YANNIS KALFOGLOU, MARCO SCHORLEMMER

THE INFORMATION FLOW APPROACH TO ONTOLOGY-BASED SEMANTIC ALIGNMENT

In this article we argue for the lack of formal foundations for ontology-based semantic alignment. We analyse and formalise the basic notions of semantic matching and alignment and we situate them in the context of ontology-based alignment in open-ended and distributed environments, like the Web. We then use the mathematical notion of information flow in a distributed system to ground three hypotheses that enable semantic alignment. We draw our exemplar applications of this work from a variety of interoperability scenarios including ontology mapping, theory of semantic interoperability, progressive ontology alignment, and situated semantic alignment.

1 INTRODUCTION

In order for two systems (databases, software agents, peers, web services, software components, etc.) to be considered semantically integrated, both will need to commit to a shared conceptualisation of the application domain. Commonly, this is achieved by providing an explicit specification of this conceptualisation—what has become to be known as an *ontology*—and by defining each system's local vocabulary in terms of the ontology's vocabulary. Thus, an ontology models the vocabulary used by knowledge engineers so that it denotes concepts and their relations, and it constrains the interpretation of this vocabulary to the meaning originally intended by knowledge engineers. As such, ontologies have been widely adopted as an enabling technology for interoperability in distributed environments, such as multi-agent systems, federated databases, or the semantic web.

This sort of interoperability is dubbed "semantic" precisely because it assumes that the ontology is some sort of structured theory T—coming thus equipped with a precise semantics for the structure it holds—and because each system's local language L_i is *interpreted* in T (e.g., in the technical sense of a theory interpretation as defined in (Enderton, 2002), when T is a theory in first-order logic). Semantic integration is therefore always relative to the theory T into which local languages are interpreted. We shall call this theory the *reference theory* of the integration.

The use of ontologies as reference theories for semantic integration, however, is more in tune with a classical codification-centred knowledge management tradition, as put forward in (Corrêa da Silva and Agustí, 2003). Such tradition comprises the efforts to define standard upper-level ontologies such as CyC (Lenat, 1995) and SUO (IEEE, 2003), or to establish public ontology repositories for specific domains to favour knowledge reuse such as the Ontolingua server (Farquhar et.al., 1997). Corrêa da Silva and Agustí remark that "centralised ontologies [...] promise to bring the control of the organisation back to what was possible under classical management techniques. The problem is that they may also bring back the rigidity of agencies organised under the classical management tenets."

Before ontologies became popular, knowledge engineers hardly ever had to work with more than one ontology at a time. Even in cases where multiple ontologies were used (see, e.g., (Borst et al., 1997)), these were mostly controlled experiments (e.g., (Uschold et al.,

1998)) in moderated environments (such as (Farquhar et al., 1997)). Nowadays, however, the practice is somewhat different. Modern trends in knowledge management dictate that we should expect to work more and more within highly distributed, open, and dynamic environments like the web. In this sort of environment it is more realistic to achieve certain levels of semantic integration by matching vocabulary on-the-fly. In addition, the proliferation of many diverse ontologies caused by different conceptualisations of even the same domain—and their subsequent specification using varying terminology—has highlighted the need of ontology matching techniques that are capable of computing semantic relationships between entities of disparate ontologies (Kalfoglou and Schorlemmer, 2003b; Shvaiko and Euzenat, 2005). Since ontologies are the result of an inter-subjective agreement among individuals about the same fragment of the objective world, they are also highly context-dependent and hardly will result to be general-purpose, regardless of how abstract and upper-level they might be.

2 ONTOLOGY-BASED SEMANTIC INTEGRATION: BASIC CONCEPTS AND DEFINITIONS

In this chapter we shall be concerned with semantic integration understood as the integration of two systems by virtue of the interpretation of their respective vocabularies into a reference theory—an ontology—expressible in some logical language. In practice, semantic integration is often carried out on subsets of first-order logic, such as description logics (DL), for which reasoning has good computational properties. This is, for instance, the approach followed by Calvanese and De Giacomo in their ontology integration system for database schemata (Calvanese and De Giacomo, 2005); W3C, too, has embraced DLs in order to develop the OWL recommendation for ontology representation (McGuinness and van Harmelen, 2004). Another example is the focus of Giunchiglia, Marchese and Zaihrayeu on propositional DLs in order to use fast SAT provers for matching taxonomically organised vocabularies (Giunchiglia et al., 2006). In contrast, the Process Specification Language (PSL) is an example of a semantic integration initiative based on full first-order logic that uses invariants to define interpretations of local vocabulary into PSL (Grüninger and Kopena, 2005).

By *vocabulary* we mean a set V of words and symbols used by a system to represent and organise its local knowledge. In a formal, logic-based representation language the vocabulary is constituted by the non-logical symbols used to form sentences and formulae (in this case it is usually referred to as *parameters* or *signature*). The *language* is then the set $L(V)$ of all well-formed formulae over a given vocabulary V. We shall also write L when we do not want to explicitly refer to the vocabulary. We call the elements of a language L, *sentences*.

