View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Sheffield Hallam University Research Archive

Some Reflections on Knowledge Representation in the Semantic Web

John Kirby

Computing and Communications Research Centre
Sheffield Hallam University
John.Kirby@shu.ac.uk

Keywords: Semantic Web, Description Logics, WordNet, Knowledge Representation

Abstract

The knowledge representation technology Description Logics (DLs) has become an important component of developments around the Semantic Web. It is suggested here that in order to be really useful, the knowledge represented in DLs should in some fundamental way reflect the way the human mind organises and structures the same knowledge. There is a short historical review of some relevant background work in cognitive psychology, including WordNet. This is followed by a brief introduction to the importance of automatic classification in DLs before considering some issues around ontologies.

1. Introduction

Twenty years ago I became involved in the PEN&PAD project (Nowlan et al, 1991, Kirby and Rector, 1996) on the user centred design of clinical data entry systems based on the formal representation of medical terminology. I also participated in the GALEN project (Rector et al, 1999) which further extended the formal representation of medical terminology using a version of the knowledge representation technology Description Logics (DLs) to develop an extensive medical ontology. At that time, I believed that the use of DLs in the development of the GALEN medical ontology was an extension of the user centred approach because in some way the representation of knowledge was related to the workings of human semantic memory.

For me a key insight of the PEN&PAD project was that a fundamental aspect of the design of a system involves developing a representation of data or knowledge that intuitively corresponds to the understanding of users. For PEN&PAD this meant the development of a system of knowledge representation for symptoms, signs and diseases necessary for clinical data entry by general medical practitioners. In my experience, user interface design, even if it is firmly focused on supporting user tasks, depends to a large extent on how well the underlying representation of knowledge or data fits with the understanding of users. So an approach to the representation of knowledge that in some way corresponds to the way users think seemed to be a significant step in the direction of real user centred system design. User centred design was not just a matter of paying attention to the surface appearance in the form of good interface design but also extended to the deep structures where knowledge representation and data in some way matched or was intuitive to the user.

The proposition of the Semantic Web is based on similar assumptions of the fit between a formal knowledge representation meaning and human understanding. For example,

> *...if the interaction between person and hypertext could be so intuitive that the machine-readable information space gave an accurate representation of the state of people's thoughts, interactions, and work patterns...*(Berners-Lee, 1996)
>
> *In the Semantic Web "information is given well-defined meaning, better enabling computers and people to work in co-operation.* (Berners-Lee et al, 2001)

With the same emphasis on capturing meaning and formal representation, it is not surprising that over the past ten years, DLs have become increasingly prominent in the developments around the Semantic

Web leading to the development of OWL or Web Ontology Language (Patel-Schneider et al, 2004; W3C, 2009).

The literature on the Semantic Web and DLs emphasises the need for formal computer representation of "meaning", "understanding" and "knowledge" all of which is, of course, implied in the word "semantic" itself. Using formal computer representation means that Semantic Web technologies would be amenable to searching and manipulating data "in ways that are useful and meaningful to the human user" (Berners-Lee et al, 2001). Ultimately, however, the final arbiter of the adequacy of the meaning, understanding and knowledge represented has to be the human end-user. This would seem to imply that the essentially human knowledge represented in DLs should in some fairly fundamental way correspond to the way this knowledge is represented in the human mind.

The paper presents a brief historical review of background work in cognitive psychology that is relevant to the development of the DLs approach to conceptual knowledge representation. The importance of automatic classification is described before the discussion of some issues around the development of DLs ontologies.

2. Cognitive Psychology Background

3.1. Hierarchy and Inheritance - Quillian's Model

Quillian (1969) proposed that human semantic memory is organised as a hierarchy of categories or nodes each of which has a set of properties that are also inherited by subordinate categories. In Figure 1, the properties of canary are "is yellow" and "can sing" with a pointer to the category of "bird" to which canary belongs. Because canary is part of the category bird it inherits the properties "has wings", "can fly" and "has feathers" The idea that properties are inherited is also referred to as "cognitive economy".

