Formal Support for Representing and Automating Semantic Interoperability

Yannis Kalfoglou[1] and Marco Schorlemmer[2]

[1] School of Electronics and Computer Science, University of Southampton
[2] E. U. de Tecn. d'Informació i Comunicació, Universitat Internacional de Catalunya

Abstract. We discuss approaches to semantic heterogeneity and propose a formalisation of semantic interoperability based on the Barwise-Seligman theory of information flow. We argue for a theoretical framework that favours the analysis and implementation of semantic interoperability scenarios relative to particular understandings of semantics. We present an example case of such a scenario where our framework has been applied as well as variations of it in the domain of ontology mapping.

1 Introduction

The problem of resolving semantic heterogeneity has been identified in the past in the field of federated databases as one of the core challenges for achieving semantic interoperability [22]. Despite collective efforts from researchers and industrialists it remains largely unsolved. Recently, the same challenge surfaced again in a different context, that of the Semantic Web. It has different characteristics though, which make it even harder to tackle, because we are dealing with a distributed and deregulated environment where the assurances of a strictly monitored database management system no longer hold.

One of the core premises of the Semantic Web vision is that systems should be able to exchange information and services with one another in semantically rich and sound manners [5]. The semantics of one system should therefore be exposed to the environment in such a way that other systems can interpret it correctly and use it to achieve interoperability, which is vital for distributed reasoning in order to support applications and services alike. However, there are numerous ways of expressing, exposing and understanding semantics, which leads to heterogeneity, more specifically, semantic heterogeneity. Lessons learned from previous attempts to resolve semantic heterogeneity—and also from peripheral areas where inconsistency has shown that semantic heterogeneity is an endemic characteristic of distributed systems and we should learn to live with it [10]—has prompted us to look at this challenge from another angle: to achieve the necessary and sufficient semantic interoperability even if it means that we will not resolve semantic heterogeneity completely.

To understand the necessary and sufficient conditions for achieving semantic interoperability we need to look what the minimal requirements for interoperability are. For two systems to interoperate there must be an established form of communication and the right means to achieve this efficiently and effectively.

An established form of communication clearly resembles the idea of agreed standards, and there has been considerable effort in the knowledge engineering community to come up with the right technologies for enforcing them. Ontologies are among the most popular ones, which act at the protocol to which systems have to adhere in order to establish interoperability. Although ontologies provide the means to establish communication efficiently there are not always effective. The crux of the problem is the increasing proliferation of domain and application ontologies on the Semantic Web, and, since they were built independently by distinct groups, they are semantically heterogeneous, hence outweighing the benefits of having an ontology in the first place. Enforcing a single standard ontology (or a set of standard ontologies) could alleviate the problem, but history of computing has taught us that this is a long process with arguable results. If we accept that ontologies are necessary for expressing and exposing semantics of systems and domains to the Semantic Web, then we have to anticipate different versions of them, semantically heterogeneous ones, which have to be shared in order to achieve interoperability.

2 Semantic Interoperability and Integration

Semantic interoperability and semantic integration are much contested and fuzzy concepts, which have been used over the past decade in a variety of contexts and works. As reported in [21], in addition, both terms are often used indistinctly, and some view these as the same thing.

The ISO/IEC 2382 Information Technology Vocabulary defines interoperability as “the capability to communicate, execute programs, or transfer data among various functional units in a manner that requires the user to have little or no knowledge of the unique characteristics of those units.” In a debate on the mailing list of the IEEE Standard Upper Ontology working group, a more formal approach to semantic interoperability was advocated: to use logic in order to guarantee that, after data were transmitted from a sender system to a receiver, all implications made by one system had to hold and be provable by the other, and that there should be a logical equivalence between those implications.[3]

With respect to integration, Uschold and Grüninger argue that “two agents are semantically integrated if they can successfully communicate with each other” and that “successful exchange of information means that the agents understand each other and there is guaranteed accuracy” [26]. According to Sowa, to integrate two ontologies means to derive a new ontology that facilitates interoperability between systems based on the original ontologies, and he distinguishes three levels of integration [23]: *Alignment*—a mapping of concepts and relations to indicate equivalence—, *partial compatibility*—an alignment that supports equivalent inferences and computations on equivalent concepts and relations—, and *unification*—a one-to-one alignment of all concepts and relations that allows any inference or computation expressed in one ontology to be mapped to an equivalent inference or computation in the other ontology.

[3] Message thread on the SUO mailing list initiated at http://suo.ieee.org/email/msg07542.html.

The above definitions reveal a common denominator, that of *communication.* As we said in the introduction, since ontologies have been established as the preferable means for supporting communication, the research issue is the following: *Having established a protocol to which communication will be based, i.e., ontologies, what is the best way to effectively make those semantically interoperable?*

A practical angle of viewing this problem is to focus on the notion of equivalence. That is, we would like to establish some sort of correspondence between the systems and, subsequently, their ontologies, to make them interoperable; this could be done by reasoning about equivalent constructs of the two ontologies. However, equivalence is not a formally and consensually agreed term, neither do we have mechanisms for doing that. Hence, if we are to provide a formal, language-independent mechanism of semantic interoperability and integration, we need to use some formal notion of equivalence. And for a precise approximation to equivalence the obvious place to look at is Logic.

In this sense first-order logic seems the natural choice: among all logics it has a special status due to its expressive power, its natural deductive systems, and its intuitive model theory based on sets. In first-order logic, equivalence is approximated via the precise model-theoretic concept of *first-order equivalence.* This is the usual approach to formal semantic interoperability and integration; see e.g., [4, 6, 20, 26] and also those based on Description Logics [1]. In Ciocoiu and Nau's treatment of the translation problem between knowledge sources that have been written in different knowledge representation languages, semantics is specified by means of a common ontology that is expressive enough to interpret the concepts in all agents' ontologies [6]. In that scenario, two concepts are equivalent if, and only if, they share exactly the same subclass of first-order models of the common ontology.

