Ontology Evaluation for Reuse in the Domain of Process Systems Engineering

Abstract

Ontologies are a useful tool for knowledge representation, sharing and reuse. Although the number of available ontologies is increasing, the concomitant reuse activities are not following respectively. This is particularly true in the domain of Process Systems Engineering where the ontology development has been proven to be a challenge and addressing their reusability is at its infancy. This paper presents an algorithm for the evaluation of ontologies for reuse purposes. The presented algorithm uses information about the ontologies, such as the terminology of ontologies and their structure, to calculate a single compatibility metric used for assessment of ontology suitability for reused and hence integration. The proposed algorithm has been verified using several Process Engineering cases.

Keywords: Ontology Reuse, Compatibility, Ontology Engineering

1. Introduction

Ontologies are a useful tool for knowledge representation, sharing and reuse. Their potential has been recognised in varieties of domains, such as biomedical (Whetzel et al. 2011, Musen et al. 2012), law and legal activities (Casellas et al. 2005, Lame 2005, Corcho et al. 2005), finance (Zhang, Zhang & San Ong 2000, Alonso et al. 2005), agricultural (Soergel et al. 2006) and the domain of chemical and process engineering (CPE).

The number of publicly available ontologies, more than 10,000 as shown by dedicated search engines (Swoogle 2009), is a good indicator of the scale of development and use of ontologies across disciplines. Ontologies are developed for and used in several areas ranging from pure knowledge representation and semantic search to data integration and web service discovery (Lee, Suh 2007, Kraines et al. 2005, Morbach, Yang & Marquardt 2007, Muñoz, Espuña & Puigjaner 2010, Raafat et al. , Muñoz et al. 2013, Trokanas et al. 2013a, Raafat et al. 2013). In the domain of chemical and process engineering efforts have been focused on process design (Zhang, Yin 2008, Bock et al. 2010, Fernandes et al. 2011), supply chain modelling and design (Muñoz, Espuña & Puigjaner 2010, Muñoz et al. 2012) and decision-making related to environmental effects and causes (Raafat et al. 2013, Trokanas et al. 2013b, Trokanas, Cecelja & Raafat 2014a) among other domains such as computer-aided process engineering (Morbach, Wiesner & Marquardt 2009a,

Abanda, Tah & Duce 2013, Panetto, Dassisti & Tursi 2012). Numerous specific applications of ontologies have been reported including applications for knowledge sharing (Morbach, Wiesner & Marquardt 2009b, Marquardt et al. 2010), (Muñoz et al. 2011, Brandt et al. 2008), standardisation of vocabularies (Muñoz, Espuña & Puigjaner 2010, Venkatasubramanian et al. 2006, Jing Ni, Jiu Yi & Suping Ni 2011), (Suresh et al. 2008) supporting process integration and interoperability (Lin, Harding 2007), (Muñoz et al. 2013, Muñoz et al. 2012, Wiesner, Morbach & Marquardt 2011, Ceccaroni, Cortés & Sànchez-Marrè 2000).

Developing ontology is a time consuming process which requires a high level of domain specific expertise (Alani, Brewster & Shadbolt 2006). By definition ontologies are conceptualisations that are aimed to be shared and reused (Gruber 1993), (Gruber 1995), (Simperl 2009) and hence the ontology reuse is expected to be a paramount activity for knowledge engineers which in turn would reduce the cost of developing (Alani, Brewster 2005), as well as promoting interoperability between applications. This is further supported by the fact that many of the existing ontologies cover complementary and/or overlapping domains (Lonsdale et al. 2010). Still, as reported (Lonsdale et al. 2010, Uschold et al. 1998, Pinto, Martins 2000, Cantador, Fernández & Castells 2007), ontology reuse is not frequently exercised activity perhaps because of the absence of robust and pragmatic methods for evaluating and identifying ontologies for reuse (Alani, Brewster 2005).

This paper presents a practical algorithm (CoRAl) for evaluating compatibility of ontologies for reuse. Quantification of semantic similarity between ontologies is used for ontology matching, aligning, merging and integrating. Although the importance of calculating semantic similarity has been identified (David, Euzenat 2008), current research efforts are directed towards calculating semantic similarity from pairwise similarity between concepts. which are then aggregated into a single measure quantifying the similarity between ontologies. This process is proven to be daunting, complex and computationally exhaustive. In contrast, we propose a method of calculating similarity between ontologies by using high-level information describing ontology, such as size, terminology, external resources and data types from which the compatibility measure between two or more ontologies is calculated. The proposed measure also addresses issues associated with granularity differences, encoding and coverage of ontologies.

2. Definitions for Ontology Reuse

Ontology is a framework for knowledge modelling and it is described as a group of terms organised in a class-subclass structure and which describes a specific domain (Trokanas, Cecelja & Raafat 2014b). Ontology is further enhanced with properties characterising terms and

restrictions on these properties. Ontology can be instantiated with instances representing specific entities of the domain.

Ontology evaluation is defined as the process of assessment of ontology quality and adequacy for the purpose of being reused in a specific context and for a specific goal (Cantador, Fernández & Castells 2007).

Ontology ranking in the context of this work is defined as a process of estimating relative standing of a set of ontologies for given evaluation criteria.

Ontology similarity is defined as the comparison of two or more ontologies returning a value ranging between 0 and 1 which indicates the level of feature correspondence between them (Ehrig 2006).

Ontology compatibility is defined as the level at which two separate ontologies are suitable to form single ontology integration and without causing any conflict or inconsistency.

Target ontology is defined as the ontology which is the basis of the process of reuse. It usually refers to the ontology that is already available and which is attempted to expand by reusing other existing ontologies.

Candidate ontology is defined as an ontology which is considered for reuse.

