Ants caught in the Semantic Web: A study in the application of description logic to animal systematics

Kerry Taylor
CSIRO ICT Centre
GPO Box 664 Canberra ACT 2601 Australia
Kerry.Taylor@csiro.au

Charles Gretton
Computer Sciences Laboratory
The Australian National University
Canberra ACT 0200 Australia
charles.gretton@csl.anu.edu.au

1. Introduction

Scientists have been organising the forms of natural life into structured hierarchical systems since Linnaeus in the 18th century. Much more recently, computer scientists have developed a class of languages, called description logics (DL), that are aimed at describing concepts so that they may be automatically classified in hierarchical structures. These languages are being adopted in recent proposals for ontology definition that underly the Semantic Web, particularly OWL-DL[1].

In this paper we study the applicability of modern description logics to the application of animal systematics. We would like to improve both the process of scientific classification itself, and the methods for communication and integration of taxonomic knowledge. As a case study, we consider a published scientific treatment of Epopostruma, a genus of Australian Formicidae (ants) [3]. We focus on expressing the morphological characters of Epopostruma, that is the features that derive from the form, structures, homologies and metamorphoses which characterise an individual. We express these characters in the description logic $\mathcal{ALCQHIO}_{\mathcal{R}^+}(D)^-$ underlying OWL-DL. RACER [2] is a readily-available reasoner for $\mathcal{ALCQHIO}_{\mathcal{R}^+}(D)^-$, and is used in this paper to support the development of the DL application to animal systematics. We have used the native syntax of RACER for DL expressions in this paper.

We find that most of the language used in a scientific description is readily adapted to the formal description logic language, with the exception of spatio-temporal elements and some higher-order constructs. We show that the reasoning capability is sufficient for consistency checking and retrieval of taxonomic knowledge. We discuss some benefits of the representation to assist the work of biological systematists.

2. Results and Experience

2.1. Modelling

Selecting the primitive, or fundamental concepts of the domain is a key modelling step in ontology design. We encountered generic primitives, such as taxa, types, components, features and publication information such as date and author. Other primitives are more specific to the description of Epopostruma, such as *anterior*, *posterior* and *dorsal* as concepts and *lighter-than* (referring to colour shade) as a role primitive. The following quote demonstrates the use of such terms, followed by its expression in DL:

> Anterior face of postpetiole continuous with dorsal face.

```
(implies |anterior| |component|)
(implies |anterior face| |face|)
(implies |anterior face| |anterior|)
...(and |anterior face|
      (some |continuous with|
         (and |dorsal face|
            (some (inv |has component|)
                  (and |postpetiole|
                       |alata component|)))))
```

Some expressions in the text required expression of background information not directly evident. For example, we found it necessary to define a *component* concept for each taxon, such as *alata component* above. This enables us to describe relationships between parts and to ensure that the description refers to parts belonging to the same taxon. We also found cases where direct appeal to the services of the DL is preferable to the creation of primitives. For example, *distinct* can be modelled via concept disjointness.

2.2. Inexpressible Concepts

We recognise that spatio-temporal concepts are often important in taxonomy and cannot be adequately represented

Proceedings of the 16th International Conference on Scientific and Statistical Database Management (SSDBM'04)

IEEE COMPUTER SOCIETY

in $\mathcal{ALCQHIO}_{\mathcal{R}^+}(D)^-$. We also found some morphological descriptions that could not be directly represented, including higher order statements such as

> In dorsal view the postpetiole are expanded laterally, the extensions generally solid and wing-like, or in one case, limited to broadly rounded posterior extensions.

2.3. Queries

Having translated the formal morphological description to DL, we can use the RACER reasoner to answer queries. We found that response times ranged up to 85 seconds for some instance queries. The following two examples, each followed by its answer, demonstrate what can be done. The first uses $\mathcal{ALCQHIO}_{\mathcal{R}^+}(D)^-$ support for concrete domains.

Type specimens What Myrmicinae have type instances with the given range of dimensions?

```
(concept−instances (and Myrmicinae
 (>= TL 4.7) (<= TL 5.6) (>= HL 1.14)
 (<= HL 1.29) (>= HW 1.21) (<= HW 1.43)
 (>= CI 103) (<= CI 111) (>= ML 0.57)
 (<= ML 0.67) (>= MI 47) (<= MI 52)
 (>= SL 0.66) (<= SL 0.74) (>= SI 52)
 (<= SI 56) (>= PW 0.64) (<= PW 0.81)
 (>= AL 1.14) (<= AL 1.36)))
```

```
(|(URL)Epopostruma lattini (holotype worker)|
 |(URL)Epopostruma frosti (holotype worker)|)
```

Identification What Alata have a postpetiole which in the dorsal view demonstrates an extension which is not solid or not wing-like?

```
(concept−children
 (and Alata
      (some |has component|
 (and |postpetiole|
      (some |in view|
            (and |dorsal view|
                 (some |demonstrates extension|
   (not (and |solid| |wing−like|)))))))))
```

```
((|angela|))
```

3. Application to Animal Systematics

First, a description logic representation and the associated reasoning capability can assist the systematist to compose and organise the character descriptions of taxa. A reasoner can assist the systematist to ensure that descriptions are logically consistent; that a description adequately distinguishes a taxon and correctly applies to sub-taxa; and that descriptions of taxa are consistent with specimen descriptions. This may be especially useful in the compilation of major collaborative reference works which cover many different areas of taxonomical expertise.

In concert with other tools for English translation and pictorial representation, description logic could be a powerful training tool for systematics. Description logic can be used to explain and relate the language terms that are used in a taxonomic discipline, and to promote the disciplined modelling required.

The formal representation can support the automatic generation of user interfaces to biological software systems. For example, automatically generated query interfaces that are specialised to taxa, ecological domains, or morphological characteristics can be used to search heterogenous collections of biological databases.

Finally, description logic enables the long term, extensible, machine-interpretable representation of some biological knowledge. The formal description of morphological characters in DLs can assist in the long term management of nomenclature, capturing the reasoning behind naming decisions and permitting automatic reclassification of specimens when nomenclature is changed. It can also be used for species identification in circumstances when traditional taxonomic keys cannot be used.

Our case study shows that it is possible to represent substantial scientific knowledge in the language of the Semantic Web. Work is needed to develop tools that incorporate description logic capability within an environment that is attuned to the work practices of systematists. We believe that adoption of the technology by the systematics community could give the science all the advantages of machine-understandable semantics attributed to the Semantic Web.

Acknowledgement The authors are grateful for scientific advice provided by Steve Shattuck and John La Salle of CSIRO, and Ian Naumann of Agriculture, Fisheries and Forestry Australia.

References

[1] S. Bechofer, F. van Harmelen, J. Hendler, I. Horrocks, D. L. McGuinness, P. F. Patel-Schneider, and L. A. Stein. OWL web ontology language reference. W3C Proposed Recommendation 15 December 2003, December 2003.

[2] V. Haarslev and R. Moller. Description of the RACER system and its applications. In *Proceedings of the International Workshop on Description Logics (DL-2001)*, pages 132–141, Stanford, USA, August 2001.

[3] S. O. Shattuck. *The epopostrumiform genus group, in B. Bolton, The ant tribe Dacetini*, volume 65 of *Memoirs of the American Entomological Institute*, pages 30–67. 2000.