Computing and Informatics, Vol. 30, 2011, 31–55

DEVOLVED ONTOLOGY IN PRACTICE FOR A SEAMLESS SEMANTIC ALIGNMENT WITHIN DYNAMIC COLLABORATION NETWORKS OF SMES

César A. Marín, Martin Carpenter, Usman Wajid
Nikolay Mehandjiev

Centre for Service Research
The University of Manchester
Booth St West
Manchester M15 6PB, UK
e-mail: cesar.marin@manchester.ac.uk

Revised manuscript received 22 October 2010

Abstract. The lack of a semantic alignment between collaborating small and medium enterprises causes frequent misinterpretations when exchanging information in the form of documents. If these companies are to achieve a seamless semantic alignment by exchanging documents, we should employ a conceptual model which does not rely on agreeing in advance on a centralised standard for document contents and format, but instead allows individual companies to maintain localised ontologies structuring their own documents allowing the companies to automatically establish a semantic alignment between pairs of collaborating companies, taking into account the ripple effects that such an alignment could trigger. In this article we demonstrate how the conceptual model of devolved ontology is engineered and tested to support such a scenario: we show how we have engineered the devolved ontology through a case study, and present experimental results on the document alignment intrinsically needed for this.

Keywords: Devolved ontology, seamless semantic alignment, document structure representation

1 INTRODUCTION

When companies, especially small and medium enterprises (SMEs), exchange information in the form of documents it is common that information misinterpretations occur. These misinterpretations emerge when a document misses information the receiver is expecting but the sender thought it irrelevant. They also happen when the document contains information the sender thought important but the receiver considers it as useless. As a result there is a lack of semantic alignment between the collaborating companies.

Semantic alignment has been addressed by others, for instance by merging ontologies (cf. [4]) or encouraging companies to use the same standard for representing their internal information (cf. [12]). Yet these solutions render problematic for SMEs because these would need to create specialised mapping rules between the companies' ontologies before any collaboration, and incur in elevated costs for standardising their internal information, respectively. Consequently there is a need for an approach to allow companies to automatically reach a semantic alignment seamlessly based on the documents they exchange.

In this article we demonstrate how an approach called devolved ontology [15] is engineered and tested to allow a seamless semantic alignment between collaborating SMEs. We evaluate our implementation in two stages: first we show experimental results on the document alignment process inherent to the approach using a pool of business documents. Then we present a case study in which we show how the semantic alignment occurs between any two SMEs even when one of them is outside the network of already semantically aligned partners. This work has been carried out within the scope of the European Commission-funded project Commius[1] (Community-based Interoperability Utility for SMEs), grant agreement No. 213876.

The rest of the article is structured as follows: the preliminaries of the ontological representation of documents and the devolved ontology are presented in Section 2. Then Section 3 illustrates how the devolved ontology has been engineered describing the necessary functionality to its support: document alignment and semantic negotiation. Sections 4 and 5 then report on the evaluation of both functionalities by means of experiments and a case study, respectively. Then a discussion is presented in Section 6 along with a brief description of the related work, before concluding the paper in Section 7.

2 PRELIMINARIES

SMEs have a different set of types of documents which they typically use and are interested in. When pairs of SMEs exchange information in the form of documents, e.g. in e-mail communication, the potential cost involved in fully aligning their documents render this unfeasible. The fact that companies are exchanging documents and not formal ontology structures does however require certain adaptation. In

[1] `http://www.commius.eu/`

particular the nature of the objects exchanged during the communication is crucial. In general there is far more information within a full e-mail, and especially any attached documents, than in a single ontology concept.

A solution for this setting therefore should purposefully avoid trying to formally describe all the contents of a document, yet it should characterise the forms of information contained within. Indeed the devolved ontology [15] is in line with this idea as described below.

2.1 Devolved Ontology

The devolved ontology [15] is an approach to deal with the problem of ensuring that multiple, dynamically changing groups of individuals or agents can successfully communicate. The problem itself has been widely studied, cf. [1, 19], yet typically the solution has been to assume that all communication between the agents uses concepts drawn from a single, fixed ontology (cf. [10]). This approach is suitable for closed systems, whilst for open systems it either limits the interactions or imposes ontology evolution costs on all parties involved in the open system to accommodate new concepts.

The devolved ontology [15] uses a shared understanding of atomic units of information, also called *information tokens*, contained in exchanged messages without the need to ensure that the parties in the conversation share a complete, formal description of the concepts they plan to use when communicating. Instead the information tokens are used and every communicating partner is given the responsibility to maintain the set of concepts in which they are interested in communicating about, thus constructing their ontology as they reach agreements on the concepts being used.

The main advantage of this approach is its utility for open systems: it offers a much lower barrier for entry for new partners wishing to interact with an existing group. In essence the approach uses the following principles:

1. There is a universally agreed set of information tokens which represent a unique, atomic meaning regardless of the values they can take. This set forms the foundation for communication between agents.
2. The agents can combine any number of information tokens to create arbitrary concepts and thus ontologies. More details about the ontologies will be given in Section 2.2.
3. Every agent is responsible for maintaining their individual ontology and can include whichever concepts they wish within this.

These principles are depicted in Figure 1. In general when an agent receives a concept from other agent, the former will try to map that concept into its own ontology to derive the semantics, driving a meaningful response. Notice that the concepts are likely to be different from agent to agent, yet they all are characterised with the information tokens.

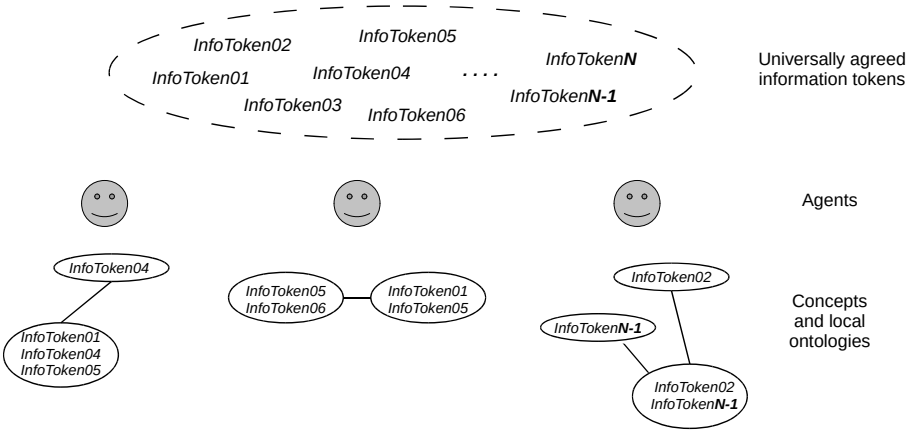


Fig. 1. Principles of the devolved ontology. The information tokens are used to characterise concepts and ontologies on an individual basis.

