e-Document Standards as Background Knowledge in Context-Based Ontology Matching

Audun Vennesland(✉)

Norwegian University of Science and Technology, Trondheim, Norway
audun.vennesland@idi.ntnu.no

Abstract. Ontology matching is the process of finding correspondence between heterogeneous ontologies and consequently support semantic interoperability between different information systems. Using contextual information relative to the ontologies being matched is referred to as context-based ontology matching and is considered one promising direction of improving the matching performance. This PhD investigates how such contextual information, often residing in disparate sources and represented by different formats, can be optimally represented to ontology matching systems and how these systems best can employ this context to produce accurate and correct correspondences. Currently we are investigating how the international e-Document standard Universal Business Language from the transport logistics domain can provide useful context when matching domain ontologies for this particular domain. Early evaluation tests and analysis of the results suggest that the current version of the Universal Business Language ontology does not impact on the matching results and that further reconfiguration and enhancements are needed.

1 Introduction

The use of external context as input to identifying correspondence between heterogenous ontologies is seen as a promising approach within ontology matching [1,2] and is referred to as context-based ontology matching. When two ontologies are to be matched, they often lack a common ground on which comparisons can be based. In context-based ontology matching the intention is to establish such a common ground using the relations between the ontologies being matched and their environment represented by external resources [3].

Often the external resources are represented by formal or less formal ontologies [4,5] or other sources of context. However, the quality of the external sources varies [6], something which threatens the validity of the identified alignments. Moreover, even if the application of semantic technologies is mature in some domains the use of semantic technologies is still limited in many other domains and formalized context is difficult to come by. A survey among ontology matching practitioners [2] states that integration of domain knowledge into alignment techniques is a significant challenge. Hence, investigation of other and reliable sources of context as well as improved techniques for exploiting such context is required.

F. Gandon et al. (Eds.): ESWC 2015, LNCS 9088, pp. 806–816, 2015.
DOI: 10.1007/978-3-319-18818-8_53

e-Documents standards specify through message specifications (including XSD schemas), business process descriptions, narratives, instance data and other material how information should be exchanged electronically.

Our approach is to identify appropriate methods for transforming such contextual information into a more formal representation, investigate how it can be optimally employed by ontology matchers, and evaluate its impact on the ontology matching process. e-Document standards is selected as a case based on their inherent qualities (presumably quality assured and sustainable information, mixture of domain-specific and more generic information elements, a combination of structured and unstructured formats, and the availability of proper instance data).

2 State of the Art

2.1 Context-Based Ontology Matching

Ontology matching is the process of identifying correspondences (alignments) between heterogeneous ontologies that enable the information systems applying the ontologies to interpret data being communicated among them. Euzenat and Shvaiko [3] distinguishes between *element-level techniques* and *structure-level techniques*. Element-level techniques focuses on the ontology entities (or instances of them) themselves while disregarding their relations with other entities (or instances of them). Examples of such techniques are *string-based similarity* measures (which might identify correspondences based on name similarity), *language-based techniques* (e.g. using NLP and lexical resources to capture conceptual similarity and hence correspondence between entities not necessarily having the same name) and *informal- or formal resource-based techniques* which employs external sources, either formal ones such as ontologies or informal sources such as web sites or documents, to improve the matching operation. Structure-level techniques on the other hand analyze how entities (or their instances) appear together in a structure. Some examples of structure-level techniques are *graph-based techniques* (such as the use of graph algorithms to identify similar neighboring entities and relations and thereby calculate correspondence), *model-based techniques* (e.g. the use of description logic reasoning in order to identify correspondence on the basis of semantic interpretation) and *instance-based techniques* (for example using statistical methods to compare sets of class instances to identify correspondence between these classes).

Context-based ontology matching uses external resources in order to help establish a common ground (contextualization) between the ontologies to be matched and is considered a promising approach [1,2]. These external resources can be formal or informal. Formal resources are typically ontological structures using a formal language such as OWL or RDF. Different levels of ontological structures have previously been applied to aid the ontology matching task, including the use of upper-level ontologies [4], a combination of many ontologies [5], use of less formal resources such as WordNet [7], and use of informal resources such as web sites to identify correspondences [8].