3. Ontology Evaluation: proposed approach

Any candidate ontology is evaluated for compatibility with a view of reusing it. The proposed approach (Figure 1) takes into account the ontology metadata. The algorithm, namely CoRAL and presented in Figure 1, takes the target and candidate ontologies as inputs and uses their metadata i.e. terminology, language, encoding, external resources and size, to calculate their compatibility. More precisely, the presented approach accounts for the following:

i) **Terminology used in the ontologies**: terminology refers to all the terms used in the ontology irrespectively of whether they describe concepts, properties or instances. The reason behind the use of terminology is in that when two ontologies share common terms, they are likely to describe similar domains. Also it is easier to combine or integrate them by using these common terms as a starting premise.

ii) **Natural language of the ontologies**: natural language refers to the languages used during ontology modelling and design. Currently, most of existing ontologies are modelled in English and then annotated in other languages. We argue that the use of one or more common languages improves the reusability potential.

iii) **Encoding information of the ontologies**: encoding information refers to the data types used during ontology modelling. This category can cover wider aspects of ontology engineering. For example, the use *integer* as the range of a data type property can cause inconsistencies, if merged with an ontology that uses *float* data type for the same property. This aspect reflects the need for minimised encoding bias during ontology engineering, as defined in ref. (Gruber 1995).

iv) **External resources used**: external resources refer to imported ontologies. Two ontologies which use the same external resources are likely to describe the same domain(s). In addition, the process of importing ontologies makes them easier to reuse.

v) **Size and breadth of the ontologies**: This aspect accounts for different types of ontologies, taking into account the number of concepts and the levels of the target and candidate ontologies. The number of levels and concepts of ontology provides an outlook of whether an ontology is detailed, top-level etc.

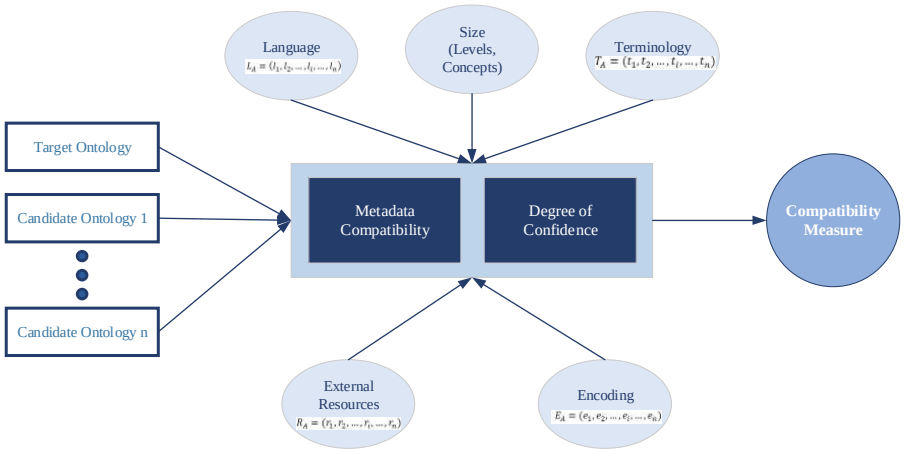

Figure 1 Structure of Measure

3.1 Acquisition of Ontology Metadata

For the assessment of the ***CoRAL*** metric, each ontology is represented by a set of vectors representing the ontology metadata. For the purpose of this work we assume that metadata specify the terms used in the ontology, the natural language, the data types and the external resources of the ontology, as expected by the 'good practice' in ontology design (Schulz et al. 2012) . A similarity score is then calculated as the vector similarity using a combination of cosine similarity and Euclidean distance, for each type of metadata (Figure 1). Similarity between

vectors of the same type, as explained in Sections 3.1.1 to 3.1.5, is aggregated into a single semantic similarity measure. Finally, the similarity measures for different types of metadata are combined to form a single similarity score.

3.1.1 Natural Language

The use of different natural languages is perhaps the biggest obstacle in ontology matching (Hawalah, Fasli 2011). Most existing approaches and techniques for ontology matching and alignment perform poorly when the two ontologies do not share a common language. This is caused by the substantial reliance of these techniques on lexical similarity, especially for the identification of initial mappings/alignments used as the basis for the matching/alignment process. The vector of natural languages L_A of ontology A is

$$L_A = (l_1, l_2, \dots, l_i, \dots, l_n) \tag{1}$$

where l_i is a boolean value representing the existence of language $i \in (0, n)$. A vector is created for each of ontologies participating in the comparison process. Length n of vector is defined by the number of elements each representing a distinct language and occurring as the union of languages used in the target ontology L_T and in the candidate ontology L_C as $L_T \cup L_C$.

3.1.2 Data types

We argue that the modelling conventions adopted during the ontology development process could be beneficially used for ontology matching. Modelling conventions include a wide range of aspects ranging from the level of detail (granularity of the ontology) and whether the ontology is instantiated or not, up to practical aspects such as the choice of data types, e.g. numerical values can be modelled as float, integers etc. or simply as undefined. Similar or identical concepts can be modelled in different ways. For example, distance can be represented in miles or kilometres. This work focuses only on the low level convention of data types, as defined by the XML Schema standard (Bray et al. 1998). Hence, the vector of data types E_A representing the data types in ontology A is defined as

$$E_A = (e_1, e_2, \dots, e_i, \dots, e_n) \tag{2}$$

where e_i is a value representing the frequency of each data type $i \in (0, n)$. Length n of vector is defined by the number of distinct data type elements resulting by the union $E_T \cup E_C$ of data types used in the target ontology E_T and in the candidate ontology E_C.

3.1.3 External Resources

External resources give an indication of the scope of the ontology. When two ontologies use the same external resources, e.g. for two ontologies both importing the same units or geographic information ontology, then they share some similarities. The vector representing the external resources of an ontology A, R_A, is formed from the ontologies which are imported in each case and is given as

$$R_A = (r_1, r_2, \dots, r_i, \dots, r_n) \tag{3}$$

where r_i is a Boolean value which represents the existence of external resource $i \in (0, n)$. Length n of vector is defined by the number of distinct external resources occurring by the union $R_T \cup R_C$ of external resources used in the target ontology R_T and in the candidate ontology R_C. All Web Ontology Language (OWL) ontologies have some default external resources which are used to define information such as the data types (xsd imports) and basic ontology structure (owl and rdf imports). These external resources are disregarded because they do not contribute to additional useful information for matching.