The above principles set the basis for creating concepts and ontologies. Yet ensuring that dynamic changing groups of agents communicate successfully, individual ontologies need to be aligned and consequently specialised. In a group of agents this will tend to a common ontology, however its creation is devolved to the individual pairs of communicating agents as described by the remaining principles below of the devolved ontology [15]:

4. An initial ontology (if any) shared by the agents allows each agent to specialise from it in a manner appropriate to the agent's interest.
5. The overlap between the ontologies of two or more communicating agents serves as a shared basis of understanding between collaborators.

An individual (owned) ontology allows an agent to specify its own concepts according to its own interests. Obviously the agent benefits from it as the agent thinks is best. Nevertheless this does not deter a successful and meaningful communication because the agents can overlap their ontologies ensuring a semantic alignment as long as they share an initial ontology based on the information tokens. In summary the generation of the final, extended ontology is devolved to independent agents by allowing them to overlap specialised versions of an initial ontology. This is shown in Figure 2.

In a practical sense the overlap between the ontologies of any set of collaborating agents will increase over time. In this way the devolved ontology approach provides a balance between the need for a low barrier to initial entry and a long term yet evolving, semantic alignment. Still there are the questions of how to represent the ontologies, and how to define the information tokens. This is explained in the following subsections.

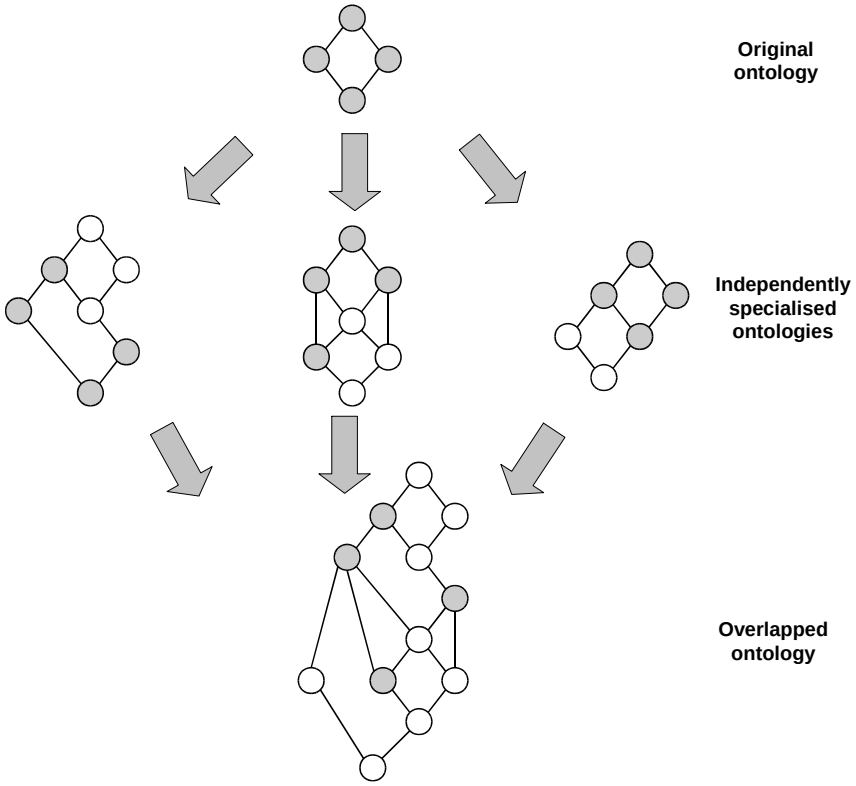


Fig. 2. Principles of the devolved ontology. Nodes in grey represent the concepts from the original ontology. The overlap between specialised ontologies function as the basis for a semantic alignment.

2.2 Representing Ontologies Using Formal Concept Analysis

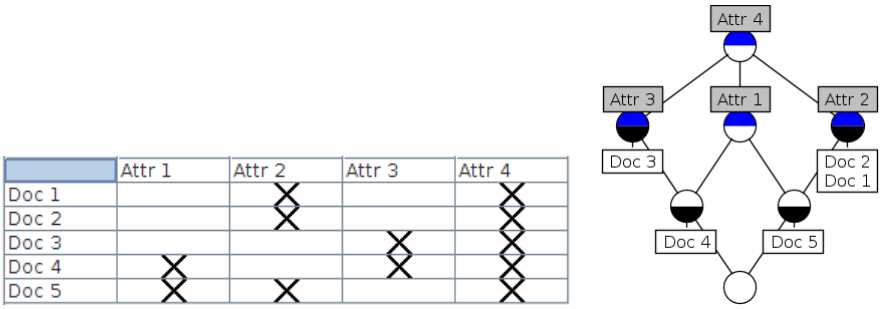
In [15] the precise technique used to map incoming concepts to the internal ontological structures in each agent is based on Formal Concept Analysis (FCA) [23]. It is a mathematical theory commonly used for representing knowledge and identifying conceptual structures such as ontologies [14]. We briefly present its main elements.

A formal context [23] is defined as a triple $K := (G, M, I)$ where G is a set of objects, M is a set of attributes, and $I \subseteq G \times M$ is the set of binary relations between the elements of the two.

A formal concept [23] within the context K as a pair (A, B) where $A \subseteq G$ is a subset of objects and $B \subseteq M$ is a subset of attributes. Then the relations $A = B^{\triangleleft}$ and $B = A^{\triangleright}$ are valid where $A^{\triangleright}$ is the set of attributes possessed by all objects in A, $B^{\triangleleft}$ represents the set of objects possessing all the attributes in B.

The sub concept - super concept relation [23] is a partial order represented as $(A_1, B_1) \leq (A_2, B_2)$ where the concept (A_1, B_1) is a sub concept of (A_2, B_2) if $A_1 \subseteq A_2$ which is equivalent to $B_2 \subseteq B_1$. Likewise, (A_2, B_2) can be called a super concept of (A_1, B_1).

A formal context K can be represented with a table depicting the object set and the attribute set in the first column and in the first row, respectively. A Boolean value at a row and column intersection (a cell) indicates that the object contains that attribute. Additionally the same context K can be depicted as a *concept lattice* [23]. Both representations are shown in Figure 3.



(a) Context table (b) Concept lattice

Fig. 3. Representations of a formal context K

Of particular interest is the concept lattice in Figure 3(b) where formal concepts are represented by nodes. The attributes contained in the path all the way up to the top node belong to the set B of a concept (A, B), whereas the objects included in the path all the way down to the bottom node belong to the set A of the same concept (A, B). Then we can say that the concept represented by the node with Doc 1 and Doc 2 is a super concept of that represented by the node with Doc 5 because the former concept is composed by Doc 1, Doc 2, and Doc 5, in contrast to the latter formed only by Doc 5.