According to [9] the use of external context can be categorized into three use cases: (i) using the external context as a reference (e.g. using linguistic resources to find synonyms that can help establish similarity among entities), (ii) as an oracle (i.e. replacing the human expert when validating suggested alignments from the matching operation by querying external background knowledge) or (iii) as a mediator (i.e. mapping entities from the source and target ontologies to an intermediate ontology and thereby identify correspondence).

2.2 e-Document Standards

Useful contextual information resides in e-Document standards (a.k.a. business document standards or e-business standards) and associated material. Although this work initially focuses on standards related to the transport logistics domain similar standards developed using similar processes and in close cooperation between standards developing organizations and domain experts exist in other domains. Some examples are general trade [10,11], public and private procurement [12], food and agriculture [13], manufacturing, and consumer electronics. So even though the focus initially is on a specific domain, the approach should be generalizable to other domains also.

In the work so far we have focused on OASIS UBL (Universal Business Language). UBL is an OASIS standard providing a library of e-Documents and information elements for the procurement and transport logistics domains. In the most recent version of the standard (version 2.1) the library consists of 698 classes (elements) with attributes and associations, encompassing both domain specific and more generic elements. The entire library is represented in XSD schemas.

Figure 1, which represents an excerpt from the UBL library, exemplifies some of the possible context data available in the XSD schemas. In this particular example, which have been compressed for the sake of brevity, we see the TransportMeans complex type, the element JourneyID, which represents a property of TransportMeans (denoted by the cbc prefix) and the element OwnerParty, which represents an association from TransportMeans (denoted by the cac prefix). In addition to the hierarchical structure the schema also includes element definitions, alternative business terms, cardinalities, and data type definition.

Besides the data represented by the XSD schemas, the standard provides XML instances for all e-Documents included in the standards, offering an opportunity to apply instance-based matching techniques.

3 Problem Statement and Contributions

The use of contextual information is claimed to improve ontology matching operations and many state of the art ontology matchers utilize different types of external sources as support in their matching operations. A preliminary literature review suggests that few research endeavors have focused on identifying exactly which features are attractive w.r.t. expressing context and how these features should

```
<xsd:complexType name="TransportMeansType">
   <xsd:annotation>
      <xsd:documentation>
         <ccts:Component>
            <ccts:Definition>A class to describe a particular vehicle or vessel used for
               the conveyance of goods or persons.</ccts:Definition>
            <ccts:ObjectClass>Transport Means</ccts:ObjectClass>
            <ccts:AlternativeBusinessTerms>Conveyance</ccts:AlternativeBusinessTerms>
         </ccts:Component>
      </xsd:documentation>
   </xsd:annotation>
   <xsd:sequence>
      <xsd:element ref="cbc:JourneyID" minOccurs="0" maxOccurs="1">
         <xsd:annotation>
            <xsd:documentation>
               <ccts:Component>
                  <ccts:Definition>An identifier for the regular service schedule of this
                     means of transport.</ccts:Definition>
                  <ccts:Cardinality>0..1</ccts:Cardinality>
                  <ccts:ObjectClass>Transport Means</ccts:ObjectClass>
                  <ccts:PropertyTerm>Journey Identifier</ccts:PropertyTerm>
                  <ccts:DataType>Identifier. Type</ccts:DataType>
                  <ccts:AlternativeBusinessTerms>Voyage Number, Scheduled Conveyance
                     Identifier (WCO ID 205), Flight Number</ccts:AlternativeBusinessTerms>
               </ccts:Component>
            </xsd:documentation>
         </xsd:annotation>
      </xsd:element>
      <xsd:element ref="cac:OwnerParty" minOccurs="0" maxOccurs="1">
         <xsd:annotation>
            <xsd:documentation>
               <ccts:Component>
                  <ccts:Definition>The party that owns this means of
                     transport.</ccts:Definition>
                  <ccts:Cardinality>0..1</ccts:Cardinality>
                  <ccts:ObjectClass>Transport Means</ccts:ObjectClass>
                  <ccts:PropertyTerm>Party</ccts:PropertyTerm>
               </ccts:Component>
            </xsd:documentation>
         </xsd:annotation>
      </xsd:element>
```

Fig. 1. Relevant information from the UBL XSD schema

be optimally represented or modeled in order to support the matching process. Furthermore, the ontology matching systems are apparently very often targeting the (bio) medical domain, and although some of them also performs well in other tracks of the OAEI benchmark campaign, preliminary analysis suggest that their performance, and the underlying matching strategies, are to some degree domain dependent (see preliminary evaluation results in Sect. 6.2).