3.1.4 Terminology

The terminological obstacles arise from the use of different terms to describe similar or the same concepts (synonyms) or the same terms to describe different concepts (homonyms). As at present, ontology is considered as a group of terms including concept and property names, annotations and instances and for that reason alone terminology is of vital significance for ontology matching. As explained before, most existing matching techniques rely on terminological information for the identification of initial mappings between two ontologies. The vector of terms T_A representing the terminology of ontology A is defined as

$$T_A = (t_1, t_2, \dots, t_i, \dots, t_n) \tag{4}$$

where t_i represents the frequency that each term $i \in (0, n)$ appears in ontology A. Length n of vector T_A is defined by the number of distinct terms of the two ontologies occurring by the union $T_T \cup T_C$ of the two terminology sets used in the target ontology T_T and in the candidate ontology T_C.

Although, the information stated in Sections 3.1.1 - 3.1.3 is comparatively easy to extract from any ontology, this is not the case with terminology. For this reason, we propose a more practical and in some way an intuitive way of extracting the terminology of an ontology. In particular, we

propose to use existing terminology extraction software using the whole ontology file (.owl) as a source. Terminology extraction software tools identifies the key terms of a document. It takes any document file as input (in this case .owl files) and generates a list with key terms along with the frequency each term appears in the file. These tools also allow for the user to define thresholds (i.e. minimum/maximum number of times a term must appear to be considered).The tool used in this work is Anchovy (Frantzi, Ananiadou & Mima 2000, Anchovy 2013) which, although one out of many similar tools available today, is found to perform well.

3.1.5 Lexical Similarity

Another difficulty in comparing concepts arises from their lexical similarity. There are two ways of addressing lexical similarity when creating vectors representing terminology. The first and perhaps the most intuitive is to only account for identical terms. This approach is faster and less complex but it is also liable to inaccuracies, i.e. in cases where the two ontologies use both plural and singular for concept names, these will not be accounted for. The second approach is to employ some lexical similarity algorithm and/or external resources such as lexicons. This approach would go beyond the focus of this work, and, instead, we employ a hybrid algorithm based on established methods for lexical similarity. For this the similarity of 60% is adopted as a threshold above which two terms are considered identical.

There are numerous algorithms existing for calculating lexical similarity (Aho, Corasick 1975, Stoilos, Stamou & Kollias 2005, Budanitsky, Hirst 2006, Jiang, Conrath 1997, Li, Bandar & McLean 2003, Maedche, Staab 2002, Mihalcea, Corley & Strapparava 2006). As reported, the process of comparing syntactic information is a heavily researched (Cheatham, Hitzler 2013). In this work we employ existing and established lexical similarity algorithms with alterations which yield better results for the terminology of the domain of chemical engineering and industrial symbiosis. To elaborate, the proposed algorithm uses *Levenshtein* distance (Levenshtein 1966) to compare the prefixes and suffixes of the compared terms and uses a modified version of the *String Metric* presented in ref. (Stoilos, Stamou & Kollias 2005) to calculate similarity $LexSim$ between two terms str_1 and str_2 as:

$$LexSim(str_1, str_2) = \left(CC + \frac{2 * LCS(str_1, str_2)}{l_1 + l_2}\right)^{MSP} \quad (5)$$

where CC is the number of common characters between the two strings, LCS is the least commom substring, l_i is the length of term i and MSP is the maximum of the Levenshtein distances of the suffixes and prefixes of the two terms.

The use of a lexical similarity allows the evaluation of compatibility using a more relaxed approach, taking into account not only identical terms but also terms that score above a certain threshold. For this specific application, threshold has been set to 0.60. Some indicative results of lexical comparison are given in Table 1, which demonstrate the performance of the lexical similarity algorithm. Specifically, it is apparent that the use of "s" to denote plural does not affect the similarity score. This is because the use of plural or singular is a mainly dictated by the knowledge engineer's preference and intuition. Another important aspect of that algorithm is that similarity is not vastly affected by the different suffixes, e.g. PET_1, PET_2, which are a very common phenomenon in the domain of CPE. Finally, the last pair presented, Article, Particle, is a pair of terms commonly used in tests for lexical similarity algorithms to test performance for two strings that are almost identical and, at the same time, semantically unrelated. In this case, it is apparent that the algorithm performs well in that aspect as well.

Table 1 Lexical Similarity

Pair	Score
Acid – Acids	1.00
Industrial – Industry	0.82
PET_1 – PET_2	0.83
HDPE – LDPE	0.88
Article - Particle	0.51

As it will be shown by the experiments in Section 4, the use of lexical similarity does not affect the results.

3.2 Structural information

3.2.1 Identifying structural information

Structure of ontologies represent tacit knowledge of the respective domains and hence *fingerprint* which can be used to assess the similarity between them. Although two ontologies can describe the same domain using the same terms, they are still not considered identical unless they share the same structure. This aspect is more closely associated with the ontology similarity rather than the ontology compatibility because reusing ontology does not require similar structure. Structural information used in this work includes the number of levels of the taxonomy and the number of concepts which indicates how much ontology expands in width (concepts per level) or breadth (number of levels). In consequence, structural information is used to assess the relative degree of confidence.

3.2.2 Calculating the degree of confidence

The degree of confidence (doc) is a metric quantifying the structural compatibility of the two or more ontologies. To this end, the structural compatibility is defined as the level at which two ontologies are considered well-suited for each other, i.e. a top level ontology is not appropriate to be a subset of a domain ontology. In order to identify the set-subset relation between two ontologies, the domain of each ontology has to be defined, the process which currently attracts numerous research activities. Instead, we assume that the target ontology is a set and the candidate ontologies are the possible subsets. In consequence, doc is calculated by following 3 steps explained below. To demonstrate the whole process, a small experiment using the Conference (Šváb et al. 2005) set of ontologies is presented in Table 2. The assumption is that the Conference ontology is our target ontology and also the set for which compatible ontologies are aimed to find in order to extend the described knowledge. Table 2 contains information about the elements of each ontology, i.e. object properties, data type properties, instances, external resources/imports, languages used and number of concepts/size.

