KNOWLEDGE MEDIA
KMi
INSTITUTE

Cognition, ontologies and description logics

Technical Report kmi-14-03
March 2014

Paul Warren

paul.warren@cantab.net

The Open University

Abstract

This report describes work undertaken and planned, the goal of which is to better understand how people comprehend and use complex ontologies, in particular those employing Description Logics (DLs). Two research questions are posed:

1. *In what way can the difficulties experienced in using Description Logics be understood in terms of an underlying theory, e.g. theories of reasoning already developed within the cognitive psychology community?*
2. *In what way could such a theory contribute to improving the usability of Description Logics?*

An initial survey of 118 ontology users has confirmed the importance of DLs, specifically the variants of OWL. The survey also looked at two particular techniques for mitigating complexity: visualization and the use of patterns. Opinions on the value of visualization were varied, with over a third of respondents finding visualization 'not at all useful' or 'useful to a small extent'. Whilst many of the respondents used patterns, the majority of the users did not import patterns from a library but merely used them as examples. Responses to a question about the purposes for using ontologies suggest that a categorisation of users into a few groups is possible. Such a categorisation could help in understanding user behaviour and requirements.

A laboratory investigation into the comprehension of commonly used DL statements revealed some common difficulties. Many users experienced difficulties with the negated conjunction; there was a common misconception concerning the inheritance of property characteristics by subproperties; and to a lesser extent there were difficulties with the functional characteristic and the existential quantifier. The mental model theory of reasoning was used to explain the difficulties experienced with negated conjunction, whilst the rule-based approach helped explain how complexity influenced accuracy and response time.

A programme of further work is proposed. The most immediate step will be to build on the existing study into DL comprehension with more controlled experiments on the effect of varying the complexity of commonly occurring DL statement structures. Subsequent work will look at the effect of alternative linguistic structures and diagrammatic representations. It is also intended to conduct interviews and focus groups with ontology users, both to gain a deeper insight into their problems and to obtain feedback to the ideas developed regarding alternative linguistic and diagrammatic representations.

Keywords: ontologies; description logics; psychological theories of reasoning; empirical studies

Contents

1 Introduction and research question

The goal of the work described in this report is to understand better how people comprehend and interact with complex ontologies. More specifically, much of the focus of the work is on Description Logics (DLs) because they have become the major paradigm for describing ontologies. Evidence of this is provided in Chapter 3.

The end goal is to improve the efficiency with which ontology users interact with DL ontologies. This might be through proposing alternative constructs to those which give particular difficulty, alternative linguistic or graphical representations, improved tools, or by indicating those aspects which require particular emphasis in training.

The intention of the work is to build on work done to understand reasoning within the field of cognitive psychology. This work has focused on 'naïve reasoners', i.e. people not trained in logic. Activities studied have included: disjunctive reasoning, e.g. Johnson-Laird et al. (2012); reasoning about ordered arrays, e.g. Newstead et al. (2006); syllogistic reasoning, e.g. Zielinski et al. (2010); and negation, e.g. Khemlani et al. (2012a) and (2012b).

Two research questions are posed:

1. *In what way can the difficulties experienced in using Description Logics be understood in terms of an underlying theory, e.g. theories of reasoning already developed within the cognitive psychology community?*
2. *In what way could such a theory contribute to improving the usability of Description Logics?*

The report is structured as follows. Chapter 2 describes some related work. Chapter 3 then gives an overview of a survey into ontology use which provides information about difficulties experienced with ontologies generally and DLs in particular, and also identifies what features of DLs are commonly used. Chapter 4 then describes an experiment to gain more insight into the difficulties users experience in comprehending DL statements and relates these difficulties to psychological theories. Finally, Chapter 5 describes proposed future work and includes a workplan for the next three years.

2 Related work

The first two subsections below discuss work which is directly related to the work described in this report, specifically in section 4. This falls into two categories: work undertaken by computer scientists to understand the comprehensibility of DLs, discussed in section 2.1; and the work of cognitive psychologists to understand the nature of reasoning in general, discussed in section 2.2. Section 2.3 then discusses visualization techniques because of their relevance to potential future work discussed in section 5. Section 2.4 briefly discusses measures of ontology complexity. This is of potential importance because the comprehension of any axioms in an ontology is likely to be influenced by the axioms' context and in particular by the complexity of the overall ontology or the particular subset of the ontology under consideration. Finally, section 2.5 discusses decidability and computational complexity in DLs and poses the question whether there is any relationship between the factors which influence decidability and computational tractability and which influence comprehensibility.

With the exception of section 2.1, where there are few relevant other studies, this section does not claim to be a comprehensive literature review. Instead it highlights a few papers which are representative of the key ideas relevant to the work reported on and planned for the future.

2.1 Comprehensibility of Description Logics

There have been few studies of the comprehensibility of Description Logics. Rector et al. (2004) describe the difficulties experienced by newcomers to OWL, based on their experience in teaching the language. They provide a set of guidelines and also English paraphrases of some OWL expressions. Horridge et al. (2011) were interested in supporting the ontology developer during the debugging process. One way to offer such support is to display the minimal subset of the ontology that generates a particular entailment. Such a subset is termed a justification. Horridge et al. investigated the cognitive complexity of justifications for entailments of OWL ontologies. They developed a complexity model and compared the predictions of this model with the difficulty experienced by computer scientists in identifying correct entailments. Their model, which was not grounded in any psychological theory, "fared reasonably well". Commenting on the study by Newstead et al. (2006), Horridge et al. identified a strong advantage of studies within the psychological literature, i.e. that the problems considered "are very constrained and comparatively easy to analyse". The difficulty in studying DLs is the need to consider a wide range of commonly occurring constructs.

Nguyen et al. (2012) had a similar interest in assisting developers to debug ontologies. Their goal was to explain, in English, why an entailment follows from an ontology. In particular, they wished to predict the comprehensibility of alternative proof trees, when expressed in English. To do this they needed to first understand the comprehensibility of the individual deduction rules comprising a proof tree. They took 51 such deduction rules, expressed in English, and tested their comprehensibility on participants obtained through a crowdsourcing service. This enabled them to generate a facility index representing the ease of comprehensibility of each deduction rule, calculated as the proportion of participants who identified the deduction rule as being correct. Since their interest was in calculating these

facility indices for future use, they did not attempt to create a model to predict and explain the indices.

2.2 Theories of human reasoning

2.2.1 Rule-based

There is a considerable psychological literature on human reasoning. One theory, the rule-based approach, is represented by the work of Rips, e.g. Rips (1983). The assumption is that the processes of human reasoning are akin to the steps executed by a logician in carrying out a proof. If this is the case, then according to Rips, accuracy might be expected to depend on the 'availability' of the rules which need to be employed. 'Availability' here equates to how obvious or natural the rule seems to a naïve reasoned. Accuracy might also be expected to depend on the number of steps in the reasoning chain.

2.2.2 Mental models

The alternative theory, the model-based approach, is represented by Johnson-Laird, e.g. Johnson-Laird and Byrne (1991), for whom mental models "have the same structure as human conceptions of the situations they represent". Johnson-Laird and his collaborators have built an extensive body of experimental evidence to support the view that at least 'naïve reasoners' (i.e. people not trained in logic), do use mental models in reasoning. It may be, though, that some individuals use a mixture of mental models and rules-based reasoning, depending upon the particular situation and their degree of training in logic. One possibility is that when logicians are constructing a proof in a rule-based way, they use mental models at each deduction step.

The mental model theorists propose that difficulties in reasoning often occur when several models need to be maintained in working memory. This may happen, for example, when a disjunction occurs. Moreover, mental model theory suggests that an inclusive disjunction will give rise to more errors than an exclusive disjunction, since the former requires three models to be held in working memory whilst the latter requires only two. Table 2.1 shows the mental models required for conjunction, exclusive disjunction, and inclusive disjunction. The relation between difficulty and number of mental models is borne out by experiment, e.g. see Johnson-Laird et al. (1992).

Table 2.1 mental models for different propositional statements

(a) conjunction: *A and B*	**(b) exclusive disjunction: *A or B but not both***	**(c) inclusive disjunction: *A or B or both***
A B	**A not B** **not A B**	**A B** **A not B** **not A B**

Mental model theory can be used to identify specific sources of error. For example, in the case of exclusive disjunctions, e.g. 'A or B but not both', then Johnson-Laird et al. (2012)

suggest that, when the intellectual demands of a task are not so great, reasoners can make use of the full models shown in table 2.1(b). However, in more demanding cases their mental models represent only those situations that are possible given an assertion. Thus, 'A or B but not both' may be represented by the two models written thus:

A
B

This leads to systematic fallacies, and the authors report experiments in which they have observed just such fallacies.

The mental model theory can also be used to explain difficulties with negation. This has been investigated by Khemlani et al. (2012a). They investigated people's comprehension of conjunction and inclusive disjunction both in affirmations and negations. In affirmation conjunction is represented by one model and inclusive disjunction by three, as shown in table 2.1(a) and 2.1(c). In negation, the situation is reversed. *not (A and B)* is represented by three models:

not A **not B**
A **not B**
not A **B**

Whilst *not (A or B)* is represented by one:

not A **not B**

The percentage accuracy of answering questions found by Khemlani et al. (2012a) is shown in Table 2.2, along with the number of mental models in each case.