In declarative representation languages, knowledge is represented and organised by means of theories. DL-based ontologies are such an example. A convenient way to abstractly characterise theories in general, is by means of a consequence relation. Given a language L, a *consequence relation* over L is, in general, a binary relation $|\!-$ on subsets of L which satisfies certain structural properties.[1] Consequence relations are also suitable to capture other sorts of mathematical structures used to organise knowledge in a systematic way, such as taxonomic hierarchies. When defined as a binary relation on L (and not on subsets of L), for instance, it

[1] These are commonly those of Identity, Weakening and Global Cut (see Definition 9)

coincides with a partial order. Furthermore, there exists a close relationship between consequence and classification relations (which play a central role in ontological knowledge organisation), which has been thoroughly studied from a mathematical perspective in (Dunn and Hardegree, 2001; Barwise and Seligman, 1997; Ganter and Wille, 1999).

We call a *theory* a tuple $T = \langle L_T, |\!-_T \rangle$, where $|\!-_T \subseteq \wp(L_T) \times \wp(L_T)$ is a consequence relation, hence capturing with this notion the formal structure of an ontology in general. Finally, in order to capture the relationship between theories, we call a *theory interpretation* a map between the underlying languages of theories that respects consequence relations. That is, a function $i: L_T \rightarrow L_{T'}$ is a theory interpretation between theories $T = \langle L_T, |\!-_T \rangle$ and $T' = \langle L_{T'}, |\!-_{T'} \rangle$ if, and only if, for all $\Gamma, \Delta \subseteq L$ we have that $\Gamma |\!-_T \Delta$ implies $i(\Gamma) |\!-_{T'} i(\Delta)$ (where $i(\Gamma)$ and $i(\Delta)$ are the set of direct images of Γ and Δ along i, respectively.[2]

2.1 Semantic Matching

We call *semantic matching* the process that takes two theories T_1 and T_2 as input (called *local theories*) and computes a third theory $T_{1\leftrightarrow 2}$ as output (called *bridge theory*) that captures the semantic relationship between T_1 and T_2's languages with respect to a reference theory T. As we shall see below, we call the output of the semantic-matching process, together with the input it relates, a *semantic alignment*. It is important to make a couple of remarks here.

First, one usually distinguishes a theory from its presentation. If the language L is infinite (as for instance in propositional or first-order languages, where the set of well-formed formulae is infinite, despite having a finite vocabulary), any consequence relations over L will also be infinite. Therefore, one deals in practice with a finite subset of $\wp(L) \times \wp(L)$, called a *presentation*, to stand for the smallest consequence relation containing this subset. A presentation may be empty, in which case the smallest consequence relation over a language L containing it, is called the *trivial theory*. We will write $Tr(L)$ for the trivial theory over L. It is easy to prove that, for all $\Gamma, \Delta \subseteq L$, $\Gamma |\!-_{Tr(L)} \Delta$ if, and only if, $\Gamma \cap \Delta \neq \varnothing$.

Rigorously speaking, current implementations of semantic matching actually take two presentations of local theories as input and compute a presentation of the bridge theory as output. But, from a conceptual perspective, we shall characterise semantic matching always in terms of the theories themselves.

Second, the reference theory T is usually *not* an explicit input to the semantic matching process (not even a presentation of it). Instead it should be understood as the background knowledge used by a semantic matcher to infer semantic relationships between the underlying languages of the respective input theories. For a manual matcher, for instance, the reference theory may be entirely dependent on user input, while a fully automatic matcher would need to rely on automatic services (either internal or external to the matcher) to infer such reference theory. It is for this reason that we talk of a *virtual* reference theory, since it is not explicitly provided to the semantic matcher, but is implicit in the way external and internal sources are brought into the matching process as background theory in order to compute a semantic alignment.

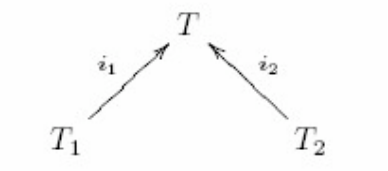
---

[2] Theories and theory interpretations as treated here can also be seen as particular cases of the more general framework provided by institution theory, which has been thoroughly studied in the field of algebraic software specification (see Goguen and Burstall, 1992).

Next, we provide precise definitions of what we mean by bridge theory to capture a semantic alignment of languages, and also what we mean by a semantic alignment underlying a semantic integration of local theories.

2.2 Integration Theory

Definition 1: Two theories T_1 and T_2 are *semantically integrated with respect to T*, if there exist theory interpretations $i_1 : T_1 \rightarrow T$ and $i_2 : T_2 \rightarrow T$.



We call $\mathcal{I} = \{i_j : T_j \rightarrow T\}_{j=1,2}$ *the semantic integration* of *local theories* T_1 and T_2 with respect to *reference theory* T. Two languages L_1 and L_2 are *semantically integrated with respect to T* if their respective trivial theories are.

In a semantic alignment we are interested in determining the semantic relationship between the languages L_{T1} and L_{T2} on which semantically integrated theories T_1 and T_2 are expressed. Therefore, a semantic integration $\mathcal{I}$ of T_1 and T_2 with respect to a reference theory T as defined above is not of direct use, yet. What we would like to have is a theory $T_{\mathcal{I}}$ over the combined language $L_{T1} \uplus L_{T2}$ (the disjoint union) expressing the semantic relationship that arises by interpreting local theories in T. We call this the *integration theory* of $\mathcal{I}$, and it is defined as the inverse image of the reference theory T under the sum of the theory interpretations in $\mathcal{I}$.