Table 2 Candidate ontologies representing conference domain

Ontology	Object Properties	Data type Properties	Instances	Imports	Languages	Size
confOf	yes	yes	no	no	en	38
conference	yes	yes	no	no	en	59
confious	yes	yes	yes	no	en	56
crs_dr	yes	yes	no	no	en	14
edas	yes	yes	yes	no	en, ru, nl	103
ekaw	yes	no	no	yes	en, fr, cn	73
iasted	yes	yes	yes	no	en	140

Step 1: extract the number of levels OL and the number of concepts OC for each of considered ontologies which accounts for two dimensions of the size of an ontology. Information regarding the conference set ontologies is presented in Table 3.

Table 3 Size information for conference ontologies

Ontology	OL	OC
conference	7	59
confious	3	56
confOf	3	38
crd_rs	2	14

edas	4	103
iasted	6	140
ekaw	6	78

Step 2: calculate the degree of confidence doc, which accounts for size of the ontologies as

$$doc = log_{10}\left[\left\|\frac{OC_{target} - OC_{candidate}}{OL_{target}}\right\|\right] \quad (6)$$

The results of calculated confidence doc for the conference ontology are shown in Table 4.

Table 4 Results for conference ontologies (*doc*)

	conference
confious	-0.37
confOf	0.48
crd_rs	0.81
edas	0.80
iasted	1.06
ekaw	0.43

As such, the metric doc accounts for the variability of granularity between the target ontology and the candidate ontologies. The resulting values vary and also include negative values, hence creating the need for normalisation.

Step 3: Normalise the degree of confidence is calculated as;

$$doc_{norm} = \frac{doc_i - doc_{min}}{doc_{max} - doc_{min}} \quad (7)$$

with results for the conference ontology shown in Table 5. Here doc_{min} and doc_{max} represent the range of calculated values including all considered ontologies i.

Table 5 Normalised *doc* for conference ontologies

	conference
confious	**0.00**
confOf	0.59
crd_rs	0.82
edas	0.81
iasted	**1.00**
ekaw	0.56

3.3 Calculating Similarity Between Target and Candidate Ontologies

A combination of cosine similarity and Euclidean distance are used for similarity calculation and are calculated for all vectors representing language vectors, terminology vectors, data types vectors and external resources vectors. The results are then aggregated into a single similarity measure Sim.

3.3.1 Cosine Similarity

Cosine similarity $CoSim$ accounts for the correlation of the values of two or more vectors. For two vectors a and b, cosine similarity is calculated as:

$$CoSim(a,b) = \frac{\sum_{i=1}^{n} a_i * b_i}{\sqrt{\sum_{i=1}^{n}(a_i)^2} * \sqrt{\sum_{i=1}^{n}(b_i)^2}} \quad (8)$$

where n is the number of elements in each vector (vectors must have the same size). For example, for the language vectors for ontologies ekaw and iasted are represented as:

$$L_{ekaw} = (1,1,1)$$

$$L_{iasted} = (1,0,0)$$

where the three dimensions of the vectors represent the three languages (en, ru, nl).

The resulting score ranges between 0 and 1. Cosine similarity is not affected by the magnitude of the two vectors, only by their direction, and for that reason a scale sensitive measure, the Euclidean distance, is also employed.

3.3.2 Euclidean Distance

Euclidean distance $Eucl$ is sensitive to the magnitude of vectors. For vectors a and b, Euclidean distance is calculated as;

$$Eucl(a,b) = \sqrt{\sum_{i=1}^{n}(a_i - b_i)^2} \quad (9)$$

where n is the number of elements in each vector (vectors must have the same size). For normalisation $NormEucl$, Euclidean distance of a vector is divided by the maximum value of the calculated Euclidean distances as;

$$NormEucl_i = \frac{Eucl_i}{Eucl_{max}} \quad (10)$$

Finally, this is subtracted from 1 to be converted to a similarity measure $EuSim$ as;

$$EuSim_i = 1 - NormEucl_i \quad (11)$$

The two similarity scores, namely the cosine similarity and the Euclidean similarity are aggregated into a single measure Sim as

$$Sim_i = \frac{CoSim_i + EuSim_i}{2}, \ i \in O \quad (12)$$

where O is the set of candidate ontologies.

3.3.3 Calculating Similarity for Conference Ontologies

Most of the ontologies are only modelled in English. Similarity of ontologies that are modelled in more than one languages differ significantly. Table 6 presents the similarity Sim calculated by observing equations (8) – (12), between the language vectors (eq. (1)) of target ontology (*conference*) and candidate ontologies.

Table 6 Natural language similarity for conference ontologies

	confOf	conference	confious	crs_dr	Edas	ekaw	iasted
conference	1.00	1.00	1.00	1.00	0.4351	0.4351	1.00

For the terminology similarity measure, all the terms have been extracted, converted into vectors, as described by eq. (4), and by observing the equations (8) – (12), which yields the results presented in Table 7.

Table 7 Terminology similarity for conference ontologies

	confOf	conference	confious	crs_dr	Edas	ekaw	iasted
conference	0.2818	1.00	0.1182	0.2859	0.0896	0.2697	0.2191

The same approach is followed for the calculation of data type similarity (eq. (2)). Results are presented in Table 8.

Table 8 Encoding (datatype) similarity for conference ontologies

	confOf	conference	confious	crs_dr	Edas	ekaw	iasted
conference	0.5516	1.00	0.6445	0.2110	0.5526	0.00	0.4560

Next, the similarity of external resources is calculated from the vectors that represent the external resources used in each ontology (eq. (3)). The limited variations on external resources used, is apparent in the similarity results presented in Table 9.