2.3 Ontological Representation of Documents

In order to use FCA for representing the ontological structures of documents we need to choose a set of well defined information tokens, as the devolved ontology dictates, to use as attributes within FCA. In the literature there are efforts trying to standardise document representations aiming at automating information exchange between companies, cf. RossetaNet[2] and ebXML [12]. In this article we simply use one of such efforts.

The Core Components standard [18] proposes a methodology for identifying and characterising key business data as semantic descriptors. It also introduces an

[2] `http://www.rossetanet.org`

initial set of these which can be written in XML. In essence a *core component* is a meaningful semantic building block used for electronic transactions [18]. There are three types of core components: basic, aggregate, and association. The first one refers to a datum with a specific meaning regardless of the business context, for instance

```
<City>London</City>
```

has a clear meaning by itself. An aggregate core component is a collection of basic core components which together have a specific meaning, for example

```
<Address>
  <BuildingID>42b</BuildingID>
  <Street>Baker St</Street>
  <City>London</City>
</Address>
```

groups three basic core components and portrays a different meaning than the basic core components by themselves. Finally, an association core component creates a link between any two aggregate core components by including one of them as part of the structure of the other as shown below

```
<Organisation>
   <Name>The Company Ltd</Name>
   <PhysicalAddress><!--Associates to Address-->
      <BuildingID>42b</BuildingID>
      <Street>Baker St</Street>
      <City>London</City>
   </PhysicalAddress>
   <PostalAddress><!--Associates to Address-->
      <POBox>1234</POBox>
      <City>London</City>
   </PostalAddress>
</Organisation>
```

Notice that the two *address* semantic descriptors have a different meaning from each other even when the aggregate core component they represent is the same. The difference relies on the structure they form part of, which in turn gives a specific meaning to each of them resembling a hierarchy.

Clearly a document represented by core components in XML already contains a structure with a unique meaning allowing the core components to function as information tokens for the devolved ontology. Yet the core components need to be used as attributes for FCA. We do this by considering the hierarchical paths from the XML root element to each of the basic core components in the XML document. Using an infix notation we convert the above XML excerpt into attributes as follows

```
Organisation.Name
Organisation.PhysicalAddress->Address.BuildingID
Organisation.PhysicalAddress->Address.Street
Organisation.PhysicalAddress->Address.City
Organisation.PostalAddress->Address.POBox
Organisation.PostalAddress->Address.City
```

where a dot ('.') represents an aggregation to a specific basic core component and a hyphen ('-') denotes an association between the connected aggregate core components. We have called this representation *CC paths* [6] which each of them portrays a specific information token contained in a document.

Notice that neither the devolved ontology nor FCA make any reference to the actual data, but only to the information tokens used. We leave the conversion of the *actual* (business) documents into XML out of the scope of this article. We assume that existing efforts address it already, cf. [9], thus we consider the CC paths as the universally agreed information tokens for the devolved ontology.

3 ENGINEERING THE DEVOLVED ONTOLOGY TO THE SEMANTIC ALIGNMENT OF SMES

The devolved ontology approach represents a powerful means of ensuring a communication within groups of agents in open systems. Within the context of collaboration networks of SMEs, the goal then is to enable a set of companies to exchange documents which they mutually understand whilst maintaining their flexibility in defining the internal structure of those documents. Therefore the devolved ontology is engineered for this purpose. In order to do this, there are some requirements to take into consideration:

1. Since SMEs are to reach a semantic alignment each of them possesses an existing collection of documents represented by CC paths.
2. An SME's interest on a set of documents with certain characteristics is represented by a document type, e.g. an invoice or a purchase order.
3. The devolved ontology assumes the agents possess the means to map or include a concept from one ontology to another. In a practical sense this has to be devised for an SME.
4. The devolved ontology also assumes that there is an individual benefit in deciding to include a concept [15] and that the mapping of concepts occurs smoothly, i.e. the agent might decide not to include a concept to its own ontology because of a low benefit, or in the opposite case it will have no problem including the concept. In practice all this requires a sophisticated, intelligent implementation even for the simplest case which is not finding a mapping of a document to any concept of an existing ontology.

To satisfy the first three requirements, first we need a repository where to put and organise the SME's collection of documents represented by CC paths. Then using

the CC paths an SME is able to define its own document types and to maintain the relationship between the document types representing which documents in the repository. We call this repository a *semantic core*. Additionally, a process for aligning exchanged documents to the most likely document type is needed. It is natural to add such a processing to the *semantic core*. We leave the process of defining document types using CC paths out of the scope of this article.

Finally, the intelligence required to decide to include a concept to an ontology calls for a component supported by the *semantic core*, and with the capability to engage in intelligent conversations with its peer on the side of another SME. We simply call this component a *negotiator*. The devised architecture is presented in the following subsection.

3.1 Architecture for the Devolved Ontology

Figure 4 shows the conceptual architecture at the SME level, i.e. the necessary components to support the devolved ontology within a company, namely a semantic core, a negotiator, users and any legacy system.

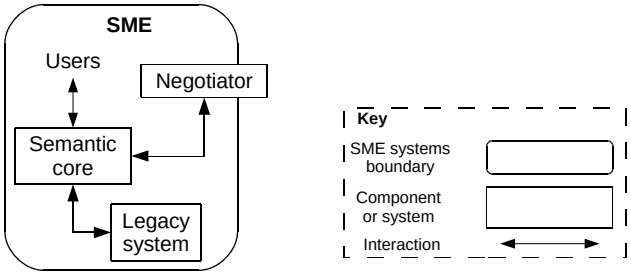


Fig. 4. Conceptual architecture at the SME level

The semantic core of the SME comprises a collection of documents the company has defined as being important for managing their information internally, and valuable to detect and process when received. These are sorted by document types according to their function in the business, e.g. a purchase order. In the context of the devolved ontology, a document type is represented by a pre-selected FCA concept (cf. [6]) defined by a set of information tokens universally agreed, i.e. the CC paths explained in Section 2.3. Consequently, the semantic core is implemented as a combination of a persistent data repository and an FCA implementation for creating and navigating through a concept lattice. When an SME receives a document and extracts its CC paths, it first attempts to align it to a document type within its own semantic core. We have presented a promising technique for this purpose called the circle of interest [6]. In Section 3.2 we present the details of it to complement the devolved ontology.

However when no similar document type can be located, the idea that either SMEs' semantic cores should be updated must be considered. For instance a company may be persistently receiving a type of document for which it has no current internal document type, but which, on consideration, it would like to include. In this case it would be useful for the software system to indicate the need for the updating of the SME's set of internal documents. In addition an SME sending documents might usefully learn to add certain additional fields when sending documents to the same company in the future. In order to help address such issues, the architecture comes with a negotiation component.

