Towards Modelling the Intended Purpose of Ontologies: A Case Study in Geography

Ronald Denaux[1], Anthony G. Cohn[1], Vania Dimitrova[1], and Glen Hart[2]

[1] School of Computing, University of Leeds, Woodhouse Lane, Leeds, LS2 9JT, UK
r.denaux@leeds.ac.uk
[2] Ordnance Survey Research, Romsey Rd, Southampton, SO16 4GU, UK

Abstract. As the Semantic Web technologies gain popularity, more organisations are providing semantic resources related to the geographical domain but are doing this based on ontologies that represent their own conceptualisations and needs. As a result: finding, understanding and evaluating relevant semantic resources is still very time consuming, cognitively demanding and often require a deep understanding of the domain and the used semantic technologies. To help address this problem, we are investigating how ontology purposes can be captured and used to facilitate ontology engineering. The paper presents an exploratory study which examines whether it is possible to identify generic categories to model the purpose of geography-related ontologies. As an output, we present an initial categorisation of ontology purposes and discuss how these categories can help us understand and reuse ontologies.

1 Introduction

As the Semantic Web technologies become more mainstream, more people and organisations are creating and publishing conceptualisations (using OWL) and descriptions of their data (using RDF). In the spirit of the World Wide Web, the technologies allow for each user to publish his own data and conceptualisations, while making it easy to interlink these semantic resources with those of other people. However, a big difference between the WWW and the SW is that in the WWW it is relatively easy to understand websites made by other people if you share the same natural language. In the SW the tasks of finding, understanding and evaluating relevant semantic resources are still very challenging. These tasks are currently time consuming, cognitively demanding and often require a deep understanding of the represented domain and the semantic technologies used. Much research is finding ways to alleviate these problems by providing semantic search [1, 2] and making it easier to understand ontologies by, for example, identifying the main concepts [3], providing vocabularies to annotate them [4] and studying their usage [5]. Our aim is to contribute to these research efforts by studying the stated purpose of ontologies.

This paper describes our first investigations in the area of ontology purposes, based on the geography domain. This domain presents a good use case for our investigation as the problem of finding and correctly re-using suitable semantic resources has already been reported [6]. These problems are partly caused by a a common practice where different organisations are producing their own geographic resources based on different conceptualisations and needs in reponse to (i) a *need for interlinking* geographic data with many different domains (e.g. agriculture, environment); (ii) the availability of large amounts of geographic data and (iii) a relative *ease of data production* as most people have at least some basic geographic knowledge[3].

The motivation for this work is to provide a high-level understanding of semantic resources by using the stated purpose. This is in line with other benefits of purpose descriptions pointed out by prominent ontology construction methodologies[7, 8], which regard the definition of the ontology purpose as a fundamental steps in the ontology development. The purpose helps to set the scope of the ontology and to evaluate whether it is fit for purpose. Furthermore, agreeing on a common ontology purpose improves collaborative ontology development by allowing ontology developers to refer to the stated ontology purpose to resolve modeling issues [9]. Thus, having a way to describe and formally represent ontology purposes could enable tool support for ontology construction.

Despite the possible advantages, we are not aware of a suitable way to formally define the purpose of ontologies. Where a purpose is defined it is done so as free text and rarely distributed with the ontology. Competency questions (CQs) [8] —questions that the ontology is required to answer— have been proposed to capture ontology requirements and tool support to formulate [10] them is available. However, CQs (i) can only refer to entities that are defined by the ontology itself, while purpose descriptions often need to refer to entities outside of the scope of the ontology in order to put the ontology into a wider context; (ii) are not commonly distributed with the ontology; (iii) can only be formalised as queries at a low granularity level, so the overall purpose of the ontology may not be clear based on the total set of CQs (especially for large ontologies) and (iv) cannot be organised in a hierarchy of purposes.

This paper presents an exploratory study which examines whether it is possible to identify generic categories to model the intended purpose of a corpus of ontologies that use geographical concepts. To find an answer to this question, we obtained a corpus of semantic resources as described in Section 2. In Section 3, we describe how the corpus was analysed in order to obtain a categorisation of the ontology purposes. We discuss some findings of our study in Section 4 and finish with a description of our future work in Section 5.

