Notions of Indistinguishability for Semantic Web Languages*

Jaap Kamps and Maarten Marx

Language and Inference Technology Group, ILLC, Universiteit van Amsterdam
{kamps,marx}@science.uva.nl

Abstract. The paper reviews the notions of expressiveness of description logics from (N. Kurtonina and M. de Rijke. Expressiveness of concept expressions in first-order description logics. *Artificial Intelligence*, 107:303–333, 1999) and exemplifies their use in the development in Semantic Web languages. The notion of bisimulation—which characterizes the description logic $\mathcal{ALC}$—provides a direct link to what's in the field of sociology called social network analysis. The perspective on data in this field—data are represented as labeled graphs—fits exactly the modeling intuitions of web languages like OIL and DAML+OIL. This is exemplified in the study of trophic networks. A further connection is established between web languages and hybrid logic, and an extension of OIL with a limited form of self reference is proposed.

1 Introduction

This paper describes foundational work which we hope benefits the further development of Semantic Web languages. The design of these languages is difficult because of the numerous imposed constraints and desires. In several cases, these pull in opposite directions. For instance, the desire to have great expressive power goes against the constraint of having reasonable inference support.

At present it looks like the eventual web language will be strongly based on description logic (as the languages[1] OIL and DAML+OIL are). Description logic provides a logical basis to the well known traditions of frame-based systems, semantic networks and KL-ONE like languages, semantic data models and type systems. Complexity issues for subsumption and consistency problems have been studied extensively (cf. the review article [7]). Relatively few papers study the expressiveness of description languages [3,4,6,12].

Here we review the results from [12] from a modeling perspective. These results can be summarized as follows. Given a domain of individuals and a set of relations and atomic concepts, a structural notion of indistinguishability between individuals is defined for a large number of languages[2] within the description logic family. This notion is such that on finite domains two individuals are

* This research was supported by the Netherlands Organization for Scientific Research (NWO, grants # 4000-20-036 and #612-000-106.)

[1] We refer to the version of OIL described in [8] and the DAML+OIL specification from www.daml.org/2001/03/daml+oil-index.

[2] To be precise, for all languages in the lattice between $\mathcal{FL}^-$ and $\mathcal{ALCNR}$.

I. Horrocks and J. Hendler (Eds.): ISWC 2002, LNCS 2342, pp. 30–38, 2002.

structurally indistinguishable if and only if no concept defined in the corresponding description language can separate them. The structural notion thus provides a semantic definition of the maximum granularity of the concepts which can be defined in a certain description logic.

We exemplify the use of this semantic-syntactic interface from two different directions. First we link the languages proposed for the Semantic Web to the field within sociology called social network analysis. With the help of the semantic-syntactic interface we are able to discover some surprisingly strong connections and similarities. Then we find a simple extension of the description logic $\mathcal{ALC}$, known as "hybrid logic" which arguably has maximum first order expressive power for Semantic Web languages.

2 Semantic Web and Social Network Analysis

In this section, we link the languages proposed for the Semantic Web to the field within sociology called social network analysis.

In social network analysis, real world data are modeled as a (labeled) graph, called a *network*. The range of applications of this modeling technique is virtually unlimited. The nodes could be published papers with vertices from paper A to paper B if A cites B [13]. Or the nodes could be web pages and the vertices denoting links between pages [11]. Another example is WordNet, in which the nodes are synsets and the vertices denote overlap [10]. In more traditional sociological or anthropological examples, the nodes are often individuals (humans, animals, animal species, organizations) and the vertices —called *ties*—indicate certain interactions (parent-of, is-friend-of, eats, is-competitor-of, etc.), cf., [17], the *Social Networks* journal, the Proceedings of the Sunbelt conferences, or the page `http://www.heinz.cmu.edu/project/INSNA/`.

