics. Additional characteristics that were not captured in the glossary can be added if suitable to the purpose and scope. The modeller is likely to use the relationship term *"has"* to create the link between a concept and its characteristic. This should be specialised where possible to explicitly describe the link. For example, we would say *"Flood Event, Has Location, Location", instead of "Flood Event, Has, Location"*.

Task 2: Identify links between different core concepts using the most suitable relationship term that explicitly defines the type of link. For the topographic domain ontology, we found these to primarily be mereologic (part of), topographic (next to), and affordance relationships.

Task 3: Using the *"equivalent to"* relationship term, add in links between synonym concepts. These concepts must share a full set of characteristics.

Task 4: Create the links between core concepts and secondary concepts. These should be relationships that describe the core concepts in terms of their relation to other things that are not their characteristics.

Task 5: Create a relations network between the relationship terms. Similarly to the concept networks the modeller can produce the relation network using graphical network diagrams where the relationship terms are the nodes. The relation network should identify which relationship terms are sub-ordinates, which have an inverse relationship (e.g. "has part" and "part of"), which are transitive (e.g. "is input of") etc. If any new relationship terms are added to the model they should be added to the glossary first. The relations network is then used to identify which relationships are missing, incomplete, or inconsistent in the concept network. It is common for ontology modellers to record relationships uni-directionally so it is likely that all inverse relationships will have to be added to the concept network.

Stage 3: Evaluation of the conceptual ontology.

The modeller should firstly check whether all information captured in the glossary has been captured as triples or restrictions and constraints in the concept network, or has been recorded as information loss. Secondly they should check that the information captured in the concept and relations networks has been captured in the glossary. If there is information missing from the glossary further checks should be made against the scope and purpose. The modeller can now evaluate their conceptual ontology against the following criteria:

- Logical consistency: Checks are made for cyclicity, repetitions, and missing triples. The competency questions can be used to identify core concepts and triples that have not been captured.
- Conceptual accuracy: The domain expert should agree with the information that has been captured as triples, in that it represents his/her own interpretation of the domain, task or application.
- Minimal ontological commitment: Only those relationships suited to the purpose and within scope have been created, i.e. the core concepts are well defined by their explicit relationships to other concepts and relations to their characteristics. Secondary concepts have only been used in the ontology to describe the core concepts.
- Clear differentiation between ontologies: The concepts and relationships captured in should be suited to the ontology type created (i.e. a domain ontology does not contain concepts more suitable to a task ontology).
- Vagueness has been handled well: the modeller has attempted to capture probability, possibility, uncertainty and fuzziness within the conceptual ontology.
- Information loss is recorded.

Stage 4: Documentation of conceptual model

The conceptual ontology documentation must include the knowledge glossary, the concept and relationship networks, recorded information loss, and any defined rules and assumptions made throughout the modelling process.

6 Analysis

Our methodology for conceptual ontologies was exposed within the European Sixth Framework project Orchestra [14] where it was accepted in November 2004 as the standard for constructing the risk management domain ontologies. Feedback from the domain experts, and our own experiences in using the methodology, has enabled us to identify obstacles within the methodology that occur in real situations outside of the academic bubble, and has subsequently been used to further develop the methodology. We discuss the main obstacles found when building the five risk management ontologies (for flooding, earthquakes, coastal zone, forest fire and systemic risk) here.

6.1 Problems with scalability

The domain experts massively underestimated the amount of time required to produce an ontology and consequently built their ontologies based on a large scope (planning and preparation phases of risk management). The resulting conceptual ontologies were consequently a mix of both domain and task ontology concepts and relationships that jumped between levels of granularity and which were incomplete and inconsistent. This identifies three major problems in the methodology: firstly, it does not provide guidelines for limiting the ontology to a small scope in order to produce smaller, more manageable ontologies; secondly, the guidelines for separating concepts into those that are suitable for either domain or task ontologies are unclear; and thirdly, there are no guidelines for modularising the ontology so that it can either be produced by various people at the same time, or broken down into sub-domains for later partial reuse. The solution to the first problem is fairly trivial and can be solved immediately by encouraging the domain expert to define a small, contained and restricted scope at the outset of the ontology modeling phase to ensure that ontology construction is manageable and is more likely to be complete. The second and third however, require further thought. We believe the processes for constructing a domain and task ontology should differ, but we have yet to produce full task ontologies through which we can refine the existing method to distinguish between these different processes or develop a new methodology specifically for task ontologies. When reviewed, the Orchestra partners' ontologies were found to contain more task based concepts than domain ones. We have begun to develop more technical approaches to solving the third problem, for example using a tool suitable for conceptual modeling that is similar to the Protégé version control system the author of the information input can be tagged. To avoid missing concepts that lie between obvious boundaries, the competency questions could be used to check whether all the required concepts and relationships have been captured.