On this basis we have devised the following research questions:

RQ1: Which external sources of context can positively impact on the ontology matching performance?
RQ2: How should this external context be modeled in order to maximize its exploitation potential in ontology matching?
RQ3: Which ontology matching strategies are best suited for exploiting such context?

4 Research Approach

The approach in this PhD can be best characterized as a mixed-strategy design [14]. On the one hand we follow a fixed design introducing an experimental strategy where we measure the effect of manipulating the variables involved, evaluate

the effect of this manipulation, and analyze why the result became as it did. On the other hand we use a flexible design in the sense that the overall process is highly exploratory and that we in the end seek to establish some theories describing why this happened given the available data and the processing performed on them. The e-Document standards represents a case study, and presumingly these standards possess context that can positively contribute to improved ontology matching results. But this is also influenced by how the context data is processed by the ontology matchers (and the choice of algorithms employed).

The experimental strategy followed is illustrated in Fig. 2. The contextual information source is transformed to a formal representation. Using this formal representation as background knowledge, the experiment takes two ontologies as input and identifies correspondences among them using an ontology matcher. The resulting alignment from the matching operation is compared against a reference alignment holding the "true" set of correspondences among the ontologies and evaluated on the basis of commonly accepted metrics (precision, recall and F-measure). The evaluation results are then analyzed and if required the external context and/or the ontology matcher is reconfigured before the next iteration in the experiment cycle.

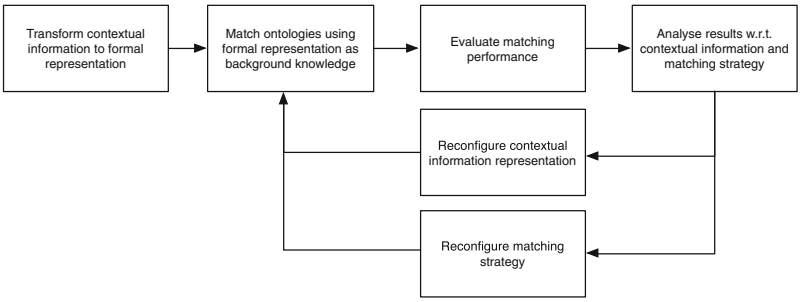

Fig. 2. Overall process

5 Preliminary Results

An initial development following the approach described in the previous chapter has been performed. The following developments are basically prerequisites for further investigation of the research questions defined in Sect. 3.

1. **Transform contextual information to formal representation:** XSLT (eXtensible Stylesheet Language Transformation) was used to transform data from UBL XSD schemas to an OWL ontology. This work extends the generic xsd2owl method [15] to fit with the characteristics of the UBL standard as well as other relevant e-Document standards in order to be generalizable to other settings.

2. **Match ontologies using formal representation (background knowledge):** The resulting OWL ontology from the previous step was used as background knowledge to support the ontology matching using an existing ontology matching system. In this initial setup we used the AgreementMakerLight (AML) ontology matcher [16,17].
3. **Evaluate matching performance and analyse results:** We manually developed a reference alignment holding the correct set of correspondences between the ontologies to be matched. This reference alignment was used as a baseline to compare the alignment from the ontology matching operation against. Currently this reference alignment is developed by the author and must be validated by domain experts. The evaluation measures used were precision, recall and F-measure (see Sect. 6 for more details about the evaluation).

Before reaching the next steps in the experiment cycle (Reconfigure contextual information representation and Reconfigure matching strategy) described in Sect. 4 a careful analysis of the evaluation scores must be conducted. This analysis must encompass an examination of the AML (and possibly other candidate ontology matchers) and how this matcher treats contextual information, the suitability (heterogeneity) of the to-be-matched ontologies, a verification of the correctness of the reference alignment, and an analysis of the UBL ontology and how this potentially could be enriched with additional semantics as well as data from supplementary material associated with the UBL 2.1 standard.