2 Corpus Collection

To obtain a corpus of ontologies, we considered a common ontology engineering scenario where a domain expert with limited knowledge engineering experience is looking for a concept to reuse in a new ontology. In particular, we decided to

[3] This is demonstrated by, for example, the OpenStreetMap project. `http://www.openstreetmap.org/`

use the concept `River` as this is a common concept which will ensure that we find a large number of ontologies that we can add to our corpus. Furthermore, `River` is a central concept in the Ordnance Survey Hydrology Ontology[4], which we can use as a baseline ontology.

2.1 Finding Semantic Resources to Create a Corpus

In order to find semantic resources based on our seed concept `River`, we followed recent research that advocates using semantic web search engines to find ontologies published on the web instead of minting new concepts [11]. As a result to our query for semantic resources based on `River`, Sindice [2] reports over 204×10^3 results, Watson [11] reports over 1500, while Swoogle [1] finds 302 results[5]. We also use used Google to find semantic resources by restricting our search to filetype `.owl` (572 results).

The large number of results for our query required further filtering until we obtained ontologies defining the concept `River`. Sindice found the biggest number of semantic resources, but most of the results were RDF documents describing instances from large datasets such as DBPedia and Geonames. For example, `http://dbpedia.org/resource/Parramatta_River` and `http://dbpedia.org/resource/Cry_Me_a_River`. Finding `River` definitions (not instances) with sindice is difficult because there is no filtering interface at `sindice.com` to navigate through the search results.

The results from Watson contained ontologies in a variety of formats (e.g. RDF(S), OWL, DAML+OIL). Watson does not allow filtering or sorting of search results either, so we browsed through the ontologies using the ontology URIs and metadata such as the size of the ontology in Kilobytes, number of classes, properties and individuals, type of ontology (e.g. RDF, OWL) and OWL-DL expressivity (e.g. $\mathcal{SHOIN}$). This information aided in discovering duplicates (the same ontology is hosted on different servers) or results which are not useful such as `foaf` user profiles, describing locations(e.g. `Fall River`) and interests (e.g. `River Dancing`).

Swoogle's results are similar to Watson's with the added advantage that Swoogle provides options for ranking the results based on ontoRank (their own ranking system), date or triple.

Google's results included several ontologies that contained a comment with string `River` but did not define a `River` concept. Also, some files with extension `owl` were not OWL files. Google results only provided text snippets and the ontology URI, but do not include OWL-specific metadata.

2.2 Finding Ontologies with a Purpose Description

Our study required that the ontologies in the corpus should have a purpose description. Initially, we searched for ontologies that define a *purpose annotation*.

[4] `http://www.ordnancesurvey.co.uk/oswebsite/ontology/Hydrology/v2.0/Hydrology.owl`

[5] data gathered in June 2009

However, the number of ontologies that contained such an annotation was very limited, so we decided to also collect generic descriptions about the ontology. In the absence of a standard way of describing ontologies, we used three additional sources of purpose descriptions; which we describe below.

First, we *searched the website hosting the ontology for an ontology purpose description*. In many cases, the website provided a description of the ontologies. This description sometimes included an explaination on why the ontology had been created which we used as a purpose description (e.g. the GWSG ontology).

We also *searched the serving website for a project purpose description*. In some cases, when the website served multiple ontologies, it did not provide a description for each separate ontology (e.g. NASA SWEET). In these cases, we looked for a more generic description of the website as a whole or of the project that produced the ontology. If this generic description included a purpose description—and it was clear that the ontology was created as a means to achieve that purpose—, we included the description for the ontology.

Finally, we *searched for publications that explain a project in more detail*. We conducted this step when the website did not contain a suitable purpose description, but was part of a project for which publications could be found. We then searched the publications for descriptions of the ontology and how it was used in the project. For example, the purpose description of the German Geo KB, was gathered from (i) a description of the ontology in [12]; (ii) a description on why the ontology was used [12] and (iii) a description of the project in [13].

2.3 Corpus Description

The corpus collected consists of 9 OWL ontologies with a free text description of their intended purposes and the provenance of the purpose description (see Table 1). The ontologies in the corpus range from small knowledge bases such as German Geo KB to relatively large ontologies like NASA SWEET.

The purpose descriptions in the corpus came from a variety of places, as described above. In the best of cases, the authors provided a succinct description of the ontology purpose (e.g. Ordnance Survey Hydrology Ontology in Table 2. In the worst cases, the purpose of the ontology has to be inferred from documents describing the project that created the ontology (e.g. [14] describes the BOEMIE project objectives and includes sentences that hint at the purpose of the GIO ontology (see Table 2).