The negotiator is implemented as a reactive software agent which exchanges messages with negotiators representing other SMEs. Such messages are in the form of unitary documents possibly sent by e-mail. The basic goal of the negotiation is to inform the sending company that the document was not understood, and to subsequently search for alternative document types more aligned with the receiver's set of internal documents. Finally it may also result in the receiver of the document adopting a new internal document type. This will only happen when there is a reason to expect to receive this document again on a frequent basis. We have presented the negotiation process in [20]. In Section 3.3 we present its details to complement the devolved ontology.

The architectural solution to the devolved ontology avoids the use of any centralised repository of document types. The set of universally agreed information tokens, including a limited set of seed document types, is instead bundled with a software distribution implementing our approach, which could be obtained by a service provider of the solution as depicted in Figure 5.

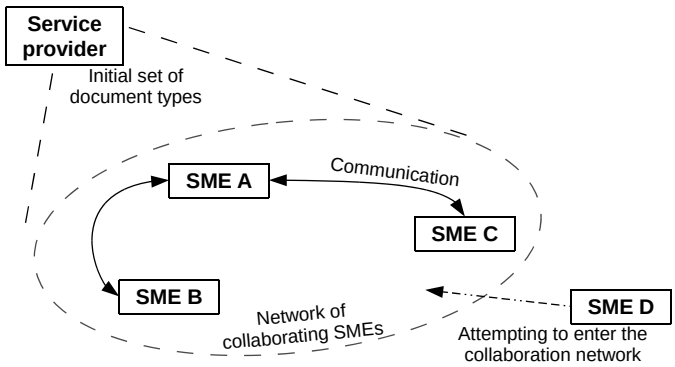


Fig. 5. Conceptual architecture at the network level

The conceptual architecture at the network level includes a set of SMEs which use or are able to extract the CC paths from documents. To create this network we follow these steps.

Step 1. Two SMEs each download independently the software distribution, together with the standard set of information tokens.

Step 2. Each company then uses the CC paths to define their own set of internal document types. When a document is received, the software of the receiving SME will then attempt to automatically align it with its set of internal document types.

Step 3. During a normal system operation, when SME A receives a document from SME B, the software in SME A will analyse the incoming document and attempt to understand it in terms of its set of internally defined document types.

Step 4. If this is not possible, for instance because the interacting SMEs have defined different document types, a negotiation process between SME A and SME B is initiated. This may result in a new document type added to the repository of SME A.

The intended result of these negotiation processes is that, over time, the sets of document types recognised by two companies in persistent communication will align to the extent required for that communication to work efficiently. This can involve the addition of new document types to either companies set of distinguished documents or more simply alterations or extensions to the definitions some of their existing ones.

3.2 Document Alignment

In [6] we presented a study on a technique called the *circle of interest* for aligning documents to document types using FCA. In essence an object in FCA represents a document and its attributes correspond to a collection of CC paths. A document type then is a pre-selected FCA formal concept. As mentioned earlier, a document type is defined by the SME interested in it using the CC paths.

Document alignment [6] is the process to determine the document type that best represents a given document. Notice however that a document could be represented by more than one document types, yet the process of document alignment gives the most appropriate one.

The circle of interest [6] of a document to align is that subset of document types allocated to the existing documents (in the semantic core) that best match the document to align.

The circle of interest uses the Rough Set Theory (RST) [13] to calculate the rough membership of an object (i.e. a document) to a "vague" concept, which in this case is a document type. For the calculations it uses a set of existing documents, which along with the document to align none can be discerned from one another when represented by an arbitrary subset of CC paths. Such a collection of documents is called the equivalence class, and the circle of interest concentrates on building it.

In order to build the equivalence class of a document to align, the circle of interest finds the minimal set of documents whose document type is likely to be the most appropriate for the alignment. The search is done in the FCA lattice when the document to align is inserted, i.e. the set of documents collected are the closest ones in the lattice, thus the most similar. Notice that the focus of this technique is on an efficient way to build the equivalence class, not on improving any similarity measure, cf. [24, 21]. Section 4 presents a comparison on the accuracy of this technique against an alternative approach.

3.3 Negotiation of Document Types

The use of CC paths to define documents and document types provides a solid basis for a semantic alignment between multiple SMEs. However the freedom of SMEs to define their own document types, or to modify the existing ones to suit their purposes, means that companies will frequently receive documents using a type which they do not currently use. In order to cater for this situation we use a negotiation approach which enables the companies to find a mutually acceptable replacement document type.

The devolved ontology [15] underpins the use of negotiation to support coordinated actions that are focused on achieving semantic alignment among different agents. The negotiation is modelled as a sequential decision making process. That is, rather than relying on predicted outcomes, the participating agents process and evaluate the information exchanged in order to decide whether to re-iterate the process for achieving a better outcome or to terminate the negotiation. Yet the actual protocol followed may differ from one another.

We have presented a negotiation protocol in [20] for reaching a semantic alignment among SMEs. We describe it here to complement the devolved ontology in terms of implementation. The protocol itself is shown in Figure 6.

The negotiation protocol is designed to feature repeated rounds where replacement concepts are proposed and replies are either sent in the form of counter proposals, an acceptance of a proposed replacement or ending the negotiation. This protocol is designed to offer agents an opportunity to reach an agreement at different levels of granularity by offering replacement concepts where a particular concept or a set of them is not understood by an agent. The different steps in which agent actions are coordinated in a negotiation are described below. It should be kept in mind that the agent negotiation process summarised in the following steps only focuses on the syntactic aspects of the negotiation i.e. how a sequence of negotiation messages can establish mutual understanding of concepts.

1. The interaction protocol starts when the Sender SME sends a document D to the Receiver SME.
2. The Receiver negotiator extracts the CC paths of the document, C_i. However it cannot find a local document type closely corresponding to the document

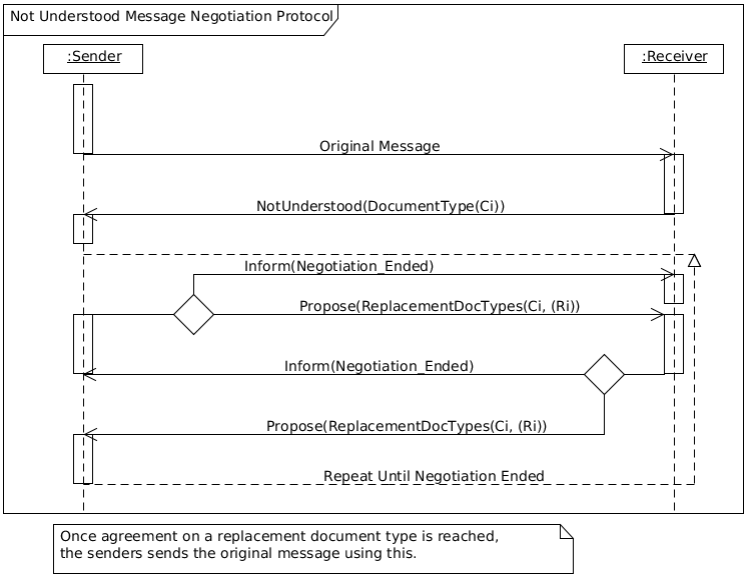


Fig. 6. A negotiation protocol

characterised by C_i. That is, the document is similar to more than one document type.