Table 2.2 Interaction between affirmation / negation and conjunction / inclusive disjunction, from Khemlani et al. (2012a)

	conjunction		inclusive disjunction	
	%age accuracy	no. mental mods	%age accuracy	no. mental mods
affirmation	86%	1	68%	3
negation	18%	3	89%	1

In the case of affirmation, we see that performance with inclusive disjunction is worse than with conjunction, as predicted by the number of mental models. In the case of negation we see a very much poorer performance in the three model case. The authors suggest that people "are likely to fail to construct the full sequence of models" and that "the order of constructing the models is unlikely to be constant, but it should usually begin with the negations of both clauses". Thus, in the case of the negation of conjunction, most people will formulate *not A and B*; some people may not get beyond this. Khemlani et al. (2012b) elaborate on their theory of negation. They propose that people "tend to formulate and to interpret negations as having small scopes" because this leads to fewer mental models. It may be that, in the case of *not (A and B)* some people restrict the scope of the negation to the individual items within the parentheses, and hence arrive at *not A and not B*. The difficulty of creating the three models in the case of negated conjunction contrasts with the case of affirmative inclusive disjunction where the three models are explicit in the words used: "A or B or both".

It also seems likely that the experimental success of the mental model theory, e.g. as described in Johnson-Laird and Byrne (1991), and Johnson-Laird et al. (1992), arises because

the individual models were of broadly equal complexity. As a consequence, situations requiring two models created more difficulty than those requiring one, and situations requiring three models were even more difficult. A situation necessitating one model, but where that one model is extremely complicated, e.g. because of its relational complexity or depth (see section 2.2.3 immediately below), could be more difficult than a situation requiring several simpler models. This would particularly be the case in situations where the models are not difficult to formulate.

Although not specific to mental model theory, it is worth noting here the possibility of error which arises when the disjunction is ambiguous, i.e. where the word 'or' is used without any further qualification. Johnson-Laird et al. (1992) cite references which suggest that the majority of people interpret this as an inclusive disjunction but "a sizeable minority prefer the exclusive interpretation". This could be a source of misunderstanding in DLs and other formal logical notations, where 'or' always means inclusive disjunction.

2.2.3 Relational complexity

Relational complexity (RC) theory offers another approach to understanding performance in reasoning. Here complexity is defined "as a function … of the number of variables that can be related in a single cognitive representation", i.e. the number of arguments of a relation (Halford & Andrews, 2004). The theory proposes that it is the maximum relational complexity in a given process of reasoning which determines the difficulty of that reasoning. RC theory could be seen as compatible with either the rule-based or the model-based approach. Zielinski et al. (2010) have attempted to reconcile the mental model and RC theories for categorical syllogisms. Goodwin and Johnson-Laird (2005) have combined mental model theory and RC theory in their study of reasoning about relations. Apart from the number of arguments, they see depth of the relation as contributing to complexity. Relations between individuals are regarded as of first-order depth; relations between relations of second-order depth; relations between relations between relations of third-order depth.

2.3 Visualization techniques

The value of diagrams in reasoning has been noted by various researchers. Larkin and Simon (1987) have analysed the benefits of diagrams in terms of support for search, recognition and inference. The importance of the design of diagrams has been pointed out by Bauer and Johnson-Laird (1993) who noted the need to avoid arbitrary symbols and make explicit alternative states of affairs. In the context of DLs, diagrams offer a strategy to overcome misconceptions and generally support reasoning. In fact, a large number of tools have been created to display ontological structures, e.g. see Katifori et al. (2007). These are chiefly aimed at viewing the structure of the overall ontology or parts of the ontology, i.e. at the subsumption relations, rather than the more cognitively difficult features of Description Logics. An exception to this is the work of Dau and Eklund (2008) and work by Howse and his collaborators, e.g. Howse et al. (2011). The former created a diagrammatic reasoning system for a particular DL. Howse et al. (2011) use concept diagrams not only to view subsumption relations but also to view and reason about role restrictions. Both papers discuss a calculus for such diagrams. Other work by Howse and his collaborators focus on the formal aspects of such a calculus, e.g. Burton et al. (2012), and procedures for creating diagrams, e.g. Rodgers et al. (2014). There does not appear to have been any empirical studies on the efficacy of such diagrams.

2.4 Ontology complexity

Ontology complexity has been the subject of a few studies, largely motivated by the desire to predict the difficulty of ontology development, reuse and modification. In this it resembles the study of complexity in software, from which it borrows certain concepts. To the author's knowledge, no attempt has been made to relate these metrics to any difficulties of comprehension of an ontology, although the difficulty of working with an ontology must in part be related to difficulties in its comprehensibility.

Yang et al. (2006) analysed the change in complexity of the Gene Ontology between December 2002 and June 2005. They used a set of metrics chiefly derived from the hierarchical structure. In particular, they identified the paths between each concept and top-level concepts. From this they calculated metrics such as the maximum and average path lengths in an ontology. Although they do not say so explicitly, this seems to be an attempt to measure the skewness of the distribution of path lengths. A more robust metric might have been the ratio of the upper quartile point to the median.

Zhang et al. (2010) provide an overview of the literature on the topic and describe its relationship to software complexity. Drawing on the literature, they propose a set of four ontology-level metrics:

- size of vocabulary, i.e. the sum of the number of classes, individuals and properties
- the ratio of edges to nodes in the ontology graph, where the edges represent subsumption, disjoint classes, and properties
- tree impurity, a measure of how far the ontology's inheritance hierarchy deviates from being a tree
- entropy, i.e. the entropy associated with the graph structure

Zhang et al. also propose a set of four class-level metrics, describing how a particular class contributes to the complexity of an ontology.

Whilst originally designed for full ontologies, all the measures discussed above could also be used for patterns, where comprehension of the whole is perhaps more relevant than for an ontology. It is an open question to what extent any of these measures can contribute to predicting the comprehensibility of patterns.

2.5 The complexity of Description Logics

The decidability and computational complexity of DLs has been extensively studied, e.g. see Baader et al. (2010). It is not clear if this form of complexity bears any relation to comprehensibility. Indeed, computational complexity describes what happens at a scale beyond human processing, and may be irrelevant to what happens at the scale of human reasoning. Problems of decidability and computational tractability share with problems of comprehensibility the characteristic that they can be created by the interaction of logical features and this may indicate an area for future investigation.

3 A survey of ontology use

This chapter provides an overview of the ontology user survey. In particular it describes results relevant to the work to be described in Chapter 4 and the planned work described in Chapter 5. A more comprehensive description of the survey results can be found in Warren (2013).

3.1 Background

The survey was organised as a web survey using *SurveyExpression*[1] and was emailed to a number of contacts in the research area and relevant mailing lists. The latter included:

- ontolog-forum (ontolog-forum@ontolog.cim3.net),
- UK Ontology Network (ontology-uk@googlegroups.com),
- two LinkedIn groups: Semantic Web for Life Sciences; Description Logic,
- lists maintained by the Open Knowledge Foundation: okfn-en@lists.okfn.org; ok-scotland@lists.okfn.org; okfn-nl@lists.okfn.org, and
- the internal mailing list within the Knowledge Media Institute at the Open University.

In all, there were 118 respondents. Whilst respondents did not in general answer all questions, there was a good response to individual questions. For example, 69 respondents identified the ontologies they chiefly used; 57 provided information about the size of the ontologies used; and 65 identified the ontology editors they use. Respondents represented a range of sectors: academic (45%); research institutes (25%); industrial (17%); besides 13% who categorised themselves as 'other'. They gave their primary application areas as: biomedical (31%); business (9%); engineering (19%); physical sciences (7%); social sciences (5%); and other (30%). Since no attempt was made to achieve a representative sampling, these breakdowns do not necessarily represent the distribution of ontology users overall. Respondents represented a range of experience with ontologies. There were a significant number of experienced users: 50% had more than 5 years' experience whilst 5% had less than one year's experience.

3.2 Reasons for using ontologies

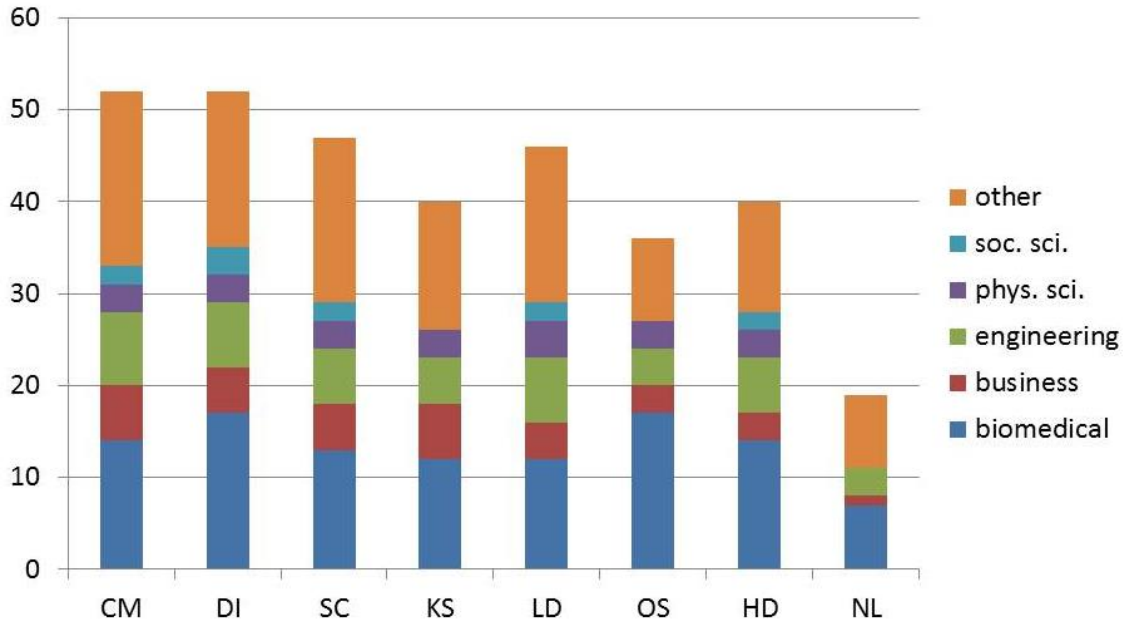
Respondents were asked for which purposes they used ontologies. There were eight options, plus 'other'. The full text used in the survey, along with a code for reference, is shown in table 3.1. Including 'other', there were 341 responses from 73 respondents, i.e. an average of 4.7 responses per person. Excluding 'other' there were 332 responses from 72 respondents, i.e. an average of 4.6 responses per respondent. A number of the 'other' responses could be seen as particular cases of one of the options.