3 Purpose Analysis, Conceptualisation and Formalisation

The presence of ontology (and project) descriptions that did not describe the ontology purpose explicitly posed a problem: modeling these implicit purpose descriptions was a subjective task that required an interpretation of the documentation based on assumptions. In order to minimise this effect and to keep track of

Table 1. Overview of the ontologies in the corpus including ontology metrics (all gathered using Protégé 4.0), type of the source of the ontology purpose description and the organisation that created the ontology.

Ontology Name	Ver.	Purpose Source	Organisation	DL	Metrics							
					Cs	OPs	DP	Inds	SubC	EqAx	DisAx	Anns
Ordnance Survey Hydrology	2	Ont. Annotation	Ordnance Survey	$\mathcal{ALCHOIQ}$	194	43	0	16	329	27	2	479
NASA Sweet HydroBody-OfWater	2	Tech. Doc. & Website	NASA	$\mathcal{SHOIN}(\mathcal{D})$	1834	130	25	139	2119	208	19	381
GIO ontology	2.5	Tech. Doc.	BOEMIE project	$\mathcal{ALCHN}(\mathcal{D})$	291	52	29	0	364	0	12	245
Germany Geo KB	n/a	Research Publication	University of Karlsruhe	$\mathcal{ALQ}(\mathcal{D})$	12	17	18	634	59	0	0	0
EnvOc	1.5	Website	envoc.org	$\mathcal{ALE}+$	1233	4	0	0	1348	0	0	5933
GWSG	1	Ont. Annotation & Website	MUSIL Univesitt Mnster	$\mathcal{SHIN}$	350	215	0	2	960	0	21	462
E-response	n/a	Ont. Annotation	e-response.org	$\mathcal{SHOIN}(\mathcal{D})$	1746	182	19	323	2147	73	1266	941
WOW	n/a	Website & Tech. Doc.	walkonweb.org	$\mathcal{ALCN}(\mathcal{D})$	55	50	50	127	46	0	0	143
hydroseek navigation	n/a	Tech. Doc.	hydroseek.org	$\mathcal{AL}$	16	0	0	0	15	0	0	15

Cs = Class Count	OPs = Object Property Count	DP = Data Properties Count
Inds = Individual Count	SubC = Subclass Axiom Count	EqAx = Equivalence Axiom Count
DisAx = Disjoint Axiom Count	Anns = Entity Annotation Count	Ver. = Ontology Version

Table 2. Examples of free text purpose descriptions, their source and corresponding purpose phrase

Description	Source	Purpose phrases	Code
Purpose: To describe in an unambiguous manner the inland hydrology feature classes surveyed by Ordnance Survey with the intention of improving the use of the surveyed data by our customers and enabling semi-automatic processing of these data.	Ontology Annotation of Ordnance Survey Hydrology Ontology	*Describe the inland hydrology feature classes surveyed by Ordnance Survey in an unambiguous manner.*	`OS1`
		Enable semi-automatic processing of the data surveyed by Ordnance Survey.	`OS3`
Driven by domain-specific multimedia ontologies, BOEMIE information extraction systems will be able to identify high-level semantic features in image, video, audio and text and fuse these features for optimal extraction. [...] The rationale behind this approach is that multimedia information extraction can benefit in many ways from the background knowledge provided by ontologies [...]	BOEMIE project description [14]	*Provide domain-specific background knowledge*	`GIO1`
		To use in information extraction tasks in multimedia	`GIO2`
		To enable identification of high-level semantic feature in image, video, audio and text	`GIO3`

the assumptions made during the analysis of the ontology purpose descriptions, we used a Grounded theory-based approach [15] for analysing the corpus. Figure 1 shows an overview of the steps taken during the analysis to arrive at a conceptualisation.

As a first step towards conceptualistion, we identify and extract *purpose phrases* based on the ontology descriptions in the corpus. These are simple statements that describe a single goal that the ontology should help to achieve. To compose the purpose phrases, sections of free text descriptions have to be rephrased (we try to stay as close to the original text as possible to minimise the introducing assumptions). As an example, Table 2 shows some purpose phrases based on purpose descriptions. The table also shows that each purpose phrase is linked to a unique code that links the purpose phrase to its originating ontology and free text description. Due to space limitations we cannot show all the purpose phrases[6].