3. The Receiver sends a Not-Understood message containing a formal object of the type *UnknownConcept*. This object specifies the set of CC paths (i.e. the document) that do not match any of the document types. Notice that the misunderstanding is at the level of a document to a document type rather than at the level of the CC paths universally agreed on.
4. Upon receiving this message the Sender negotiator creates a set of different ways in which it could resend the basic information of the original message. This might include adding new CC paths, removing certain CC paths or simply using an entirely different structure by moving an aggregate core component to the internal structure of another aggregate through an association core component (see Section 2.3).
5. The Sender negotiator then sends a Propose message containing a set of objects of the form $ReplacementDocTypes(C_i, R_i)$ where C_i is the set of CC paths in the original, not understood message and R_i is a set of CC paths defining those document types which the Sender negotiator considers to be acceptable replacements.
6. Upon receiving a set of replacement concepts, the Receiver negotiator evaluates them to determine which ones are acceptable, and replies with the set Q_i listing those replacements considered acceptable.

7. Steps 5 and 6 iterate until the Sender negotiator has a replacement format in which both partners are happy for the message to be sent.
8. The Sender then resends the information from their original message with the new, acceptable CC paths. Note that one option in the negotiation above was for the Receiver negotiator to decide that the original CC paths in which the message was sent was in fact optimal. In this case the message is not resent.

Overall the negotiation fulfills its goal of facilitating an accurate long term alignment of the documents exchanged between SMEs who are in persistent communication, while also allowing every SME freedom to use the internal document types that they desire. A case study on the negotiation within the devolved ontology is presented in Section 5 as a validation for achieving a semantic alignment in a network of collaborating SMEs.

4 EXPERIMENTS ON ALIGNING DOCUMENTS

This section reports on the results of the experiments carried out on the document alignment process. In [6] we reported on the feasibility of the circle of interest technique for aligning documents. In this article we complement those results with a comparison between the circle of interest and another well-known similar technique: case based, explained later on in this section. The purpose of the experiments presented here is to demonstrate the accuracy and thus utility of the circle of interest for a seamless semantic alignment among a network of collaborating SMEs.

The experiments consist of aligning a new document to a document type. For this purpose we use a set of "seed" documents from which we generate new documents to align. The seed documents were obtained from companies within the consortium of the EC-funded project Commius[3] and consist of six authorisations of invoice, five delivery notes, nineteen invoices, five job assignments, and six purchase orders, each of them slightly different from the others within the same type. This natural variation among the seed documents helps make the experiments closer to an every day setting.

Initially all documents are manually converted to collections of CC paths. In a more automated version this process can be carried out by existing approaches, e.g. [9]. Then a seed document from each type is selected randomly and is treated as the definition of a document type. Afterwards a document to align is generated by randomly selecting a seed document and introducing a 10 % noise to it as a way to make each generated document slightly different from the seed documents yet still realistic. The noise is introduced by replacing, with a probability of 0.1, each CC path with another one randomly selected. Removing and adding extra CC paths is

[3] http://www.commius.eu/

also possible. The noise introduced reflects the typical difference on contents from one document to another even when they are of the same type.

Furthermore, we collected twenty six "chatter" e-mails which are typically exchanged between SMEs but which should not match to any document type. These sort of e-mails refer to personal communication, non-work related discussions, opinions, simple questions, announcements, etc. We consider chatter to occur with a 0.2 probability in addition to the other documents. Thus chatter is also a document type and new chatter emails are generated from the collected seeds e-mails with a 10 % noise as well.

We compare the circle of interest technique with a well-known technique based on past cases: case based. The latter is chosen because of being parallel to the circle of interest. In the latter the equivalence class is selected by comparing the document to align with other documents close in the FCA lattice. In turn the case based technique searches through all the aligned documents to find the most similar one to the document to align, once it is found the document to align is simply allocated to the same document type as the most similar existing document.

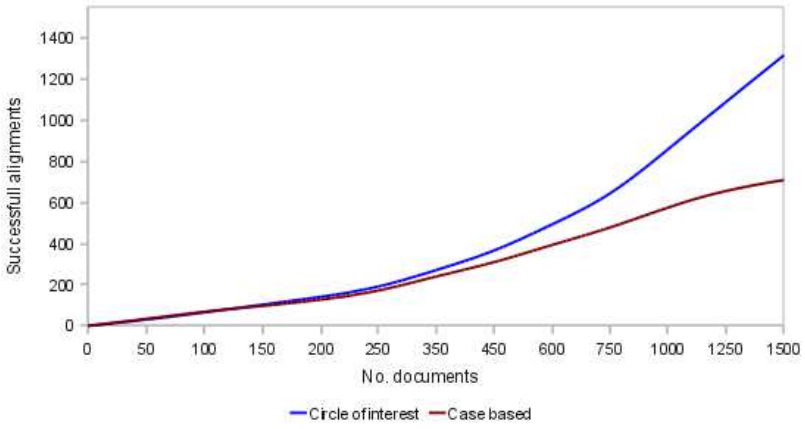
In both cases the comparison at the structural level (i.e. semantics) consists of determining the presence of a CC path in the documents being compared. The more CC path any two documents share, the more similar they are. The same applies for having extra or missing CC paths.

The experiments consist of automatically generating a document and let each algorithm to try aligning it to a document type. Because we already know the seed document used, we also know the document type for the alignment. If an algorithm selects the right one, it counts as a successful case for that algorithm. After the document alignment, the document is placed in the semantic core with the right document type the document belongs to regardless of the algorithms' output. That is, for these experiments the learning process of the algorithms is fully supervised.

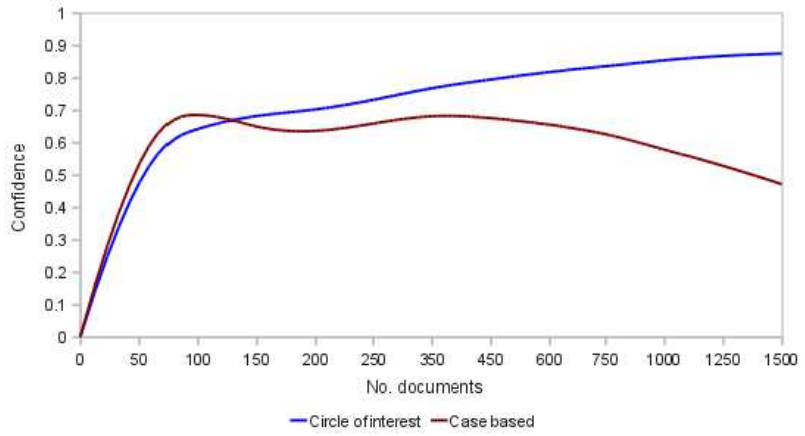
Figure 7(a) depicts the number of successful alignments for a total of 1 500 generated documents. As can be appreciated, both algorithms initially take some time to increase the number of successful cases due to their dependence on past cases on which the internal comparison is based. In the same figure it can be appreciated that the circle of interest eventually starts to produce a considerably higher number of successful cases than the case based technique. This behaviour indicates that the circle of interest is more accurate than the case based technique.