The key idea behind this way of modeling data—and now we come to the link with web-languages—is that structure in the data can be discovered by inspecting the structure of the network. The assumption is that a *position* in the network is structurally determined: that is, only by its links to other elements in the network. A position in a network is most naturally thought of as a subset of its nodes. For instance, in an organization chart (organigram) of an organization (where the nodes are individuals and the vertices denote the hierarchical structure) typical positions are the CEO, the managerial level, the support staff, the technical core and the workforce. In such a chart, two individuals occupy the same (e.g., managerial) position, not because they have ties to and from the same individuals, but because they have ties to and from individuals *in the same position.* Formally,

(∗) two individuals occupy the same position in a network if they have similar ties to and from individuals in the same position.[3]

The view of a position as a subset of individuals in a network of relations is exactly the same as the semantic meaning of a concept in the web languages OIL

[1] In social network jargon, this means that they are regular equivalent.

and DAML+OIL. As the logicians and computer scientists might have remarked, (*) is nothing but the definition of bisimilarity (disregarding atomic properties in (*)):

(**) For $(N, R_1, \ldots, R_k)$ a labeled graph, we say that nodes $a, b \in N$ are bisimilar (notation: aBb) if
(1). aR_ic implies the existence of a $c' \in N$ such that bR_ic' and cBc';
(2). cR_ia implies the existence of a $c' \in N$ such that $c'R_ib$ and cBc'.
If the graph also contains a set of unary properties $P_1, \ldots, P_m$ it is also required that
(0). P_ia holds if and only if P_ib holds.

Now logic, in particular the work of Kurtonina and de Rijke [12], comes in to create the strong connection with the semantic web languages. They have extended the work of Hennesy–Milner and van Benthem on the connection between bisimilarity and modal logic to the hierarchy of description logics between $\mathcal{FL}^-$ and $\mathcal{ALCNR}$. These languages are the logical basis behind OIL and DAML+OIL and several weaker frame–based languages. The strength of these results lies in the fact that they relate the purely semantic notion of "the sameness" or "indistinguishability" to the purely syntactic notion of being definable in a certain language. The surprising and remarkable thing now is that the core language[4] behind OIL, the description logic $\mathcal{ALCI}$ ($\mathcal{ALC}$ with inverse roles), is *exactly* the right language to describe positions—as defined semantically in (*) and (**)—in a network. This strong claim is based on the following facts:[5]

(1) If two elements occupy the same position in a network, they cannot be distinguished by an $\mathcal{ALCI}$ concept.

(2) In finite networks, two elements which occupy different positions in a network, can be distinguished by an $\mathcal{ALCI}$ concept.

Moreover the language $\mathcal{ALCI}$ is, at least for first order definable concepts, complete:

(3) Every position which is first order definable is definable by an $\mathcal{ALCI}$ concept.

[4] The description logic $\mathcal{ALCI}$ corresponds to the following OIL fragment:

- all OIL **class-expressions** are permitted except those which have **slot-constraints** with cardinality restrictions.
- all components of OIL **slot-def**'s are permitted except **subslot-of** and **properties** (this last component is used to specify transitivity or symmetry of a slot).

[5] These facts are just the translation to the terminology of the present paper of the well known characterization theorem of modal logic, cf., [12].

We view this as strong support for the claim that web languages like OIL and DAML+OIL are well designed. We find this support especially promising because the range of applications of the two fields shows such a clear and vast overlap.

Indistinguishability notions like bisimulation provide an upper bound on the grain-size of the definable concepts: bisimilar individuals are not distinguished. Results of the form (1) and (3) are then very useful: (1) says that individuals which are indistinguishable with a certain grain-size cannot but be classified in the same way if a certain language is used. This is a safety criterion: you cannot differentiate in the language what should be considered the same. (3) states the reassuring fact that all concepts with a certain grain-size can be defined in a certain language.

A recent study in the field of ecological network analysis [9] uses the notion of a position as defined in (*) to derive a foodweb from a data set.[6] A foodweb or trophic network describes the energy flow between species (in particular who eats who). Of interest for the Semantic Web community is the data-mining perspective. In [9] a foodweb is constructed from a set of noisy data using existing software.[7] A semantic network containing four distinct classes is found, here reproduced in Figure 1.