Table 9 External resources similarity for conference ontologies

	confOf	conference	confious	crs_dr	edas	ekaw	iasted
conference	1.00	1.00	1.00	1.00	1.00	0.3536	1.00

3.4 Aggregating Similarity

3.4.1 Aggregation Process

Finally, the similarity measures of each aspect, the Sim metric for each of the elements presented in eq. (1) – (4), are aggregated (Table 10) into a single measure S between two ontologies A and B as:

$$S_{AB} = \frac{\sum_i^O Sim_i}{\sum_i^O w_i} \tag{13}$$

where $i \in O$, and O is the set of candidate ontologies, represents the ontology features as described in Section 3.1, Sim is the similarity measure from eq. (12) and w the weighting factor of each feature i. At present, weighting factors are set to 1 for all elements; natural language L, terminology T, data types E and external resources R.

Table 10 1st Aggregation step - Aggregation of different features

	confOf	conference	confious	crs_dr	edas	ekaw	iasted
conference	0.8195	1.0000	0.7930	0.9070	0.6864	0.3727	0.8264

Most of the high scoring ontologies are general ontologies with granularity level ranging from very generic, e.g. *person*, *event* ontologies, to very specific, e.g.. *early_paid_applicant*, *workshop* ontologies. On the other hand, low scoring ontologies, are mainly ontologies that represent specific conferences. These ontologies contain scientific specific terminology, e.g. *MultimediaTopic* ontology, different natural languages and instances of conference details, e.g. place of conference and currency.

3.5 Aggregating Similarity and Degree of Confidence

Finally, the degree of confidence doc is aggregated with the similarity score S to give the compatibility score C of the ontologies as:

$$C_{ij} = 0.7 * S_{ij} + 0.3 * doc_{ij} \tag{14}$$

where S_{ij} is the similarity between ontologies i and j and doc_{ij}, is the degree of confidence between ontologies i and j. It is evident from eq. (14) that the similarity S carries a higher weight

than the degree of confidence doc. This is because the degree of confidence merely provides an indication of the structural compatibility.

The results of aggregated values for the conference ontology in Table 11 show that the most compatible ontology for the target ontology *conference*, is *iasted* ontology and the second highest is *crs_dr* ontology. It is apparent that general ontologies with granularity level ranging from very generic, i.e. *person*, *event*, to very specific i.e. *early_paid_applicant*, *workshop*, are considered more suitable for reuse. On the other hand, low scoring ontologies, are mainly ontologies that represent specific conferences. These ontologies contain scientific specific terminology, i.e. *MultimediaTopic*, different natural languages and instances of conference details (place of conference and currency).

Table 11 2nd Aggregation step (C) - Aggregating *doc* and *Similarity*

	confOf	confious	crs_dr	Edas	ekaw	iasted
conference	0.70475	0.3965	0.8635	0.7482	0.46635	0.9132

4. The Compatibility for Reuse Algorithm (CoRAl) Outlined

The ontology compatibility algorithm (**CoRAl**) presented in Figure 1, is outlined in the following steps.

Step R.1: Extract ontology metadata such as terminology, languages, datatypes, imports and size and form vectors observing as defined by eq. (1) - (4);

Step R.2: Calculate compatibility measure C by using eq. (5) and(14);

Step R.3: Rank results obtained from Step R.2 by their values;

The process of ontology development affects its reusability potential. By observing the following guidelines can improve the reuse process:

Step D.1: Modularise the ontology given your needs by e.g. domain, level, property.

Step D.2: Follow "good practice" conventions by considering naming conventions, minimised bias, consistent encoding, annotations.

Step D.3: Reuse existing ontologies and existing vocabularies, lexicons etc.

5. eSymbiosis Ontology Experiment

The experiment with eSymbiosis ontology is presented. It involves ontologies developed for materials and chemical substances. The eSymbiosis ontology (Trokanas et al. 2012) has been developed for the eSymbiosis project (Cecelja et al. 2014). It represents knowledge about Industrial Symbiosis, including waste, materials, energy and processing technologies (Raafat et al. 2013). The ontology is used for user registration and for formation of symbiotic networks by input/output matching (Trokanas, Cecelja & Raafat 2014a).

Three candidate ontologies have been considered for the integration andreuse; the *chemElem* ontology, the *substance* ontology both part of the SWEET (Raskin, Pan 2005, Raskin, Pan 2003) ontologies developed by NASA and the *substance_class* ontology which is part of the OntoCAPE ontology (Morbach, Yang & Marquardt 2007, Marquardt et al. 2010), all denoted as C_1, C_2 and C_3, respectively. The eSymbiosis ontology is the target ontology, denoted as T.

The *chemElem* ontology represents chemical elements which are defined in this ontology as "*pure chemical substances consisting of one type of atom distinguished by its atomic number*". In terms of number of concepts, *chemElem* is the largest of the candidate ontologies, consisting of 2363 concepts. *Substance* ontology represents non-living building blocks of nature including particles and chemical compounds (Raskin, Pan 2003). Finally, *substance_class* of OntoCAPE ontology (Wiesner, Morbach & Marquardt 2007), a subclass of *Material* class, represents pure substances and mixtures.

The first step involves the extraction of the metadata from each ontology, as described in Section 3.1 with focus on extraction of the terminology. As mentioned earlier, this work attempts to overcome this challenge with the use of existing term extraction software. This, however, requires a certain trade-off between accuracy and simplicity.

The similarity score calculated from the vectors formed from the metadata of the ontology (presented in Section 3.1) is presented in Table 12, which contains aggregated similarity score S, for each candidate ontology, obtained from eq. (13). This aggregated metric S consists of the similarity of all types of metadata including terminology, data types, languages and external resources.

