Contexts and ontologies in schema matching

Paolo Bouquet*

Department of Information and Communication Technology, University of Trento
Via Sommarive, 14 – 38050 Trento (Italy)

Abstract. In this paper, we propose a general model of schema matching based on the following ideas: on the one hand, a schema is viewed as a *context* (namely as a partial and approximate representation of the world from an agent's perspective); on the other hand, a schema cannot be assigned any arbitrary interpretation, as the meaning of the expressions used to label nodes (and possibly arcs) may be constrained by shared social conventions or agreements expressed in some *lexical* or *domain ontologies*. Accordingly, the proposed schema matching method can be viewed as an attempt of coordinating intrinsically context-dependent representations by exploiting socially negotiated constraints on the acceptable intepretations of the labels as codified in shared artifacts like lexicons or ontologies.

1 Introduction

In the literature, we find many different approaches to the problem of schema matching, and each of them reflects a theoretical view on what a schema is (a graph, a linguistic structure, a data model, ...). In this paper, we propose a general model based on the following ideas: on the one hand, a schema can be viewed as a context in the sense defined in [**?**] (a partial and approximate representation of the world from an agent's perspective); on the other hand, a schema cannot be assigned any arbitrary interpretation, as the meaning of the expressions used to label nodes (and possibly arcs) may be constrained by shared social conventions or agreements expressed in some lexical or domain ontologies. Accordingly, a schema matching method can be viewed as an attempt of coordinating intrinsically context-dependent representations by exploiting socially negotiated constraints on the acceptable intepretations of the labels as codified in shared artifacts like lexicons or ontologies.

Our claim is that this type of approach may also provide a general view on the relation between contexts and ontologies. The idea is the following: contexts are representations which encode an agent's point of view; to be shared or communicated, these representations need to be linguistically expressed; however, this linguistic representation cannot be arbitrary, otherwise agents would never succeed in cokmmunication; lexical and domain ontologies are the reification

* Part of the material for this paper was developed in collaboration with Stefano Zanobini as part of his PhD work.

of partial and evolving agreements achieved in a linguistic or in other types of communities on the use of terms used for communication (perhaps in limited domains).

The paper is structured as follows. First, we present the intuitions underlying our approach. Then we show how these intuitions are captured in a formal model. Finally we argue that this model can also be used to explan what other approaches to schema matching do.

2 The building blocks

Let us start with a few general definitions.

Definition 1 (Schema). *Let L_{ext} be a set of labels. A* schema $\mathcal{S}$ *is a 4-tuple $\langle N, E, \text{lab}_N, \text{lab}_E \rangle$, where $\langle N, E \rangle$ is a graph, $\text{lab}_N : N \rightarrow L_{ext}$ is a function that associates each node with a label in L_{ext}, and $\text{lab}_E : E \rightarrow \mathcal{L}_{NL}$ is a function that associates each edge to a label in $L_{ext} \cup \emptyset$.*

In this definition, L_{ext} is the language used to externalize a schema outside an agent's mind (e.g. for publishing the schema and sharing it with other agents). In many real situations, it may be a subset of some natural language. For example, in most web directories, the communication language is basically the portion of English which is needed to label the directory categories, and the sign `>` to denote the sub-category relation (e.g. `Music > Baroque > Europe` is a subcategory of `Music > Baroque`)

Now we need to capture the intuition that an agent a may associate a set of objects to a schema element e based on her understanding of the meaning of e, and that different agents may have a different understanding of e. Let a_1 and a_2 be two agents and L_{ext} a suitable communication language. We now introduce the notion of a representation language for an agent, namely the internal (mental?) language which represents agents know about their environment. Let L^i be the representation language of the agent a_i (where $i = 1, 2$), W a set of worlds, and C a set of contexts of use. Intuitively, W is the set of all possible interpretations of L^j, and C represents a collection of distinct contexts of use of expressions belonging to L_{ext} (C is necessary to model the fact that many communication languages, including natural languages, are polysemous, namely the same word may have a different meaning in different contexts of use). We do not make any special assumption on L_{ext} and L^j; the only important requirement is that they are distinct – and possibly different – languages. For the sake of this paper, we will assume that the representation languages are some sort of Description Logic (DL) language (see [1] for an introduction to DL languages).