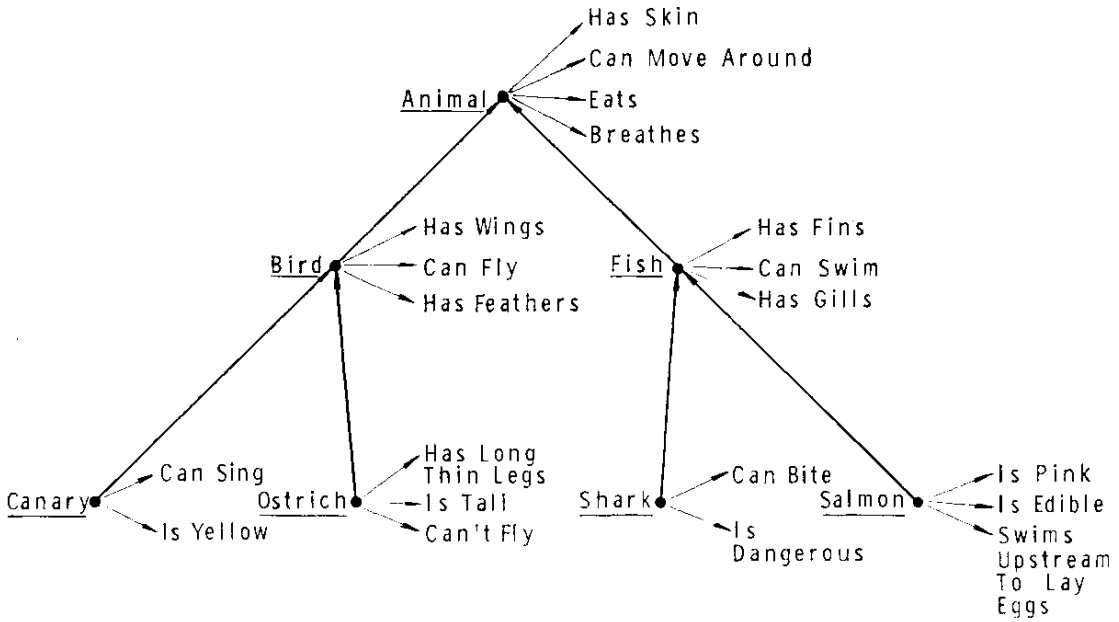

Figure 1: Illustration of the hypothetical memory structure for a 3-level hierarchy (Collins & Quillian, 1969)

Collins and Quillian (1969) tested this hypothesis in reaction time experiments in which subjects were presented with short sentences of two types. In one case subjects were asked to confirm whether or not a word was a member of another category, for example, "a canary is a bird" or "a canary is an animal". In the other case they were asked to confirm whether or not a word possessed a property either directly or inherited from a higher level category, for example, "a canary is yellow", "a canary

has wings" or "a canary has skin". It was found that the reaction times of subjects increased in direct proportion to the number of levels in the hierarchy that would have to be traversed.

3.2. Experimental Counter Evidence

A number of subsequent experimental studies cast considerable doubt on both of Quillian's main propositions on semantic memory: its hierarchical organisation and the inheritance of properties.

> **Familiarity**: Conrad (1972) argued that longer reaction times were due to the lack of familiarity of subjects statements such as "a canary has skin". No difference in reaction times was found when the original experiments were repeated with controls for familiarity, suggesting that there is no evidence for the inheritance of properties.
>
> **Typicality**: Reaction times for verification of categories have shown a marked tendency for typical members of categories to have faster response times than atypical ones, for example, categorising a robin as a bird compared with an ostrich or a chicken. Rosch (1973) concluded that categorisation might be by property matching rather than being derived from the hierarchical organisation of memory.
>
> **Arbitrary Categories**: Rips et al (1973) found that subjects verified instances of mammals, such as Cat, Goat and Mouse, more rapidly as animals than they verified them as mammals which would give rise to the conclusion that mammal was the super-ordinate of animal, something that is dismissed as being nonsensical. The conclusion drawn is that the network categories are essentially arbitrary and assigned on a logical rather than empirical basis.
>
> **Fuzzy Categories:** McCloskey and Glucksbery (1978) found that some items were categorised consistently while other items were categorised inconsistently both between different subjects and by the same subject at different times. For example, tomato would be sometimes categorised as a fruit and sometimes as a vegetable. The conclusion is that natural categories have fuzzy boundaries.