6.2 Recording triples

The domain ontologies produced in Orchestra included many concept-concept relationships, but included limited numbers of concept-characteristic relationships where core concepts are described by their wholes and parts. In most cases the level of explicit detail required by the conceptual ontology was not captured within the risk management conceptual ontology triples. The types of relationships recorded were generic and ambiguous; for example, "Rainfall causes Flood", from which the logical ontology would not then infer that it was specifically that it is *"heavy rainfall that causes a river to burst its banks which then causes a flood",* which is the true logical relationship. The tabular format for recording the triples was not the most effective means of encouraging the level of detail required from a conceptual ontology. It also proved difficult to identify loops of iterated relationships, repeated triples, or missing triples. This has prompted us to develop more efficient and effective means of capturing and structuring this information which include the use of text mining tools to extract concepts from documents, along with developing our own tools to facilitate the authoring of the domain ontology "triples". The intention is that the tool would take the domain expert through the steps of the conceptual ontology methodology up to the triples stage. The triples could be stored as either RDF or as simplified OWL concepts, whilst retaining the distance between the domain expert and the restrictions of OWL. This would of course not be full OWL as most of the knowledge would still be in natural language in annotation which would require further methods for transforming it into a complete logical representation. We are also developing a toolset of common ontologies that describe spatial relations, shapes (e.g. lines and polygons), time, and other relationship terms that can be reused to produce the Ordnance Survey full topographic ontology, or by others producing geographic ontologies.

6.3 Dealing with information loss

We encouraged domain experts to record any information that they could not model as triples either against the relevant triples in a column labelled "restrictions" in the triples table, or as semi-structured sentences. We evaluated the information loss to identify common areas across domains where information could not be captured as triples.

The primary cause of information loss was in the recording of fuzzy or uncertain relationships. It is common to find that domain experts do not have an explicit model of the conditions under which a relationship is true. This is part of the well-known knowledge elicitation problem and therefore it is difficult for domain experts to record information at the level of detail required. Our solutions to common issues are:

1. Quantified uncertainty and probability (e.g. one flood in 100 years). In these circumstances we record the probability as a concept within the ontology.
2. Where an instance has characteristics of more than one class (e.g. a section of a floodplain containing a number of different vegetation types). In the conceptual ontology we record *"Floodplain, has cover, Grassland and Shrubs and…"*, which would be added to the logical ontology as *"Floodplain, has cover, a number of: grassland, shrubs…"*.
3. Where there is a lack of information (e.g. a flood is less likely to occur when the river banks are high). The solution to this is to use a scale of categories that can be assigned meanings (e.g. high – low; less likely – probable – more likely).

Another common area of information loss occurred in domains which attempt to model comparisons that are numerical and based on inexact relationships. For example within the earthquake risk domain, many of the concepts in "risk assessment" require comparisons to be made between the hazard (the demand) and the vulnerability/resistance of the elements at risk (the capacity). This type of relationship cannot be modelled in the triples format. Similarly, the occurrence of induced events depends on inexact relationships between the causative and consequent event. In addition, information loss occurred when domain experts attempted to model triples that have conditions (e.g. an “if…then…else” statement) and tasks and processes. These issues suggest a conceptual ontology should comprise more than a glossary and a set of triples.

6.4 Evaluating ontologies

Although we have identified the domain ontologies produced within Orchestra as being incomplete and inconsistent, our set of criteria was insufficient for a robust evaluation, as we have no means of formally testing the logical consistency of the conceptual ontologies using the competency questions. We intend to incorporated this feature into the tools we are developing for recording the triples more effectively. We have since identified that the evaluation criteria will also vary depending on who is using the ontology. The ontology producer would want their conceptual ontology to be logically consistent, agree with purpose and scope, have well defined concepts, and contain reused concepts and relationships only originating from authoritative sources; and in these cases a logical ontology modeller is often required to second the evaluation to ensure logical consistency, until there are more formal means of testing this. Someone who intends to reuse an ontology, in addition to looking for the producer’s requirements, would want to reuse an ontology produced using the de facto standard, in a format compatible with theirs, and would perform checks to ascertain whether the ontology has reused ontologies from credible sources or from companies with similar interests to their own, hence, evaluation would be suited to check for this criteria.