But this approach also has its drawbacks. First, such formal notion of equivalence requires the entire machinery of first-order model theory, which includes set theory, first-order structures, interpretation, and satisfaction. This appears to be heavyweight for certain interoperability scenarios. Madhavan et al. define the semantics in terms of instances in the domain [16]. This is also the case, for example, in Stumme and Maedche's ontology merging method, FCA-Merge [24], where the semantics of a concept symbol is captured through the instances classified to that symbol. These instances are documents, and a document is classified to a concept symbol if it contains a reference that is relevant to the concept. For FCA-Merge, two concepts are considered equivalent if, and only if, they classify exactly the same set of documents. Menzel makes similar objections to the use of first-order equivalence and proposes an axiomatic approach instead, inspired on property theory [25], where entailment and equivalence are not model-theoretically defined, but axiomatised in a logical language for ontology theory [19].

Second, since model-theory does not provide proof mechanisms for checking model equivalence, this has to be done indirectly via those theories that specify the models. This assumes that the logical theories captured in the ontologies are complete descriptions of the intended models (Uschold and Grüninger call these *verified ontologies* [26]), which will seldom be the case in practice. Furthermore, Corrêa da Silva et al. have shown situations in which even a common verified ontology is not enough, for example when a knowledge base whose inference engine is based on linear logic poses a query to a knowledge base with the same ontology, but whose inference engine is based on relevance logic [7]. The former should not accept answers as valid if the inference carried out in order to answer the query was using the contraction inference rule, which is not allowed in linear logic. Here, two concepts will be equivalent if, and only if, we can infer exactly the same set of consequences on their distinct inference engines.

A careful look at the several formal approaches to semantic integration mentioned above reveals many different understandings of semantics depending on the interoperability scenario under consideration. Hence, what we need in order to successfully tackle the problem of semantic interoperability is not so much a framework that establishes a particular semantic perspective (model-theoretic, property-theoretic, instance-based, etc.), but instead we need a framework that successfully captures semantic interoperability despite the different treatments of semantics.

3 An Approach Based on Information-Flow Theory

We observe that, in order for two systems to be semantically interoperable (or semantically integrated) we need to align and map their respective ontologies such that *the information can flow*. Consequently, we believe that a satisfactory formalisation of semantic interoperability can be built upon a mathematical theory capable of describing under which circumstances information flow occurs.

Although there is no such theory yet, there have been many notable efforts [9, 8, 3]. A good place to start establishing a foundation for formalising semantic interoperability is Barwise and Seligman's channel theory, a mathematical model that aims at establishing the laws that govern the flow of information. It is a general model that attempts to describe the information flow in any kind of distributed system, ranging form actual physical systems like a flashlight connecting a bulb to a switch and a battery, to abstract systems such as a mathematical proof connecting premises and hypothesis with inference steps and conclusions.

A significant effort to develop a framework around the issues of organising and relating ontologies based on channel theory is Kent's Information Flow Framework (IFF) [14], which is currently developed by the IEEE Standard Upper Ontology working group as a meta-level foundation for the development of upper ontologies[13].

3.1 IF Classification, Infomorphism, and Channel

In channel theory, each component of a distributed system is represented by an *IF classification* $\mathbf{A} = \langle tok(\mathbf{A}), typ(\mathbf{A}), \models_{\mathbf{A}} \rangle$, consisting 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.[4] It is a very simple mathematical structure that effectively captures the local syntax and semantics of a community for the purpose of semantic interoperability.

For the problem that concerns us here the components of the distributed systems are the ontologies of the communities that desire to communicate. We model them as IF classifications, such that the syntactic expressions that a community uses to communicate constitute the types of the IF classification, and the meaning that these expressions take within the context of the community are represented by the way tokens are classified to types. Hence, *the semantics is characterised by what we choose to be the tokens of the IF classification*, and depending on the particular semantic interoperability scenario we want to model, types, tokens, and its classification relation will vary. For example, in FCA-Merge [24], types are concept symbols and tokens particular documents, while in Ciocoiu and Nau's scenario [6] types are expressions of knowledge representation languages and tokens are first-order structures. The crucial point is that *the semantics of the interoperability scenario crucially depends on our choice of types, tokens and their classification relation for each community.*

The flow of information between components in a distributed system is modelled in channel theory by the way the various IF classifications that represent the vocabulary and context of each component are connected with each other through *infomorphisms*. An infomorphism $f = \langle f\hat{}, f\check{} \rangle : \mathbf{A} \rightleftarrows \mathbf{B}$ from IF classifications $\mathbf{A}$ to $\mathbf{B}$ is a contravariant pair of functions $f\hat{} : typ(\mathbf{A}) \to typ(\mathbf{B})$ and $f\check{} : tok(\mathbf{B}) \to tok(\mathbf{A})$ satisfying, for each type $\alpha \in typ(\mathbf{A})$ and token $b \in tok(\mathbf{B})$, the fundamental property that $f\check{}(b) \models_{\mathbf{A}} \alpha$ iff $b \models_{\mathbf{B}} f\hat{}(\alpha)$:[5]

$$\begin{tikzcd}
\alpha \arrow[r, mapsto, "f\hat{}"] \arrow[d, dash, "\models_{\mathbf{A}}"'] & f\hat{}(\alpha) \arrow[d, dash, "\models_{\mathbf{B}}"] \\
f\check{}(b) & b \arrow[l, mapsto, "f\check{}"]
\end{tikzcd}$$

A *distributed IF system* $\mathcal{A}$ consists then of an indexed family $cla(\mathcal{A}) = \{\mathbf{A}_i\}_{i \in I}$ of IF classifications together with a set $inf(\mathcal{A})$ of infomorphisms all having both domain and codomain in $cla(\mathcal{A})$.

A basic construct of channel theory is that of an *IF channel*—two IF classifications $\mathbf{A}_1$ and $\mathbf{A}_2$ connected through a core IF classification $\mathbf{C}$ via two infomorphisms f_1 and f_2:

$$\begin{tikzcd}
 & typ(\mathbf{C}) \arrow[dd, dash, "\models_{\mathbf{C}}"] & \\
typ(\mathbf{A}_1) \arrow[ur, "f\hat{}_1"] \arrow[dd, dash, "\models_{\mathbf{A}_1}"'] & & typ(\mathbf{A}_2) \arrow[ul, "f\hat{}_2"'] \arrow[dd, dash, "\models_{\mathbf{A}_2}"] \\
 & tok(\mathbf{C}) \arrow[dl, "f\check{}_1"] \arrow[dr, "f\check{}_2"'] & \\
tok(\mathbf{A}_1) & & tok(\mathbf{A}_2)
\end{tikzcd}$$

[4] We are using the prefix 'IF' (information flow) in front of some channel-theoretic constructions to distinguish them from their usual meaning.