Figure 3.1 shows the responses, excluding 'other', broken down by application area of respondent. There appears to be no obvious relationship between application area and purpose of use. The maximal predictive classification technique was used to cluster the data. This technique is suitable for binary data and seeks to maximise the total number of agreements between data points in a cluster and a predictor for that cluster.

[1] http://www.surveyexpression.com

Table 3.1 Purposes for using ontologies

Code	Text in survey
CM	Conceptual modelling – e.g. formally defining a domain
DI	Data integration – i.e. merging a number of databases
SC	Defining knowledgebase schemas – e.g. as a means of storing and retrieving information
KS	Knowledge sharing – e.g. between individuals in an organisation
LD	Linked data integration – e.g. linking data from different public knowledgebases
OS	Ontology-based search – i.e. using ontologies to refine search
HD	Providing common access to heterogeneous data – i.e. providing a common schema for data access
NL	Supporting natural language processing



72 respondents; multiple responses permitted; 'other' responses ignored

Figure 3.1 Purposes for using ontologies, by application area

The technique was applied as far as four clusters, giving:

- A cluster of 33 respondents who all indicated a large number of separate purposes. The average number of purposes ticked was 6.2. With the exception of NL (33%) the lowest score on any category was 73%. These users might be termed *multipurpose users*.
- A cluster of 16 users who indicated very few purposes; the average was 2.2 per cluster member. A majority (75%) indicated conceptual modelling. The highest score on any other category was 38% (for NL). This cluster might be called the *conceptualizers*.
- A cluster of 11 users, who averaged 3.9 responses, and the majority (73%) of whom also indicated CM. There was also a majority for LD (91%), OS (73%) and SC (64%). This group might be called *searchers*.

- A cluster of 12 users, averaging 4.3 responses, for whom, unlike in the case of the three other groups, only a minority (25%) indicated CM. A majority did indicate DI (100%), HD (92%), LD (83%) and SC (83%). This group could be called *integrators*.

Note that a high proportion of the searchers and integrators indicated LD and SC. The differentiation between searchers and integrators is that a high proportion of the searchers indicated CM and OS, whereas the integrators were more likely to indicate DI and HD.

The above represents one particular way to categorize the respondents according to their use of ontologies; there are clearly other ways. The key point is that there are a number of different motivations in using ontologies and, in future work, an insight into the motivation of particular users may be helpful in understanding their difficulties.

3.3 Ontologies

The five most commonly used ontologies were (in decreasing order): Dublin Core, FOAF, Dbpedia, the Gene Ontology, and SKOS. The next most common response was 'own'. There were then a number of ontologies relating to biology, medicine and chemistry; and also some generic ontologies, e.g. the W3C provenance ontology.

The size of the ontologies being used varied considerably. 35% of the ontologies reported had no more than ten classes and 55% had no more than 30 classes, whilst two had more than a million classes; the latter were both in the biomedical area. A similar variation existed for the number of individuals and the number of properties, although in the latter case there was more concentration at the lower end with 69% having no more than 30 properties.

There was also a considerable variation in the shape of the ontologies. 59% had had no more than five top-level classes whilst 5% had over one 100 top-level classes. Most of the ontologies were relatively shallow: 40% had a depth of no more than two and 71% a depth of no more than five. However, 14% had a depth of greater than ten.

3.4 Ontology languages and editors

65 people responded to a question about which ontology languages they use. Of these respondents, 58 used OWL, 56 RDF and 45 RDFS. 13 respondents ticked the 'other' box; although a number of the responses in the 'other' category were not languages at all, e.g. some were ontologies; others were query languages. The 'other' responses did include two references to OBO, which, as one of the respondents noted, "should be considered an eccentric dialect of OWL". In any case, response to the question made clear the dominant position of OWL.

The dominant role of OWL was also apparent from a question about ontology editors. Respondents were asked which ontology editors they used. They were given a choice of 12 editors, and there was also an 'other' option. Multiple responses were permitted. 63 respondents replied to the question. Figure 3.2 shows all the tools for which there was more than one response. All the tools shown were amongst the predefined category except for OBO-Edit and Neurolex. The light blue editors are OWL editors, the dark blue are non-OWL. In fact, OBO-Edit uses a variant of OWL developed for biological applications, and the respondent citing Neurolex noted that it "gets translated into OWL from RDF". What is perhaps surprising is how few respondents used the frame version of Protégé.

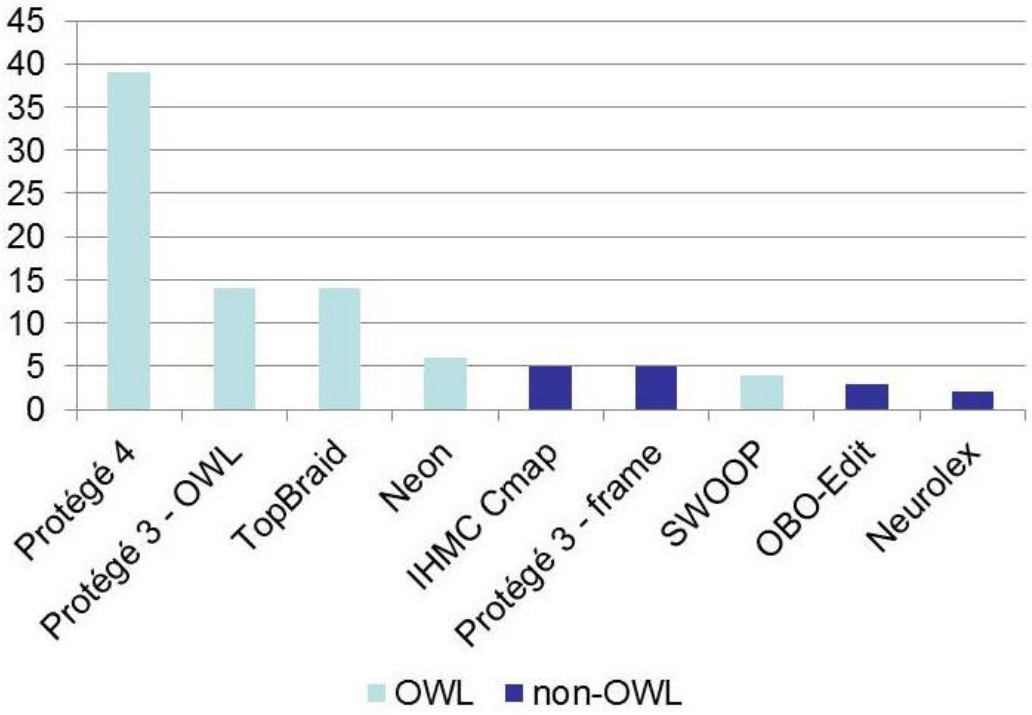


Figure 3.2 Usage of ontology editors

There were a number of general comments about ontology tools. Many of these related to desired features and to performance. A few were more fundamental, relating in particular to the difficulties experienced by people who are not expert ontologists. One respondent commented: "No tool I know is able to abstract from the technical details and allows non-experts to model useful and correct ontologies". Two other respondents noted the need for different tools for domain experts, e.g. for "content creation, addition and editing", than for more expert ontologists. Whilst ostensibly relating to tools, these comments are also pertinent to language design.

3.5 Description Logic features

Respondents were asked which of 23 DL features they used. The features were taken from those available in Protégé 4. There were 47 responses to this question and the results are shown in table 3.2.

Table 3.2 Usage of DL features
- showing the number using each feature, out of 47 respondents

DL feature	
object property domain	37
object property range	36
disjoint classes	35
datatype properties	34
intersection of classes	33
transitive object properties	32
object subproperties	32
union of classes	31
existential restrictions	31
inverse object properties	30
functional object properties	30
universal restrictions	28

DL feature	
hasValue restrictions	28
cardinality restrictions	24
symmetric object property	24
functional datatype property	24
datatype subproperties	23
complement of a class	22
qualified cardinality restriction	20
inverse functional object prop	17
reflexive object property	14
asymmetric object property	12
irreflexive object property	8

The features at the top of this list are as one might expect. Perhaps more surprising are the number of respondents using the more specialist features, e.g. the four object property characteristics at the bottom of the list (inverse functional, reflexive, asymmetric and irreflexive). It might be thought that only a small number of respondents used the less common features. However, this was not borne out by inspection of the data. For example, there were 40 respondents out of the 47 who used at least one of the 8 least common features and 45 respondents who used at least one of the 12 less common features. Two caveats need to be applied. Firstly, the respondents were not asked about the frequency with which they used these features. It seems likely that those who used the more sophisticated features would use them less frequently than they use the more common ones. Secondly, 47 was a relatively low number of responses. This should be compared, for example, with the 115 respondents who provided information about the length of time they had been using ontologies. It seems likely that these 47 represented the more sophisticated users and that respondents who did not answer this question might in general only use a subset of the features. Nevertheless, the response to this question indicates the usage of a wide range of features.

Particular points to note are that the most common object property characteristics are transitive, functional and symmetric. Amongst the restrictions, the existential was used by a slightly larger group of respondents than the universal. A slightly smaller group used the unqualified cardinality restriction, and a smaller group still the qualified cardinality restriction. Finally, the complement operator was used by only 23 respondents, i.e. slightly less than one-half of the total respondents to this question.

3.6 Visualization and visualization tools

Respondents were asked how useful they found visualization. There were 56 responses to this question, with the following breakdown:

- essential 18%
- very useful 16%
- quite useful 29%
- useful to a small extent 32%
- not at all useful 5%

There was no apparent relationship between the answer to this question and the characteristics of the respondent, e.g. the application area.