The domain experts reported that the methodology was very systematic. This assisted them in consistently recording the required information in a structure that was common across the five risk management domains, which enhanced the potential for interoperability. Although not all were complete and consistent (primarily caused by the problems with scalability) the risk management conceptual ontologies reflected the domain experts’ true interpretation of their own domains. The information was captured without being constrained by the description logic representation of ontology languages such as OWL, a common limitation of promoting codification in early stages of ontology development. Our approach clearly demonstrated the benefits of separating conceptualisation of the domain, which is captured in a conceptual ontology, to the stages of formalising the domain in a logical ontology. The mere process of capturing their knowledge more formally has also enlightened the domain experts about details within their data. Previously undocumented relationships and assumptions have become explicit, and areas of similarity across the five risk management domains have been identified, which will facilitate future interoperability research.

7 Conclusions and recommendations for further work

The primary output of this research is the robust testing of our proposed methodology for assisting domain experts to construct ontologies themselves: an exercise which has not been reported in the literature before. Our approach successfully demonstrated the benefits of splitting ontology construction into two separate stages: conceptual ontology modelling and logical ontology modelling. As a consequence, the resulting domain ontologies for risk management and hydrology reflected the domain experts' interpretation of their own domain within a structure suitable for transformations into a logical ontology but without the common restrictions and compromises forced by description logic formalities. The ontologies were also found to be more expressive (that is, they were more than hierarchies or taxonomies) than many previous attempts by domain experts to develop ontologies described in the literature. Evaluation of the ontologies and feedback on the domain experts' experiences was useful for identifying future developments in the methodology. It firstly illustrated where further detailed explanation was needed and secondly it identified the areas for further research. These include the development of tools for assisting the domain expert in recording the conceptual triples, for example, to identify cyclicity and facilitate formal testing through the use of competency questions. Another area of further research concerns ontology modularity, in order to facilitate scalability and conceptual and logical ontology reuse.

References

1. Kuhn, W. Implementing Semantic Reference Systems, In, Proceedings of the 6Th Agile conference on Geographic Information Science, pages 63-72, 2003.

2. Ordnance Survey MasterMap ®
http://www.ordnancesurvey.co.uk/oswebsite/products/osmastermap/

3. Kidd, A. Knowledge Acquisition for Expert Systems, A Practical Handbook, Plenum Press. 1987.

4. Cristiani, M., Methodologies for the Semantic Web, In, AIS SIGSEMIS Bulletin, Vol.1, No.2, p. 102-136. 2004.

5. Uschold, M. and King, M., A Methodology for Building Ontologies, In IJCA195 Workshop on Basic Ontological Issues in Knowledge Sharing, Montreal, 1995.

6. Fernàndez-López, M., Gómez-Pérez, A., and Jurista, N., METHONTOLOGY: From ontological art towards ontological engineering, Stanford, 1997.

7. Grüninger, M., and Fox, M. Methodology for the Design and Evaluation of Ontologies In IJCA195 Workshop on Basic Ontological Issues in Knowledge Sharing, Montreal, 1995.

8. Noy, N. F., and McGuinness, D.L., Ontology Development 101: A guide to creating your first ontology, 2001.

9. Gómez-Pérez, A., Fernàndez-López M., and De Vicente, M., Towards a method to conceptualize domain ontologies. In Working notes of the workshop on Ontological Engineering, ECAI'96, pages 41-52. ECCAI, 1996.

10. Hemsley-Flint, F. Towards interoperability between ecological and topographical data through the application of ontologies, First year report, Unpublished, 2005.

11. Bainbridge, L. Asking Questions and Accessing Knowledge, In Future Computing Systems, Vol. 1, pp 143-150, 1986.

12. Gruber T. Towards principles for the design of ontologies used for knowledge sharing. International Journal of Human-Computer Studies. 43, p. 907-928. 1995.

13. Rector A, Drummond, N, Horridge M, Rogers, J, Knublauch,,H., Stevens, R, Wang, H and Wroe, C. OWL Pizzas: Common errors & common patterns from practical experience of teaching OWL-DL Proceedings of the Engineering and Knowledge Management Workshop 2004

14. Orchestra Open Architecture and Spatial Data Infrastructure for Risk Management IST FP6-511678 http://www.eu-orchestra.org/ 2004.

Acknowledgements

This work has been partially funded by the European IST FP6 project Orchestra (IST FP6-511678).

This article has been prepared for information purposes only. It is not designed to constitute definitive advice on the topics covered and any reliance placed on the contents of this article is at the sole risk of the reader.