Definition 2: Let $i : T \rightarrow T'$ be a theory interpretation. The *inverse image* of T' under i, denoted $i^{-1}[T']$, is the theory over the language of T such that $\Gamma \mid\!\sim_{i^{-1}[T']} \Delta$ if, and only if, $i(\Gamma) \mid\!\sim_{T'} i(\Delta)$.

It is easy to prove that, for every theory interpretation $i : T \rightarrow T'$, T is a *subtheory* of $i^{-1}[T']$, i.e., $\mid\!\sim_T \subseteq \mid\!\sim_{i^{-1}[T']}$

Definition 3: Given theories $T_1 = \langle L_{T1}, \mid\!\sim_{T1}\rangle$ and $T_2 = \langle L_{T2}, \mid\!\sim_{T2}\rangle$, the sum $T_1 + T_2$ of theories is the theory over the sum of language (i.e., the disjoint union of languages) $L_{T1} \uplus L_{T2}$ such that $\mid\!\sim_{T1+T2}$ is the smallest consequence relation such that $\mid\!\sim_{T1} \subseteq \mid\!\sim_{T1+T2}$ and $\mid\!\sim_{T2} \subseteq \mid\!\sim_{T1+T2}$.

Given theory interpretations $i_1 : T_1 \rightarrow T$ and $i_2 : T_2 \rightarrow T$, the sum $i_1 + i_2 : T_1 + T_2 \rightarrow T$ of theory interpretations is just the sum of their underlying map of languages.

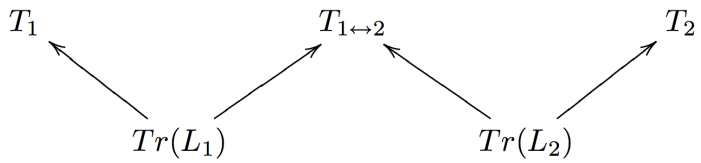
Definition 4: Let $\mathcal{I} = \{i_j : T_j \rightarrow T\}_{j=1,2}$ be a semantic integration of T_1 and T_2 with respect to T. The *integration theory* $T_{\mathcal{I}}$ of the semantic integration $\mathcal{I}$ is the inverse image of T under the sum of interpretations $i_1 + i_2$, i.e. $T_{\mathcal{I}} = (i_1+i_2)^{-1}[T]$.

The integration theory faithfully captures the semantic relationships between sentences in L_{T1} and L_{T2} as determined by their respective interpretation into T, but expressed as a theory over the combined language $L_{T1} \uplus L_{T2}$. The sum of local theories $T_1 + T_2$ is therefore always a subtheory of the integration theory T_I, because it is through the interpretations in T where we get the semantic relationship between languages. It captures and formalises the intuitive idea that an integration is more than just the sum of its parts.

2.3 Semantic Alignment

In semantic matching one usually isolates as output to the matching process the bit that makes T_I genuinely a super theory of $T_1 + T_2$. The idea is to characterise a theory $T_{1\leftrightarrow 2}$ over the disjoint union of subsets $L_1 \subseteq L_{T1}$ and $L_2 \subseteq L_{T2}$, called bridge theory, which, together with T_1 and T_2, uniquely determines the integration theory T_I. To keep everything characterised uniformly in the same conceptual framework, the bridge theory, together with its relationship to the local theories T_1 and T_2, can be expressed by a diagram of theory interpretations as follows.

Definition 5: *A semantic alignment* $\mathcal{A}$ of T_1 with T_2 is a diagram



in the category of theories and theory interpretations, where $L_i \subseteq L_{Ti}$ and $T_{1\leftrightarrow 2}$ is a theory whose underlying language $L_{T1\leftrightarrow 2} = L_1 \uplus L_2$, and where all arrows are theory inclusions. We shall also write $T_1 \longleftarrow^{\mathcal{A}} \longrightarrow T_2$ as shorthand of an alignment.

We say that a semantic alignment A underlies a semantic integration I when the colimit of A in the category of theories and theory interpretations (which always exists) is the integration theory of I, i.e., $colim(\mathcal{A}) = T_I$.

This representation of semantic alignment as a system of objects and morphisms in a category, and of semantic integration by means of a colimit of such a diagram, bears a close relationship to the notion of *W-alignment diagram* described in (Zimmermann et al., 2006). This is so because both notions share the same categorical approach to semantic alignment. But, unlike in (Zimmermann et al., 2006), we further take a dual "type-token" structure of semantic integration into account, and we define an alignment with respect to this two-tier model. We claim that in this way we better capture Barwise and Seligman's basic insight that "information flow involves both types and their particulars" (Barwise and Seligman, 1997). This will become clearer next when we describe the role of tokens in semantic alignment scenarios.

3 SEMANTIC ALIGNMENT THROUGH MEANING COORDINATION

We shall consider a scenario in which two agents A_1 and A_2 want to interoperate, but each agent A_i has its knowledge represented according to its own conceptualisation, which we assume is explicitly specified by means of its own ontology O_i. Any expression α_i using the vocabulary O_i will be considered semantically distinct a priori from any expression α_j using vocabulary O_j (with $j \neq i$), even if they happen to be syntactically equal, unless the semantic evidence unveiled by an ontology-matching process of the kind described below makes them mean the same to A_1 and A_2. Furthermore, we assume that the agents' ontologies are not open for inspection, so that semantic heterogeneity cannot be solved by semantically matching the ontologies beforehand.