Respondents were also asked about their use of visualization tools. Figure 3.3 shows all those tools for which there was more than one response. These consist partly of tools specified in the question and tools noted amongst the 'other' responses. The figure indicates which tools are available in one of the versions of Protégé. OWLViz is available in both Protégé 3 and 4; OntoViz and Jambalaya in Protégé 3, and OntoGraf in Protégé 4. IsaViz is a standalone visualizer which can import and export RDF/XML and hence can be used with Protégé. The figure reflects the dominance of Protégé shown in figure 3.2.

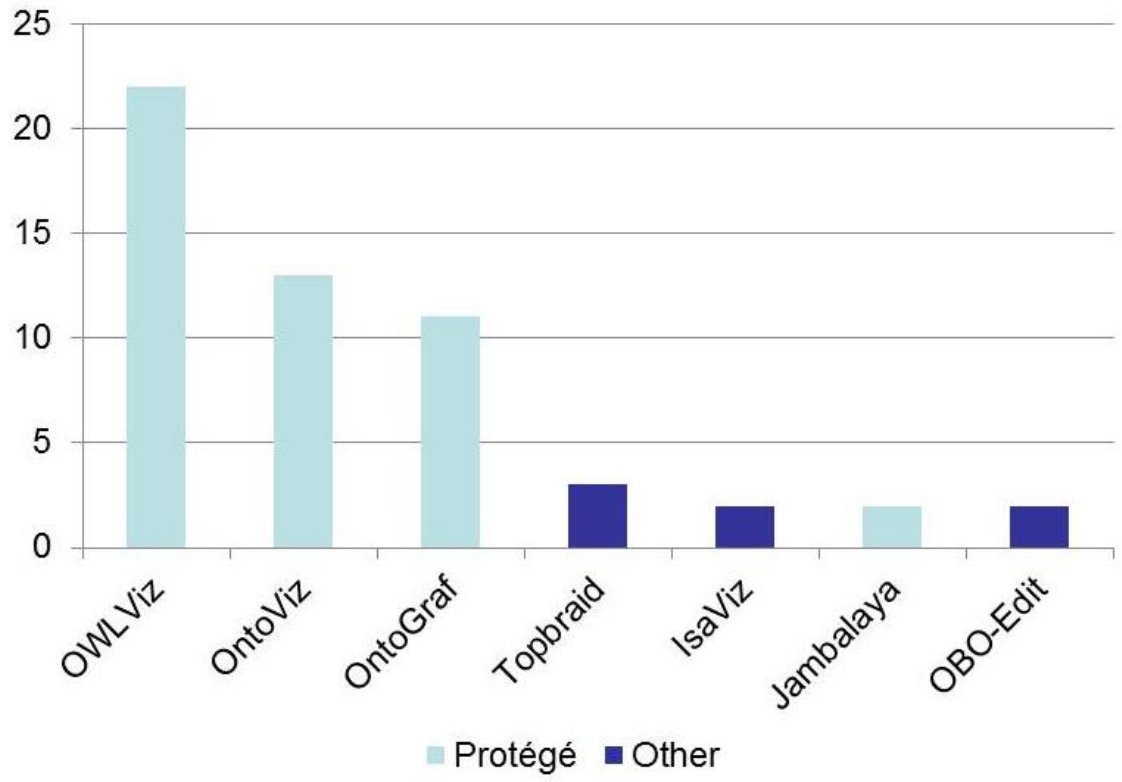


Figure 3.3 Usage of ontology visualization tools

3.7 Ontology patterns

The questionnaire contained a section specifically for those who use patterns. They were firstly asked from where they obtained their patterns. Five possible sources were listed, plus 'other'. There were 35 respondents, with multiple responses permitted. The wording of the options is shown in table 3.3. Responses in the 'other' category included two references to the Open Biomedical and Biological Ontologies[2], one to 'domain specialists', and one to a very small more generic pattern library[3]. Figure 3.4 shows the results, broken down by application area. The figure suggests that biomedical specialists are more likely to use their own or colleagues collections, or their own mental models, than libraries or the Protégé wizard. Biomedical specialists' patterns appear mostly specific, whereas in the other application areas there appears to be a tendency to use the general and the specific. Supporting this view, a biomedical specialist commented that there are "seldom some available patterns out there for us to use".

A Pearson χ^2 test was undertaken by dividing:

- the sources into two groups consisting of the two catalogues and the Protégé wizard on the one hand; and own or colleagues' collections, own mental models and 'other' on the other hand (i.e. the leftmost three categories versus the rightmost three categories in figure 3.4);
- the application areas into biomedical versus the remainder.

This confirmed that these two factors were not independent ($p = 0.039$), i.e. the biomedical specialists were significantly less likely than the other specialists to use the generic catalogues or the Protégé wizard.

Table 3.3

Code	Text in survey
MAN	The ODP public catalog (i.e. http://www.gong.manchester.ac.uk/odp/html/index.html)
DES	OntologyDesignPatterns.org (i.e. http://ontologydesignpatterns.org)
WIZ	Protege patterns wizards (under 'Tools' in Protégé 3)
COL	Pattern collections created by yourself or colleagues
OMM	Your own frequently used mental models (i.e. not written down)

[2] http://www.obofoundry.org/
[3] http://www.essepuntato.it/2012/04/tvc/

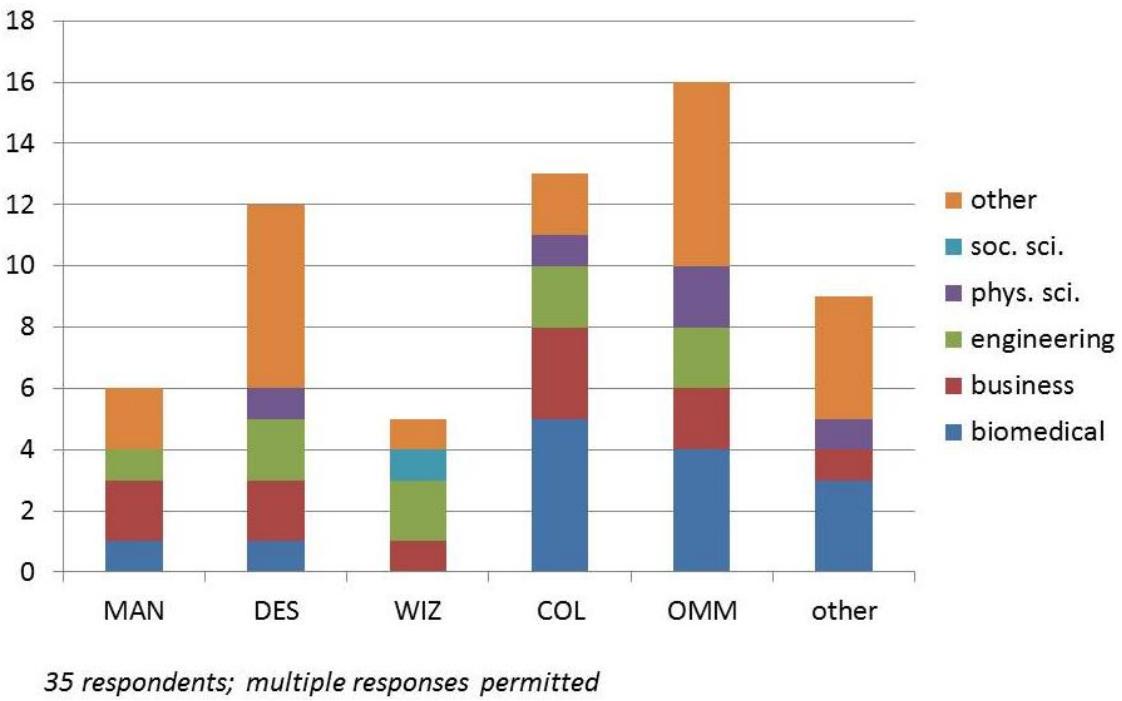


Figure 3.4 Sources of patterns, by application area

Respondents were asked how they used patterns, i.e. specifically whether they imported patterns or whether they used them as examples and recreated them. 32 people responded to this question, with multiple responses permitted. Only 28% of the respondents imported patterns. The remainder solely used patterns as examples.

One respondent noted the difficulty of understanding patterns: "initially hard to learn, but provide required functionalities". This supports the general thesis of this report that the representation of ontology languages is in need of improvement.

3.8 Respondents' comments

There were a range of comments from respondents. Some comments, like the two relating to patterns already quoted, were specific to particular sections of the survey. Others were more general and provided after completing the whole survey. A number related to specific issues such as functionality of ontology editors. However, others were directly relevant to the thesis of this report. In the section on ontology languages, one respondent wrote "… the complexity of it all is way beyond what we can hope to hold in our minds at any given time, but I have yet to use a tool that makes this complexity easily understood, or even easily workable with". In the same section, another respondent wrote "the rigor of the languages exceeds the rigor of the typical user by a wide margin". At the end of the survey there were two comments relating to the difficulty of designing ontologies. One respondent noted the difficulty of defining classes, which had taken "many years of learning". Another respondent noted that domain scientists are "not too comfortable with ontology relationships".

4 An investigation into the comprehension of Description Logics statements

This section describes a study into the usability of DLs, specifically to understand the comprehensibility of the most commonly used features of DLs as implemented in OWL. The study has revealed a number of misconceptions and the report makes suggestions as to how these can be overcome. The study also attempts to relate comprehensibility of DL features to the theories of reasoning discussed in section 2.2. This theoretical approach has helped to explain the most significant of these misconceptions and also explained the time to confirm inferences.