Figure 7(b) presents the confidence level of the same experiment, i.e. it shows the variation of the percentage of successful cases. As can be seen, the case based algorithm is more confident initially than the circle of interest. However just after a hundred documents it becomes less confident until dropping to 0.47, whereas the circle of interest gains more confidence for the rest of the experiments, up to 0.88. This complements the previous graph indicating that the circle of interest is more accurate and confident than the case based algorithm.

The reason for the case based technique to show such a behaviour is because it considers a single point for aligning a document, i.e. the most similar document



(a) Successful cases.



(b) Confidence level.

Fig. 7. Experimental results on document alignment. Notice that the X-axis is not on scale, yet this does not affect the presented results.

to the one to align. In contrast, the circle of interest calculates a membership to a document type based on a set of the most similar documents. In addition, the membership function can be seen as a function for reaching a consensus among the most similar documents.

In summary, the circle of interest is a promising approach for aligning documents as part of reaching a semantic alignment as indicated by the devolved ontology [15]. This renders suitable for application in network of collaborating SMEs where documents are exchanged.

5 A CASE STUDY ON NEGOTIATION

In this section we present a scenario which illustrates the operation of the architecture, especially the negotiation process, proposed within this paper and described in Section 3. We refer to the small network of SMEs shown in Figure 5 for this scenario mainly from the point of view of the SME A. We also concentrate on the evolution of the semantic core of SME A after an alignment triggered by an incoming document of an unknown type. For the sake of simplicity the FCA concept lattices we show portray the objects as "Doc01" (document) and the attributes as "CCP01" (CC path).

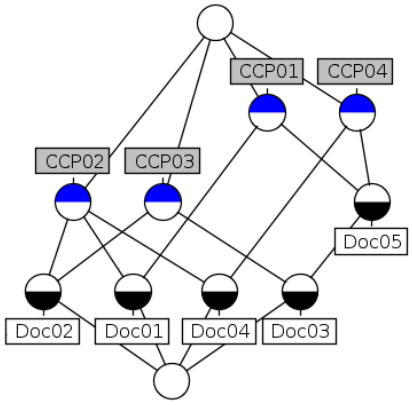


Fig. 8. Initial semantic core

Initially let us say that SME A downloads and installs a software distribution of the devolved ontology containing an initial set of documents and document types depicted in Figure 8. In the same figure we see five documents belonging each to a different document type (i.e. formal concept) represented by the nodes with a document attached to it. For example, the document type representing Doc01 contain the CC paths CCP01 and CCP02, the latter shared by two other document types representing Doc02 and Doc04. Notice that the document type representing Doc03 is a sub concept of the document type representing Doc05.

Elsewhere, SME B and SME C independently download and install the same software distribution, and proceed to customise it to their own needs and legacy systems. SME B, for example, adds a new document type to its semantic core; this is a concept defined by the document instance Doc07 with the CC paths CCP01, CCP03 and CCP05. CCP05 is a new information token not used in the initial set of document types. On the other hand, SME C specialises its own semantic core by creating a new document type defined by Doc06 and the CC paths CCP01, CCP02 and CCP03.

After the initial setup is complete, both SME B and SME C interact with SME A. SME B sends a message using Doc6 and SME C sends a document using Doc7. In both cases this triggers the negotiation protocol described previously in this paper.

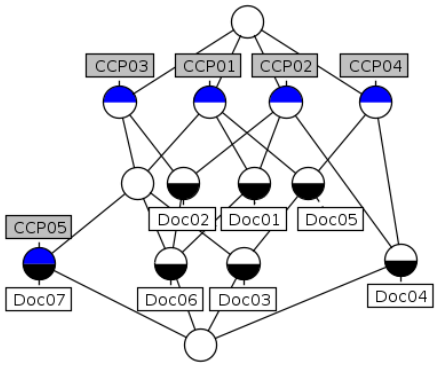


Fig. 9. Semantic core at SME A after collaborating with SME B and SME C

Since the three companies plan to interact extensively in the future, both new document types are of interest to SME A. This results in both negotiations concluding that SME A should add the documents in question to their set of internal documents. The resulting set of internal documents at SME A is depicted in Figure 9 which makes explicit the subconcept – superconcept relationship between the document type with Doc06 and the two document types with Doc01 and Doc02. Doc07, which uses a different set of information tokens, can be seen to one side.

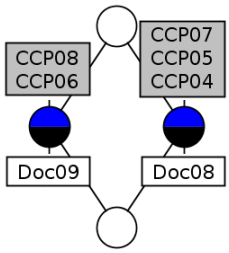


Fig. 10. Semantic core at SME D

Now we consider the cases where another company, SME D, has the internal document types shown in Figure 10. Note that SME D uses some information tokens not featured in SME A's existing document types. SME D then sends a document of type Doc08 to SME A and, after negotiation, SME A again adds it to their set of internal document types. The resultant set of internal document types can be seen in Figure 11. This case illustrates how the shared set of information tokens allows a level of interoperability to be achieved even where there is little overlap in terms of actual document types.

In addition to the purely additive way in which the set of document types has been developed here, the system can support other forms of interoperability, for instance:

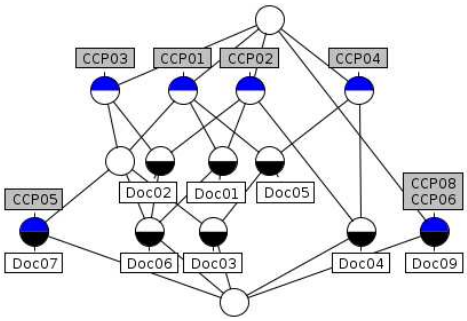


Fig. 11. Semantic core at SME A after collaborating with SME D

- If SME A received a document containing CCP01, CCP02 and CCP08 then they might choose to disregard the document type specialisation role of CCP08 and treat the document as if it is the same type as Doc01. The presence of CCP08 is thus not critical for the correct treatment of this document, and SME A do not need to alter their internally chosen set of documents. For instance the purchase order sent by SME D may contain some specialised delivery information in addition to the delivery address. SME A would still despatch the order as usual, and leave the transporting company to make use this piece of information.
- If SME A were to be sent a document containing CCP01 and CCP03 then they might contact the company who sent it (say SME D), informing that they can process it as Doc07 if only it also contained CCP05. Say SME A receives a general business enquiry, which they can treat as a "quite request" but only if it contained a quantity or each item (CCP05). The sender will provide the quantity and SME A will process the document.