An agent may learn about the ontology of another agent only through meaning coordination. Thus, we assume that agent A_i is capable of requesting from agent A_j to explain the intended meaning of an expression α_j that is in a message from A_j to A_i and uses the vocabulary O_j. Agent A_i might request such an explanation with the intention of determining the semantic relationship of the fragment of O_j used in α_j with respect to its local vocabulary O_i. Correspondingly, we assume that agent A_j is capable of explaining to A_i the meaning of expression α_j by means of a *token* of this expression.

The formal framework we describe in the next section is neutral with respect to the syntactic form of expressions and, more importantly, to what tokens might be, giving an interesting level of generality to ontology alignment. The Oxford Dictionary of English defines a token as "a thing serving as a visible or tangible representation of something abstract." In our scenario a token will be something agent A_i is capable of processing and putting into relationship with its own local ontology O_i.

Take for instance the ontology negotiation process described in (Bailin and Truszkowski, 2002). There, an agent A_i, upon the reception from another agent A_j of a message containing a list of keywords, either sends to A_j an interpretation of the keywords in the form of WordNet synonyms in order to check that it has interpreted A_j's vocabulary correctly, or else requests A_j for a clarification of the interpretation of unknown keywords, also in form of WordNet synonyms. Thus, in this scenario, the role of tokens is played by WordNet synonyms of those keywords whose interpretation needs to be confirmed or clarified.

Looking at another ontology alignment scenario, (Wang and Gasser, 2002) present an ontology-matching algorithm for open multi-agent systems, where ontologies are partitions of domain instances into categories, based on the K-means algorithm, a typical partition-based clustering method. The alignment is computed out in an online fashion by exchanging instances between two agents, rather than by exchanging abstract concepts. When an agent plans to express some concept or category to other agents it uses an instance belonging to that category to represent this concept. In this scenario it is particular domain instances who play the role of tokens of a concept or category. Wang and Gasser further note, that "unless a set of agents already has a compatible and verified shared ontology, it is difficult to see how they could specify categories to each other in another way." The capability of a set of agents to process and classify tokens according to their own local ontologies is what underlies the ontology-matching process. (van Diggelen et al., 2007) also describe an ontology matching protocol pointing to instances for concept explication. One agent communicates a number of positive and negative examples of the concept to the other agent, which in turn, classifies these examples using the concept classifier from its own ontology.

Finally, in other scenarios, (Giunchiglia and Shvaiko, 2004) and (Bouquet et al., 2003) use mappings of concepts in a tree hierarchy to propositional expressions using WordNet synsets in order to check, by means of a SAT prover (a software program that checks the satisfiability of the propositions supplied to it), the semantic relationships between concepts occurring in two different hierarchies. In this scenario, a concept is represented by a propositional formula, playing the role of the token for this concept, which can then be processed by each agent with the SAT prover.

4 SEMANTIC ALIGNMENT HYPOTHESES

We have described a process by which agents compute an ontology alignment by making the intended meaning of syntactic expressions explicit to each other through the use of tokens for these expressions. We deliberately have left unspecified what these tokens actually are, and have only briefly mentioned that we shall consider tokens as something agents are capable of processing and putting into relationship with their own local vocabulary. This view of a semantic alignment is the result of the research initiated by (Kent, 2000) on conceptual knowledge organization, and applied to ontology alignment by (Schorlemmer and Kalfoglou, 2003; Kalfoglou and Schorlemmer, 2004) aiming at a formal foundation for semantic interoperability and integration based on channel theory—Barwise and Seligman's proposal for a mathematical theory of information (Barwise and Seligman, 1997).

In this section we introduce the main channel-theoretic constructs required for our formal foundation for ontology alignment, motivating them by means of three *Semantic Alignment Hypotheses*.

Channel theory takes the idea of a classification as the fundamental notion for modelling the local context by which tokens relate to types:

Definition 6: A *classification* $\mathbf{A} = \langle tok(\mathbf{A}), typ(\mathbf{A}), \models_{\mathbf{A}} \rangle$ consists of a set of *tokens* $tok(\mathbf{A})$, a set of *types* $typ(\mathbf{A})$ and a *classification relation* $\models_{\mathbf{A}} \subseteq tok(\mathbf{A}) \times typ(\mathbf{A})$ that classifies tokens to types.

Although a very simple notion, classifications have recently been used, under varying terminology, in many related fields of formal knowledge representation and theoretical computer science (e.g., in algebraic logic (Dunn and Hardegree, 2001), categorical logic (Barr, 1996), formal concept analysis (Ganter and Wille, 1999), and process algebra (Pratt, 2001)).