[5] Such contravariance is a recurrent theme in logic and mathematics and has been thoroughly studied within the context of Chu spaces [2, 12]; it also underlies the mathematical theory of concept formation [11].

This basic construct captures the information flow between components $\mathbf{A}_1$ and $\mathbf{A}_2$. Note that, in Barwise and Seligman's model it is the particular tokens that carry information and that information flow crucially involves both types and tokens.

In fact, our approach uses this model to approximate the intuitive notion of equivalence necessary for achieving semantic interoperability with the precise notion of a type equivalence that is supported by the connection of tokens from $\mathbf{A}_1$ with tokens from $\mathbf{A}_2$ through the tokens of the core IF classification $\mathbf{C}$. This provides us with the general framework of semantic interoperability we are after, one that accommodates different understandings of semantics—depending on the particularities of the interoperability scenario—whilst retaining the core aspect that will allow communication among communities: a connection through their semantic tokens.

The key channel-theoretic construct we are going to exploit in order to outline our formal framework for semantic interoperability is that of a *distributed IF logic*. This is the logic that represents the information flow occurring in a distributed system. In particular we will be interested in a restriction of this logic to the language of those communities we are attempting to integrate. As we proceed, we will hint at the intuitions lying behind the channel-theoretical notions we are going to use; for a more in-depth understanding of channel theory we point the interested reader to [3].

3.2 IF Theory and Logic

Suppose two communities $\mathbf{A}_1$ and $\mathbf{A}_2$ need to interoperate, but are using different ontologies. To have them semantically interoperating will mean to know the semantic relationship in which they stand to each other. In terms of the channel-theoretic context, this means to know an *IF theory* that describes how the different types from $\mathbf{A}_1$ and $\mathbf{A}_2$ are logically related to each other.

Channel theory has been developed based on the understanding that information flow results from regularities in a distributed system: information of some components of a system carries information of other components because of the regularities among the connections. These regularities are implicit in the representation of the systems' components and its connections as IF classifications and infomorphisms, but in order to derive a notion of equivalence on the type-level of the system we need to capture this regularity in a logical fashion. This is achieved with IF theories and IF logics in channel theory.

An *IF theory* $T = \langle typ(T), \vdash \rangle$ consists of a set $typ(T)$ of types, and a binary relation $\vdash$ between subsets of $typ(T)$. Pairs $\langle \Gamma, \Delta \rangle$ of subsets of $typ(T)$ are called *sequents*. If $\Gamma \vdash \Delta$, for $\Gamma, \Delta \subseteq typ(T)$, then the sequent $\Gamma \vdash \Delta$ is called a *constraint*. T is *regular* if for all $\alpha \in typ(T)$ and all sets $\Gamma, \Gamma', \Delta, \Delta', \Sigma', \Sigma_0, \Sigma_1$ of types:

1. *Identity:* $\alpha \vdash \alpha$
2. *Weakening:* If $\Gamma \vdash \Delta$, then $\Gamma, \Gamma' \vdash \Delta, \Delta'$

3. *Global Cut:* If $\Gamma, \Sigma_0 \vdash \Delta, \Sigma_1$ for each partition $\langle \Sigma_0, \Sigma_1 \rangle$ of Σ', then $\Gamma \vdash \Delta$.[6]

Regularity arises from the observation that, given any classification of tokens to types, the set of all sequents that are satisfied[7] by all tokens always fulfill these three properties. In addition, given a regular IF theory T we can generate a classification $Cla(T)$ that captures the regularity specified in its constraints. Its tokens are partitions $\langle \Gamma, \Delta \rangle$ of $typ(T)$ that are *not* constraints of T, and types are the types of T, such that $\langle \Gamma, \Delta \rangle \models_{Cla(T)} \alpha$ iff $\alpha \in \Gamma$.[8]

The IF theory we are after in order to capture the semantic interoperability between communities $\mathbf{A}_1$ and $\mathbf{A}_2$ is an IF theory on the union of types $typ(\mathbf{A}_1) \cup typ(\mathbf{A}_2)$ that respects the local IF classification systems of each community—the meaning each community attaches to its expressions—but also interrelates types whenever there is a similar semantic pattern, i.e., a similar way communities classify related tokens. This is the type language we speak in a semantic interoperability scenario, because we want to know when type α of one component corresponds to a type β of another component. In such an IF theory a sequent like $\alpha \vdash \beta$, with $\alpha \in typ(\mathbf{A}_1)$ and $\beta \in typ(\mathbf{A}_2)$, would represent an implication of types among communities that is in accordance to how the tokens of different communities are connected between each other. Hence, a constraint $\alpha \vdash \beta$ will represent that every α is a β, together with a constraint $\beta \vdash \alpha$ we obtain type equivalence.

Putting the idea of an IF classification with that of an IF theory together we get an *IF logic* $\mathfrak{L} = \langle tok(\mathfrak{L}), typ(\mathfrak{L}), \models_{\mathfrak{L}}, \vdash_{\mathfrak{L}} \rangle$. It consists of an IF classification $cla(\mathfrak{L}) = \langle tok(\mathfrak{L}), typ(\mathfrak{L}), \models_{\mathfrak{L}} \rangle$ and a regular IF theory $th(\mathfrak{L}) = \langle typ(\mathfrak{L}), \vdash_{\mathfrak{L}} \rangle$, such that all tokens $tok(\mathfrak{L})$ satisfy all constraints of $th(\mathfrak{L})$;[9] a token $a \in tok(\mathfrak{L})$ satisfies a constraint $\Gamma \vdash \Delta$ of $th(\mathfrak{L})$ if, when a is of all types in Γ, a is of some type in Δ.