The generated UBL ontology is quite large, counting 1338 classes, 821 object properties and 1314 data properties. In addition to these declarations the ontology contains the following axioms: Sub Class, Object Property Domain, Object Property Range, Data Property Domain, Data Property Range, Functional Object Properties, and Functional Data Properties.

Figures 3, 4 and 5 shows the structures used by the three involved ontologies for describing the means of transport used in a transport logistics operation. As can be seen the structures are quite similar (as perceived by humans), but the naming conventions differ.

A difference between the generated UBL ontology and the other two is that while the Common Framework and LogiCO ontologies use sub class relations between classes, the relations between these entities in the UBL ontology are represented as object properties (as indicated by the dotted associations). As the UBL XSD schemas do not differentiate between associations representing what conceptually could be interpreted as sub classes (e.g. that MaritimeTransportMeans could be a sub class of TransportMeans) and other associations (e.g. that there is an association from TransportMeans to the MeasurementDimension element that enable a specification of the dimensions of the TransportMeans) we have treated these associations as object properties rather than identified those that are true sub class relations.

6 Preliminary Evaluation

Two ontologies from the transport logistics domain is being matched using an open source ontology matching system. The matching is performed both on

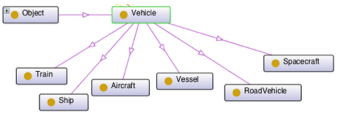

Fig. 3. Vehicle structure in the Common Framework ontology

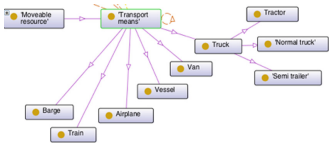

Fig. 4. Means of transport structure in the LogiCO ontology

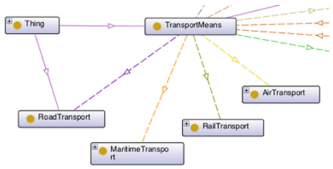

Fig. 5. Transport means structure in the UBL ontology

classes and properties, but only equivalence correspondences are identified. The ontologies to be matched are the LogiCO ontology [18] and the Common Framework ontology. The LogiCO ontology contains 153 classes, 96 properties and 14 individuals. The Common Framework ontology contains 331 classes, 283 properties and 1384 individuals. The Common Framework ontology imports the PROTON upper-layer ontology [19] for modeling generic concepts while LogiCO relies on DOLCE+DNS Ultralite [20].

In this first evaluation we are using the AgreementMakerLight ontology matching system. Although AML is primarily focused on matching ontologies for the biomedical domain it was chosen as an evaluation testbed since it specializes in the use of background knowledge, it is easy to reconfigure and extend, and has received top scores in the latest OAEI benchmarks which also includes matching ontologies outside of the biomedical domain [21].

6.1 Evaluation Scenarios

The following evaluation scenarios are run:

1. A comparative evaluation of one ontology matching operation including the two ontologies where one run is using the constructed ontology from e-Document standards and the other run is performed without any context information
2. A comparative evaluation of one ontology matching operation including the two evaluation ontologies using other sources of contextual information (e.g. WordNet)

The evaluation measures applied are precision, recall and F-measure. Precision measures the ratio of correctly found correspondences over the total number of

found correspondences. Recall measures the ratio of correctly found correspondences over the total number of expected correspondences. F-measure represents the harmonic mean of precision and recall and balances the importance of the other two evaluation measures [3].

6.2 Preliminary Evaluation Results

With a very limited verification of the UBL ontology and the reference alignment, the first evaluation run showed, as illustrated in Fig. 6, that when matching the two ontologies with no background knowledge using the default confidence level of 0.6 the precision was 77.1 %, the recall 57 % and the F-measure 66 %. The confidence level basically states that the ontology matcher trusts with 60 % certainty or above that the identified correspondence is correct.

When matching also the properties, this significantly decreased the scores with a precision of only 6.2 %, a recall of 57.1 % and an F-measure of 11.2 %. The highest scores were achieved when tuning the confidence degree from the default .6 to .9. When matching classes only this yielded a precision of 100 %, a recall of 52.4 % and an F-measure of 68.8 %.

