Complex Alignments Between Ontology Universes

Alex Borgida
Dept. of Computer Science
Rutgers University
New Brunswick, USA
borgida@cs.rutgers.edu

The theme of this workshop is the synthesis of database and AI views of semantic integration.

We started working on issues of integrating Description Logic ontologies a few years ago, by examining to extent to which such formalisms are used as advertised: to *define* in a precise and formal way the meaning of every concept used in the terminology, rather than encoding meaning in the names of the terms [2]. Interestingly, that preliminary work was instigated by the definition of "conflict free schema integration" introduced in databases by Biskup and Convent [1].

However, let us start here by contrasting the treatment of *individuals* in approaches to semantic integration in the two fields.

In AI research on ontologies, it is almost always assumed (whether explicitly or not) that the areas of overlap between two ontologies being integrated concern the same individuals. In other words, if two concepts, C_1 from ontology O_1 and C_2 from ontology O_2, are to be related, then this is viewed as being based on a direct set-theoretic relationship — subset, equality, disjointness, overlap — between the sets of individuals denoted by C_1 and C_2. This still permits quite complex mappings, by allowing C_1 and C_2 to be complex concepts defined in terms of the terminologies O_1 and O_2 (e.g., [5]), but it does not make it possible to capture *systematic* relationships between individuals, such as the fact that the objects in one ontology (e.g., households in a census) correspond to sets of objects (e.g., persons living at that address) in another one.

In contrast, researchers in database integration, have recognized for a long time the need for complex translations between values in the databases being integrated: the early work of Kent [6] is replete with a variety of such examples, including the need to convert currencies, and convert different notions of income (before and after tax, net vs. gross, etc.). Kent's solution relies on functions expressed in a what is close to a general purpose programming language, equipped with loops and conditionals. He also provides examples where the relationship is not functional, such as the case when letter grades would need to be mapped to numeric values. This focus on complex mappings between individuals, evident in other work, such as the Clio project [7] for example, may also be due, in part, to the nature of the relational data model, which "dematerializes" individuals into tuples of values (integers and strings), and by the availability of powerful query languages to reconstruct values in the new, integrated database.

A natural question is to what extent complex correspondences between individuals are of interest to ontology integration and reasoning, or only in translating *database facts* (e.g., "Joan's salary_1 is 3000 dollars per week" to "Joan's salary_2 is 5000 Euros bi-weekly").

Consider the following example: One information source, IS_1, concerns literary works — novels, plays, poems, articles, authors, etc.; another information source, IS_2, is an entertainment guide for Southern Ontario, and maintains information about current and forthcoming events, such as sports events, performances of music, plays, etc. In integrating these two sources, we will want to match a literary work to its performances. Note that this correspondence is certainly not the identity function, and is not even a function: prefaces to plays and theater reviews are literary works that do not receive performances, and some plays are not being performed, while others are being performed on multiple nights (or receive multiple stagings). Suppose that as part of the process of semantic integration, we can be told information about this correspondence. For example, in this region plays are always performed in theaters, and all events occurring in Niagara on the Lake are performances of works by G.B. Shaw[1]. From this,

[1]Truth in advertising: Although this town does host a

one should be able to deduce that all events in Niagara on the Lake are theatrical performances.

Since Description Logics appear to be favored both as ontology representation languages, and as semantic representations for database schemas (they are more expressive than Entity Relationship diagrams), we have extended in [3, 4] the framework of Description Logics to allow such general binary relationships between individuals in the local domains of the information sources being integrated. (In fact, because the mapping is directional, we prefer to think of the resulting system as a federation of independent agents that import information to conduct local reasoning, rather than a single integrated entity.)

A central question in studying the resulting so-called "distributed description logics" is the language for expressing the properties of the correspondence R bbetween local domains. As usual, the choice of language affects the nature and complexity of reasoning in the resulting formalism.

It is obvious that allowing the mapping R to be represented by an arbitrary computable function prevents any kind of meaningful automatic reasoning in the resulting system. The papers mentioned above concentrate on simple restrictions of the form $\mathrm{R}(A) \subseteq D$ and $E \subseteq \mathrm{R}(B)$. But it is possible to view R as a *Description Logic role* (e.g., an OWL property [8]), in which case one can consider using the DL formalism to constrain it! In fact, a theorem proven in [4] shows that for a large class of description logic families this can be done using axioms involving property restrictions on R and its inverse, and then performing standard DL reasoning in a merged theory. For example, if we want to say that the mapping R is a bijection between the individuals in concepts A and D, we can assert

$A \sqsubseteq = 1\,\mathrm{R}$
$D \sqsubseteq = 1\,\mathrm{R}^-$
$A \sqsubseteq \forall \mathrm{R}.D$
$D \sqsubseteq \forall \mathrm{R}^-.A$

There are numerous open research problems concerning the extended formalism for expressing ontologies and mappings between them. These include

- problems introduced by the presence of datatypes in OWL
- the treatment of constraints on mappings, such as "$\mathrm{R}(A)$ and D overlap", which cannot be represented directly as subsumption axioms.
- the problem of expressing correspondences between complex objects in two ontologies, including the case when more than one element in one domain determines an individual in the other

Shaw festival, there are also other performances in town!

REFERENCES

[1] J. Biskup and B. Convent. "A Formal View Integration Method", *Proc. SIGMOD'86*

[2] A. Borgida and R. Ksters "What's not in a name: Some Properties of a Purely Structural Approach to Integrating Large DL Knowledge Bases", *Proc. Int. Workshop on Description Logics (DL'00)*, 2000.

[3] A. Borgida and L. Serafini, " Distributed Description Logics: Directed Domain Correspondences in Federated Information Sources", *Proc. CoopIS'02*, pp.36-53 (2002).

[4] A. Borgida and L. Serafini, "Distributed Description Logics: Assimilating Information from Peer Sources", *J. of Data Semantics*, to appear.

[5] T. Catarci and M. Lenzerini: "Representing and Using Interschema Knowledge in Cooperative Information Systems." *International Journal of Intelligent and Cooperative Information Systems 2(4)*, pp.375–398 (1993).

[6] W. Kent. "Solving Domain Mismatch and Schema Mismatch Problems with an Object-Oriented Database Programming Language." *Proc. VLDB'91*, Barcelona, Spain, pp. 147-160 (1991).

[7] R.J. Miller, L.M. Haas and M. Hernndez. "Schema Mapping as Query Discovery". *Proc VLDB'00*, Cairo, Egypt, pp.77-88 (2000).

[8] Peter F. Patel-Schneider, Patrick Hayes, Ian Horrocks (eds), "Web Ontology Language (OWL) Abstract Syntax and Semantics", `http://www.w3.org/TR/owl-semantics/` , August 2003