The connection between schema elements and data happens in two steps: in the first, we take the specification of a schema element e in L_{ext} and build the representation of its meaning in L^j (given a context of use c); in the second, we provide an interpretation function from the resulting expression of L^j into a set of objects in the domain of W.

The first step is formalized by the following *translation function.*

Definition 2 (Translation function). *Let L^j be a representation language, L_{ext} a communication language and c a context in C. The* translation function *$\mathcal{T}_c^j : L_{ext} \rightarrow L^j$ is the function which associates an expression of L^j to an expression of L_{ext} when used in c.*

We notice that the translation function is indexed with an agent's name, as it reflects the way such an agent assign a (subjective) meaning to the (public) expressions of a communication language. In the following, we will use the notation $\mathcal{T}^j$ to denote the family of functions $\{\mathcal{T}_c^j \mid c \in C\}$.

The second step is formalized by the following *projection function.*

Definition 3 (Projection function). *Let L^j be a representation language and $w \in W$. The* projection function *$\mathcal{P}_w^j : L^j \rightarrow 2^w$ is a function which, for any possible world w, associates an extension to any term of L^j.*

In the following, we will use the notation $\mathcal{P}^j$ to denote the family of functions $\{\mathcal{P}_w^j \mid w \in W\}$.

To model the fact that agents may have domain knowledge about the concepts which they associates to schema elements, we introduce the notion of an agent's ontology, expressed as a set of axioms in the agent's representation language: $\mathcal{O}^j = \{t_i \sqsubseteq t_k \mid t_i, t_k \in L^j\}$[1].

Finally, an *agent* a_j which uses L_{ext} as a communication language is defined as follows:

Definition 4 (Agent). *An* agent *$a_j^{L_{ext}} = \langle \mathcal{T}^j, \mathcal{P}^j, L^j, \mathcal{O}^j \rangle$ is a quadruple, where $\mathcal{T}^j$ is a family of translation functions, $\mathcal{P}^j$ is a family of projection functions, L^j is the agent's representation language, $\mathcal{O}^j$ is the agent's knowledge, and the following holds:*

$$\forall w \in W \;\; t_i \sqsubseteq t_k \in \mathcal{O}^j \Rightarrow \mathcal{P}_w^j(t_i) \subseteq \mathcal{P}_w^j(t_k)$$

Now we have all the necessary building blocks for defining the formal object of schema matching in this model.

3 Semantic Coordination

The main idea of our model is that in no real world sitation one can guarantee that two agents share meanings just because they share a communication language. Indeed, the notion of shared meaning is not available (meaning is always mediated through concepts, and therefore partially private). Therefore, to model schema matching, we need to introduce a notion of agreement which does not

[1] For the sake of simplicity, we assume that the agent knowledge can be represented as a set of entailment axioms between concepts. Notice that this formalization of the knowledge basis is not a novelty, but it is the standard one used in Description Logics [1]. Following this approach, for expressing the *Is-A* relation between the concepts, we use the DL symbol '$\sqsubseteq$'.

presuppose that shared meanings are available and which we call *semantic coordination*: two agents are semantically coordinated on the use of two expressions t_1 and t_2 of a communication language L_{ext}, with respect to a relation R and in some context of use c, when the interpretation they give to the two linguistic expressions is *compatible*[2]:

Definition 5 (Semantic Coordination). *Let* $a_1^{L_{ext}} = \langle \mathcal{T}^1, \mathcal{P}^1, L^1, \mathcal{O}^1 \rangle$ *and* $a_2^{L_{ext}} = \langle \mathcal{T}^2, \mathcal{P}^2, L^2, \mathcal{O}^2 \rangle$ *be two agents,* t_1 *and* t_2 *two expressions of a communication language* L_{ext}*, and* R *be any set–theoretical relation. We say that* a_1 *and* a_2 *are* semantically coordinated *on* t_1 *and* t_2*, with respect to* R *in some context of use* c*, if the following holds:*

$$\forall w \in W \quad \mathcal{P}_w^1(\mathcal{T}_c^1(t_1)) \; R \; \mathcal{P}_w^2(\mathcal{T}_c^2(t_2))$$

Imagine, for example, that R is an equivalence relation and $t_1 = t_2 =$ 'cat'; then the definition above says that two agents are coordinated with respect to the use of the word 'cat' (in English) in a context of use c if, for any world $w \in W$, they associate to it the set of objects belonging to the domain of w (via translation and projection).