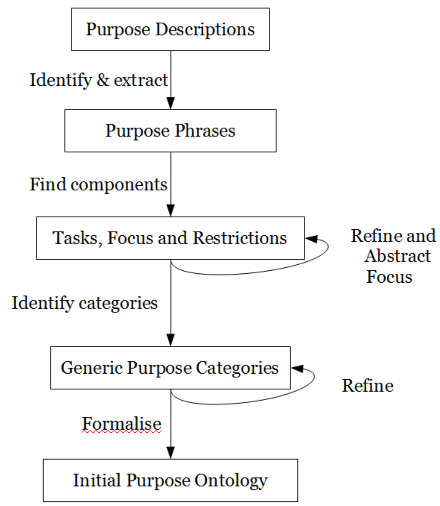

Fig. 1. Overview of steps to obtain an initial ontology of purposes for geographic ontologies based on free text purpose descriptions

The next step in the analysis is to make sure that the purpose phrases have a common structure that represents an ontology purpose. This structure has been inspired by existing work describing goals in the medical context [16]. Although there is no standard terminology for describing goals, [16] states that most goals can be decomposed into a tuple consisting of a *task* and a *focus*. This tuple can be extended with optional components such as *restrictions*, a *situation description*, a *policy definition* (e.g. whether goal is obligatory) and a *rationale* (i.e. clarification of the goal). For our purpose phrases (i) the situation description is always the same: the creation of an ontology; (ii) no policy definition is necessary and (iii) the rationale is the purpose description source. So we can decompose our purpose phrases (e.g. `OS3` in Table 2) into a task (e.g. *Enable*), a focus (e.g. *semi-automatic processing of data*) and zero or more restrictions(e.g. *the data has been surveyed by Ordnance Survey*).

The *task* is the verb phrase at the beginning of each purpose phrase. For example, the task for `GIO1` is *Provide* and the task for `OS3` is *enable* (see Table 2). Defining a single best focus and a set of restrictions is not always easy. For

[6] At the moment of writing we have identified 63 purpose phrases. A full table can be found at `http://spreadsheets.google.com/pub?key=tkeez029qTGOVZ1Rd41JbTw&gid=2`

purpose phrase OS3, the focus can be *semi-automatic processing of data* (plus some restrictions), or it can be *processing of data* (where the fact that this processing is semi-automatic is a restriction)[7].

We found generic concepts to use in our categorisation by making the focus of each purpose phrase as abstract as possible and expressing as many restrictions as possible. The restrictions of the purpose phrases define specialisations of the generic concepts in the categorisation. After refining these generic categories we arrived at 7 types of purposes:

1. **Domain Defining**: these purposes occur when there is a need to conceptualise a domain. They usually restrict the scope of the ontology and may also impose restrictions on how concepts are represented. An example of this is gwsg7:*Find an ontological consistent description of the qualities of a data set of observations.*
2. **Ontology process related**: these purposes state that the ontology should influence (enable, improve, etc.) ontology processes such as the use, re-use, merging and alignment of ontologies. An example: seres3:*Enable comparison between conceptualisations.*
3. **Data process related**: the ontology influences a data process: for example the creation, editing, navigation, annotation and publishing of data. These type of purposes also can impose restrictions on the scope of the domain, but differ from *Domain Defining* and *Ontology process related* purposes in that the focus of the purpose lies on the data that can be represented by the ontology instead of the conceptualisation itself. An example is EnvO3:*support the annotation of the environment of any organism or biological sample.*
4. **Investigative**: when the ontology is created to study a system (e.g. GermanGeo1 *Use as a knowledge base to study the ORAKEL system*), the Ontology Engineering process itself (e.g. gwsg5:*Investigate how to convert the German classification schema for ecological assessment of watercourse structure into a DOLCE-aligned domain ontology*) or specific ways to formulate and encode knowledge (e.g. gwsg4:*Investigate the possible combinations of basic qualities into complex qualities*).
5. **Collaboration enhancing**: when the ontology should enhance social processes. As social processes may include ontology and data related processes, this category can overlap those two categories. For example: Pont9:*Enabling collaborative discussion of ontologies by practitioners.*
6. **External application**: are purposes where an ontology is used by an external application to perform a task. E.g. GIO2:*Optimise information extraction tasks in multi-media.*
7. **Analogous**: these are purposes that refer to a similar ontology in a different domain that is being emulated. E.g. EnvO1:*Provide similar benefits as the Gene Ontology.*

[7] All the purpose phrase decompositions can be found at http://spreadsheets.google.com/pub?key=tkeez029qTGOVZ1Rd41JbTw&gid=3

Formalisation A lightweight ontology has been created based on the shown conceptualisation. This ontology is currently only suitable for annotating purpose descriptions. In the future we would like to produce an ontology that enables authors to easily formalise their purpose for an ontology and enables services such as purpose comparison and classification.