Each class consists of a (often huge) number of species. The arrows indicate who is eaten by who. Now obviously this ontology can be described in a web language. The description of it in OIL is given in the same Figure. To get an impression of the contents of the concepts, the top predators contain e.g., screech owls, boa snakes and parasitic insects, the intermediate consumers contain specialist herbivores and detritivores such as decomposers and various insects, the basals contain primarily generalist omnivores such as insects, spiders and birds, and the primary producers contain plants, algae, nectar, dead wood and detritus.

3 Delimiting the Design Space, a Case for Hybrid Logic

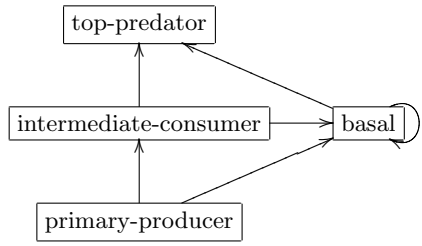
Within the literature of social network analysis one can also find (semantic) definitions of positions which differ from the one in (*). (Again these can be tightly linked to concept-definition languages, using the technique of [12].) They all agree on the following principle though:

(#) a position is determined by the properties of its elements and their ties to other elements in the network. In particular, elements in the network which cannot be reached by a path of ties (in forward or backward direction) are irrelevant.

For instance, to describe positions in an organization *only* the organizational members occurring in the organizational chart are relevant. We note that this

[6] The data consisted of 156 compartments, each consisting of various levels of species aggregations (compartments could have from 1 up to 429 different species). The relations between the compartments were obtained by direct observation and from the literature.

[7] The REGE algorithm, incorporated in the package UCINET V, has been used [15].



slot-def *eats*
inverse *is-eaten-by*

class-def defined primary-producer
subclass-of species
slot-constraint *eats*
value-type ⊥
slot-constraint*is-eaten-by*
value-type intermediate-consumer **OR** basal

class-def defined intermediate-consumer
subclass-of species
slot-constraint *eats*
value-type primary-producer
slot-constraint*is-eaten-by*
value-type top-predator **OR** basal

class-def defined basal
subclass-of species
slot-constraint *eats*
value-type primary-producer **OR** basal **OR** intermediate-consumer
slot-constraint*is-eaten-by*
value-type top-predator **OR** basal

class-def defined top-predator
subclass-of species
slot-constraint *eats*
value-type basal **OR** intermediate-consumer
slot-constraint*is-eaten-by*
value-type ⊥

disjoint-with top-predator, basal, intermediate-consumer, primary-producer

Fig. 1. A foodweb from [9] and its description in OIL.

principle is also behind description logic[8] and OIL. Principle (#) implies that first order logic is too expressive as a position definition language. For, consider the two models (or networks) below. Here elements are indicated by points and the relation (named R) by the arrow. Element x is related to y (notation: Rxy) if there is an arrow from x to y. According to principle (#), the element a should occupy the same position in both networks. but the formula $\exists y(\neg Rxy \wedge x \neq y)$ distinguishes them.

There exists a sub language of first order logic which exactly captures this principle and which is very close to the description logic $\mathcal{ALC}$. It is called *hybrid logic*[9]. It extends $\mathcal{ALC}$ with a mechanism for naming and referring to individuals as follows: a new set of primitive concepts, called nominals[10], are introduced. These nominals can be bound by a binder $\downarrow$. So if C is a concept and w a nominal, then also $\downarrow w.C$ is a concept. The meaning of $\downarrow w.C$ consists of all elements d which form the interpretation of C under the assumption that all occurrences of w in C denote the set $\{d\}$. For instance, in a domain of web pages, the concept $\downarrow w.\exists$ `has_link` w denotes all pages with a link to themselves; the concept $\downarrow w.\forall$ `has_link` $\exists$ `has_link` w denotes all pages d which only link to pages which have a link back to d.