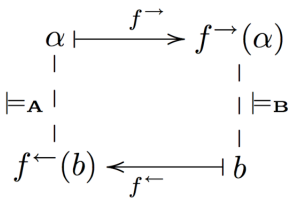
> **Hypothesis 2**: *Semantic alignment presupposes a flow of information between expressions (i.e., types) of separate agents that happens by virtue of shared tokens for these expressions. This flow of information can be accurately described by means of an information channel (Definition 8).*

A fundamental construct of channel theory is that of an information channel between two classifications. It models the information flow between components. First, though, we need to describe how classifications are connected with each other through infomorphisms:

Definition 7: An *infomorphism* $f = \langle f^{\rightarrow}, f^{\leftarrow} \rangle$: $\mathbf{A} \rightarrow \mathbf{B}$ from classifications $\mathbf{A}$ to $\mathbf{B}$ is a

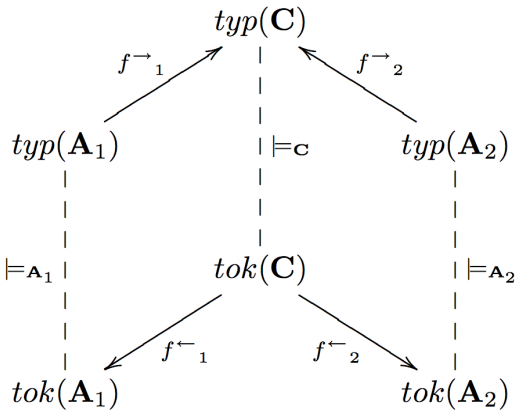
contravariant pair of functions $f^{\rightarrow}$: $typ(\mathbf{A}) \rightarrow typ(\mathbf{B})$ and $f^{\leftarrow}$: $tok(\mathbf{B}) \rightarrow tok(\mathbf{A})$ satisfying the following fundamental property, for each type $\alpha \in typ(\mathbf{A})$ and token $b \in tok(\mathbf{B})$:

$$f^{\leftarrow}(b) \models_{\mathbf{A}} \alpha \quad \textit{iff} \quad b \models_{\mathbf{B}} f^{\rightarrow}(\alpha)$$



As with classifications, infomorphisms have been around in the literature for a long time, and its contra-variance between the type- and token- level is recurrent in many fields. They would correspond to *interpretations* when translating between logical languages (Enderton, 2002), or to *Chu transforms* in the context of *Chu spaces* (Pratt, 1995). Channel theory makes use of this contra variance to model the flow of information at type-level because of the particular connections that happen at the token-level:

Definition 8: An *information channel* consists of two classifications $\mathbf{A}_1$ and $\mathbf{A}_2$ connected through a core classification C via two infomorphisms f_1 and f_2:



Hypothesis 3: *Semantic alignment is formally characterised by a consequence relation between expressions (i.e., types) of separate agents. This consequence relation can be faithfully captured by the natural logic (Definition 11) of the core of the information channel underlying the integration.*

Channel theory is based on the understanding that information flow is the result of regularities in distributed systems. These regularities are implicit in the representation of systems as interconnected classifications. However, one can make these regularities explicit in a logical fashion by means of theories and local logics:

Definition 9: A *theory* $T = \langle typ(T), \vdash_T \rangle$ consists of a set $typ(T)$ of types, and a binary relation between subsets of $typ(T)$. Pairs $\langle \Gamma, \Delta \rangle$ of subsets of $typ(T)$ are called *sequents*. If $\Gamma \vdash_T \Delta$, for

$\Gamma, \Delta \subseteq \text{typ}(T)$, then the sequent $\Gamma \vdash_T \Delta$ is called a *constraint*. T is regular if for all $\alpha \in typ(T)$ and all $\Gamma, \Gamma', \Delta, \Delta', \Sigma \subseteq typ(T)$:

1. *Identity*: $\alpha \vdash_T \alpha$
2. *Weakening*: If $\Gamma \vdash_T \Delta$, then $\Gamma, \Gamma' \vdash_T \Delta, \Delta'$
3. *Global Cut*: If $\Gamma, \Sigma_0 \vdash_T \Delta, \Sigma_1$ for each partition $\langle \Sigma_0, \Sigma_1 \rangle$ of Σ, then $\Gamma \vdash_T \Delta$

Note that, as is usual with sequents and constraints, we write α instead of $\{\alpha\}$ and Γ, Γ' instead of $\Gamma \cup \Gamma'$. Also, a partition of Σ is a pair $\langle \Sigma_0, \Sigma_1 \rangle$ of subsets of Σ, such that $\Sigma_0 \cup \Sigma_1 = \Sigma$ and $\Sigma_0 \cap \Sigma_1 = \varnothing$; Σ_0 and Σ_1 may themselves be empty (hence it is actually a quasi-partition). Note that Global Cut is implied by the usual (Finitary) Cut only if the binary relation is *compact*, i.e., $\Gamma \vdash_T \Delta$ implies the existence of finite subsets $\Gamma_0 \subseteq \Gamma$ and $\Delta_0 \subseteq \Delta$ such that $\Gamma_0 \vdash_T \Delta_0$.