In what follows, we will use the notation $coord(a_1, a_2, t_1, t_2, R, c)$ to denote that the agents a_1 and a_2 are semantically coordinated on t_1, t_2 with respect to R in the context c.

We now introduce a syntactic notion of mapping across schemas:

Definition 6 (Mapping). *Let* $\mathcal{S}_1 = \langle N_1, E_1, \text{lab}_N^1, \text{lab}_E^1 \rangle$ *and* $\mathcal{S}_2 = \langle N_2, E_2, \text{lab}_N^2, \text{lab}_E^2 \rangle$ *be two schemas and* $\Re = \{r_1, \ldots, r_n\}$ *be a set of binary relations which may hold between elements of the two schemas. A* mapping $\mathcal{M}_{\mathcal{S}_1^P \to \mathcal{S}_2^P}$ *is a set of* mapping elements $\langle n_1, n_2, r_i \rangle$*, where* $n_1 \in N_1$*,* $n_2 \in N_2$*, and* $r \in \Re$*.*

We say that a mapping element $m = \langle n_1, n_2, R, q \rangle$ is correct if the two agents a_1 and a_2 are semantically coordinated with respect to n_1, n_2 and R, in a context c:

Definition 7 (Correct Mapping). *Let* $\mathcal{M}_{\mathcal{S}_1 \to \mathcal{S}_2}$ *be a mapping between the schemas* $\mathcal{S}_1$ *and* $\mathcal{S}_2$*.* $\mathcal{M}_{\mathcal{S}_1 \to \mathcal{S}_2}$ *is correct if and only if, for any mapping element* $m = \langle n_1, n_2, r \rangle \in \mathcal{M}_{\mathcal{S}_1 \to \mathcal{S}_2}$*, it holds that:*

$$coord(a_1, a_2, n_1, n_2, r, c)$$

To illustrate the generality of the model, consider the two following cases. In the first, we imagine that a_1 and a_2 are the same agent; in this case, semantic

[2] Compatibility here refers to a precise formal notion which was defined in [15] as part of a logic of contextual reasoning. For lack of space, we will not try event to summarize this notion in any detail. We only stress that compatibility captures the idea of logical constraints holding between two distinct logical languages, and therefore seems especially suitable in this paper, where we imagine that agents have distinct representation languages.

coordination boils down to the translation of elements of different schemas into the same representation language, and to checking whether the relation r holds between the two expressions of the representation language itself. In the second, imagine that a_1 and a_2 are different, but $\mathcal{S}_1$ and $\mathcal{S}_2$ are the same schema; then, a_1 and a_2 might be not coordinated even on the same schema element n, as they might assign a different meaning to n and therefore r might not hold between them.

4 Default rules for semantic coordination

Many theoretical results can be used to prove that semantic coordination (and therefore the correctness of a mapping) can't be directly checked, as it would require knowledge on what any agent really means by a word in a context, and this would be equivalent to look into an agent's mind. However, in this section we show that the condition of semantic coordination can be (and actually is) approximated by three types of *default rules* which are used by agents to "jump to the conclusion" that semantic coordination holds.

4.1 Syntactic default rules

The first type of default rule has to do with the used of a communication language L_{ext} in a community of agents. The idea is the following. Take any two expressions t_1 and t_2 of L_{ext}, and a family R^+ of relations which connect some syntactic features of any two expressions of L_{ext} (e.g. string identity, substring, permutations of strings, and so on). Now suppose that we have a "table" associating the elements of R^+ with a family of set–theoretic relations R. The syntactic default rule (SDR) says that, whenever $r^+ \in R^+$ holds between t_1 and t_2, then $r \in R$ holds between the meaning of t_1 and t_2.

Definition 8 (Syntactic Default Rule). *Let let a_1 and a_2 be two agents, t_1 and t_2 be two expressions of the language L_{ext}, r^+ a relation between expressions of L_{ext}, and r the set-theoretical relation which corresponds to the syntactic relation r^+. Then:*

$$\begin{array}{rl} \textit{if} & t_1\ r^+\ t_2 \quad \textit{in a context } c \\ \textit{then} & \textit{coord}(a_1, a_2, t_1, t_2, r, c) \end{array}$$

As an example, let t_1 be the phrase 'black and white images', t_2 be the word 'images' and r^+ a relation holding two strings when one *contains* the other. Imagine that this relation is associated with set inclusion ($\subseteq$). Then one would be allowed to conjecture that an agent a_1 using the expressions 'black and white images' is semantically coordinated with an agent a_2 using the expression 'images' with respect to set inclusion.