The $\downarrow$ binder provides self-reference not available in OIL and DAML+OIL. This feature can be useful when the graph like nature of the network is important; e.g., in the network of papers with citation ties from [13] it is important to separate the self-citations (a citation to a paper with the same author). We give further examples in the next section.

The principle that non-reachable elements should not contribute to the meaning of concepts should in our opinion also be behind semantic web languages. We note again that this principle is already endorsed by description logic and OIL and DAML+OIL. A result from [2] then sets a frontier to these languages:

(##) each first order semantic web language should be a fragment of hybrid logic.

This bold claim is based on a semantic characterization of hybrid logic similar to the ones described in the previous section ([2], Theorem 3.11). It says that a concept whose meaning is not affected by non–reachable elements is first order definable if and only if it is definable in the hybrid language.

[8] In technical terms: DL is preserved under generated submodels. This means that if in a DL model an individual d belongs to some DL concept C, it still belongs to C if all individuals which are not reachable by a path of (forward and backward) slot relations from d are removed from the model.

[9] Cf. the hybrid logic page: `www.hylo.net`.

[10] Nominals are closely related to the **ONE-OF** constructor: the interpretation in a model of **ONE-OF**$\{d\}$, for d a name of an element in the model, is the singleton set $\{d\}$. The interpretation of a nominal is always a singleton set.

The close connection between hybrid logic and description logic is described in [1]. The formal properties of hybrid logic are well investigated, cf. for instance [2]. The full language is undecidable but [14] contains a useful decidable fragment, called $\mathcal{ALCI}\mathsf{self}^{\exists}$, which extends $\mathcal{ALCI}$ with a form of self-reference. The next section exemplifies this.

4 A Self Referential Web Language

In this section we discuss an extension of OIL which allows for self reference in concept definitions. This extension is based on the hybrid logic discussed in the previous section, but presented here in a limited decidable format. Instead of using variables, we decided to use the pronouns "I" and "me". This example is an indication that the discussed semantic constraints are very useful in guiding the search for and design of future web languages.

The example is discussed in the pseudo-XML syntax of OIL. The following constructions are added to the language:

- ME is a predefined class name;
- within each class-definition the component

  I.**slot-constraint** *relation*

  followed by any of the OIL fields **has-value, value-type** or any of the cardinality restrictions, may occur, for any slotname *relation*.

As an example consider the class of narcissist web pages: web pages which have a link to themselves:

class-def defined narcissist-webpage
  subclass-of webpage
  I. **slot-constraint** *has-link*
    has-value ME

The semantics of I and ME is exactly the same as that of $\downarrow x$ and x, respectively. So an element d is in the interpretation of slot constraint I.ϕ, if d is in the interpretation of ϕ assuming that every occurrence of ME in ϕ denotes $\{d\}$.

The second example comes from a paper describing the annotation of photographs using semantic web languages [16]. One of the concepts defined there is a "monkey scratching his head". This concept can be defined in the extension of OIL as

class-def defined head-scratching-monkey
  subclass-of monkey
  I. **slot-constraint** *scratch*
    has-value head **AND**
      slot-constraint *part-of*
        has-value ME

Such definitions are not possible[11] in OIL or DAML+OIL without the use of I and ME. One of the examples in [16] describes a user who wants to find a picture of a monkey doing something with its head. In OIL this query can be represented as

subclass-of monkey
I. **slot-constraint** *action*
has-value head **AND**
slot-constraint *part-of*
has-value ME

With the slot-definition specifying that scratching is an action, this query subsumes the class head-scratching-monkey, which will cause that photographs thus annotated are given as an answer. Without the I, me apparatus, one can only specify that a monkey is scratching some head. The query then cannot be represented in the specific way as it is stated, leading to possibly wrong answers (pictures of monkeys scratching the head of their spouse, for instance).

In [14] this expansion is discussed in more detail, and a tractable version of the language is presented.