Regularity arises from the observation that, given any classification of tokens to types, the set of all sequents that are satisfied by all tokens always fulfills Identity, Weakening, and Global Cut. Hence, the notion of a local logic:

Definition 10: A *local logic* $\mathcal{L} = \langle tok(\mathcal{L}), typ(\mathcal{L}), \models_{\mathcal{L}}, \vdash_{\mathcal{L}}, N_{\mathcal{L}} \rangle$ consists of a classification $cla(\mathcal{L}) = \langle tok(\mathcal{L}), typ(\mathcal{L}), \models_{\mathcal{L}} \rangle$, a regular theory $th(\mathcal{L}) = \langle typ(\mathcal{L}), \vdash_{\mathcal{L}} \rangle$ and a subset of $N_{\mathcal{L}} \subseteq tok(\mathcal{L})$ of *normal tokens*, which satisfy all the constraints of $th(\mathcal{L})$; a token $a \in tok(\mathcal{L})$ satisfies a constraint $\Gamma \vdash_{\mathcal{L}} \Delta$ of $th(\mathcal{L})$ if, when a is of all types in Γ, a is of some type in Δ.

Finally, every classification determines a natural logic, which captures the regularities of the classification in a logical fashion, and which we shall use in order model the semantic interoperability between agents with different ontologies:

Definition 11: The *natural logic* is the local logic $Log(\mathbf{C})$ generated from a classification $\mathbf{C}$, and has as classification $\mathbf{C}$, as regular theory the theory whose constraints are the sequents satisfied by all tokens, and whose tokens are all normal.

The three Semantic Alignment Hypotheses above comprise the core of what we call *the information-flow approach to ontology-based semantic alignment*. The basic concepts and definitions of Section 2 characterise semantic alignment in terms of theory interpretations, which amount to maps of languages, actually maps of types. Hypotheses 1 and 2, however, make the role of tokens explicit in the characterisation of a semantic integration. The natural logic then determines the integration theory of Section 2 entirely through the way tokens are classified to types in the core of an information channel, thus playing the role of the reference theory of the integration. In the next section we summarise how we have been applying this view of semantic integration in order to successfully tackle the semantic heterogeneity problem in a variety of different scenarios.

5 APPLICATIONS AND EXPLORATIONS

Ontology Mapping: A thorough survey on existing ontology mapping techniques in this domain revealed a surprising scarcity of formal, theoretically sound approaches to the

problem (Kalfoglou and Schorlemmer, 2003b). Consequently, we set out to explore information-flow theoretic ways to tackle the problem. In (Kalfoglou and Schorlemmer, 2003a) we describe a novel ontology mapping method and a system that implements it, IF-Map, which aims to (semi-)automatically map ontologies by representing ontologies as IF classifications and automatically generate infomorphisms between them. We demonstrated this approach by using the IF-Map system to map ontologies in the domain of computer science departments from five UK universities. The underlying philosophy of IF-Map follows the assumption that the way communities classify their instances with respect to local types reveals the semantics that could be used to guide the mapping process. The method is operationalised in a system that includes harvesting mechanisms for acquiring ontologies from online resources, translators for processing different ontology representation formalisms, and APIs for web-enabled access of the generated mappings, all in the form of infomorphisms which are encoded in RDF/OWL formats.

Theory of Semantic Interoperability: We have also explored the suitability of the information flow theory to define a framework that captures semantic interoperability without committing to any particular semantic perspective (model-theoretic, property-theoretic, proof-theoretic, etc.), but which accommodates different understandings of semantics (Kalfoglou and Schorlemmer, 2004). We articulated this framework around four steps that, starting from a characterisation of an interoperability scenario in terms of IF classifications of tokens to types, define an information channel that faithfully captures the scenario's semantic interoperability. We used this framework in an e-government alignment scenario, where we used our four-step methodology to align UK and US Governmental departments using their ministerial units as types and their respective set of responsibilities as tokens, which were classified against those types.

Progressive Ontology Alignment: More recently, we applied information-flow theory to address the issues arising during ontology coordination (Schorlemmer and Kalfoglou, 2004; Schorlemmer and Kalfoglou, 2005). We have been modelling ontology coordination with the concept of a coordinated information channel, which is an IF channel that states how ontologies are progressively coordinated, and which represents the semantic integration achieved through interaction between two agents. It is a mathematical model of ontology coordination that captures the degree of participation of an agent at any stage of the coordination process, and is determined both, at the type and at the token level. Although not yet a fully-fledged theory of ontology coordination, nor an ontology coordination methodology or procedure, we have illustrated our ideas in a scenario taken from (Sowa, 2000) where one needs to coordinate different conceptualisations in the English and French language of the concepts of 'river' and 'stream' on one side, and 'fleuve' and 'reivière' on the other side.

Situated Semantic Alignment: Most ontology matching mechanisms developed so far have taken a classical functional approach to the semantic heterogeneity problem, in which ontology matching is seen as a process taking two or more ontologies as input and producing a semantic alignment of ontological entities as output (Giunchiglia and Shvaiko, 2004). Furthermore, matching often has been carried out at design-time, before integrating knowledge-based systems or making them interoperate. But, multi-agent communication, peer-to-peer information sharing, and web-service composition are all of a decentralised, dynamic, and open-ended nature, and they require ontology matching to be locally performed

during run-time. In addition, in many situations peer ontologies are not even open for inspection (e.g., when they are based on commercially confidential information). (Atencia and Schorlemmer, 2007) claim that a semantic alignment of ontological terminology is ultimately relative to the particular situation in which the alignment is computed, and that this situation should be made explicit and brought into the alignment mechanism. Even two agents with identical conceptualisation capabilities, and using exactly the same vocabulary to specify their respective conceptualisations, may fail to interoperate in a concrete situation because of their differing perception of the domain. They address the case in which agents are already endowed with a top-down engineered ontology (it can even be the same one), which they do not adapt or refine, but for which they want to find the semantic relationships with separate ontologies of other agents on the grounds of their communication within a specific situation. In particular, they provide a formal model that formalises situated semantic alignment as a sequence of information-channel refinements capturing the flow of information occurring in distributed systems due to the particular situations—or tokens—that carry information. Analogously, the semantic alignment that will allow information to flow ultimately will be carried by the particular situation agents are acting in (Atencia and Schorlemmer, 2008).