Table 12 Metadata Similarity

	Metadata Similarity
chemElem	0.3748
Substance	0.3801
substance_class	0.3259

The second step involves the extraction of the structural information from each ontology, required for the calculation of the degree of confidence (Section 3.2.2). This information is presented in Table 13. By observing eq. (6), the number of levels OL of the target ontology and the number of concepts OC for the target and candidate ontologies are extracted.

Table 13 Structural information for candidate ontologies

Ontology	Number of Concepts	Number of Levels
eSymbiosis	2250	10
chemElem	2363	N/A

Substance	483	N/A
substance_class	94	N/A

The degree of confidence is calculated for each of the candidate ontologies by using eq. (6) – (7), yielding the results presented in Table 14. The results in this table present the normalised doc values for each of the candidate ontologies.

Table 14 Degree of Confidence for Materials

Degree of Confidence (doc)	
chemElem	0.4513
Substance	0.9600
substance_class	1.0000

Finally, the two scores (Table 12 and Table 14) are aggregated (Table 15) by using the eq. (14). Table 15 presents the final results of the compatibility C between the target and candidate ontologies. In specific, *Substance* ontology is identified as the most compatible between the three candidates, while the *chemElem* ontology is the least compatible. *chemElem* ontology scored low in both doc and S measures. Not only it does not share many commonalities in terms of terminology or other metadata with the target ontology, but it is also bigger than the target ontology thus scoring low in doc. Although *substance_class* ontology had the highest doc score (Table 14), its low metadata similarity S (Table 12) affected the final compatibility score C.

Table 15 Aggregated Scores

Aggregated Compatibility Score	
chemElem	0.39780
Substance	0.55490
substance_class	0.52810

With the use of lexical similarity for the comparison of terminology (equation (5)), the results are not significantly affected (Table 16), leading to the conclusion that the extraction of terminology using existing software and the identification of only identical terms is sufficient for the high-level comparison that this work proposes.

Table 16 Aggregated Scores (with Lexical Similarity)

Aggregated Compatibility Score	
chemElem	0.4020
Substance	0.5616
substance_class	0.5330

6. Conclusions

This paper presents an algorithm for the evaluation of ontologies for the purpose of reusing. The presented algorithm benefits from information about ontologies, such as the terminology of ontologies and their structure, to enables calculation of a single compatibility metric. For this, the

algorithm relies on the ontology high level information which is readily available, easy to extract and does not require any special expertise in ontology matching. It also takes advantage of existing terminology extraction tools in an effort to simplify the process. The use and performance of the algorithm has been validated with two experiments; one, the Conference ontologies, is a well-elaborated benchmark for ontology matching and alignment, whereas the second experiment is based on the eSymbiosis ontology, which is an application ontology that supports an Industrial Symbiosis web service.

7. Acknowledgements

This work has been partly funded by the European Commission (LIFE09 ENV/GR/000300) and the UK Engineering and Physical Sciences Research Council (EPSRC).

8. References

Abanda, F.H., Tah, J.H.M. & Duce, D. 2013, "PV-TONS: A photovoltaic technology ontology system for the design of PV-systems", *Engineering Applications of Artificial Intelligence*, vol. 26, no. 4, pp. 1399-1412.

Aho, A.V. & Corasick, M.J. 1975, "Efficient string matching: an aid to bibliographic search", *Communications of the ACM*, vol. 18, no. 6, pp. 333-340.

Alani, H. & Brewster, C. 2005, "Ontology ranking based on the analysis of concept structures", *Proceedings of the 3rd international conference on Knowledge capture*ACM, , pp. 51.

Alani, H., Brewster, C. & Shadbolt, N. 2006, "Ranking ontologies with AKTiveRank" in *The Semantic Web-ISWC 2006* Springer, , pp. 1-15.

Alonso, L., Bas, L., Bellido, S., Contreras, J., Benjamins, R. & Gomez, M. 2005, "WP10: Case Study eBanking D10. 7 Financial Ontology", *Data, Information and Process Integration with Semantic Web Services, FP6-507483*, .

Anchovy, 2013, , *Cross-platform tools for translators* [Homepage of Maxprograms], [Online]. Available: http://www.maxprograms.com/products/anchovy.html [2014, 10/30].

Bock, C., Zha, X., Suh, H. & Lee, J. 2010, "Ontological product modeling for collaborative design", *Advanced Engineering Informatics*, vol. 24, no. 4, pp. 510-524.

Brandt, S.C., Morbach, J., Miatidis, M., Theißen, M., Jarke, M. & Marquardt, W. 2008, "An ontology-based approach to knowledge management in design processes", *Computers & Chemical Engineering*, vol. 32, no. 1, pp. 320-342.

Bray, T., Paoli, J., Sperberg-McQueen, C.M., Maler, E. & Yergeau, F. 1998, "Extensible markup language (XML)", *World Wide Web Consortium Recommendation REC-xml-19980210.http://www.w3.org/TR/1998/REC-xml-19980210*, .

Budanitsky, A. & Hirst, G. 2006, "Evaluating wordnet-based measures of lexical semantic relatedness", *Computational Linguistics*, vol. 32, no. 1, pp. 13-47.

Cantador, I., Fernández, M. & Castells, P. 2007, "Improving ontology recommendation and reuse in WebCORE by collaborative assessments", .

Casellas, N., Blázquez, M., Kiryakov, A., Casanovas, P., Poblet, M. & Benjamins, R. 2005, "OPJK into PROTON: Legal domain ontology integration into an upper-level ontology", *On the Move to Meaningful Internet Systems 2005: OTM 2005 Workshops*Springer, , pp. 846.

Ceccaroni, L., Cortés, U. & Sànchez-Marrè, M. 2000, "WaWO-An ontology embedded into an environmental decision-support system for wastewater treatment plant management", .

Cecelja, F., Raafat, T., Trokanas, N., Innes, S., Smith, M., Yang, A., Zorgios, Y., Korkofygas, A. & Kokossis, A. 2014, "e-Symbiosis: technology-enabled support for industrial symbiosis targeting SMEs and innovation", *Journal of Cleaner Production*, vol. To appear.