Section 4.1 describes how the most commonly used DL features were identified. The study focused on the features that are commonly used, rather than those features which, whilst useful in particular domains, are not extensively used. Section 4.2 explains how the study was designed and conducted. Section 4.3 presents the questions used and discusses the five which were found most difficult by participants. Some potential remedies for these problems are suggested. Section 4.4 discusses participants' feedback, which confirms the usefulness of some of these suggestions. Section 4.5 then provides a more detailed analysis, including some results relating the accuracy and response time to psychological theories of reasoning. Finally, section 4.6 summarises the key findings of the study

4.1 Identifying the commonly used features

Four sources were employed to identify the most commonly used features. Power and Third (2010) provide a list of the most commonly used OWL functors based on an analysis of ontologies in the TONES Ontology Repository[4]. Power (2010) also used TONES to identify common axiom patterns. This identified the frequency of use of Boolean operators such as intersection, and also of the existential and universal restrictions. Khan and Blomqvist (2010) searched 682 online ontologies to determine the frequency of occurrence of content patterns from the ODP (ontology design pattern) portal[5]; content patterns are essentially small autonomous solutions to particular design problems. I then analyzed the 20 most frequent patterns that they detected, to determine the commonly occurring OWL features. In addition, Warren (2013) undertook a survey of ontology users which included a question about the usage of OWL features, as explained in section 3.5 above. The resultant ranking of features is also shown in table 3.1 above.

The resultant lists were then compared. They identified broadly the same set of commonly occurring features. There were a few differences, e.g. Power and Third found class equivalence to be the second most commonly used functor; analysis of the common content patterns from Khan and Blomqvist identified class equivalence as the thirteenth most commonly used OWL feature, whilst it was not included in the survey by Warren. The set of features used in this study consisted of all those features which were relatively common in at least one of these lists, with two exceptions. The reason for these exceptions was that the study participants would not necessarily be familiar with OWL and they would need to be given information about the language which should be kept brief. Firstly, all features relating to datatype properties were ignored. It was felt that datatype properties would present no cognitive challenges that could not be represented with object properties. This is not to say

[4] http://rpc295.cs.man.ac.uk:8080/repository/
[5] http://ontologydesignpatterns.org/wiki/Main_Page

that there might not be challenges arising from datatype properties during the learning process, but rather that subsequently, when working with ontologies, they do not give rise to any specific problems of cognition. Secondly, cardinality restrictions were not included. In fact, these did not occur in the list of most commonly used patterns in Power and were ranked relatively low in the survey by Warren. The 'min 1' cardinality restriction did occur moderately frequently in the patterns identified by Khan and Blomqvist. This states that a particular individual is the subject of at least one instance of a particular property. It is equivalent to an existential restriction, which was included in the study.

Table 4.1 shows the set of OWL features chosen for the study. In each case the Manchester OWL Syntax (MOS) representation of the language feature is also shown (Horridge et al., 2006). The features are grouped into those relevant to classes, those relevant to properties and the existential and universal restrictions. In each of these groupings, they are listed broadly in order of occurrence (i.e. with most commonly occurring at the top), although, as already noted, rankings differed across the various sources.

Table 4.1 Commonly used OWL features investigated in the study

Class features

language feature	MOS
subsumption	SubClassOf
class equivalence	EquivalentTo
disjoint classes	DisjointWith
class assertion	Type
conjunction	and
disjunction	or
complement	not

Property features

language feature	MOS
property range	Range
property domain	Domain
property hierarchy	SubPropertyOf
inverse object properties	InverseOf
transitive object property	Characteristics: Transitive
functional object property	Characteristics: Functional
symmetric object property	Characteristics: Symmetric

Restrictions

language feature	MOS
qualified existential restriction	some
universal restriction	only

4.2 The study

In order to test comprehension of these OWL features, they were incorporated into a set of twenty-one questions based on three patterns from the the ODP portal (see previous section). The three patterns used were: Componency, Coparticipation, and Types of Entities; the second and third were modified to enable all the features in table 4.1 to be tested, with some simplification of the second to remove unnecessary statements. The three modified patterns were associated with ten, six and five questions. Each of the questions consisted of a set of statements and a proposed inference. The participant was required to indicate, by clicking on a button, whether the inference was or was not valid. In all there were thirteen questions with valid inferences and eight where the inference was not valid. The patterns and question statements were expressed in a simplified form of the Manchester OWL Syntax. Classes and properties were defined in the patterns and had intuitive names. Individuals were defined in the questions and were named A, B, C, D.

These patterns and questions were incorporated in a survey, using the tool SurveyExpression[6]. The three patterns were ordered in all six permutations; each of these permutations existed in a form with the 'yes' option first and with the 'no' option first, to safeguard against any bias from the order of the possible answers. Thus there were twelve variants of the survey. There were also twelve participants, i.e. one participant per variant.

[6] http://www.surveyexpression.com

Screen capture software was used to record the user's behaviour and in particular to provide the precise times spent in each question. The participants were observed as they took part in the study and any comments they made were noted.

The study was organized into five sections. In the first section participants were asked to rate their knowledge of logic and of OWL. Specifically, they were asked in each case whether they had: no knowledge at all, a little knowledge, some knowledge, or expert knowledge. In neither case did any respondents reply in the first of these categories.

The next three sections contained the three patterns and the questions. Each section began with a webpage displaying the pattern and then a series of pages, with each page repeating the pattern and containing one question. Participants were able to move through the pages at their own speed. The final section was an opportunity for the participants to provide feedback.

Participants were provided with a handout containing all the necessary information about the OWL features and notation used. They were asked to read this at the beginning and it was available to them for reference during the session.

4.3 Survey questions and the difficulties

Of the 21 questions, eight were answered correctly by all of the participants, four were answered correctly by all but one of the participants, and a further four were answered correctly by all but two of the participants. The remaining five are discussed here in decreasing order of difficulty. Tables 4.2, 4.3 and 4.4 show the three patterns and the associated questions. The columns headed 'yes/no', 'MM' and 'RC' represent the correct answer, the maximum number of mental models and the maximum relational complexity associated with the question. The column headed 'num steps' shows the number of steps to arrive at a correct deduction for questions with answer 'yes'. The remaining two columns show the percentage of correct responses and the average time for each question.

4.3.1 Complementing the *and* operation – table 4.2: Q2

(answer: no; 25% correct responses)

The most direct way of arriving at a 'no' conclusion for this question is to note that, since Event and Quality are disjoint classes, then Event and Quality must be Nothing (⊥). Hence not (Event and Quality) is Thing (⊤) and the statement A Type not (Event and Quality) is tautological.

Table 4.2 Questions based on the modified entity types pattern

Class Entity	EquivalentTo Event or Abstract or Quality or Object
Class Event	SubClassOf Entity DisjointWith Abstract, Quality, Object
Class Abstract	SubClassOf Entity DisjointWith Event, Quality, Object
Class Quality	SubClassOf Entity DisjointWith Event, Abstract, Object
Class Object	SubClassOf Entity DisjointWith Event, Abstract, Quality
Class Nonconceptual	EquivalentTo Event or Object
Class Nontemporal	EquivalentTo Abstract or Quality or Object
Property represents	Characteristic Functional

	Question	yes/ no	MM	RC	num steps	%age corr	av. time (secs)
1	A represents B; C represents D => A DifferentFrom C	no	1	2	n/a	83%	91.5
2	A Type Entity; A Type not (Event and Quality) => A Type (Abstract or Object)	no	4	3	n/a	25%	75.1
3	A represents B; C represent D; B Type Object; D Type Event => A DifferentFrom C	yes	1	4	2	50%	75.8
4	A Type Entity; A Type not (Event or Quality) => A Type (Abstract or Object)	yes	4	4	2	92%	44.0
5	A Type (Nonconceptual and Nontemporal) => A Type Object	yes	3	3	3	75%	63.1

The question should be contrasted with question Q4 in table 4.2, which is identical in form except for the replacement of the *and* with *or*, and has correct answer 'yes'. Q4 had 92% correct responses and was answered much more quickly; the average response time was 44 seconds for Q4 and 75 seconds for Q2. It is interesting to compare these results with those of Khemlani et al. (2012a), reporting on an experiment with 'naïve reasoners'. As noted in section 2.2.2, they investigated the comprehension of compound sentences of the form *not (A and B)* and *not (A or B)* and found a similar wide gap in accuracy of answering questions: 18% correct for the negated conjunction and 89% correct for the negated disjunction. They interpret these results in terms of the mental model theory and the expansion of *not (A and B)* into three models. In fact, in both Q2 and Q4, arguably four mental models are required, representing the decomposition of Entity *(Event or Abstract or Quality or Object)*. The problem is not simply one of managing a number of different models, but of the difficulty of creating the full set of models in the negation process. In Q4 all the participant has to do is to erase Event and Quality from the decomposition of Entity, leaving Abstract and Object. In Q2, rather than evaluate *Event and Quality* and then *not (Event and Quality*) as proposed in the paragraph above, participants attempt to expand *not (Event and Quality)* and many arrive at a single mental model corresponding to the term *not Event and not Quality*.

Apart from emphasis during training, potential solutions to this problem are:

- automatic expansion of *not (A and B)* into its three atomic constituents;
- an automatically generated graphical representation.

There is also the additional possibility of confusion between the everyday use of *and* and its logical use. It may be that, when faced with the difficulty of negating a conjunction,

participants take the easy option by interpreting *and* as equivalent to *or* (e.g. as in the English statement "the car is available in blue and silver"). The use of an alternative keyword to *and*, e.g. *intersection* or *int*, could avoid this linguistic confusion.

Whatever the reason for the erroneous treatment of Q2, it is striking that the results for naïve reasoners were so similar to those found amongst our participants. This suggests that both groups were using the same mental processes.

The complement operation was used by 22 of the 47 respondents to the question on DL feature usage in the survey by Warren (2013). However, it was not identified by Power (2010) as being commonly used and was not in the commonly used patterns identified by Khan and Blomqvist (2010). This relatively low usage may account for the low proportion of correct responses, since some participants may not have been familiar with the use of the complement operation.