These results have led many cognitive psychologies to conclude that Quillian's basic propositions of hierarchy and inheritance were essentially flawed, for example, Eysenck and Keane (2010) and Baddeley et al (2009). Collin who collaborated with Quillian later produced the alternative "Spreading Activation" model of semantic memory (Collins and Loftus, 1975). Despite being more successful in explaining experimental findings, Eysenck and Keane (2009) conclude that it is difficult to assess the adequacy of the spreading activation theory because it does not make precise predictions. For this same reason, the spreading activation theory would not appear to be amenable to the development of computerised models or knowledge representation.

3.3. The Psycho-linguistics - WordNet

However, not all workers in the field accepted the conclusions from this experimental work, in particular those who developed the "psycho-linguistic" WordNet project:

> *An alternative conclusion - the conclusion on which WordNet is based - is that the inheritance assumption is correct, but that reaction times do not measure what Collins and Quillian, and other experimentalists, assumed they did. Perhaps reaction times indicate a pragmatic rather than a semantic distance - a difference in word use, rather than a difference in word meaning.* (Miller, 1990)

Consequently, in WordNet around 100,000 English nouns[1] are organised into sets of synonymous words (synsets) that are hierarchically organised. Each synset is therefore defined in relation to its parent synset plus distinguishing features which is basically in line with Quillian's original theory of semantic memory. In addition, a synset may also include links to other synsets that capture whole-part relationships.

[1] Although WordNet now also contains verbs, adjectives and adverbs, they are beyond the scope of this paper .

It is important to note that its authors have emphasised that WordNet is a "lexical database" organised on "psycho-linguistic" principles but more recently they have also described it as an "ontology" (Miller & Fellbaum, 2007). However, these psycholinguistic principles are not incorporated in any form of automatic classification system which means that semantic inconsistencies are possible. Nevertheless, WordNet has emerged as a significant element in the developments around the Semantic Web and has been used by many workers to construct, compare and merge ontologies.

3. Description Logics

Description logics (DLs) are a family of computerised knowledge representation systems that have their roots in earlier semantic networks (Woods, 1975) and frame based approaches (Minsky, 1975). The proponents of DLs such as Nardi and Brachman (2007) emphasise the importance of the development of logical systems based on sound computational algorithms and in the process move away from any explicit reference to human semantic memory. They conceded that "owning to their more human-centred origins, the network-based systems were often considered more appealing , and more effective from a practical viewpoint than logical systems." However, they conclude that such systems are "not fully satisfactory because of their lack of precise semantic characterization."

2.1. Automatic Classification

Like WordNet, the knowledge represented in DLs are ontologies consisting of concepts, also referred to as classes, organised in a strict hierarchy where each concept "is-a-kind-of" its parent. Also like WordNet, concepts may have properties, referred to as roles, that are themselves organised in a subsumption hierarchy. As with object-oriented programming languages, properties of concepts are inherited, so that, for example, all of the properties of animal are inherited by bird and fish - see Figure 1.