We should recognize that this default rule is extremely powerful and widely used even in human communication, in particular in the special case of a single

term (we typically assume that, unless there is any evidence to the contrary, what other people mean by a word t is what we mean by the same word in a given context). The intuitive correctness and the completeness of this default rule essentially depends on the fact that a syntactical relation r^+ be an appropriate representation of a set–theoretical relation r, and vice–versa, that a set–theoretical relation r is appropriately represented by some syntactical relation r^+. But, in general, polysemy does not allow to guarantee the correctned and completeness of this rule even in the trivial case when $t_1 = t_2$; and the existence of synonyms makes it quite hard to ensure completeness.

4.2 Pragmatic default rule

The second type of default rules says that agents tend to induce that they are semantically coordinated with other agents from a very small number of cases in which the agreed with another agent upon the use of a word. For example, from the fact that they agreed upon a few examples of objects called "laptops", they tend to induce a much stronger form of coordination on the meaning of the word "laptop". Formally, this pragmatic default rule can be expressed as follows:

Definition 9 (Pragmatic Default Rule). *Let t_1 and t_2 be two expressions of the language L_{ext}, W a set of worlds, $c \in C$ a context of use and r a set–theoretical relation. Furthermore, let $w_A \in W$ be a finite world. Then:*

$$\begin{aligned} &\textit{if } \mathcal{P}^1_{w_A}(\mathcal{T}^1_c(t_1))\ R\ \mathcal{P}^2_{w_A}(\mathcal{T}^2_c(t_2)) \\ &\textit{then } \quad coord(a_1, a_2, t_1, t_2, r, c) \end{aligned}$$

For exmaple, if $t_1 = t_2 = t$ and R is the equality symbol ('='), then the pragmatic default rule says that the restricted notion of semantic coordination can be inferred when two agents associate the same subset of the current world to t.

Pragmatic default rules are a very strong form of induction from the particular to the universal, and it is well-known that this not a valid pattern of reasoning. Indeed, if the positive examples are taken from a very small domain, then the two agents may happen to induce their coordination on equivalence simply because they never hit a negative example (lack of correctness); or vice versa they may fail to recognize their coordination simply because they could not find any positive example (lack of completeness).

4.3 Conceptual default rule

Finally, we discuss a third type of default rule, which can be stated as follows:

$$\text{if } \mathcal{T}^i_c(t_1)\ r^*\ \mathcal{T}^j_c(t_2), \text{ then } \forall w \in W\ \ \mathcal{P}^i_w(\mathcal{T}^i_c(t_1))\ r\ \mathcal{P}^j_w(\mathcal{T}^j_c(t_2))$$

where r^* would be any relation between concepts.

However, there are two major problems with this definition:

– first of all, the premise of the rule does not make sense, as we do not know how to check in practice whether $\mathcal{T}_c^i(t_1)\ R^*\ \mathcal{T}_c^j(t_2)$ holds or not, as the two concepts belong to different (and semantically autonomous) representation languages;
– second, it not clear how to determine a relation between concepts. Indeed, the syntactic and pragmatic default rule are based on conditions (relation between strings, and relations between sets of objects) which can be externally verified. But how do we check whether the concept of "cat" is subsumed by the concept of "mammal"?

A possible way out for the second issue may be that agents infer that such a relation holds (or does not hold) from what they know about the two concepts. This, of course, introduces an essential directionality in this rule, as it may be that two agents have different knowledge about the two concepts to be compared. So we need two distinct checks:

$$\begin{aligned} &\text{if } \ \mathcal{O}^i \models \mathcal{T}_c^i(t_1)\ R^*\ \mathcal{T}_c^j(t_2) \\ &\text{then } \ \forall w \in W \ \ \mathcal{P}_w^i(\mathcal{T}_c^i(t_1))\ R\ \mathcal{P}_w^j(\mathcal{T}_c^j(t_2)) \end{aligned}$$

and

$$\begin{aligned} &\text{if } \ \mathcal{O}^j \models \mathcal{T}_c^i(t_1)\ R^*\ \mathcal{T}_c^j(t_2) \\ &\text{then } \ \forall w \in W \ \ \mathcal{P}_w^i(\mathcal{T}_c^i(t_1))\ R\ \mathcal{P}_w^j(\mathcal{T}_c^j(t_2)) \end{aligned}$$

Suppose we accept this asymmetry. How can we address the first problem?