When using the current version of the UBL ontology as background knowledge this did not influence the results at all, leaving the precision, recall and F-measure measures as they were when no background knowledge was employed. This was the case regardless of confidence level setting.

We also made an attempt using WordNet as contextual support in order to see how this compared to using the UBL ontology as background knowledge. When matching only the classes this produced the same scores as with the UBL ontology. When including the properties this actually lowered the scores resulting in a precision of 10.7 %, a recall of 33.9 % and an F-measure of 16.3 %.

These results indicatively show that the classes of the two ontologies are to some extent homogeneous, but that the properties of the two ontologies are very differently structured and named. Examining the resulting alignment from the matching operation manually we see that all identified correspondences are all exact string matches (e.g. 'Train' = 'Train'), while other (humanly) intuitive correspondences (e.g. 'CoordinateSystemName' vs. 'GeoCoordinateSystem') have

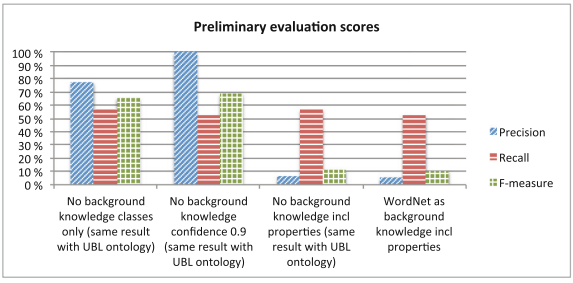

Fig. 6. Preliminary evaluation scores

not been captured by the matching system. Another observation is that the entities imported from the upper-layer ontologies being used (DOLCE+DNS Ultralite in the LogiCO ontology and PROTON in Common Framework ontology) are correctly matched. These entities are typically generic elements such as 'object' and 'event'.

When examining the three ontologies more in-depth we observe that if the ontology matching had utilized the object properties of the UBL ontology as anchors and from this derived that two classes using the same object property as anchors corresponds to each other the results could improve. An example of this is the 'TransportMeans' entity in the LogiCO ontology vs the 'Vehicle' entity in the Common Framework ontology (see Figs. 3, 4 and 5). The UBL ontology, similar to the LogiCO ontology, specifies the 'TransportMeans' class. In the UBL ontology this class has an object property 'hasAirTransport'. The range of this object property is the class 'AirTransport'. Further this 'AirTransport' class has an object property called 'hasAircraftID'. In the Common Framework ontology there is a class 'Aircraft' which is a sub class to 'Vehicle'. It should be possible to derive a correspondence between the 'Aircraft' class in the Common Framework ontology and the 'hasAircraftID' object property in the UBL ontology and hence that 'Vehicle' corresponds to 'TransportMeans' in the UBL ontology. Following this trail a correspondence between 'Vehicle' in the Common Framework ontology and 'TransportMeans' in the LogiCO ontology could be deduced.

Being preliminary work there are some obvious threats to both validity and reliability. We need to investigate more in-depth the matching techniques used by the matching system and possibly investigate how different techniques perform with the current setup. Further, the reference alignment should be more thoroughly assessed by domain expertise from the transport logistics domain. Such an assessment is planned, but not performed yet. Last but not least, the UBL ontology requires further validation and enrichment in order to make sure that the domain knowledge possessed by the standard is appropriately maintained in full scale. Primary candidate enhancements are to include the element definitions from the UBL XSD schemas (see Fig. 1 depicting among other things the element definitions) and instance data to see how this affects the evaluation scores.

7 Conclusions and Future Work

This PhD investigates how contextual information, often residing in disparate sources and represented by different formats, can be optimally represented to ontology matching systems and how these systems best can exploit this context to produce accurate and correct alignments. Using the Universal Business Language as a case, we have investigated how international e-Document standards in the transport logistics domain can provide background knowledge to support the matching of domain ontologies for this particular domain. The current developments include prerequisite artifacts required for performing the rest of the study. Early evaluation tests involving two ontologies from the transport logistics domain suggest that using the current version of the generated UBL

ontology as background knowledge does not influence the matching performance and that several reconfigurations and enhancements are required. However, manual analysis of the two ontologies being matched and the developed UBL ontology suggest that the UBL standard can provide useful context to support a matching between the two transport logistics ontologies. Further work include a deeper analysis of the evaluation results, domain expertise assessment of the UBL ontology developed and the reference alignment, a more in-depth investigation of the ontology matcher being used in the evaluation and enriching the UBL ontology using additional contextual information from the UBL standard.