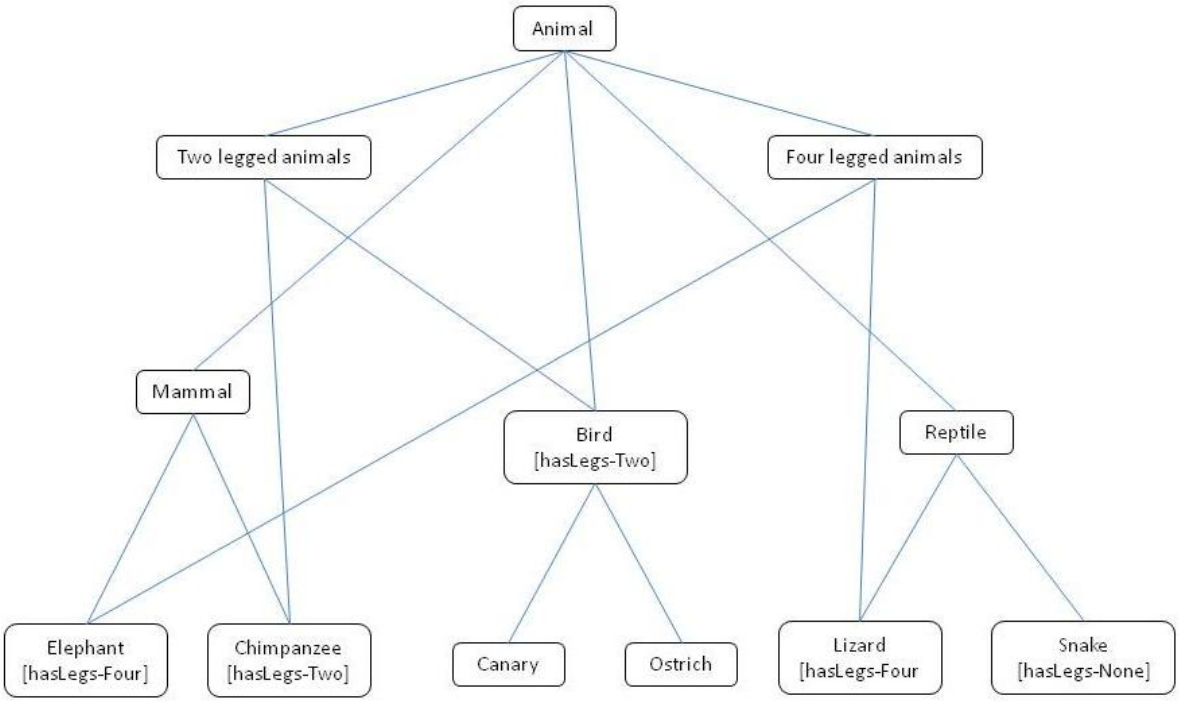

Figure 2: Hierarchy of animals with properties

The key difference with WordNet is that DLs carry out automatic classification. In addition to "atomic" concepts - such as bird and canary - DLs also allow for concept descriptions to be composed from atomic concepts and properties. These "complex" concepts are formed by creating a description that consists of a base concept and one or more property plus value pairs.

Consider the example of a role or property of "hasLegs" with possible values of Two and Four and these properties are applied to Animal concepts as shown in Figure 2. The complex concepts of "Two legged animal" - Animal: hasLegs-Two - and "Four legged animal" - Animal: hasLegs-Four - may then be constructed. When these complex concepts are created in DLs they are automatically inserted into the correct position hierarchy, that is, they are automatically classified. Figure 2 shows that the children of Two legged animal include Chimpanzee and Bird, and by inheritance Canary and Ostrich. In DLs, automatic classification is used both in the construction of DLs ontologies to ensure that they are logically and semantically consistent and for searching and reasoning using DLs.

3.2. Ontologies

The usefulness of DLs depends on the creation of comprehensive ontologies. In his review of progress on the Semantic Web, Horrocks (2007) cites examples of the development of specialist ontologies for biology, medicine, geography, geology, astronomy, agriculture and defence. However, Shadbolt et al (2006) appear to accept that the development of comprehensive and usable ontologies has been relatively slow and go on to suggest that too much effort may have been expended on specialist, or "deep", ontologies rather than on "shallow" ontologies representing more everyday concepts such as customer, account number and overdraft.

3.2.1. Empirical Findings Revisited

Berners-Lee et al (2001) assert that the Semantic Web will "improve the accuracy of web searches [by using] precise concepts instead of ... ambiguous keywords". However, I am not convinced that formally and logically correct ontologies will necessarily enable the development of systems providing improved user experience. In spite of being rejected by the developers of WordNet, and essentially ignored by advocates of DLs, there may still be issues arising from the early experimental studies mentioned earlier.

It seems likely that most of us have a mental definition of a "typical bird" that has wings and feathers and, in particular, can fly. Ostriches and penguins are not "typical birds" because they do not fall into that definition of because neither flies which may mean that we put them into one or more different categories of "atypical birds". However, when asked most people know that "technically" ostriches and penguins are types of birds albeit atypical ones. An alternative explanation of the typicality effect reported by Rosch (1973) might be that the categories used by human beings are not necessarily always the scientifically or logically correct ones assumed in the experiments.