3.3 Distributed IF Logic

The sought after IF theory is the IF theory of the distributed IF logic of an IF channel

$$\begin{tikzcd} & \mathbf{C} & \\ \mathbf{A}_1 \arrow[ur, "f_1"] & & \mathbf{A}_2 \arrow[ul, "f_2"'] \end{tikzcd}$$

that represents the information flow between $\mathbf{A}_1$ and $\mathbf{A}_2$. This channel can either be stated directly, or indirectly by some sort of partial alignment of $\mathbf{A}_1$ and $\mathbf{A}_2$ (as we show, e.g., in Section 4.2).

The logic we are after is the one we get from *moving* a logic on the core $\mathbf{C}$ of the channel to the sum of components $\mathbf{A}_1 + \mathbf{A}_2$: The IF theory will be induced

[6] 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 $\Sigma_0 \cap \Sigma_1 = \emptyset$; Σ_0 and Σ_1 may themselves be empty (hence it is actually a quasi-partition).

[7] Defined further below.

[8] These tokens may not seem obvious, but these sequents code the content of the classification table: The left-hand sides of the these sequents indicate to which types they are classified, while the right-hand sides indicate to which they are not.

[9] Properly speaking this is the definition of a *sound* IF logic. Channel theory has room for unsound IF logics, but they are not needed for the purpose of this paper.

at the core of the channel; this is crucial. The distributed IF logic is the *inverse image* of the IF logic at the core.

Given an infomorphism $f : \mathbf{A} \rightleftarrows \mathbf{B}$ and an IF logic $\mathfrak{L}$ on $\mathbf{B}$, the *inverse image* $f^{-1}[\mathfrak{L}]$ of $\mathfrak{L}$ under f is the IF logic on $\mathbf{A}$, whose theory is such that $\Gamma \vdash \Delta$ is a constraint of $th(f^{-1}[\mathfrak{L}])$ iff $f\hat{}[\Gamma] \vdash f\hat{}[\Delta]$ is a constraint of $th(\mathfrak{L})$.

The type and tokens system at the core and the IF classification of tokens to types will determine the IF logic at this core. We usually take the *natural IF logic* as the IF logic of the core, which is the IF logic $Log(\mathbf{C})$ generated from an IF classification $\mathbf{C}$: its classification is $\mathbf{C}$ and its regular theory is the theory whose constraints are the sequents satisfied by all tokens. This seems natural, and is also what happens in the various interoperability scenarios we have been investigating.

Given an IF channel $\mathcal{C} = \{f_{1,2} : \mathbf{A}_{1,2} \rightleftarrows \mathbf{C}\}$ and an IF logic $\mathfrak{L}$ on its core $\mathbf{C}$, the *distributed IF logic*, $DLog_{\mathcal{C}}(\mathfrak{L})$, is the inverse image of $\mathfrak{L}$ under the sum infomorphisms $f_1 + f_2 : \mathbf{A}_1 + \mathbf{A}_2 \rightleftarrows \mathbf{C}$. This sum is defined as follows: $\mathbf{A}_1 + \mathbf{A}_2$ has as set of tokens the Cartesian product of $tok(\mathbf{A}_1)$ and $tok(\mathbf{A}_2)$ and as set of types the disjoint union of $typ(\mathbf{A}_1)$ and $typ(\mathbf{A}_2)$, such that for $\alpha \in typ(\mathbf{A}_1)$ and $\beta \in typ(\mathbf{A}_2)$, $\langle a, b\rangle \models_{\mathbf{A}_1+\mathbf{A}_2} \alpha$ iff $a \models_{\mathbf{A}_1} \alpha$, and $\langle a, b\rangle \models_{\mathbf{A}_1+\mathbf{A}_2} \beta$ iff $b \models_{\mathbf{A}_2} \beta$. Given two infomorphisms $f_{1,2} : \mathbf{A}_{1,2} \rightleftarrows \mathbf{C}$, the sum $f_1 + f_2 : \mathbf{A}_1 + \mathbf{A}_2 \rightleftarrows \mathbf{C}$ is defined by $(f_1 + f_2)\hat{}(\alpha) = f_i(\alpha)$ if $\alpha \in \mathbf{A}_i$ and $(f_1 + f_2)\check{}(c) = \langle f\check{}_1(c), f\check{}_2(c)\rangle$, for $c \in tok(\mathbf{C})$.

4 Representing Semantic Interoperability

In this section we illustrate, by means of an example, our approach to semantic interoperability via IF channels. Suppose that we are dealing with a situation where an agent or a group of agents (human or artificial) are faced with the task of aligning organisational structures and responsibilities of ministries across different governments. This is a realistic scenario set out in the domain of e-governments. Our agents have to align UK and US governments, by focusing on governmental organisations, like ministries. The focal point of this alignment is not only the structural and taxonomic differences of these ministries but the way in which responsibilities are allocated in different departments and offices within these ministries. This constitutes the semantics of our interoperability scenario, and consequently this will determine our choice of types, tokens and their classification relation for each community, as already pointed out in Section 3.1.

For the sake of brevity and space reasons, we only describe here four ministries: The UK Foreign and Commonwealth Office, the UK Home Office, the US Department of State, the US Department of Justice (hereafter, FCO, HO, DoS and DoJ, respectively). We gathered information related to these ministries from their web sites[10] where we focused on their organisational structures, assuming that the meaning of these structures is in accordance to the separation

[10] Accessible from www.homeoffice.gov.uk, www.fco.gov.uk, www.state.gov and www.usdoj.gov.

of responsibilities. These structures were trivial to extract, either from the hierarchical lists of departments, agencies, bureau, directorates, divisions, offices (which we shall commonly refer to as *units*) within these ministries, or organisational charts and organograms publicly available on the Web. The extraction of responsibilities and their units though, requires an intensive manual knowledge acquisition exercise. At the time of our experiments, the ministries' taxonomies ranged from 38 units comprising the US DoJ to 109 units for the UK HO.

In order to capture semantic interoperability via IF channels we devised the following four steps:

1. define the various contexts of each community by means of a distributed IF system of IF classifications;
2. define an IF channel—its core and infomorphisms—connecting the IF classifications of the various communities;
3. define an IF logic on the core IF classification of the IF channel that represents the information flow between communities;
4. distribute the IF logic to the sum of community IF classifications to obtain the IF theory that describes the desired semantic interoperability.