3.1 Application

Currently, we are planning two extensions to this study where we investigate whether we can apply the model of ontology purposes to compare the ontologies in the corpus. In the first study extension we will study whether we can use the purpose categories to determine the similarity between two ontology purposes. This similarity can then be used to compare the ontologies. For example, our corpus analysis shows that the NASA SWEET ontology states its purpose in terms of ontology engineering, community benefits and domain description while ignoring data processes such as data creation, publishing or annotations. The Hydroseek navigation ontology has an opposite approach where the purpose focuses on data search, mapping and navigation, while not much is said about ontology engineering.

The second study extension will investigate whether we can find a correlation between ontology metrics and ontology purpose. The intuition is that the stated purpose of the ontology should have an impact on the design decisions made when building the ontology, so there might be a relation between ontology characteristics and ontology purpose. For example, an ontology that has as a *Data process related purpose* (e.g. German Geo KB), may be more likely to provide instances of the modelled concepts than an ontology that does not specify this type of purpose. Another example is the Hydroseek Navigation ontology, that aims to make navigation tasks through data easier. This purpose may be reflected in its small number of concepts and shallow taxonomic depth.

4 Discussion

The categorisation identified in our study suggests that it is possible to identify generic categories to model the purpose of geographical ontologies. In fact, the presented categorisation does not rely on geography specific terminology, which suggest that the model could be applicable to ontologies outside of the geographic domain. Indeed, the corpus already contains ontologies that are not strictly geographical such as E-response(emergencies).

The dependence on the geographic domain is only apparent when we look at the set of restrictions, where we find geographic domain restriction (e.g restriction for `wow5` to *geographic objects valuable in tourism*) and data restrictions that are strongly related to the geographical domain (e.g `wow4`:*Allow automatic creation of multilingual hiking path descriptions*).

An unexpected result of this study is that geography-specific external processes were not found: ontologies were built to aid *information extraction in* *multi-media* and *to study semantic reference systems*, but none of the geographic ontologies mentions specific usages in geography related tasks. The OS Hydrology ontology has a generic data-process (*use of the surveyed data by Ordnance Survey's customers*). NASA SWEET does not mention how the data described by the ontology will be used and only refers to the conceptualisation in terms of ontology engineering tasks such as evolution and alignment.

A clarification for the small number of specific geography-related processes may be the relatively small corpus of ontologies and the difficulty of finding good purpose descriptions of the ontologies. Another clarification may follow the argument by C.M. Keet [17] who says that many ontology authors do not define a particular purpose for an ontology because their ultimate aim is to build an ontology that is application independent—even when the ontology is initially built for a particular purpose. This may be the case in the geographic domain, when ontologies encode well-established classification schemas that already are used by many applications (e.g. OS Hydrology ontology), so that ennumerating the specific use cases is not desireable.

While the presented study suggests that it is possible to find generic categories of ontology purposes, the approach that we used has several limitations: (i) the introduction of subjective interpretation of purpose descriptions; (ii) only a subset of the geographical ontologies was used as we restricted ourselves to ontologies that defined a single concept (i.e. we missed ontologies that define geographical features, but do not define the `River` concept); (iii) finding appropriate purpose descriptions for ontologies was time consuming; so that, even when a suitable ontology is found, no purpose description can be found by searching the (semantic) web (e.g. the Mooney ontology of geographic data[8]).

5 Future Work

To address the limitations of the current study, we hope to be able to contact some of the authors of these ontologies to (i) verify the ontology purpose, (ii) get feedback on their personal goals when constructing the ontology, (iii) validate and improve the vocabulary to describe ontology purposes and (iv) investigate whether providing the vocabulary helps to make ontology purposes explicit, which were previously implicit and not published.