The scenario presented here depicts the functionality and utility of the negotiation, which along with the document alignment process within the architecture demonstrate the applicability of the devolved ontology engineered for a seamless semantic alignment in a network of SMEs.

6 DISCUSSION AND RELATED WORK

Existing work on semantic alignment mostly focuses on fixing a unique meaning (semantics) through a commitment to a common ontology, which is used as metadata that is shared by all collaborating entities. The predominant use of ontology to foster semantic alignment is reflected by the numerous research efforts in the area of ontology modelling. This has also given rise to standards like Resource Description Framework[4] (RDF) and its elaboration in Web Ontology Language[5] (OWL)

[4] `http://www.w3.org/RDF/`
[5] `http://www.w3.org/TR/owl-features/`

along with other related standards, namely the Web Service Modelling Ontology[6] (WSMO) and Web Service Modelling Language[7] (WSML). A common aspect of these standards is to establish a universal medium for information exchange based upon XML syntax.

To further complement the approaches to ontological modelling a number of tools for ontology editing, storage and querying are now available. These include several semantic frameworks for accessing and manipulating documents [5, 11]. The proponents of creating domain specific ontologies and in particular their extension as common ontologies envision (achieving) semantic alignment through agreement over the use of common ontologies. We have seen that with the engineering of the devolved ontology this is not necessary.

Moreover, an agreement on common semantic model (ontology) means that the approach for semantic alignment will be typically centralised. In this respect, many have recognised that the use of single ontology is untenable in a distributed environment [7]. This has led to research in ontology alignment, which is typically realised by mapping one ontology to another or by creating some meta-structure to relate ontologies, cf. the KRAFT project[8].

In particular, OntoMorph [2] is another approach consisting of a rule based system that "re-writes" the syntax and semantic when merging ontologies represented in symbolic knowledge bases. A similar approach is called OntoMerge [4]. Yet another approach, FCA-merge [17], combines ontologies acquired from documents and merges them in an FCA lattice. Nevertheless this approach requires a knowledge engineer. In contrast, with the devolved ontology only using the CC paths is necessary to achieve a seamless semantic alignment.

The approaches for merging and mapping ontologies aim to come up with a single global ontology that combines the information sources from various ontologies [8]. However, both mapping and merging approaches are prone to scalability issues and require huge efforts when it comes to adding or mapping new ontologies in the global ontology, e.g. when a new company/partner wants to join a consortium or existing network of collaborating companies. Moreover, in both cases, i.e. merging and mapping, one must learn the local structures and create appropriate mapping rules or meta-structures, which can be complex and cumbersome.

Also, mapping rules typically cater for one-to-one mappings, e.g. *contact = phone_number*. However, many concepts in business documents involve relationships requiring complex mapping rules, such as *address = building_number, street_name, city, postcode, country*. Hence the development of techniques for automatically constructing complex mapping rules is an important prerequisite for the application of these approaches in real scenarios, such as the one presented in this article.

Another approach that addresses the semantic alignment issue advocates the use of partial shared ontologies to enable on-demand translations of specific con-

[6] http://www.w3.org/Submission/WSMO/

[7] http://www.w3.org/Submission/WSML/

[8] http://www.csd.abdn.ac.uk/research/kraft

cepts [16]. The partial shared approach tries to combine the benefits of merging and mapping without overlooking the scalability issues, i.e. new information sources can be added in a shared ontology, which is maintained as a separate resource from the private ontologies of collaborating entities.

However, like the previous approaches this approach also uses mapping rules to translate the concepts from the private ontologies into the shared one. Establishing such mapping rules may require sophisticated mechanisms to interpret and even re-write the shared ontology. Moreover, this approach does not represent a suitable solution when multiple entities are accessing a shared ontology because an agreement between a private ontology and the shared ontology may not represent an ideal fit/preference for all parties that are sharing the ontology.

Finally, document classification can be related to semantic alignment if a specific group or class is looked for for a given document. [22] considers a model of neural networks to classify Reuters news articles. Initially classes of news have to be defined, cf. defining document types, so that the neural network could be trained with a specific set of pre-classified news before setting the neural network to work. Regardless of the incompatibility between document classification and document alignment, our approach using the circle of interest does not require a training stage as can be appreciated in the experiments presented in Section 4. Additionally, [3] presents an algorithm to cluster documents automatically by being modelled as flocking birds. The result is that birds (i.e. documents) cluster according to their similarities. Nevertheless the approach does not consider adding new documents.

7 CONCLUSIONS

In this article we engineered the devolved ontology approach to support a seamless semantic alignment among a network of SMEs. We described the devolved ontology approach, how the ontologies are represented, and how to semantically describe documents using information tokens. All these function as the basis for the conceptual architecture supporting a semantic alignment at two levels: within an SME and at a network of collaborating SMEs. A document alignment process underpinned by the circle of interest technique is then identified as necessary to support the SME level whereas a negotiation protocol describes how any two SMEs can reach an alignment at the network level.

Experiments on the accuracy and confidence of the circle of interest technique for document alignment is presented. The results show that the technique is more accurate and with a higher confidence than a well-known, similar case based technique. This validates the suitability of the circle of interest for document alignment within the devolved ontology.

Then a case study is presented exemplifying how the negotiation protocol helps reach a seamless semantic alignment. Initially the semantic cores of collaborating SMEs are specialised according to their specific (business) use, yet they are still able to reach an alignment which does not compromise the structures of their own

document types. Such an alignment occurs even when an SME does not contain a initial shared set of document types. This maintains the claim that such an implementation of the devolved ontology supports a semantic alignment seamlessly among SMEs.

REFERENCES

[1] Chaib-Draa, B.—Dignum, F.: Trends in Agent Communication Language. Computational Intelligence, Vol. 18, 2002, pp. 89–101.

[2] Chalupsky, H.: OntoMorph: A Translation System for Symbolic Knowledge. Principles of Knowledge Representation and Reasoning, 2000, pp. 471–482.

[3] Cui, X.—Potok, T. E.: A Distributed Agent Implementation of Multiple Species Flocking Model for Document Partitioning Clustering. In Cooperative Information Agents, LNAI Vol. 4149, Springer-Verlag, 2006, pp. 124–137.

[4] Dou, D.—McDermott, D.—Qi, P.: Ontology Translation by Ontology Merging and Automated Reasoning. In V. Tamma, S. Cranefield, T. W. Finin, and S. Willmott (Eds.): Ontology for Agents: Theory and Experiences, Whitestein Series in Software Agent Technologies and Autonomic Computing, Birkhäuser, 2006, pp 73–94.

[5] Janssen, D.—Lins, A.—Schlegel, T.—Kühner, M.—Wanner, G.: A Framework for Semantic Web Service Retrieval. In 3rd Nordic Conference on Web Services Proceedings, Mathematical modelling in physics, engineering and cognitive sciences, Växjö University Press, 2004, pp. 1–14.