Cheatham, M. & Hitzler, P. 2013, "String similarity metrics for ontology alignment" in *The Semantic Web–ISWC 2013* Springer, , pp. 294-309.

Corcho, O., Fernández-López, M., Gómez-Pérez, A. & López-Cima, A. 2005, "Building legal ontologies with METHONTOLOGY and WebODE" in *Law and the semantic web* Springer, , pp. 142-157.

David, J. & Euzenat, J. 2008, "Comparison between ontology distances (preliminary results)" in *The Semantic Web-ISWC 2008* Springer, , pp. 245-260.

Ehrig, M. 2006, *Ontology alignment: bridging the semantic gap*, Springer.

Fernandes, R.P., Grosse, I.R., Krishnamurty, S., Witherell, P. & Wileden, J.C. 2011, "Semantic methods supporting engineering design innovation", *Advanced Engineering Informatics*, vol. 25, no. 2, pp. 185-192.

Frantzi, K., Ananiadou, S. & Mima, H. 2000, "Automatic recognition of multi-word terms:. the c-value/nc-value method", *International Journal on Digital Libraries*, vol. 3, no. 2, pp. 115-130.

Gruber, T.R. 1995, "Toward principles for the design of ontologies used for knowledge sharing", *International journal of human computer studies*, vol. 43, no. 5, pp. 907-928.

Gruber, T.R. 1993, "A translation approach to portable ontology specifications", *Knowledge acquisition*, vol. 5, no. 2, pp. 199-220.

Hawalah, A. & Fasli, M. 2011, "A graph-based approach to measuring semantic relatedness in ontologies", *Proceedings of the International Conference on Web Intelligence, Mining and Semantics*ACM, , pp. 29.

Jiang, J.J. & Conrath, D.W. 1997, "Semantic similarity based on corpus statistics and lexical taxonomy", *arXiv preprint cmp-lg/9709008*, .

Jing Ni, Jiu Yi & Suping Ni 2011, "A Practical Development of Knowledge Management Model for Petrochemical Product Family", *Information Management, Innovation Management and Industrial Engineering (ICIII), 2011 International Conference on*, pp. 197.

Kraines, S., Batres, R., Koyama, M., Wallace, D. & Komiyama, H. 2005, "Internet-Based Integrated Environmental Assessment Using Ontologies to Share Computational Models", *Journal of Industrial Ecology*, vol. 9, no. 3, pp. 31-50.

Lame, G. 2005, "Using NLP techniques to identify legal ontology components: concepts and relations" in *Law and the Semantic Web* Springer, , pp. 169-184.

Lee, J. & Suh, H. 2007, "Owl-based product ontology architecture and representation for sharing product knowledge on a web", ASME, .

Levenshtein, V.I. 1966, "Binary codes capable of correcting deletions, insertions and reversals", *Soviet physics doklady*, pp. 707.

Li, Y., Bandar, Z.A. & McLean, D. 2003, "An approach for measuring semantic similarity between words using multiple information sources", *Knowledge and Data Engineering, IEEE Transactions on*, vol. 15, no. 4, pp. 871-882.

Lin, H. & Harding, J.A. 2007, "A manufacturing system engineering ontology model on the semantic web for inter-enterprise collaboration", *Computers in Industry*, vol. 58, no. 5, pp. 428-437.

Lonsdale, D., Embley, D.W., Ding, Y., Xu, L. & Hepp, M. 2010, "Reusing ontologies and language components for ontology generation", *Data & Knowledge Engineering*, vol. 69, no. 4, pp. 318-330.

Maedche, A. & Staab, S. 2002, "Measuring similarity between ontologies" in *Knowledge engineering and knowledge management: Ontologies and the semantic web* Springer, , pp. 251-263.

Marquardt, W., Morbach, J., Wiesner, A. & Yang, A. 2010, *OntoCAPE: a re-usable ontology for chemical process engineering*, Springer.

Mihalcea, R., Corley, C. & Strapparava, C. 2006, "Corpus-based and knowledge-based measures of text semantic similarity", *AAAI*, pp. 775.

Morbach, J., Wiesner, A. & Marquardt, W. 2009a, "OntoCAPE—A (re) usable ontology for computer-aided process engineering", *Computers & Chemical Engineering*, vol. 33, no. 10, pp. 1546-1556.

Morbach, J., Wiesner, A. & Marquardt, W. 2009b, "OntoCAPE—A (re) usable ontology for computer-aided process engineering", *Computers & Chemical Engineering,* vol. 33, no. 10, pp. 1546-1556.

Morbach, J., Yang, A. & Marquardt, W. 2007, "OntoCAPE—A large-scale ontology for chemical process engineering", *Engineering Applications of Artificial Intelligence,* vol. 20, no. 2, pp. 147-161.

Muñoz, E., Capón-García, E., Moreno-Benito, M., Espuña, A. & Puigjaner, L. 2011, "Scheduling and control decision-making under an integrated information environment", *Computers & Chemical Engineering,* vol. 35, no. 5, pp. 774-786.

Muñoz, E., Espuña, A. & Puigjaner, L. 2010, "Towards an ontological infrastructure for chemical batch process management", *Computers & Chemical Engineering,* vol. 34, no. 5, pp. 668-682.

Muñoz, E., Capón-García, E., Espuña, A. & Puigjaner, L. 2012, "Ontological framework for enterprise-wide integrated decision-making at operational level", *Computers & Chemical Engineering,* vol. 42, no. 0, pp. 217-234.

Muñoz, E., Capón-García, E., Laínez, J.M., Espuña, A. & Puigjaner, L. 2013, "Integration of enterprise levels based on an ontological framework", *Chemical Engineering Research and Design,* vol. 91, no. 8, pp. 1542-1556.

Musen, M.A., Noy, N.F., Shah, N.H., Whetzel, P.L., Chute, C.G., Story, M.A., Smith, B. & NCBO team 2012, "The National Center for Biomedical Ontology", *Journal of the American Medical Informatics Association : JAMIA,* vol. 19, no. 2, pp. 190-195.