These steps illustrate a theoretical framework and need not to correspond to actual engineering steps; but we claim that a sensible implementation of semantic interoperability can be achieved following this framework, as it constitutes the theoretical foundation of a semantic interoperability scenario. In fact, we have proposed an IF-based method to assist in ontology mapping [15], and in Section 5 we briefly discuss how it relates to this framework.

4.1 Community IF Classifications

UK and US governments use different ontologies to represent their respective ministries; therefore, we shall be dealing with two separate sets of types, $typ(\mathbf{UK}) = \{\mathsf{FCO},\mathsf{HO}\}$ and $typ(\mathbf{US}) = \{\mathsf{DoS},\mathsf{DoJ}\}$. We model the interoperability scenario using a separate IF classification for each government, **UK** and **US**, whose types are ministries.

To have UK and US ministries semantically interoperable will mean to know the semantic relationship in which they stand to each other, which we take, in this particular scenario, to be their set of responsibilities. It is sensible to assume that there will be no obvious one-to-one correspondence between ministries of two governments because responsibilities of a ministry in one government may be spread across many ministries of the other, and vice versa. But we can attempt to derive an IF theory that describes how the different ministry types are logically related to each other—an IF theory on the union of ministry types $typ(\mathbf{UK}) \cup typ(\mathbf{US})$ in which a constraint like $\mathsf{FCO} \vdash \mathsf{DoS}$ would represent the fact that a responsibility of the UK Foreign and Commonwealth Office is also a responsibility of the US Department of State.

We shall construct the IF channel that will allow us to derive the desired IF theory using the hierarchical structure of units shown in Figure 1. Within the context of one government, different ministries represent already the top-level

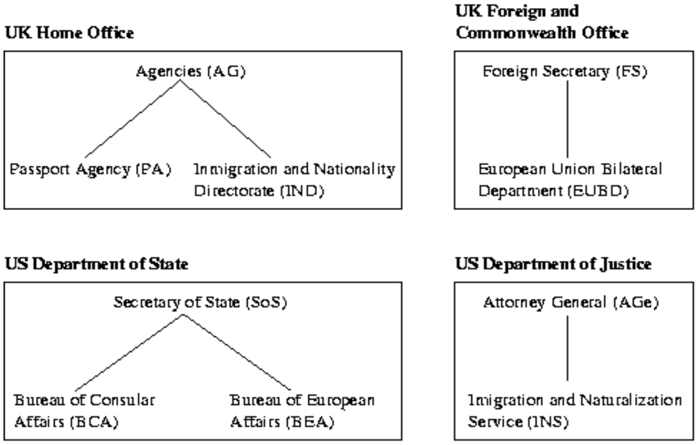
UK Home Office
Agencies (AG)
Passport Agency (PA)
Inmigration and Nationality Directorate (IND)
UK Foreign and Commonwealth Office
Foreign Secretary (FS)
European Union Bilateral Department (EUBD)
US Department of State
Secretary of State (SoS)
Bureau of Consular Affairs (BCA)
Bureau of European Affairs (BEA)
US Department of Justice
Attorney General (AGe)
Imigration and Naturalization Service (INS)

Fig. 1. Hierarchical structures of government ministries

separation of responsibilities. From the hierarchical structures we extract an IF theory on unit types for each government. Following are the two IF theories of UK and US units, respectively:

$$\begin{array}{rrcrr} \vdash \mathsf{AG,FS} & \mathsf{IND} \vdash \mathsf{AG} & \quad & \vdash \mathsf{SoS,AGe} & \mathsf{BEA} \vdash \mathsf{SoS} \\ \mathsf{AG,FS} \vdash & \mathsf{PA,IND} \vdash & & \mathsf{SoS,AGe} \vdash & \mathsf{BCA,BEA} \vdash \\ \mathsf{PA} \vdash \mathsf{AG} & \mathsf{EUE} \vdash \mathsf{FS} & & \mathsf{BCA} \vdash \mathsf{SoS} & \mathsf{INS} \vdash \mathsf{AGe} \end{array}$$

By extracting responsibilities from the units' web sites we are able to define an IF classification for each government whose tokens are responsibilities and types are ministry units, and then classify responsibilities to their respective units. In the table below, we list the extracted responsibilities for both UK and US ministries along with their IDs, which we will use in sequel for the sake of brevity.

ID	UK responsibilities	ID	US responsibilities
r_1	issues UK passports	s_1	US passport services and information
r_2	regulate entry and settlement in the UK	s_2	promotes US interests in the region
r_3	executive services of the HO	s_3	heading the DoS
r_4	promote productive relations	s_4	facilitate entry to the US
r_5	responsible for the work of FCO	s_5	supervise and direct the DoJ

The IF classifications will have to be in accordance to the hierarchy as represented in the IF theories. That is, if a responsibility is classified to a unit, it shall also be classified to all its supra-units. This can be done automatically. The IF classifications $\mathbf{A}_{UK}$ and $\mathbf{A}_{US}$ for UK and US units, respectively, along with their abbreviated responsibilities is as follows:

	AG	PA	IND	FS	EUE
r_1	1	1	0	0	0
r_2	1	0	1	0	0
r_3	1	0	0	0	0
r_4	0	0	0	1	1
r_5	0	0	0	1	0

	SoS	BCA	BEA	AGe	INS
s_1	1	1	0	0	0
s_2	1	0	1	0	0
s_3	1	0	0	0	0
s_4	0	0	0	1	1
s_5	0	0	0	1	0

To represent how ministry types (like FCO,HO, etc.) from the IF classification **UK** relate to the IF classification $\mathbf{A}_{UK}$ of ministerial units, we will use the *flip* $\mathbf{A}_{UK}^{\perp}$[11] of the IF classification table and its *disjunctive power* $\vee\mathbf{A}_{UK}^{\perp}$[12]. The flip classifies ministerial units to responsibilities, and for the UK case is shown in Figure 2 (a). The disjunctive power of this flip classifies ministerial units to sets of responsibilities, whenever at least one of its responsibilities are among those in the set. A fragment of this IF classification is shown in Figure 2 (b).