[6] Joseph, D.—Marín, C. A.: A Study on Aligning Documents Using the Circle of Interest Technique. In J. Cordeiro, M. Virvou, and B. Shishkov (Eds.): 5th International Conference on Software and Data Technologies, SciTePress, 2010, pp. 374–383.

[7] Kalfoglou, Y.—Schorlemmer, M.: Ontology Mapping: The State of the Art. Knowledge Engineering Review, Vol. 18, 2003, No. 1, pp. 1–31.

[8] Klein, M.: Combining and Relating Ontologies: An Analysis of Problems and Solutions. In G. A. Pérez, M. Gruninger, H. Stuckenschmidt, and M. Uschold (Eds.): Workshop on Ontologies and Information Sharing, IJCAI '01, 2001.

[9] Laclavík, M.—Dlugolinský, Š.—Šeleng, M.—Kvassay, M.—Hluchý, L.: Email Analysis and Information Extraction for Enterprise Benefit. Computing and Informatics, Vol. 30, 2011 No. 1, pp. 57–87.

[10] McGuinness, D. L.—Fikes, R.—Steve Wilder, J. R.: An Environment for Merging and Testing Large Ontologies. In Proceedings of the Seventh International Conference on Principles of Knowledge Representation and Reasoning KR2000, 2000, pp. 483–493.

[11] Moscato, F.—Di Martino, B.—Venticinque, S.—Martone, A.: A Collaborative Framework for the Semantic Annotation of Documents and Websites. International Journal on Web Grid Services, Vol. 5, 2009, No. 1, pp. 30–45.

[12] OASIS: ebXML Technical Architecture Specification. Technical report, ebXML, 2001.

[13] Pawlak, Z.: Rough Sets. International Journal of Information and Computer Sciences, Vol. 11, 1982, pp. 341–356.

[14] Priss, U.: Formal Concept Analysis in Information Science. Annual Review of Information Science and Technology, Vol. 40, 2006.

[15] Stalker, I. D—Mehandjiev, N.: A Devolved Ontology Model for the Pragmatic Web. In M. Schoop, A. de Moor, and J. L. G. Dietz (Eds.): Proceedings of the First International Conference on the Pragmatic Web, Vol. 89 LNI, 2006, GI, pp. 38–52.

[16] Stuckenschmidt, H.: Exploiting Partially Shared Ontologies for Multi-Agent Communication. In M. Klusch, S. Ossowski, and O. Shehory (Eds.): Proceedings of the 6th International Workshop on Cooperative Information Agents, LNCS Vol. 2446, Springer, 2002, pp. 249–263.

[17] Stumme, G.—Maedche, A.: FCA-MERGE: Bottom-Up Merging of Ontologies. In IJCAI, 2001, pp. 225–234.

[18] UN/CEFACT: Core Components Technical Specification – Part 8 of the ebXML Framework. Technical report, UN/CEFACT, 2003.

[19] Wache, H.—Vögele, T.—Visser, U.—Stuckenschmidt, H.—Schuster, G.—Neumann, H.—Hübner, S.: Ontology-Based Integration of Information – A Survey of Existing Approaches. In IJCAI'01 Workshop on Ontologies and Information Sharing, 2001.

[20] Wajid, U.—Marín, C. A.: Enhancing Enterprise Collaboration Using a Protocol for Semantic Alignment. In S. Reddy (Eds.): Workshops on Enabling Technologies: Infrastructures for Collaborative Enterprises, 18th IEEE WETICE 2009, IEEE Computer Society, 2009, pp. 13–18.

[21] Wang, L.—Liu, X.: A New Model of Evaluating Concept Similarity. Knowledge-Based Systems, Vol. 21, 2008, No. 8, pp. 842–846.

[22] Wermter, S.—Hung, C.: Selforganizing Classification on the Reuters News Corpus. In Proceedings of the 19th international conference on computational linguistics, Association for Computational Linguistics, 2002, pp. 1–7.

[23] Wille, R.: Formal Concept Analysis As Mathematical Theory of Concepts and Concept Hierarchies. In B. Ganter, G. Stumme, and R. Wille (Eds.): Formal Concept Analysis: Foundations and Applications, LNAI Vol. 3626, Springer-Verlag, 2005.

[24] Zhao, Y.—Wang, X.—Halang, W.: Ontology Mapping Based on Rough Formal Concept Analysis. In Proceedings of the Advanced International Conference on Telecommunications and International Conference on Internet and Web Applications and Services, IEEE, 2006.

César A. Marín is a Research Associate at the Centre for Service Research at the University of Manchester. His research is mainly about the applications of adaptive and self-organising approaches to service systems engineering. He completed a Ph. D. in Informatics at The University of Manchester, focused on adaptation to unexpected changes in business ecosystems. He obtained both an M. Sc. and a Diploma in Intelligent Systems, and a B. Sc. in Computer Systems Engineering from the Monterrey Institute of Technology, Mexico. Before joining The University of Manchester, he was a Research Assistant in the Centre for Intelligent Systems at the Monterrey Institute of Technology, where he worked in research areas such as decentralised agent-based workflow systems, information and knowledge distribution, and behavioural adaptive modelling. Currently he is coordinating the contributions to the EC-funded project Commius.

Martin Carpenter is a Research Associate at the Centre for Service Research at the University of Manchester. He finished a Ph. D. in Informatics at the University of Manchester, where he became specialist on techniques related to dynamic team formation and composition. He holds an M. Sc. in Computation from the University of York, an M. Sc. in Mathematical Logic from the University of Manchester and a first degree on Mathematics from the University of Warwick. At the University of Manchester he has been working on a number of European projects, namely MaBE, Crosswork, SUddEN, and Commius.

Usman Wajid is a Researcher at Centre for Service Research at The University of Manchester. He holds a Ph. D. in Informatics (University of Manchester), an M. Sc. in E-Commerce (Middlesex University London) and an M. Sc. in Computer Science (University of Arid-Agriculture, Pakistan). His interests lie in the development of intelligent solutions for dynamic business environments. He is particularly interested in the application of software agents in different domains, e.g. enabling enterprise collaboration and web-service composition. He coordinated the University of Manchester's contribution to the successful EC-funded project SUddEN. He is currently working on European Union projects Commius and SOA4ALL funded by European Commission 7^{th} Framework Programme.

Nikolay Mehandjiev is a Reader at the Centre for Service Research of the Manchester Business School. He obtained his Ph. D. for research in user-adaptable office information systems. His current research is focused on approaches and models which enable non-technical audience to design dynamic service systems. Key techniques used are intelligent software systems and formalised domain knowledge, including their use to achieve dynamic interoperability at semantic level between collaborating software agents. He has published two books, more than 100 refereed papers, and has acted as a guest editor for three special issues of international journals.