Panetto, H., Dassisti, M. & Tursi, A. 2012, "ONTO-PDM: Product-driven ONTOlogy for Product Data Management interoperability within manufacturing process environment", *Advanced Engineering Informatics,* vol. 26, no. 2, pp. 334-348.

Pinto, H.S. & Martins, J. 2000, "Reusing ontologies", *AAAI 2000 Spring Symposium on Bringing Knowledge to Business Processes*AAAI Press, , pp. 7.

Raafat, T., Trokanas, N., Cecelja, F. & Bimi, X. 2013, "An ontological approach towards enabling processing technologies participation in industrial symbiosis", *Computers & Chemical Engineering,* vol. 59, no. 0, pp. 33-46.

Raafat, T., Cecelja, F., Yang, A. & Trokanas, N. "Semantic Support for Industrial Symbiosis Process" in *Computer Aided Chemical Engineering* Elsevier, , pp. 452-456.

Raskin, R. & Pan, M. 2003, "Semantic web for earth and environmental terminology (sweet)", *Workshop on Semantic Web Technologies for Searching and Retrieving Scientific Data at ISWC 2003*.

Raskin, R.G. & Pan, M.J. 2005, "Knowledge representation in the semantic web for Earth and environmental terminology (SWEET)", *Computers & Geosciences,* vol. 31, no. 9, pp. 1119-1125.

Schulz, S., Seddig-Raufie, D., Grewe, N., Röhl, J., Schober, D., Boeker, M. & Jansen, L. 2012, "Guideline on Developing Good Ontologies in the Biomedical Domain with Description Logics", .

Simperl, E. 2009, "Reusing ontologies on the Semantic Web: A feasibility study", *Data & Knowledge Engineering*, vol. 68, no. 10, pp. 905-925.

Soergel, D., Lauser, B., Liang, A., Fisseha, F., Keizer, J. & Katz, S. 2006, "Reengineering thesauri for new applications: the AGROVOC example", *Journal of digital information*, vol. 4, no. 4.

Stoilos, G., Stamou, G. & Kollias, S. 2005, "A string metric for ontology alignment", *The Semantic Web–ISWC 2005*, , pp. 624-637.

Suresh, P., Joglekar, G., Hsu, S., Akkisetty, P., Hailemariam, L., Jain, A., Reklaitis, G. & Venkatasubramanian, V. 2008, "Onto MODEL: Ontological mathematical modeling knowledge management" in *Computer Aided Chemical Engineering* Elsevier, , pp. 985-990.

Šváb, O., Svátek, V., Berka, P., Rak, D. & Tomášek, P. 2005, "Ontofarm: Towards an experimental collection of parallel ontologies", *Poster Track of ISWC*, vol. 2005.

Swoogle, 2009, [Homepage of University of Maryland], [Online]. Available: http://swoogle.umbc.edu/ [2014, 10/28].

Trokanas, N., Cecelja, F., Raafat, T. & Innes, P. 2013a, "An Ontological Approach Enabling "a priori" Quantitative Assessment of Industrial Symbiosis Networks", AIChE Annual Meeting 2013, , November 2013.

Trokanas, N., Raafat, T., Cecelja, F. & Kokossis, A. 2013b, "OFIS - Ontological Framework for Industrial Symbiosis", *Computer Aided Chemical Engineering*, vol. 32, no. 23rd European Symposium on Computer Aided Process Engineering, pp. 523-528.

Trokanas, N., Raafat, T., Cecelja, F., Kokossis, A. & Yang, A. 2012, "Semantic Formalism for Waste and Processing Technology Classifications Using Ontology Models", *Computer-Aided Chemical Engineering*, vol. 30, pp. 167-171.

Trokanas, N., Cecelja, F. & Raafat, T. 2014a, "Semantic input/output matching for waste processing in industrial symbiosis", *Computers & Chemical Engineering*, vol. 66, pp. 259-268.

Trokanas, N., Cecelja, F. & Raafat, T. 2014b, "Towards a Re-Usable Ontology for Waste Processing", *Computer Aided Chemical Engineering*, vol. 33, no. 0, pp. 841-846.

Uschold, M., Healy, M., Williamson, K., Clark, P. & Woods, S. 1998, "Ontology reuse and application", *Formal ontology in information systems*, pp. 192.

Venkatasubramanian, V., Zhao, C., Joglekar, G., Jain, A., Hailemariam, L., Suresh, P., Akkisetty, P., Morris, K. & Reklaitis, G.V. 2006, "Ontological informatics infrastructure

for pharmaceutical product development and manufacturing", *Computers & Chemical Engineering*, vol. 30, no. 10, pp. 1482-1496.

Whetzel, P.L., Noy, N.F., Shah, N.H., Alexander, P.R., Nyulas, C., Tudorache, T. & Musen, M.A. 2011, "BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications", *Nucleic acids research*, vol. 39, no. Web Server issue, pp. W541-5.

Wiesner, A., Morbach, J. & Marquardt, W. 2011, "Information integration in chemical process engineering based on semantic technologies", *Computers & Chemical Engineering*, vol. 35, no. 4, pp. 692-708.

Wiesner, A., Morbach, J. & Marquardt, W. 2007, "An overview on OntoCAPE and its latest applications", *Proceedings of the 2007 AIChE Annual Meeting*.

Zhang, W. & Yin, J. 2008, "Exploring Semantic Web technologies for ontology-based modeling in collaborative engineering design", *The International Journal of Advanced Manufacturing Technology*, vol. 36, no. 9-10, pp. 833-843.

Zhang, Z., Zhang, C. & San Ong, S. 2000, "Building an ontology for financial investment" in *Intelligent Data Engineering and Automated Learning—IDEAL 2000. Data Mining, Financial Engineering, and Intelligent Agents* Springer, , pp. 308-313.