	r_1	r_2	r_3	r_4	r_5
AG	1	1	1	0	0
PA	1	0	0	0	0
IND	0	1	0	0	0
FS	0	0	0	1	1
EUE	0	0	0	1	0

(a)

	$\{r_1, r_2, r_3, r_4, r_5\}$	$\cdots$	$\{r_1, r_2, r_3\}$	$\cdots$	$\{r_4, r_5\}$
AG	1		1		0
PA	1		1		0
IND	1		1		0
FS	1		0		1
EUE	1		0		1

(b)

Fig. 2. Flip and disjunctive power of a classification

The way ministries relate to these sets of responsibilities can then be represented with an infomorphism $h_{UK} : \mathbf{UK} \rightleftarrows \vee\mathbf{A}_{UK}^{\perp}$; and each context for a government, with its ministries, their respective units, and hierarchy captured by an IF theory, is then represented as a distributed IF system of IF classifications. For the UK government this distributed system is $\mathbf{UK} \xrightarrow{h_{UK}} \vee\mathbf{A}_{UK}^{\perp} \xleftarrow{\eta_{\mathbf{A}_{UK}^{\perp}}} \mathbf{A}_{UK}^{\perp}$, with $h_{UK}(\mathsf{HO}) = \{r_1, r_2, r_3\}$ and $h_{UK}(\mathsf{FCO}) = \{r_4, r_5\}$.

4.2 The IF Channel

We construct an IF channel from a partial alignment of some of the responsibilities extracted from the ministerial units' web sites. This is the crucial aspect of the semantic interoperability, since it is the point where relations in meaning are established. We assume a partial alignment, that is, one where not all responsibilities r_1 to r_5 are related to responsibilities s_1 to s_5. In particular we shall assume the alignment of UK responsibilities r_1, r_2 and r_4 with US responsibilities s_1, s_4 and s_2. An agreed description of these responsibilities is the following:

- (a) passport services: $r_1 \longleftrightarrow s_1$
- (b) immigration control: $r_2 \longleftrightarrow s_4$
- (c) promote productive relations: $r_4 \longleftrightarrow s_2$

[11] The flip $\mathbf{A}^{\perp}$ of an IF classification $\mathbf{A}$ is the classification whose tokens are $typ(\mathbf{A})$ and types are $tok(\mathbf{A})$, such that $\alpha \models_{\mathbf{A}^{\perp}} a$ iff $a \models_{\mathbf{A}} \alpha$

[12] The disjunctive power $\vee\mathbf{A}$ of an IF classification $\mathbf{A}$ is the classification whose tokens are the same as $\mathbf{A}$, types are subsets of $typ(\mathbf{A})$, and given $a \in tok(\mathbf{A})$ and $\Phi \subseteq typ(\mathbf{A})$, $a \models_{\vee\mathbf{A}} \Phi$ iff $a \models_{\mathbf{A}} \sigma$ for some $\sigma \in \Phi$. There exists a natural embedding $\eta_{\mathbf{A}} : \mathbf{A} \rightleftarrows \vee\mathbf{A}$ defined by $\eta\hat{}_{\mathbf{A}}(\alpha) = \{\alpha\}$ and $\eta\check{}_{\mathbf{A}}(a) = a$, for each $\alpha \in typ(\mathbf{A})$ and $a \in tok(\vee\mathbf{A})$

The focus of this paper is not how this partial alignment is established; various heuristic mechanisms have been proposed in the literature (see e.g., [20]), as well as mapping methods based on information-flow theory (see [15] and Section 5). We assume that we have already applied one of those heuristics or methods and come up with the agreed descriptions given above.

The above partial alignment is a binary relation between $typ(\mathbf{A}^{\perp}_{UK})$ and $typ(\mathbf{A}^{\perp}_{US})$. In order to represent this alignment as a distributed IF system in channel theory, we decompose the binary relation into a couple of total functions $g\hat{}_{UK}, g\hat{}_{US}$ from a common domain $typ(\mathbf{A}) = \{a, b, c\}$. (For example $g\hat{}_{UK}(b) = r_2$ and $g\hat{}_{US}(b) = s_4$.) This will constitute the type-level of a couple of infomorphisms. We complete the alignment to a system of IF classifications $\mathbf{A}^{\perp}_{UK} \xleftarrow{g_{UK}} \mathbf{A} \xrightarrow{g_{US}} \mathbf{A}^{\perp}_{US}$ by generating the IF classification on $typ(\mathbf{A})$ with all possible tokens, which we generate formally, and their classification. To satisfy the fundamental property of infomorphisms, the token-level of g_{UK}, g_{US} must be as follows:

	a	b	c
n_0	0	0	0
n_1	0	0	1
n_2	0	1	0
n_3	0	1	1
n_4	1	0	0
n_5	1	0	1
n_6	1	1	0
n_7	1	1	1

$$\begin{array}{rr} g\check{}_{UK}(\mathsf{AG}) = n_6 & g\check{}_{US}(\mathsf{SoS}) = n_5 \\ g\check{}_{UK}(\mathsf{PA}) = n_4 & g\check{}_{US}(\mathsf{BCA}) = n_4 \\ g\check{}_{UK}(\mathsf{IND}) = n_2 & g\check{}_{US}(\mathsf{BEA}) = n_1 \\ g\check{}_{UK}(\mathsf{FS}) = n_1 & g\check{}_{US}(\mathsf{AGe}) = n_2 \\ g\check{}_{UK}(\mathsf{EUE}) = n_1 & g\check{}_{US}(\mathsf{INS}) = n_2 \end{array}$$

Obviously, not all tokens of $\mathbf{A}$ will be in the images of $g\check{}_{UK}$ and $g\check{}_{US}$.