6 CONCLUSIONS

We have approached the limits of ontology-based semantic alignment from its mathematical foundations and in the context of alignment scenarios in open and distributed environments, like the Web, and its extension, the Semantic Web. We argued for the need to address what we believe is still a lack of sound mathematical models of information, semantics, and interoperability for multi-agent systems, and distributed knowledge models on the Semantic Web (Kalfoglou et al., 2004). We showed that we needed to go beyond the usual approach, which models semantic alignment as the first-order interpretation of dissimilar vocabularies into a common ontology.

We propose a general theory of semantic integration that uses a logic-independent formulation of language, ontology, and ontological commitment that can cope with the variety of logics and understandings of semantics occurring in highly decentralised and distributed environments. Furthermore, our proposed theory defines semantic alignment on top of this logic-independent formulation by means of channel theory. In particular we have shown that the natural logic of the core of an information channel adequately and faithfully captures the intuitive consequence relation lying behind semantically aligned systems. This led us to advocate for a channel-theoretic characterisation of semantic alignment that we stated in the form of three *Semantic Alignment Hypotheses*. Such channel-theoretic characterisation allowed us to look beyond the standard ontology-based approach to semantic alignment, and we illustrated this by means of interaction-based meaning coordination between agents.

By providing a sound theoretical ground upon which we base our three hypotheses for enabling semantic alignment, we enable the use of our framework to model semantic-alignment as it occurs in semantic heterogeneity scenarios by applying a variety of technologies. Instead of exploring concrete instantiations of the formal model to particular alignment technologies—wandering into the discussion of particular choice methods, termination criteria, and alignment algorithms—we decided to shift our attention to what basic capability an agent should have to be able to engage in an ontology-alignment interaction. Choice of tokens and types, interaction termination criteria, and concrete matching algorithms will play a central role when grounding the formal model in concrete

domains. This has been explored in two exemplar uses of our work: progressive ontology alignment and situated semantic alignment.

Acknowledgements: This work is supported under the Advanced Knowledge Technologies (AKT) Interdisciplinary Research Collaboration (IRC), sponsored by the UK Engineering and Physical Sciences Research Council under grant number GR/N15764/01; under the UPIC project, sponsored by Spain's Ministry of Education and Science under grant number TIN2004-07461-C02-02; and under the OpenKnowledge Specific Targeted Research Project (STREP), sponsored by the European Commission under contract number FP6-027253. M. Schorlemmer is also supported by a *Ramon y Cajal* Research Fellowship from Spain's Ministry of Education and Science, partially funded by the European Social Fund.

References

Atencia, M., and Schorlemmer, M. (2007). *A formal model for situated semantic alignment*. In Durfee, E. H., and Yokoo, M., editors, Proceedings of the Sixth International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS 2007), pp. 1270-1277.

Atencia, M., and Schorlemmer, M. (2008). *I-SSA: Interaction-situated semantic alignmnet*. Proceedings of the Sixteenth International Conference on Cooperative Information Systems (CoopIS 2008), Lecture Notes in Computer Science. Springer.

Bailin, S., and Truszkowski, W. (2002). *Ontology negotiation between intelligent information agents.* The Knowledge Engineering Review, 17(1):7–19.

Barr, M. (1996). *The Chu construction*. Theory and Applications of Categories, 2(2):17–35.

Barwise, J., and Seligman, J. (1997). *Information Flow: The Logic of Distributed Systems*, volume 44 of Cambridge Tracts in Theoretical Computer Science. Cambridge University Press.

Borst, P., Akkermans, H., and Top, J. (1997). *Engineering ontologies.* International Journal of Human-Computer Studies, 46(2–3):365–406.

Bouquet, P., Giunchiglia, F., van Harmelen, F., Serafini, L., and Stuckenschmidt, H. (2003). *C-OWL: Contextualizing ontologies*. In Fensel, D., et al., editors, The Semantic Web – ISWC 2003, volume 2870 of Lecture Notes in Computer Science, pages 164–179. Springer.

Calvanese, D. and De Giacomo, G. (2005). *Data integration: A logic-based perspective.* AI Magazine, 26(1):59–70.

Corrêa da Silva, F., and Agusti., J. (2003). *Knowledge Coordination*. Wiley.

Dunn, J. M., and Hardegree, G. M. (2001). *Algebraic Methods in Philosophical Logic*. Oxford University Press.

Enderton, H. (2002). *A Mathematical Introduction to Logic*. Academic Press, 2nd edition.

Farquhar, A., Fikes, R., and Rice, J. (1997). *The Ontolingua Server: a tool for collaborative ontology construction*. International Journal of Human-Computer Studies, 46(6):707–727.