Senses and Dictionaries Here is where we introduce the notion of socially negotiated meanings. Indeed, most communication languages (for example, natural languages) provide dictionaries which list all accepted *senses* of a word (WordNet [13] is a well-known example of an electronic dictionary). A sense can be viewed as a tentative bridge between syntax and semantics: its goal is to list possible meanings (semantics) of a word (syntax), but this is done by providing definitions which are given in the same language which the dictionary is supposed to define. So, dictionaries have two interesting properties:

– on the one hand, they provide a *publicly accessible* and *socially negotiated* list of acceptable senses for a word;
– however, senses cannot *ipso facto* be equated with a list of shared meanings for the speakers of that language, as senses are (circularly) defined through other words, and do not contain the concept itself.

However, we believe that dictionaries are crucial tools for communication languages, and indeed a linguistic community can be defined as a group of speakers which agree on a common dictionary. Let us show how this idea can be used to define a surrogate of a conceptual default rule.

Definition 10 (Lexical Default Rule). *Let t be an expression of the language L_{ext}, W a set of worlds, and $c \in C$ a context of use. Furthermore, let $a_1^{L_{ext}}$ and $a_1^{L_{ext}}$ be to agents of the same linguistic community. Then:*

$$\begin{aligned} &\textit{if } SR_c^1(t) = SR_c^2(t) \\ &\textit{then } \quad coord(a_1, a_2, t, t, \equiv, c) \end{aligned}$$

where $SR_c^i(t)$ is a function that, given a context c and a term $t \in L_{ext}$, returns a suitable sense for t from a dictionary (notice that this function is again parametric on agents).

Such default rules overcomes the first issue of an the ideal conceptual default rule, i.e. comparing terms from different representation languages. Indeed, it is a verifiable condition whether two agents refer to the same dictionary sense of a word in a given context of use. However, as it is, it can only be used to infer the restricted form of semantic coordination, as it applies only to a single term t of L_{ext}. The general default rule should say that two agents a_1 and a_2 are semantically coordinated with respect to two expressions t_1 and t_2 of the communication language L_{ext}, and with respect to a relation r, when the dictionary senses individuated by the sense retrieval function are r^*–related, where r^* is a relation between senses corresponding to the relation r between concepts. As we said at the beginning of this section, the relation r^* can be determined only with respect to an agent's knowledge about the relation between concepts corresponding to senses. To capture this aspect, we need to make a further assumption, namely that there is a mapping from the concepts is an agent's ontology $\mathcal{O}^j$ and dictionary senses. For example, if a_1's ontology $\mathcal{O}^1$ contains the axiom '$\mathsf{cat} \sqsubseteq \mathsf{animal}$' (where cat and animal are expressions of L^1), then $[\mathcal{O}^1]$ contains the axiom '$s_g \sqsubseteq s_h$', where s_g is the dictionary sense 'feline mammal' associated to the word 'cat' and s_h is the dictionary sense 'a living organism' associated to the word 'animal'[3]. Now, we can introduce an extended lexical default rule.

Definition 11 (Lexical Default Rule Extended). *Let t_1 and t_2 be two expressions of the language L_{ext}, W a set of worlds, $c \in C$ a context of use and r a set–theoretical relation. Furthermore, let a_1 and a_1 be two agents belonging to the same linguistic community. Then:*

$$\begin{aligned} &\textit{if } [\mathcal{O}^1] \models SR_c^1(t_1)\ r^*\ SR_c^2(t_2) \\ &\textit{then } \quad coord(a_1, a_2, t_1, t_2, r, c) \quad w.r.t.\ [\mathcal{O}^1] \end{aligned}$$

and

$$\begin{aligned} &\textit{if } [\mathcal{O}^2] \models SR_c^1(t_1)\ r^*\ SR_c^2(t_2) \\ &\textit{then } \quad coord(a_2, a_1, t_1, t_2, r, c) \quad w.r.t.\ [\mathcal{O}^2] \end{aligned}$$

[3] The problem of lexicalizing the ontologies with respect to some dictionary is not completely new. In computer science area, a lot of studies are dedicated to this problem. Among them, in our opinion the most relevant approaches are described in [3, 14, 30].

where r^* is the relation between senses corresponding to r with respect to the conceptual level of meaning.