5 Wrap Up

We have emphasized the importance of semantic characterizations of Web languages. The characterization of $\mathcal{ALC}$ in terms of bisimulation showed a surprising connection with the field of social network analysis. Web research can learn a lot from this field because its datastructures—networks—are everywhere in Web applications. As an example, Google's successful Pagerank measure goes back to centrality measures in [5].

The second contribution of the paper consists of the connection between hybrid logic and web languages. There are two good reasons to consider hybrid logic as an upper expressivity bound for web languages and as a guide in the design process. Firstly, its extremely simple syntactic structure which is a very intuitive extension of the description logic $\mathcal{ALC}$. Secondly, its semantic characterization as the fragment of first order logic whose truth is unaffected by unreachable elements, a natural semantic invariance for web languages. We illustrated how easy hybrid ideas combine with the web language OIL in an example about photo annotation.

[11] Of course a concept own-head can be defined in OIL, which is subsumed by head. But not all of the meaning of own-head is captured in this way. Moreover, all concepts which can be used in self referential expressions then need to be duplicated, and logical relations which could be inferred in the I–me set up have to be explicitly stated as well (e.g., that own-mouth is *part-of* own-head).

References

1. C. Areces. *Logic Engineering*. ILLC–DS–00–8, Institute for Logic, Language and Computation, University of Amsterdam, 2000.
2. C. Areces, P. Blackburn, and M. Marx. Hybrid logics: Characterization, interpolation and complexity. *Journal of Symbolic Logic*, 66(3):977–1010, 2001.
3. F. Baader. A formal definition for the expressive power of terminological knowledge representation languages. *Journal of Logic and Computation*, 7:33–54, 1997.
4. A. Borgida. On the relative expressiveness of description logics and predicate logics. *Artificial Intelligence*, 82:353–367, 1996.
5. R. Burt. *Toward a Structural Theory of Action: Network Models of Social Structure, Perception and Action*. Academic Press, 1982.
6. M. Cadoli, L. Palopoli, and M. Lenzerini. Datalog and description logics: Expressive power. In *Proc. International workshop on Database Programming Languages*, number 1369 in LNCS. Springer, Berlin, 1998.
7. D. Calvanese, G. De Giacomo, D. Nardi, and M. Lenzerini. Reasoning in expressive description logics. In A. Robinson and A. Voronkov, editors, *Handbook of Automated Reasoning*. Elsevier Science Publishers, 1999.
8. S. Bechhofer et. al. An informal description of standard OIL and instance OIL (OIL whitepaper). Available at the OIL page `www.ontoknowledge.org/oil/`.
9. J. Johnson, S. Borgatti, J. Luczkovich, and M. Everett. Network role analysis in the study of food webs: An application of regular role coloration. *Journal of Social Structure*, 2(3), 2001.
10. J. Kamps and M. Marx. Words with attitude. In *Proceedings of the 1st International Conference on Global WordNet*, pages 332–341. CIIL, Mysore India, 2002.
11. J. Kleinberg and S. Lawrence. The structure of the web. *Science*, 294:1849–1850, 2001.
12. N. Kurtonina and M. de Rijke. Expresiveness of concept expressions in first-order description logics. *Artificial Intelligence*, 107:303–333, 1999.
13. S. Lawrence. Online or invisible. *Nature*, 411(6837):521, 2001.
14. M. Marx. Narcissists, stepmothers and spies. Proceedings of the 2002 International Workshop on Description Logic workshop, Toulouse, 2002.
15. L. Freeman S. Borgatti, M. Everett. *UCINET V. Software for Social Network Analysis*. Natick: Analytic Technologies, 1999.
16. A. Th. Schreiber, B. Dubbeldam, J. Wielemaker, and B. J. Wielinga. Ontology-based photo annotation. *IEEE Intelligent Systems*, pages 2–10, May/June 2001.
17. S. Wassermann and K. Faust. *Social Network Analysis*. Cambridge University Press, Cambridge, UK, 1994.