4.3.2 Non-inheritance of transitivity – table 4.3: Q4

(answer: no; 33% correct responses)

In the pattern, *has_component* is defined as a subproperty of *has_part*, which is defined to be transitive. For the deduction to be true it would be necessary for *has_component* to also be transitive. There are a number of reasons why this question might be answered incorrectly. It may be that participants forget which is the parent, transitive property, and which is the subproperty, i.e. that they confuse the two names. It might also be that the name *has_component* suggests transitivity. Alternatively, people may assume that property characteristics are necessarily inherited by subproperties. This would be natural for people coming from an object oriented background, or those chiefly used to thinking about class subsumption relations in ontologies. That this is not the case for transitivity was noted in the handout, which cited the example of the property *is_descendant_*of and its subproperty *is_child_of*[7]. In fact, a different choice of property name might guard against all these problems; *has_direct_part*, in place of *has_component*, could better convey the required meaning. Subproperties appear to be relatively frequently used, and the transitive characteristic is one of the most commonly used characteristics. When training ontology users, attention needs to be drawn to the fact that not all characteristics are inherited, perhaps spelling out those which are and those which are not.

[7] Similarly, the characteristic of symmetry is not inherited, as can be seen from the property *is_sibling_of* and its subproperty *is_brother_of*. On the other hand, functionality is inherited, since if a subproperty has two values for the same subject, then so will its superproperty.

Table 4.3 Questions based on the componency pattern

Class Object	SubClassOf has_component only Object SubClassOf is_component_of only Object
Property has_part	Characteristic Transitive
Property is_part_of	Characteristic Transitive InverseOf has_part
Property has_component	SubPropertyOf has_part
Property is_component_of	SubPropertyOf is_part_of InverseOf has_component

	Question	yes/ no	MM	RC	num steps	%age corr	av. time (secs)
1	A is_part_of B; C is_part_of B => A is_part_of C	no	1	2	n/a	100%	62.3
2	A is_part_of B; B is_part_of C => A is_part_of C	yes	1	2	1	100%	20.3
3	B is_part_of C; A is_part_of B => A is_part_of C	yes	1	3	1	100%	30.7
4	A has_component B; B has_component C => A has_component C	no	1	2	n/a	33%	62.8
5	A has_component B; B has_component C => A has part C	yes	1	2	3	83%	29.0
6	A has_component B; B is_part_of C => A has_part C	no	1	2	n/a	83%	57.9
7	A has_component B; C is_part_of B => A has_part C	yes	1	4	3	100%	37.4
8	A Type Object; A has_component B; C Type not Object => B DifferentFrom C	yes	1	3	2	100%	49.9
9	A Type Object; A has_part B; C Type Not Object => B DifferentFrom C	no	1	2	n/a	83%	47.5
10	A has_component B; C is_component_of B => C is_part_of A	yes	1	4	4	100%	54.2

4.3.3 The functional characteristic – table 4.2: Q3

(answer: yes; 50% correct responses)

Since Object and Event are disjoint, B and D must be different. The functionality of represents then ensures that A and C are different. It may be that those who answered this question incorrectly did not fully understand the nature of a functional characteristic, despite it being explained in the handout. It may also be that the high relational complexity (i.e. RC = 4) of the question contributed to its difficulty. Here again, a diagrammatic representation would aid comprehension.

4.3.4 The existential qualifier – table 4.4: Q6

(answer: yes; 67% correct responses)

Each member of the class Game *has_participant some Player*; hence this is true of A. Since Player is a subclass of Object, *A has_participant some Object*. Moreover, every individual that *has_participant some Object* is in Event, because of the EquivalentTo statement. Therefore, A is in Event. The fact that a relatively large number of participants got this question right, despite its apparent complexity, may be due to the frequency of use of the existential quantifier, e.g. see Khan and Blomqvist (2010), Power (2010) and Warren (2013).

Table 4.4 Questions based on the modified coparticipation pattern

Class Event	EquivalentTo has_participant some Object DisjointWith Object
Class Object	DisjointWith Event
Class Player	SubClassOf Object
Class Game	SubClassOf has_participant some Player
Property coparticipates_with	Domain Object, Range Object Characteristics Symmetric, Transitive
Property has_participant	Domain Event, Range Object InverseOf is_participant_in

	Question	yes/ no	MM	RC	num steps	%age corr	av. time (secs)
1	A coparticipates_with B => A Type not Event	yes	1	2	2	92%	54.9
2	A is_participant_in B; C coparticipates_with D => A DifferentFrom C	no	1	2	n/a	92%	68.8
3	A is_participant_in B; C is_participant_in B => A is participant in C	no	1	2	n/a	100%	43.6
4	A has_participant B; C is_particpant_in D => B DifferentFrom D	yes	1	2	3	92%	44.6
5	B coparticipates_with A; B coparticipates_with C => C coparticipates_with A	yes	1	3	2	100%	34.8
6	A Type Game => A Type Event	yes	1	3	3	67%	47.6

4.3.5 Superclasses – table 4.2: Q5

(answer: yes; 75% correct responses)

Participants did relatively well on the question, only two providing incorrect answers and one not responding. In all the analyses the non-response is treated as an incorrect answer. The question is discussed here in part because it provides an example of a different approach to the use of mental models. Entity is composed of four disjoint subclasses. Nonconceptual and Nontemporal comprise two and three of these disjoint subclasses respectively. The only one they have in common is Object; hence their conjunction is equivalent to Object. A straightforward application of the mental model approach suggests that the maximum number of mental models is three, since Nontemporal is comprised of three disjoint classes. There is, however, little difficulty in formulating these models, unlike in the case of negation of a conjunction in table 2, Q2. In fact, a quite natural way to think about this is as two overlapping superclasses, with Object constituting the overlapping portion. This also lends itself naturally to a graphical representation; some participants might even have visualized it. This only requires that two models be held in working memory, one representing Nonconceptual, and the other Nontemporal.

4.4 Participants' feedback

After completing the questions, participants were able to provide written feedback about what they found difficult and what they found easy, and to make general comments. Some participants also made comments verbally. The most common theme was the use of intuition, in particular relating to names. There were conflicting views. One participant (p1) commented that "using named individuals instead of capital letters would have been easier" whilst another (p2) held the opinion that it was "easy to reason with anonymous things", since this safeguarded against the danger of using intuition rather than relying on the formal axioms. The contrasting views were also present when considering class and property names.

One participant (p3) commented that because the class and property names were familiar it was necessary to check whether the meaning in the OWL expression was similar to the normal English usage; another (p4) stated that "the axioms were realistic so one could rely to some extent on common sense". Participant p1 also commented favourably on the lack of use of formal logic symbols, which is a feature of the Manchester OWL syntax.

Four participants commented on the value of diagrams. Here there were no conflicting views but a consensus that diagrams are useful, e.g. participant p3 stated: "perhaps I would have done better if I'd drawn diagrams on paper" and another participant (p5) commented: "a pictorial representation of the relationships would have been easier to use". Indeed, the automatic generation of diagrams is likely to have helped comprehension in all the questions discussed above. One participant (p6) expressed a related view that colour-coding for OWL entity types and font weights and styles for keywords would be useful.

There were some interesting comments about OWL features, including the difficulty of using the existential and universal quantifiers (participant p1); confusion between *and* and *or* (participant p3, see Q2 from table 4.2 discussed in section 5.1); and (participant p7) the effect of users' legacy, e.g. that of a database background. Each of these comments is relevant to one of the questions discussed in section 4.3.

4.5 Statistical analysis

4.5.1 The participants

The majority of participants achieved high scores; two achieved twenty out of twenty-one, whilst the lowest score was thirteen. Ranking on knowledge of logic and OWL significantly correlates with ranking on accuracy. The Spearman rank correlation coefficient between knowledge of logic and accuracy was 0.53, corresponding to $p = 0.038$ on a one-tailed t test. For the correlation between knowledge of OWL and accuracy, the coefficient was 0.54, corresponding to $p = 0.036$ on a one-tailed t test. The effect was greater when we consider just the questions with correct answer 'yes'. For these questions, the correlation factor was 0.57 ($p = 0.027$) for knowledge of logic and 0.60 ($p = 0.019$) for knowledge of OWL. For the 'no' questions the rank correlation with knowledge of logic and knowledge of OWL were no longer significant ($p = 0.052$ and $p = 0.102$ respectively).

There was considerable variation in the total time taken to answer the questions, ranging from around thirteen minutes to around forty-two minutes. For our participants, knowledge of OWL had a much greater effect on the total time taken than did knowledge of logic. For the former the Spearman rank correlation coefficient was -0.65, with $p = 0.011$ on a one-tailed t test; for the latter the coefficient was -0.29, with $p = 0.178$. The low correlation in the case of knowledge of logic may have occurred because the majority (67%) of our participants ranked themselves in the same category ('some knowledge').

4.5.2 The questions

Most of the questions were answered correctly by all or most of the participants, with an apparent tendency to achieve greater accuracy on the questions with correct answer 'yes'. Table 4.5 provides a breakdown of the responses showing how many were correct and incorrect for the two categories of questions; it also shows the average times for each combination. A Pearson $\chi 2$ test confirmed the greater accuracy on the 'yes' questions ($p =$

0.005). The greater accuracy for the 'yes' questions occurs despite the fact that the average maximum relational complexity for these questions is greater than that for the 'no' questions, i.e. they might have been expected on average to be harder.