In arguing that hierarchical categories seem to be arbitrary, Rips et al (1973) cite the seemingly anomalous finding that instances of mammals such as horse or elephant were categorised as Animal faster than as Mammal. They go on to reject as "most implausible" the idea that people would categorise such instances of mammal being immediate subordinates of Animal. However, anecdotal evidence suggests that many people do indeed consider such instances of mammals to be direct sub-ordinates of the term Animal. This includes my crossword puzzle dictionary in which the overwhelming majority of entries in the Animal section are mammals (Bailie, 1998). If asked directly, it seems likely that most people would be able to confirm the scientific and logically correct position that such mammal instances were technically mammals and that not all animals are mammals. This may suggest that many of us can simultaneously use different classification hierarchies, in this case an everyday one in which cats, goats and mice are animals and another more logical and scientific one in which they are mammals.

The finding of McCloskey and Glucksbery (1978) that tomato would be sometimes categorised as a fruit and sometimes as a vegetable may also be explained by proposing different hierarchies or at least branches of hierarchy. In the strictly scientific sense tomato is a fruit whilst in relation to food it is a vegetable because it is used in savoury sauces, salads etc rather than in desserts where we would normally think of consuming more typical fruit such as apples, pears or mangos. This would not suggest that natural categories have fuzzy boundaries but that human beings are capable of classifying the same thing in a number of different, but equally valid, ways.

These tentative alternative explanations of experimental findings support the idea of a hierarchically organised semantic memory with the inheritance of properties as proposed by Quillian (1969) and supported by Miller (1990). However, it is suggested that human semantic memory might consist of several possibly contradictory or inconsistent hierarchies involving the same concepts even within the same individual. It is not clear what impact this would have for creating DLs ontologies that more intuitively correspond to the way human beings represent conceptual knowledge.

3.2.2. Developing Ontologies

Shadbolt et al (2006) believe that the Semantic Web needs ontologies that are "developed, managed, and endorsed by committed practice communities". However, the development and management of ontologies is a time consuming and labour intensive process. Because ontologies are hand-crafted, the hierarchical structure and properties reflect the knowledge, understanding and even the values and prejudices of their authors. In addition, ontologies may be created for specific purposes in the same subject area which may result in mean important differences that may be difficult to reconcile. It may therefore be difficult to develop an ontology in each subject area that is endorsed by specialists in that area.

Furthermore, there is no guarantee that endorsement of an ontology by specialists in a subject area will provide the basis for producing useful and usable systems easily. For example, in the early days of the PEN&PAD project the clinical data entry user interface had been driven directly from the underlying medical ontology. The later version of PEN&PAD that I was involved in with used the more general and re-usable GALEN ontology that had been developed and endorsed by a number of physicians and surgeon. However, this broader ontology contained much that was not relevant for clinical data entry. For example, the fact that arteries and veins were modelled as hollow tubes was of little use to general practitioners entering data about symptoms, signs and diseases of the circulatory system. The solution was to develop an additional processing layer between the ontology and the user interface using pragmatic knowledge of what aspects of the underlying ontology were clinically relevant.

In general, the development of an ontology for a specific purpose in a particular subject area is likely to meet the requirements of that purpose but is unlikely to meet other requirements as in the above PEN&PAD example. Instead of creating a multitude of potentially contradictory single purpose ontologies in the same subject area, it would seem desirable to build general and re-usable ontologies. However, to gain acceptance of correctness across a potentially diverse community of practice such ontologies are likely to become abstract and divorced from any particular purpose. In order to address this issue within the GALEN project, Rector et al (2001) described the development of a layered architecture.

5. Conclusions

DLs and WordNet are major components in developments around the Semantic Web. Despite their differences, both approaches are firmly based on the idea that human knowledge is represented in a hierarchical fashion with the inheritance of properties. This basic proposition seems to lead to the development of simple hierarchies and more extensive ontologies that are "correct" in some strictly logical or accepted scientific sense. It is speculated here that such representations of knowledge may not correspond to the way human beings organise these concepts for everyday purposes.

7. References

Baddeley A, Eysenck MW & Anderson MC (2009) Memory. The Psychology Press, Hove & New York.