References

1. Shvaiko, P., Euzenat, J.: Ontology matching: state of the art and future challenges. IEEE Trans. Knowl. Data Eng. **25**(1), 158–176 (2013)
2. Otero-Cerdeira, L., Rodríguez-Martínez, F.J., Gómez-Rodríguez, A.: Ontology matching: a literature review. Expert Syst. Appl. **42**(2), 949–971 (2015)
3. Euzenat, J., Shvaiko, P., et al.: Ontology Matching. Springer, Heidelberg (2013)
4. Mascardi, V., Locoro, A., Rosso, P.: Automatic ontology matching via upper ontologies: a systematic evaluation. IEEE Trans. Knowl. Data Eng. **22**(5), 609–623 (2010)
5. Locoro, A., David, J., Euzenat, J.: Context-based matching: design of a flexible framework and experiment. J. Data Semant. **3**(1), 25–46 (2013)
6. Thayasivam, U., Doshi, P.: On the utility of WordNet for ontology alignment: is it really worth it? In: Proceedings of the Fifth IEEE International Conference on Semantic Computing (ICSC), pp.267–274. IEEE (2011)
7. Jean-Mary, Y.R., Shironoshita, E.P., Kabuka, M.R.: Ontology matching with semantic verification. Web Semant. Sci. Serv. Agents World Wide Web **7**(3), 235–251 (2009)
8. Paulheim, H., Hertling, S.: WeSeE-match results for OAEI 2013. In: Proceedings of the ISWC Workshop OM-2013, pp.197–202 (2013)
9. Chen, X.: Exploiting BioPortal as background knowledge in ontology alignment. Master thesis, Miami University (2014)
10. GS1 Global: GS1 XML Business Message Standards (BMS) (2014)
11. UN/CEFACT: Core Components Technical Specification - Part 8 of the ebXML Framework. Technical report (2003)
12. CEN BII/WS: CEN BII (CEN Workshop on Business Interoperability Interfaces for Public Procurement in Europe) (2015)
13. KTBL: About AgroXML (2012)
14. Robson, C.: Real World Research, 3rd edn. Wiley, Chichester (2011)
15. García, R., Celma, O.: Semantic integration and retrieval of multimedia metadata. In: Proceedings of the 5th International Workshop on Knowledge Markup and Semantic Annotation, pp. 69–80 (2005)
16. Faria, D., Pesquita, C., Santos, E., Cruz, I.F., Couto, F.M.: AgreementMakerLight: a scalable automated ontology matching system. In: Proceedings from the 10th International Conference on Data Integration in the Life Sciences (DILS 2014), July 2014
17. Faria, D., Martins, C., Nanavaty, A., Taheri, A., Pesquita, C., Santos, E., Cruz, I.F., Couto, F.M.: AgreementMakerLight Results for OAEI 2014 (2014)

18. Daniele, L., Pires, L.F.: An ontological approach to logistics. In: Enterprise Interoperability: Research and Applications in Service-Oriented Ecosystem (Proceedings of the 5th International IFIP Working Conference IWIE 2013). Wiley (2014)
19. Damova, M., Kiryakov, A., Simov, K., Petrov, S.: Mapping the central LOD ontologies to PROTON upper-level ontology. In: Proceedings of the Fifth International Workshop on Ontology Matching (OM-2010), pp.61–72 (2010)
20. Mascardi, V., Cordì, V., Rosso, P.: A Comparison of Upper Ontologies (Technical report DISI-TR-06-21). Technical report, University of Geova, Italy (2007)
21. Dragisic, Z., Eckert, K., Euzenat, J., Faria, D., Ferrara, A., Granada, R., Ivanova, V.: Results of the ontology alignment evaluation initiative 2014. In: Proceedings of the 9th International Workshop on Ontology Matching Collocated with the 13th International Semantic Web Conference (ISWC 2014) (2014)