Table 4.5 Breakdown of responses, also showing average times in each category

	Yes	No
Correct	138; 43.4 secs	72; 59.5 secs
Incorrect	18; 58.4 secs	24; 76.3 secs

The time spent by any participant answering a single question varied from 9 seconds to 208 seconds, the average time across all participants for each of the twenty-two question varied from 20.3 seconds to 91.5 seconds. A two sample one-sided unpaired t test indicated that the 'yes' questions were answered on average significantly more quickly than the 'no' questions ($p < 0.001$). This may represent a tendency to initially attempt to prove the validity of the deduction. After first such attempts fail, the participant then has two possible strategies: either to continue such attempts until convinced that a proof is not possible or to attempt to prove explicitly that the deduction does not hold. The strategy adopted is likely to depend upon the person and the particular question. A one-sided unpaired t test also indicated that the correct responses were arrived at significantly more quickly than the incorrect responses ($p < 0.001$). A two-way ANOVA indicated that the two factors did not interact ($p = 0.884$). Thus the correct responses to the 'yes' questions averaged the least time (43.4 seconds) whilst the incorrect responses to the 'no' questions averaged the greatest time (76.3 seconds).

A simple linear regression showed that, overall, questions with a large number of correct responses were answered more quickly than questions with fewer correct responses ($p < 0.001$). This was also true when analysis was restricted to those questions with correct answer 'yes' ($p < 0.001$) and also to all those questions correctly answered ($p = 0.001$). However, there was no significant relationship between time and number of correct responses for those questions where the correct answer was 'no' ($p = 0.349$), nor for the incorrectly answered questions ($p = 0.947$).

4.5.3 Theories of reasoning

An objective was to determine whether any of the psychological theories could be used to predict, in terms of accuracy and time, the behaviour of our participants, and thus whether any of these theories would be useful in understanding how people reason about DLs. Each question was analyzed to determine the maximum number of mental models and maximum relational complexity which it would entail; as shown in tables 4.2, 4.3 and 4.4. For the questions with correct answer 'yes', this was done by constructing the proof of the deduction, and determining the number of mental models and the relational complexity at each stage. The questions with correct answer 'no' were examined to determine the maximum number of mental models and maximum relational complexity which would be met in thinking about them.

Only three questions required more than one mental model, making any statistical analysis impossible. One (table 4.2, Q2) was a 'no' question requiring four mental models and was the least well answered, with only three correct responses; this question is discussed in

section 4.3.1. The other two were 'yes' questions requiring four (table 4.2, Q4) and three mental models (table 4.2, Q5) and with 11 and 9 correct responses; the second of these is discussed in section 4.3.5. Moreover, whilst these three questions might be regarded as requiring more than one mental model, a participant with some knowledge of logic might well have used an alternative approach. The prevalence of questions requiring only one mental model seems to arise from the way in which Description Logics are used. The complexity is often in the relations, rather than in the existence of numerous possibilities.

This last statement might lead one to expect that relational complexity would be a better predictor of performance. However, a logistic regression of accuracy against maximum relational complexity did not provide a significant result. This was also the case for a linear regression of time to answer each question versus maximum relational complexity. When the latter regression was limited to the 'yes' questions, then there was a significant result ($p = 0.009$). The difference between 'yes' and 'no' questions may be a feature of the way people approach 'no' questions, or it may arise from the design of questions; all but one 'no' question had relational complexity of two.

The rule-based theory leads one to expect that, for the 'yes' questions, performance might be predicted by the number of steps in the reasoning chain. Whatever the validity of this theory is for naïve reasoners in everyday life, it could have some relevance to our participants, all of whom had at least a little knowledge of logic. This was investigated by looking at the 13 'yes' questions. It might be expected that, as the number of steps in the reasoning chain increases, the accuracy of answering will decline, as the possibility of error multiplies and fatigue sets in. A logistic regression of accuracy against number of steps did not provide a significant result ($p = 0.355$). However, all questions had one, two or three steps, with the exception of Q10 of table 4.3 which had four steps. This last question was correctly answered by all participants. When this question is removed from the analysis, a significant result is achieved ($p = 0.046$).

A linear regression of time for each response against number of steps also provided a significant result ($p = 0.036$). This effect was slightly more significant when only the correct responses were analyzed ($p = 0.028$), but not significant for the incorrect responses ($p = 0.360$). In fact, the mean times for one, two, three and four step questions were 25.5, 51.9, 44.3 and 54.2 seconds respectively. A Tukey range test at the 95% level indicated a significant difference between the means of only the first two groups.

4.6 Key findings

This study represents an attempt to understand the cognitive difficulties of using Description Logics. A key message is that, despite training, users are prone to certain misconceptions. These include confusion about the combined use of *not* and *and*; about the inheritance of property characteristics; and to a lesser extent about the functional characteristic and also the existential quantifier. Confusion may also arise through choice of names, a point taken up in the comments made by participants. The use of realistic names can lead to erroneous intuitions; however the mnemonic advantage is likely to outweigh this disadvantage. The important thing is to use names which are not likely to create incorrect intuitions.

In the study, maximal relational complexity did not significantly affect accuracy but did significantly affect the time to confirm an inference. The number of steps in a reasoning chain affected the time to reason and also, when one question was removed from the analysis,

affected the accuracy of reasoning. Given the participants' background and the nature of the questions, it seems likely that they did at the conscious level adopt a rule-based approach, as evidenced by the effect of number of steps on accuracy and time. The fact that one-third of the respondents commented on the value of drawing a diagram indicates that they were also thinking in terms of models.

5 Future work and workplan

5.1 Discussion

In section 1, two research questions were posed:
1. *In what way can the difficulties experienced in using Description Logics be understood in terms of an underlying theory, e.g. theories of reasoning already developed within the cognitive psychology community?*
2. *In what way could such a theory contribute to improving the usability of Description Logics?*

The ontology user survey discussed in section 3 has confirmed the importance of these questions. The survey showed both the importance of DLs and the difficulties users experience in creating and maintaining ontologies. Those difficulties frequently relate to the formality of the languages used and to the complexity of the information being described.

The study described in section 4 has illustrated how psychological theories of reasoning can make a contribution to understanding difficulties with DLs. This is seen in section 4.3.1, where the work of Khemlani et al. (2012a) and (2012b) helps explain the problem of negating conjunction in DLs. It is also seen in section 4.5.3 where the number of steps in a reasoning chain was shown to be a predictor of accuracy and time to reason, and relational complexity was shown to be a predictor of time. As already noted, there is a considerable psychological literature on reasoning, including studies of the use of diagrams. There is scope for utilising the results of this work to design better DL representations, be they linguistic, diagrammatic, or some combination of the two.

It is worth commenting at this stage on whether the focus of future work should be on DLs in general or OWL in particular. To a large extent this is a false dichotomy. DLs are fragments of first-order logic[8]. The variants of OWL are standardised fragments of first-order logic possessing particular computational properties suited to particular applications. For example, OWL2EL "is designed for representing large and moderately complex ontologies", e.g. biomedical ontologies, and OWL2QL is designed for applications with a simple ontology and large amounts of instance data, e.g. see Horridge et al. (2012). The essence of the OWL standardisation is the choice of these subsets of first-order logic. Since the aim of this work is to study the difficulties of those features which are relatively commonly used, any combination of features will generally be restricted to those occurring together in a commonly used (decidable) OWL variant. There are certain representations associated with OWL, e.g. the Manchester OWL Syntax (MOS) (Horridge et al., 2006). However, these are not intrinsic to OWL. An aim of the proposed work is to determine alternative representations which are easier to understand and work with.

There are a number of areas for further study. In particular, these include:

- further investigation of the cognitive complexity of DL statements;
- investigation of alternative linguistic and visual representations;
- further investigation into the difficulties experienced by ontology users in their working environments, and the contribution of the ideas developed in the previous activities.

[8] In particular, as noted by Krötsch et al. (2012), "most DLs are decidable fragments of first-order logic".

The next three subsections consider each of these areas in turn. The final subsection provides a workplan for the next three years.

5.2 Cognitive complexity of DL deductions

Section 4 has described how psychological theories of reasoning can be used to explain problems in interpreting DL deductions and in particular how relational complexity and number of statements in a reasoning chain can be used as measures of cognitive complexity to help predict accuracy and time to reason. In the work reported these effects were confounded with, for example, the varying difficulties of different DL features. A more rigorous study would need to vary the complexity of deductions constructed from a minimum possible set of features. In this way comparisons could be made between deductions of varying complexity without the confounding effect of different features. For example, the deduction rules examined by Nguyen et al. (2012) include nested existential quantifiers[9]. In their work the complexity of such deductions is confounded with the effect of other terms in the deduction rule. Understanding the cognitive effect of such complexity would assist the design of proof trees.

5.3 Alternative linguistic and visual representations

Linguistic and visual representations are considered together here both because it is valid to compare them and because they may be used together. This latter strategy would accommodate those with a preference for visual representations and those with a preference for linguistic representations. It may be that some people can profit from both kinds of representation simultaneously; the one complementing the other.

A starting point would be the OWL 'patterns'[10] identified as commonly used by Power and Third (2010). These patterns can be expressed in a variety of different ways, using different quantifiers and different sentence orders. The optimum form may depend on the context. For example, Johnson-Laird and Byrne (1989) demonstrated that the use of 'all' rather than 'only' led to more correct inferences in the case of modus ponens and fewer in the case of modus tollens. It could be the case that, in the context of the work of Nguyen et al. discussed above, the optimum expression of a DL statement might depend on the other statements in a deduction rule.

Visualization can be helpful, but the data presented in section 3.6 showed that this is not true for all users. Visualization is, in any case, chiefly limited to viewing taxonomic relations rather than, for example, the use of existential, universal and cardinality restrictions. Exceptions to this are the work on DL visualization by, e.g. Dau and Eklund (2008) and Howse et al. (2011). There have been many studies of various visualization schemes for taxonomic relations. However, as noted in section 2.3, there does not seem to have been any empirical studies into the efficacy of visualizations of DL statements, as envisaged in the cited papers. The OWL patterns identified by Power and Third could also be translated into diagrammatic representations and used as the basis for empirical studies.