This alignment allows us to generate the desired channel between **UK** and **US** that captures the information flow according to the aligned responsibilities. This is done by constructing a classification $\mathbf{C}$ and a couple of infomorphisms $f_{UK} : \vee\mathbf{A}^{\perp}_{UK} \rightleftarrows \mathbf{C}$ and $f_{US} : \vee\mathbf{A}^{\perp}_{US} \rightleftarrows \mathbf{C}$ that correspond to a category-theoretic colimit [18] of the following distributed IF system, which includes the alignment and the contexts of each government:

$$\begin{tikzcd}
 & & & \mathbf{C} & & & \\
 & \vee\mathbf{A}^{\perp}_{UK} \arrow[urr, "f_{UK}"] & & & & \vee\mathbf{A}^{\perp}_{US} \arrow[ull, "f_{US}"'] & \\
\mathbf{UK} \arrow[ur, "h_{UK}"] & & \mathbf{A}^{\perp}_{UK} \arrow[ul, "\eta_{\mathbf{A}_{UK}}"'] & \mathbf{A} \arrow[l, "g_{UK}"] \arrow[r, "g_{US}"'] & \mathbf{A}^{\perp}_{US} \arrow[ur, "\eta_{\mathbf{A}_{US}}"] & & \mathbf{US} \arrow[ul, "h_{US}"']
\end{tikzcd}$$

4.3 The IF Logic on the Core

This is how colimit $\mathbf{C}$ is constructed: its set of types $typ(\mathbf{C})$ is the disjoint union of types of $\vee\mathbf{A}^{\perp}_{UK}$ and $\vee\mathbf{A}^{\perp}_{US}$; its tokens are connections—pairs of tokens—that connect a token a of $\vee\mathbf{A}^{\perp}_{UK}$ with a token b of $\vee\mathbf{A}^{\perp}_{US}$ only when a and b are send by the alignment infomorphisms g_{UK} and g_{US} to tokens of the alignment IF classification $\mathbf{A}$ that are classified as of the same type. For example, the core $\mathbf{C}$ will have a token $\langle\mathsf{AG},\mathsf{SoS}\rangle$ connecting $\vee\mathbf{A}^{\perp}_{UK}$-token AG with $\vee\mathbf{A}^{\perp}_{US}$-token SoS, because $g\check{}_{UK}(\mathsf{AG}) = n_6$ and $g\check{}_{US}(\mathsf{SoS}) = n_5$, and both n_5 and n_6 are of type a in $\mathbf{A}$.

The following is a fragment of the IF classification on the core (not all types are listed, but all tokens are):

	$\{r_1, r_2, r_3\}$	$\{r_4, r_5\}$	$\{s_1, s_2, s_3\}$	$\{s_4, s_5\}$
⟨FS,BEA⟩	0	1	1	0
⟨EUE,BEA⟩	0	1	1	0
⟨FS,SoS⟩	0	1	1	0
⟨EUE,SoS⟩	0	1	1	0
⟨IND,AGe⟩	1	0	0	1
⟨IND,INS⟩	1	0	0	1
⟨AG,AGe⟩	1	0	0	1
⟨PA,BCA⟩	1	0	1	0
⟨PA,SoS⟩	1	0	1	0
⟨AG,BCA⟩	1	0	1	0
⟨AG,SoS⟩	1	0	1	0

It shows the IF classification of all connections to those types of the core that are in the image of $f_{UK} \circ h_{UK}$ and $f_{US} \circ h_{US}$, which are the infomorphisms we will use in the next step to distribute the IF logic on the core to the IF classifications **UK** and **US**.

As the IF logic on the core we will take the natural IF logic of the IF classification **C**, whose constraints are:

$$\begin{array}{rlrl} \{r_4, r_5\} \vdash \{s_1, s_2, s_3\} & & \{s_4, s_5\} \vdash \{r_1, r_2, r_3\} & \\ \{r_1, r_2, r_3\}, \{r_4, r_5\} \vdash & & & \vdash \{r_1, r_2, r_3\}, \{r_4, r_5\} \\ \{s_1, s_2, s_3\}, \{s_4, s_5\} \vdash & & & \vdash \{s_1, s_2, s_3\}, \{s_4, s_5\} \end{array}$$

The natural IF logic is the one that captures in its constraints a complete knowledge of the IF classification. Since we have constructed the IF classification from those in the distributed system—which captured the contexts of governments together with the alignment of certain responsibilities—the natural IF logic will have as its IF theory all those sequents that conform to the government's contexts as well as to the alignment, which is what we desire for semantic interoperability.

4.4 The Distributed IF Logic

The natural IF logic has an IF theory whose types are sets of responsibilities taken from UK or US web sites, but we want to know how this theory translates to government ministries, by virtue of what responsibilities each ministry has. Hence we take the IF theory of the distributed IF logic of the IF channel $\mathbf{UK} \xrightarrow{f_{UK}\circ h_{UK}} \mathbf{C} \xleftarrow{f_{US}\circ h_{US}} \mathbf{US}$:

$$\begin{array}{rl} \mathsf{FCO} \vdash \mathsf{DoS} & \mathsf{DoJ} \vdash \mathsf{HO} \\ \mathsf{HO,FCO} \vdash & \qquad\vdash \mathsf{HO,FCO} \\ \mathsf{DoS,DoJ} \vdash & \qquad\vdash \mathsf{DoS,DoJ} \end{array}$$

which is the inverse image along $(f_{UK} \circ h_{UK}) + (f_{US} \circ h_{US})$ of the natural IF logic $Log(\mathbf{C})$ generated from the core IF classification. Its theory has the constraints shown above and captures the semantic interoperability between all ministries in our scenario.

5 Toward Automating Semantic Interoperability

The case described above showed the four steps of the proposed framework for representing semantic interoperability through an example scenario. As these steps exemplify the application of a theoretical framework to a test case, they do not correspond to actual engineering processes. Furthermore, when it comes to implementation we do not impose any specific requirements as to what formalisms or inference engine will be used, or how it will be deployed on the Web. It depends on the interoperability scenario at question. For example, in our previous work we focused on ontology mapping and devised the IF-Map method, which comprises four phases: *acquisition*, *translation*, *infomorphism generation*, and *map projection*. The IF-Map method is described in detail in [15], but here we recapitulate on some key parts and draw an analogy with the generic framework proposed above.