Essentially, this rule says that, if the sets of senses associated to two expressions t_1 and t_2 by the agents a_1 and a_2 are in some relation r^* with respect to the (lexicalized) knowledge of the agent a_i (for $i = 1, 2$), then they are semantically coordinated with respect to t_1, t_2 and r. An example will better clarify the situation. Suppose t_1 is 'images of cats', and t_2 is 'images of animals'. Furthermore, imagine that the senses that a_1 associates to t_1 are $\langle s_q, s_w, s_e\rangle$ ($SR^1_c(t_1) = \langle s_q, s_w, s_e\rangle$), and that the senses that a_2 associates to t_2 are $\langle s_q, s_w, s_r\rangle$ ($SR^2_c(t_2) = \langle s_q, s_w, s_r\rangle$), where s_q = 'a visual representation', s_w = 'concerning', s_e = 'feline mammal' and s_r = 'a living organism'. Imagine now that the (lexicalized) ontology $[\mathcal{O}^1]$ contains the following two axioms: (i) 'for each pair of concepts c, d, if $c \sqsubseteq d$, then 'a visual representation' 'concerning' c is less general than 'a visual representation' 'concerning' d; (ii) 'feline mammal' $\sqsubseteq$ 'a living organism'. In this case, we can deduce that $\langle s_q, s_w, s_e\rangle$ 'less general than' $\langle s_q, s_w, s_e\rangle$, and, by applying the default rule, that the two agents are coordinated with respect to the relation 'less general than' ($\subseteq$). The same considerations can be done if we take into account the agent knowledge $[\mathcal{O}^2]$ of the other agent.

As we announced at the beginning of the section, the extended lexical default rule introduces a form of *directionality* in the notion of semantic coordination, as the computation of the relation r^* between (lexicalized) concepts relies on an agent's knowledge about them, and such knowledge may lack, or be different in two different agents. However, as it was proved in [2], this directionality effect can be weakened. Indeed, the relation computed by a_1 is guarantee to be correct also for a_2, if we can prove that $\mathcal{O}^1 \sqsubseteq \mathcal{O}^2$.

The correctness and completeness of the lexical default essentially depends on the condition that two agents use the same function to associate dictionary senses to concepts in their internal representation and vice versa. Clearly, this condition cannot be guaranteed, as we can always conceive a situation where two agents point to the same dictionary sense s_k for a term t, but then their internal representation of s_k is different. However, this type of rule makes an essential use of socially negotiated tools, which provide a powerful extension to purely syntactic or pragmatic methods.

5 Conclusions

The model we propose leds to two general results. The first is negative, as it says that the condition required to prove that a mapping is correct (even in the weak sense of semantic coordination) can never be formally proved. The second, however, is that all the proposed matching methods can be classified into three broad families, one for each default rule[4]:

[4] Wever, see [25, 27] for a classification of matching methods based on different principles.

Syntactic methods: methods which try to determine a mapping by a purely syntactic analysis of the linguistic expressions occurring in different schemas, namely by comparing the properties of the strings that are used to label nodes, and reasoning on their arrangement into a schema. Examples can be found in [35, 32, 24, 6, 12, 34, 7, 19];

Pragmatic methods: methods which assume that the relation between schema elements can be inferred from the relation between the data associated to them. Examples can be found in [33, 8, 31, 7, 9, 29, 11, 5, 10, 20, 28];

Conceptual methods: methods which try to compute a mapping by comparing the (lexical representation of the) concepts associated by the schema creators to the schema elements. Examples can be found in [26, 4, 16].