Ganter, B., and Wille, R. (1999). *Formal Concept Analysis*. Springer.

Giunchiglia, F., Marchese, M., and Zaihrayeu, I. (2006). *Encoding classifications into lightweight ontologies.* In Sure, Y. and Domingue, J., editors, The Semantic Web: Research and Applications, volume 4011 of Lecture Notes in Computer Science, pages 80–94. Springer.

Giunchiglia, F., and Shvaiko,P. (2004). *Semantic matching*. The Knowledge Engineering Review, 18(3):265–280.

Goguen, J., and Burstall, R. (1992). *Institutions: Abstract model theory for specification and programming*. Journal of the ACM, 39(1):95–146.

Grüninger, M. and Kopena, J. B. (2005). *Semantic integration through invariants.* AI Magazine, 26(1):11–20.

IEEE (2003, December 18). *Standard Upper Ontology Working Group (SUO WG)*. Last retrieved on September 15, 2007, from http://suo.ieee.org.

Kalfoglou, Y., Alani, H., Schorlemmer, M., and Walton, C. (2004). *On the emergent semantic web and overlooked issues.* In, McIlraith, S., et al., editors, The Semantic Web – ISWC 2004, volume 3298 of Lecture Notes in Computer Science, pages 576–590. Springer

Kalfoglou, Y., and Schorlemmer, M. (2003a). *IF-Map: An ontology-mapping method based on information-flow theory*. In Spaccapietra, S., et al., editors, Journal on Data Semantics I, volume 2800 of Lecture Notes in Computer Science, pages 98–127. Springer.

Kalfoglou, Y., and Schorlemmer, M. (2003b). *Ontology mapping: The sate of the art*. The Knowledge Engineering Review, 18(1):1–31.

Kalfoglou, Y., and Schorlemmer, M. (2004). *Formal support for representing and automating semantic inter-operability*. In Bussler, C., et al., editors, The Semantic Web: Research and Applications, volume 3053 of Lecture Notes in Computer Science, pages 45–60. Springer.

Kent, R. E. (2000). *The information flow foundation for conceptual knowledge organization*. In Proceedings of the Sixth International Conference of the International Society for Knowledge Organization.

Lenat, D. (1995). *CyC: A large-scale investment in knowledge infrastructure*. Communications of the ACM, 38(11).

McGuinness, D. and van Harmelen, F. (2004). *OWL web ontology language overview*. Technical report, World Wide Web Consortium (W3C). Last retrieved on September 15, 2008, from http://www.w3.org/TR/2004/REC-owl-features-20040210/.

Pratt, V. (1995). *The Stone gamut: A coordintization of mathematics*. In Proceedings of the Tenth Annual Symposium on Logic in Computer Science, pages 444–454. IEEE Computer Society Press.

Pratt, V. (2001). *Orthocurrence as both interaction and observation*. In Rodriguez, R. and Anger, F., editors, Proceedings of the IJCAI'01Workshop on Spatial and Temporal Reasoning.

Schorlemmer, M., and Kalfoglou, Y. (2003). *On semantic interoperability and the flow of information*. In Doan, A., Halevy, A., and Noy, N., editors, Semantic Integration, volume 82 of CEUR Workshop Proceedings.

Schorlemmer, M., and Kalfoglou, Y. (2004). *A Channel-theoretic foundation for ontology coordination.* In Proceedings of the ISWC'04 Workshop on Meaning Coordination and Negotiation.

Schorlemmer, M., and Kalfoglou,Y. (2005). *Progressive ontology alignment for meaning coordination: An information-theoretic foundation*. In Dignum, F. et al., editors, Proceedings of the Fourth International Joint Conference on Autonomous Agents and Multiagent Systems, pages 737–744. ACM Press.

Shvaiko, P., and Euzenat, J. (2005). *A survey of schema-based matching approaches*. In Spaccapietra, S., et al., editors, Journal on Data Semantics IV, volume 3730 of Lecture Notes in Computer Science, pages 146–171. Springer.

Sowa, J. F. (2000). *Knowledge Representation: Logical, Philosophical, and Computational Foundations*. Brooks/Cole.

Uschold, M., Healy, M., Williamson, K., Clark, P., and Woods, S. (1998). *Onology reuse and application.* In Guarino, N. , editor, Formal Ontology in Information Systems, volume 46 of Frontiers in Artificial Intelligence and Applications, pages 179–192. IOS Press.

van Diggelen, J., Beun, R.-J., Dignum, F., van Eijk, R. M., and Meyer, J.-J. (2007). *Ontology negotiation: goals, requirements and implementation*. International Journal of Agent-Oriented Software Engineering 1(1):63-90.

Wang, J., and Gasser, L. (2002). *Mutual online ontology alignment.* In OAS'02 Ontologies in Agent Systems, Proceedings of the AAMAS 2002 Workshop, volume 66 of CEUR Workshop Proceedings.

Zimmermann, A., Krötzsch, M., Euzenat, J., and Hitzler, P. (2006). *Formalizing ontology alignment and its operations with category theory.* In Bennett, B. and Fellbaum C., editors, Formal Ontology in Information Systems, Proceedings of the Fourth International Conference (FOIS 2006), volume 150 of Frontiers in Artificial Intelligence and Applications. IOS Press.