The *acquisition* and *translation* phases of IF-Map fall into the first step of our framework. In particular, they support the definition of the contexts of each community by representing source ontologies as IF classifications. The *acquisition* phase actually supports the harvesting of ontologies from various sources when these are not immediately available. IF-Map's next phase, *infomorphism generation*, supports the generation of the IF channel, which constitutes the second step in our framework. In the example of Section 4 we used an alignment structure to generate the desired channel between IF classifications **UK** and **US**. The IF-Map method is able to support and automate the generation of the necessary infomorphisms of this alignment structure, and also of the infomorphisms of the IF channel. The third and fourth steps of our framework—the generation of the IF logic at the core and its distribution to the sum of communities in order to obtain the distributed IF logic—do not have a direct counterpart in the IF-Map method as it would have been if we were interested in representing the integration of the two ontologies. Finally, the last phase of IF-Map, *map projection*, projects and stores the generated infomorphisms into RDF stores, which lies outside the scope of the theoretical framework presented here. We currently represent infomorphisms as custom-made RDF statements but we could have also used the OWL construct `owl:sameAs`. As it is reported in [17], `owl:sameAs` constructs could be used to represent links from one individual to another individual, and in more expressive versions of the language, like OWL Full, `owl:sameAs` could be used to define class equality, thus indicating that two concepts have the same intentional meaning. As the semantics of `owl:sameAs` do not impose a particular form of equality—only indicating individuals which share the same identity or sets of individuals (classes) that are interpreted intentionally as equal—we could see them as candidates for representing equivalence between types (a.k.a. classes).

6 Conclusions

We elaborated on the efforts been made to formalise and to provide automated support to semantic interoperability. We argued for the need to represent semantic interoperability in such a way that different understandings of semantics can be accommodate and potentially automated. We presented a theoretical framework for achieving this based on Information-Flow theory and illustrated an example scenario. Variations of this framework have been used in our recent work on mapping for Semantic Web ontologies. In the future, we plan to apply this framework to different semantic interoperability scenarios and to focus on semantic integration of distinct ontologies on the Semantic Web.

Acknowledgments. Y. Kalfoglou is supported by the Advanced Knowledge Technologies IRC, which is sponsored by UK EPSRC grant GR/N15764/01 and comprises the Universities of Aberdeen, Edinburgh, Sheffield, Southampton and the Open University. M. Schorlemmer is supported by a 'Ramón y Cajal' Fellowship from the Spanish Ministry of Science and Technology.

References

1. F. Baader, D. Calvanese, D. McGuinness, D. Nardi, and P. Pater-Schneider. *The Description Logic Handbook*. Cambridge University Press, 2003.
2. M. Barr. The Chu construction. *Theory and Applications of Categories*, 2(2):17–35, 1996.
3. J. Barwise and J. Seligman. *Information Flow*. Cambridge University Press, 1997.
4. T. Bench-Capon and G. Malcolm. Formalising ontologies and their relations. *Database and Expert Systems Applications, Proc. 10th Int. Conf.*, LNCS 1677, pp. 250–259, Springer, 1999.
5. T. Berners-Lee, J. Hendler, and O. Lassila. The Semantic Web. *Scientific American*, May 2001.
6. M. Ciocoiu and D. Nau. Ontology-based semantics. *Proc. 7th International Conference on the Principles of Knowledge Representation and Reasoning*, pp. 539–548, 2000.
7. F. Corrêa da Silva, W. Vasconcelos, D. Robertson, V. Brilhante, A. de Melo, M. Finger, and J. Agustí. On the insufficiency of ontologies: Problems in knowledge sharing and alternative solutions. *Knowledge Based Systems*, 15(3):147–167, 2002.
8. K. Devlin. *Logic and Information*. Cambridge University Press, 1991.
9. F. Dretske. *Knowledge and the Flow of Information*. MIT Press, 1981.
10. A. Finkelstein, D. Gabbay, A. Hunter, J. Kramer, and B. Nuseibeh. Inconsistency handling in multi-perspective specifications. *IEEE Trans. on Software Engineering*, 20(8):569–578, 1994.
11. B. Ganter and R. Wile. *Formal Concept Analysis*. Springer, 1999.
12. V. Gupta. *Chu Spaces: A Model of Concurrency*. PhD thesis, Stanford University, 1994.
13. R. Kent. A KIF formalization of the IFF category theory ontology. *Proc. IJCAI'01 Workshop on the IEEE Standard Upper Ontology*, 2001.
14. R. Kent. The IFF foundation for ontological knowledge organization. *Knowledge Organization and Classification in International Information Retrieval*, Cataloging and Classification Quarterly, The Haworth Press Inc., 2003.
15. Y. Kalfoglou and M. Schorlemmer. IF-Map: an ontology-mapping method based on information-flow theory. *Journal on Data Semantics I*, LNCS 2800, pp. 98–127, Springer, 2003
16. J. Madhavan, P. Bernstein, P. Domingos, and A. Halevy. Representing and reasoning about mappings between domain models. *Proc. 18th Nat. Conf. on AI*, 2002.
17. D. McGuinness and F. van Harmelen, eds. OWL Web Ontology Language. W3C Recommendation, 10 February 2004. http://www.w3.org/TR/2004/REC-owl-reatures-20040210/
18. S. McLane. *Categories for the Working Mathematician*. Springer, 2nd edition, 1998.
19. C. Menzel. Ontology theory. *Ontologies and Semantic Interoperability, Proc. ECAI-02 Workshop*, CEUR-WS 64, 2002.
20. P. Mitra and G. Wiederhold. Resolving terminological heterogeneity in ontologies. *Ontologies and Semantic Interoperability, Proc. ECAI-02 Workshop*, CEUR-WS 64, 2002.
21. J. Pollock. The Web Services Scandal: How data semantics have been overlooked in integration solutions. *eAI Journal*, pp. 20–23, August 2002.
22. A. Sheth and J. Larson. Federated database systems for managing distributed, heterogeneous, and autonomous databases. *ACM Computing Surveys*, 22(3):183–230, 1990.
23. J. Sowa. *Knowledge Representation*. Brooks/Cole, 2000.
24. G. Stumme and A. Maedche. FCA-Merge: Bottom-up merging of ontologies. *Proc. 17th International Joint Conference on Artificial Intelligence*, pp. 225–230, 2001.
25. R. Turner. *Properties, propositions and semantic theory*. Computational linguistics and formal semantics, chapter 5. Cambridge University Press, 1992.
26. M. Uschold. Creating semantically integrated communities on the World Wide Web. *WWW'02 Semantic Web Workshop*, 2002.