The three types of methods have their own pros and cons. For example, *syntactic methods* are highly effective, as they exploit and reuse very efficient techniques from graph matching; however, their meaningfulness is quite low, as they take into account a very superficial level of meaning which disregards potential ambiguities and cannot capture semantic relations between concepts which are not reflected in the pure syntax (for example, the relation between "cats" and "mammals"). *Pragmatic methods* are extremely meaningful (if two schemas were used to classify the same data set, and if the data set was representative of the domain of discourse, then the outcome of pragmatic methods would be always correct – independently from the analysis of labels and from the arrangement of nodes); however, the necessary preconditions (same data set and appropriate coverage) are extremely hard to match, and in practice we can never know when they are matched (we can only rely on statistical methods). Finally, *conceptual methods* – like pragmatic methods – have the advantage of being independent from the syntactical structure of the schemas, as in principle they can find the correct relation between two schema elements even if labels are (syntactically) very different and nodes are arranged in different orders (when such an order is inessential from a semantic point of view). In addition, like syntactic methods, they have the advantage of being independent from the data contained into the schema elements. However, conceptual methods may fail in two crucial steps. First of all, the function which returns a sense for a word in a context of use is quite complex, and inherits most well-known issues related to word sense disambiguation in NLP. Second, an agent might lack part of the relevant knowledge to compute a mapping between two concepts. However, we should add that these methods are the only ones which are semantically incremental: one can always know why a mapping was not found (or why a wrong match was computed) and fix the problem is a general way (and not, for example, by tuning some parameters, which may have bad effects on the performance on different schemas).

Probaby to overcome some of these limitations, most actual methods are indeed hybrid, as they use techniques which are based on more than one default rule (for example, syntactic methods use lexical information from thesauri, and some conceptual methods use string matching techniques for improving the quality of their results).

References

1. F. Baader, D. Calvanese, D. McGuinness, D. Nardi, and P. Patel-Schneider, editors. *The Description Logic Handbook. Theory, Implementation and Applications*. Cambridge University Press, January 2003.
2. M. Benerecetti, P. Bouquet, and S. Zanobini. Soundness of schema matching methods. In L. Torenvliet A. Gòmez-Pèrez, J. Euzenat, editor, *second European Semantic Web Conference (ESWC 2005)*, volume 3532 of *LNCS*, Heraklion, Crete, Greece, May 29–June 1, 2005 2005. Springer.
3. S. Bergamaschi, S. Castano, and M. Vincini. Semantic integration of semistructured and structured data sources. *SIGMOD Record*, 28(1):54–59, 1999.
4. P. Bouquet, L. Serafini, and S. Zanobini. Semantic coordination: a new approach and an application. In D. Fensel, K. P. Sycara, and J. Mylopoulos, editors, *The Semantic Web – 2nd international semantic web conference (ISWC 2003)*, volume 2870 of *Lecture Notes in Computer Science (LNCS)*, pages 130–145, Sanibel Island, Fla., USA, 20-23 October 2003. Springer Verlag.
5. J. Broekstra, M. Ehrig, P. Haase, F. van Harmelen, M. Menken, P. Mika, B. Schnizler, and R. Siebes. Bibster - a semantics-based bibliographic peer-to-peer system. In *In Proceedings of the SemPGrid 04 Workshop*, New York, USA, May 2004.
6. J. Carroll and HP. Matching rdf graphs. In *Proc. in the first International Semantic Web Conference - ISWC 2002*, pages 5–15, 2002.
7. A. Doan, P. Domingos, and A. Y. Halevy. Reconciling schemas of disparate data sources: A machine-learning approach. In *SIGMOD Conference*, 2001.
8. A. Doan, P. Domingos, and A. Y. Levy. Learning source description for data integration. In *WebDB (Informal Proceedings)*, pages 81–86, 2000.
9. A. Doan, J. Madhavan, P. Domingos, and A. Halevy. Learning to map between ontologies on the semantic web. In *Proceedings of WWW-2002, 11th International WWW Conference, Hawaii*, 2002.
10. M. Ehrig and S. Staab. Qom - quick ontology mapping. In *In Proceedings of the 3rd ISWC*, Hiroshima (JP), November 2004.
11. J. Euzenat. Brief overview of t-tree: the tropes taxonomy building tool. In *In Proceedings of the 4th ASIS SIG/CR workshop on classification research*, pages 69–87, Columbus (OH), USA, 1994.
12. J. Euzenat and P. Valtchev. An integrativive proximity measure for ontology alignment. Proceedings of the ISWC2003 workshop on *Semantic Integration*, Sanibel Island (Florida, USA), October 2003.
13. Christiane Fellbaum, editor. *WordNet: An Electronic Lexical Database*. The MIT Press, Cambridge, US, 1998.
14. A. Gangemi. *Comparative and Functional Genomics*, volume 4, Some tools and methodologies for domain ontology building, pages 104–110. John Wiley & Sons, Ltd., 2003.
15. C. Ghidini and F. Giunchiglia. Local Models Semantics, or Contextual Reasoning = Locality + Compatibility. *Artificial Intelligence*, 127(2):221–259, April 2001.
16. F. Giunchiglia and P. Shvaiko. Semantic matching. *The Knowledge Engineering Review Journal*, 18(3):265–280, 2003.
17. P. Hitzler, J. Euzenat, M. Krotzsch, L. Serafini, H. Stuckenschmidt, H. Wache, and A. Zimmermann. D2.2.5 – Integrated view and comparison of alignment semantics. Deliverable of the EU funded network of excellence *KnowlegdeWeb*, 2005.
18. P. Bouquet, L. Serafini, and S. Zanobini. Bootstraping semantics on the web: meaning elicitation from schemas. Proceedings of *WWW2006*, Edinburgh, Scotland, 22nd – 26th May 2006.