Further future work include (i) investigating whether there are links between stated ontology purposes and features of the ontology such as number of classes and ontology expressivity; (ii) designing and investigating the effect of tool support based on ontology purposes; (iii) researching ways to represent the ontology purpose and the contributor's goals to be able to model conflicts of interests to support multi-perspective ontology development; (iv) investigate the link between ontology purpose and usage; specifically, whether we can design a shared vocabulary to describe both purpose and usage of an ontology and whether we can use this vocabulary to evaluate the fitness-for-purpose of an ontology.

[8] `http://www.ifi.uzh.ch/ddis/fileadmin/ont/nli/geography.owl`

References

1. Ding, L., Finin, T., Joshi, A., Pan, R., Cost, R., Peng, Y., Reddivari, P., Doshi, V., Sachs, J.: Swoogle: A search and metadata engine for the semantic web. In: Proceedings of the 13th International Conference on Information and Knowledge Management, ACM (2004) 652–659
2. Oren, E., Delbru, R., Catasta, M., Cyganiak, R., Stenzhorn, H., Tummarello, G.: Sindice. com: A document-oriented lookup index for open linked data. International Journal of Metadata, Semantics and Ontologies **3**(1) (2008) 37–52
3. Peroni, S., Motta, E., d'Aquin, M.: Identifying key concepts in an ontology, through the integration of cognitive principles with statistical and topological measures. In: Proceedings of the 3rd Asian Semantic Web Conference. (2008) 242–256
4. Hartmann, J., Sure, Y., Haase, P., Palma, R., Suarez-Figueroa, M.: Omv–ontology metadata vocabulary. In: Workshop on Ontology Patterns for the Semantic Web at ISWC2005. (2005)
5. Alberts, R., Calvanese, D., Giacomo, G.D., et al.: Analysis of test results on usage scenarios. Deliverable D27, FP6-7603 Thinking ONtologiES (TONES) (2008)
6. Goodwin, J., Dolbear, C., Hart, G.: Geographical linked data: The administrative geography of great britain on the semantic web. Transactions in GIS **12**(s1) (2008) 19–30
7. Fernandez-Lopez, M., Gomez-Perez, A., Juristo, N.: Methontology: from ontological art towards ontological engineering. In: Proceedings of the AAAI97 Spring Symposium Series on Ontological Engineering. (1997) 33–40
8. Uschold, M., King, M.: Towards a methodology for building ontologies. In: IJCAI95 Workshop on Basic Ontological Issues in Knowledge Sharing, Montreal (1995)
9. Dellschaft, K., Engelbrecht, H., Barreto, J., Rutenbeck, S., Staab, S.: Cicero: Tracking design rationale in collaborative ontology engineering. In: Proceedings of the 5th European Semantic Web Conference, Springer (2008) 782–787
10. Sure, Y., Erdmann, M., Angele, J., Staab, S., Studer, R., Wenke, D.: Ontoedit: Collaborative ontology development for the semantic web. In: Proceedings of the International Semantic Web Conference, Springer (2002) 221–235
11. d'Aquin, M., Motta, E., Dzbor, M., Gridinoc, L., Heath, T., Sabou, M.: Collaborative semantic authoring. IEEE Intelligent Systems **23**(3) (2008) 80–83
12. Cimiano, P., Haase, P., Heizmann, J., Mantel, M., Studer, R.: Towards portable natural language interfaces to knowledge bases – the case of the orakel system. Data & Knowledge Engineering **65**(2) (2008) 325–354
13. Cimiano, P.: The orakel natural language interface (2009) `http://www.aifb.uni-karlsruhe.de/WBS/pci/home/doku.php?id=orakel` accessed on 29/07/2009.
14. Spyropoulos, C., Paliouras, G., Tsakonas, A.: Project description and website. Deliverable D1.1, Bootstrapping Ontology Evolution with Multimedia Information Extraction (2006)
15. Strauss, A., Corbin, J., Lynch, M.: Basics of qualitative research: Grounded theory procedures and techniques. Sage Newbury Park, CA (1991)
16. Fox, J., Alabassi, A., Patkar, V., Rose, T., Black, E.: An ontological approach to modelling tasks and goals. Computers in Biology and Medicine **36**(7-8) (2006) 837–856
17. Keet, C.M.: Ontology design parameters for aligning agri-informatics with the semantic web. In: Proceedings of 3rd International Conference on Metadata and Semantic, Milan, Italy (2009) to be published.