---

[9] In the terminology of Nguyen et al: *SubClaOf(X,ObjSomValF(r0,ObjSomValF(r0,Y)))*

[10] The use of 'patterns' here differs from that in section 3.7. In the terminology of Power and Third, 'pattern' means a statement structure, e.g. *SubClassOf(Class,ObjectSomeValuesFrom(ObjectProperty,Class))*

A proposed next step is to conduct laboratory experiments to compare comprehension of the most common patterns of Power and Third when expressed in Manchester OWL Syntax and with alternative linguistic forms. Subsequent studies should include diagrammatic representations.

5.4 User feedback – interviews and focus groups

Any empirical studies of users need to be informed by an understanding of how real users behave and what problems they face when actually performing their work. The laboratory study described in section 4, and the studies proposed in sections 5.2 and 5.3 focus on ontology patterns known to be occur in practice. This needs to be complemented with an understanding of how the patterns are used in an actual working environment. The survey described in section 3 has gained some understanding of the needs and behaviour of actual users. It is proposed to conduct interviews and focus groups with ontology users to obtain richer insight into their problems and ways of working, and to gain feedback regarding some of the alternative linguistic and diagrammatic representations being developed in the work discussed in section 5.3. This work needs to take account of the different objectives of ontology users, e.g. using a categorisation similar to that of section 3.2.

5.5 Workplan

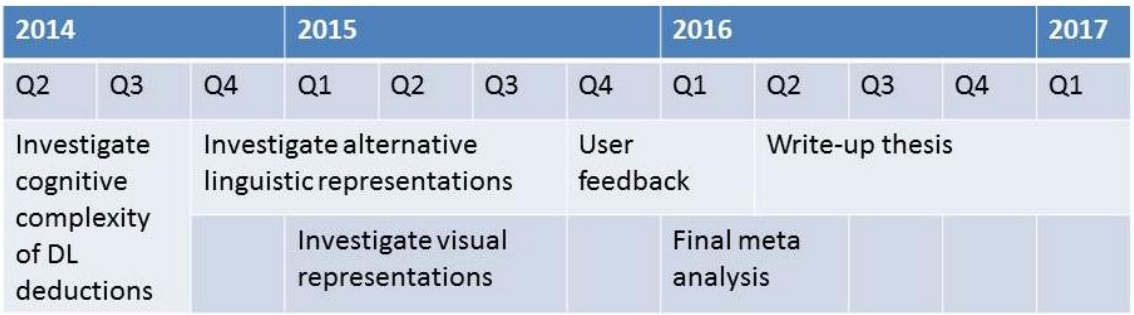


Figure 5.1 Timeline for April 2014 to March 2017

Figure 5.1 shows the workplan for the next three years, i.e. from April 2014 until March 2017. The first six months, i.e. Q2 and Q3 of 2014, will be concerned with the work on cognitive complexity of DL deductions discussed in section 5.2. This is essentially a more controlled investigation of the issues discussed in section 4. Work will then move on to investigating alternative linguistic and visual representatives, as discussed in section 5.3. This will begin with linguistic representations in Q4 2014. Work on visual representations will begin in Q1 2015 and the two related streams will continue until the end of Q3 2015. Interviews and focus groups, described in section 5.4, are planned to take place during Q4 2015 and Q1 2017. During the second half of this period, work on a final overall analysis is planned. The intention of this work is to draw

together the findings of the various studies and relate them to continuing work in the domain of psychological reasoning. This work will be concluded in Q2 2017 when, at the same time, writing up of the thesis will begin.

It is intended that each of the activities described in sections 5.2, 5.3 and 5.4 will be reported in a peer-reviewed conference or other publication.

References

Baader, F., Calvanese, D., McGuiness, D., Nardi, D., & Patel-Schneider, P. F. (2010). *The Description Logic Handbook: Theory, Implementation and Applications*. Cambridge University Press.

Bauer, M. I., & Johnson-Laird, P. N. (1993). How diagrams can improve reasoning. *Psychological Science*, *4*(6), 372–378.

BURTON, J., Stapleton, G., & Howse, J. (2012). Completeness proof strategies for Euler diagram logics. Retrieved from http://eprints.brighton.ac.uk/11378/

Dau, F., & Eklund, P. (2008). A diagrammatic reasoning system for the description logic ALC. *Journal of Visual Languages & Computing*, *19*(5), 539–573.

Goodwin, G. P., & Johnson-Laird, P. N. (2005). Reasoning about relations. *Psychological Review*, *112*(2), 468.

Halford, G. S., & Andrews, G. (2004). : The development of deductive reasoning: How important is complexity? *Thinking & Reasoning*, *10*(2), 123–145.

Horridge, M., Aranguren, M. E., Mortensen, J., Musen, M., & Noy, N. F. (2012). Ontology Design Pattern Language Expressivity Requirements (Vol. 929). Presented at the WOP 2012 workshop on ontology patterns, Boston, MA. Retrieved from http://ceur-ws.org/Vol-929/

Horridge, M., Bail, S., Parsia, B., & Sattler, U. (2011). The cognitive complexity of OWL justifications. *The Semantic Web–ISWC 2011*, 241–256.

Horridge, M., Drummond, N., Goodwin, J., Rector, A., Stevens, R., & Wang, H. H. (2006). The manchester owl syntax. *OWL: Experiences and Directions*, 10–11.

Howse, J., Stapleton, G., Taylor, K., & Chapman, P. (2011). Visualizing ontologies: a case study. *The Semantic Web–ISWC 2011*, 257–272.

Johnson-Laird, P. N., & Byrne, R. M. (1989). Only Reasoning. *Journal of Memory and Language*, *28*(3), 313–330.

Johnson-Laird, P. N., & Byrne, R. M. (1991). *Deduction*. Lawrence Erlbaum Associates, Inc. Retrieved from http://psycnet.apa.org/psycinfo/1991-97828-000

Johnson-Laird, P. N., Byrne, R. M., & Schaeken, W. (1992). Propositional reasoning by model. *Psychological Review*, *99*(3), 418.

Johnson-Laird, P. N., Lotstein, M., & Byrne, R. M. (2012). The consistency of disjunctive assertions. *Memory & Cognition*, 1–10.

Katifori, A., Halatsis, C., Lepouras, G., Vassilakis, C., & Giannopoulou, E. (2007). Ontology visualization methods—a survey. *ACM Computing Surveys (CSUR)*, *39*(4), 10.

Khan, M. T., & Blomqvist, E. (2010). Ontology design pattern detection-initial method and usage scenarios. In *SEMAPRO 2010, The Fourth International Conference on Advances in Semantic Processing* (pp. 19–24). Retrieved from http://www.thinkmind.org/index.php?view=article&articleid=semapro_2010_1_40_50071

Khemlani, S., Orenes, I., & Johnson-Laird, P. N. (2012a). *Negating compound sentences*. NAVAL RESEARCH LAB WASHINGTON DC NAVY CENTER FOR APPLIED RESEARCH IN ARTIFICIAL INTELLIGENCE. Retrieved from http://mindmodeling.org/cogsci2012/papers/0110/paper0110.pdf

Khemlani, S., Orenes, I., & Johnson-Laird, P. N. (2012b). Negation: A theory of its meaning, representation, and use. *Journal of Cognitive Psychology*, *24*(5), 541–559.

Krötzsch, M., Simancik, F., & Horrocks, I. (2012). A description logic primer. *arXiv Preprint arXiv:1201.4089*. Retrieved from http://arxiv.org/abs/1201.4089

Larkin, J. H., & Simon, H. A. (1987). Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science*, *11*(1), 65–100.

Newstead, S. E., Bradon, P., Handley, S. J., Dennis, I., & Evans, J. S. B. (2006). Predicting the difficulty of complex logical reasoning problems. *Thinking & Reasoning*, *12*(1), 62–90.

Nguyen, Power, Piwek, & Williams. (2012). Measuring the understandability of deduction rules for OWL. Presented at the First international workshop on debugging ontologies and ontology mappings, Galway, Ireland. Retrieved from http://oro.open.ac.uk/34591/

Power, R. (2010). Complexity assumptions in ontology verbalisation. In *Proceedings of the ACL 2010 Conference Short Papers* (pp. 132–136). Retrieved from http://dl.acm.org/citation.cfm?id=1858866

Power, R., & Third, A. (2010). Expressing OWL axioms by English sentences: dubious in theory, feasible in practice. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters* (pp. 1006–1013). Retrieved from http://dl.acm.org/citation.cfm?id=1944682

Rector, A., Drummond, N., Horridge, M., Rogers, J., Knublauch, H., Stevens, R., … Wroe, C. (2004). OWL pizzas: Practical experience of teaching OWL-DL: Common errors & common patterns. In *Engineering Knowledge in the Age of the Semantic Web* (pp. 63–81). Springer. Retrieved from http://link.springer.com/chapter/10.1007/978-3-540-30202-5_5

Rips, L. J. (1983). Cognitive processes in propositional reasoning. *Psychological Review*, *90*(1), 38.

Rodgers, P., Stapleton, G., Flower, J., & Howse, J. (2014). Drawing area-proportional Euler diagrams representing up to three sets. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6570477

Warren, P. (2013). *Ontology Users' Survey - Summary of Results* (KMi Tech Report No. kmi-13-01). Retrieved from http://kmi.open.ac.uk/publications/pdf/kmi-13-01.pdf

Yang, Z., Zhang, D., & Ye, C. (2006). Ontology analysis on complexity and evolution based on conceptual model. In *Data Integration in the Life Sciences* (pp. 216–223). Retrieved from http://link.springer.com/chapter/10.1007/11799511_19

Zhang, H., Li, Y.-F., & Tan, H. B. K. (2010). Measuring design complexity of semantic web ontologies. *Journal of Systems and Software*, *83*(5), 803–814.

Zielinski, T. A., Goodwin, G. P., & Halford, G. S. (2010). Complexity of categorical syllogisms: An integration of two metrics. *European Journal of Cognitive Psychology*, *22*(3), 391–421.