19. W. Hu, N. Jian, Y. Qu, and Y. Wang. Gmo: A graph matching for ontologies. In *Integrating ontologies Workshop Proceedings, K-CAP 2005*, pages 41–48, 2005.
20. Ryutaro Ichisem, Hiedeaki Takeda, and Shinichi Honiden. Integrating multiple internet directories by instance–base learning. In *AI AND DATA INTEGRATION*, pages 22–28, 2003.
21. N. Jian, W. Hu, G. Chen, and Y. Qu. Falcon-ao: Aligning ontologies with Falcon. In *Integrating ontologies Workshop Proceedings, K-CAP 2005*, pages 85–91, 2005.
22. J. Madhavan, P. A. Bernstein, and E. Rahm. Generic schema matching with CUPID. In *The VLDB Journal*, pages 49–58, 2001.
23. T. Milo and S. Zohar. Using schema matching to simplify heterogeneous data translation. In *Proc. 24th Int. Conf. Very Large Data Bases, VLDB*, pages 122–133, 24–27 1998.
24. M. Pelillo, K. Siddiqi, and S. W. Zucker. Matching hierarchical structures using association graphs. *Lecture Notes in Computer Science*, 1407, 1998.
25. E. Rahm and P.A. Bernstein. A survey of approaches to automatic schema matching. *VLDB Journal*, 10(4), 2001.
26. L. Serafini, S. Zanobini, S. Sceffer, and P. Bouquet. Matching hierarchical classifications with attributes. In *European Semantic Web Conference (ESWC2006)*, 2006.
27. P. Shvaiko. A classification of schema-based matching approaches. Technical Report DIT-04-093, Dipartimento di Informatica e Telecomunicazioni, University of Trento, 2004.
28. U. Straccia and R. Troncy. OMAP: Results of the ontology alignment context. In *Integrating ontologies Workshop Proceedings, K-CAP 2005*, pages 92–96, 2005.
29. Gerd Stumme and Alexander Maedche. FCA-merge: bottom-up merging of ontologies. In *In Proceedings of 17th IJCAI*, pages 225–230, Seattle (WA), USA, 2001.
30. W. R. van Hage, S. Katrenko, and G. Schreiber. A method to combine linguistic ontology-mapping techniques. In Y. Gil, E. Motta, V. R. Benjamins, and M. A. Musen, editors, *The Semantic Wec – ISWC 2005*, volume 3729 of *LNCS*. Springer, 2005.
31. C. Clifton W. Li. Semint: A tool for identifying attribute correspondences in heterogeneous databases using neural networks. In *Data & Knowledge Engineering*, volume 33(1), pages 49–84, 2000.
32. J. T. Wang, K. Zhang, K. Jeong, and D. Shasha. A system for approximate tree matching. *Knowledge and Data Engineering*, 6(4):559–571, 1994.
33. Q. Y. Wang, J. X. Yu, and K. Wong. Approximate graph schema extraction for semi-structured data. In *Proceedings of the 7th International Conference on Extending Database Technology*, pages 302–316. Springer-Verlag, 2000.
34. L. Xu and D. W. Embley. Using domain ontologies to discover direct and indirect matches for schema elements. In *Proceedings of the workshop on Semantic Integration*, Sanibel Island (Florida, USA), October 2003.
35. K. Zhang, J. T. L. Wang, and D. Shasha. On the editing distance between undirected acyclic graphs and related problems. In Z. Galil and E. Ukkonen, editors, *Proceedings of the 6th Annual Symposium on Combinatorial Pattern Matching*, volume 937, pages 395–407, Espoo, Finland, 1995. Springer Verlag.