Bailie J (1988) Pocket Crossword Dictionary. Hamlyn, London.

Berners-Lee T (1996) The World Wide Web: Past, Present and Future. http://www.w3.org/People/Berners-Lee/1996/ppf.html

Berners-Lee T, Hendler J and Lassila O (2001) The Semantic Web. Scientific American, 284(5): 35–43, May 2001.

Collins AM & Loftus FL (1975) A Spreading-Activation Theory of Semantic Processing. Psychological Review Vol. 82, No. 6, 407-428

Collins AM & Quillian MR (1969) Retrieval time from semantic memory. Journal of Verbal Learning and Verbal Behaviour, 8, 240-247.

Conrad C (1972) Cognitive economy in semantic memory. Journal of Experimental Psychology, 92, 149-154.

Eysenck MW & Keane MT (2010) Cognitive Psychology: a student handbook - 6th edition. The Psychology Press, Hove & New York.

Horrocks I (2007) Semantic Web: The Story So Far. Proceedings of the 2007 International Cross-Disciplinary Conference on Web Accessibility (W4A2007), Banff, Canada, May 7–8, 225. 120–125. ACM Press, NY, USA.

Kirby J and Rector AL (1996) The PEN&PAD data entry system: from prototype to practical system. In: Cimino J, editor, AMIA Full Symposium. Washington DC: Hanley and Belfus, 709-13.

Kintsch W (1980) Semantic Memory: A tutorial. In: Nickerson RS (Ed), Attention and Performance VIII. Hillsdale, NJ. Lawrence Erlbaum Associates Inc. 595-620.

McCloskey ME & Glucksberg S (1978) Natural categories: Well defined or fuzzy sets. Memory and Cognition, 6, 462-472.

Miller GA (1990) Nouns in WordNet: A Lexical Inheritance System. International Journal of Lexicography, Vol. 3 No. 4, 245-264.

Minsky M (1975) A framework for representing knowledge. In: Winston PH (Ed), The psychology of computer vision. McGraw Hill, New York. 211-277.

Nardi D and Brachman RJ (2007) An Introduction to Description Logics. In Baader F, Calvanese D, McGuinness D, Nardi D and Patel-Schneider PF (Eds) The Description Logic Handbook: Theory, Implementation and Applications. Cambridge University Press.

Nowlan WA, Rector AL, Kay S, Horan B and Wilson A (1991) A patient care workstation based on a user centred design and a formal theory of medical terminology: PEN&PAD and the SMK formalism. In Clayton P, editor, Proceedings of the Fifteenth Annual Symposium on Computer Applications in Medical Care. SCAMC-91. Washington, DC. McGraw-Hill, 855-7.

Patel-Schneider PF, Hayes P and Horrocks I (2004) OWL Web Ontology Language semantics and abstract syntax. W3C Recommendation, 10 February 2004. Available at http://www.w3.org/TR/owl-semantics/.

Quillian MR (1969) The teachable language comprehender: A simulation program and theory of language. Communication of the ACM, 12, 459-476.

Rector AL, Zanstra PE, Solomon WD, Rogers JE, Baud R, Ceusters W, Claassen W, Kirby J, Rodrigues J-M, Mori AR, van der Haring EJ, and Wagner J. (1999) Reconciling Users' Needs and Formal Requirements: Issues in Developing a Reusable Ontology for Medicine. IEEE Transactions on Information Technology in Biomedicine, Vol 2, No 4, 229-242.

Rips LJ, Shoben EJ & Smith EE (1973) Semantic Distance and the Verification of Semantic Relations. Journal of Verbal Learning and Verbal Behaviour, 12, 1-20.

Shadbolt N, Hall W & Berners-Lee T (2006) The Semantic Web Revisited. IEEE Intelligent Systems, May/June 2006, 96-101.

W3C (2009) OWL 2 Web Ontology Language, Document Overview, W3C Recommendation, 27 October 2009. Available at http://www.w3.org/TR/owl2-overview/

Woods WA (1975) What's in a link: Foundations for semantic networks. Reprinted in: Brachman RJ and Levesque H.J (Eds), Readings in Knowledge Representation. Morgan Kaufmann Publishers, San Francisco, California, 1985. 217–241.