Tackling Resource Interoperability: Principles, Strategies and Models

Wim Peters
Department of Computer Science
University of Sheffield
UK
E-mail: w.peters@sheffield.ac.uk

Abstract

In order to accommodate the flexible exploitation and creation of knowledge resources in text and data mining (TDM) workflows, the TDM architecture will need to enable the re-use of resources encoding linguistic/terminological/ontological knowledge, such as ontologies, thesauri, lexical databases and the output of linguistic annotation tools. For this purpose resource interoperability is required in order to enable text mining tools to uniformly handle these knowledge resources and operationalise interoperable workflows. The Open Mining Infrastructure for Text and Data (OpenMinTeD) aims at defining this interoperability by adhering to standards for modelling and knowledge representation, and by defining a mapping structure for the harmonisation of information contained in heterogeneous resources.

Keywords: resource interoperability, standards, linked data

1. Introduction

The Open Mining Infrastructure for Text and Data (OpenMinTeD) is a new European initiative which seeks to promote the cause of text and data mining (TDM). OpenMinTeD will promote collaboration between the providers of TDM infrastructures as well as working outside of the field to encourage uptake in other areas which may benefit from TDM. Service providers will benefit from this project through the standardisation of formats for TDM as well as the creation of a new interoperable TDM workflow, which will seek to standardise existing content and allow previously incompatible services to work together.

In order to accommodate the flexible exploitation and creation of knowledge resources, the architecture will need to enable the re-use of resources encoding linguistic/terminological/ontological knowledge, such as ontologies, thesauri, lexical databases and linguistic annotation tools by means of uniform access and query techniques.

A key text mining interoperability challenge is that linguistic descriptions come from heterogeneous and distributed knowledge resources. Individual linguistic and terminological resources greatly differ in the explicit linguistic information they capture, which may vary in format, content granularity and the motivation for their creation, such as the immediate needs of the intended user. In order to accommodate these factors, we need to be able to integrate information coming from heterogeneous knowledge resources and text mining applications, at the levels of both representation format and conceptual structure (see section 2). For this purpose, we need to make use of linked standards for resource data category classification.

2. Linked Data

Our strategy to enable interoperability is to adhere to existing standards and best practices. Our principal choice for data modelling is to adopt the Linked Data paradigm (Bizer et al., 2009). The semantic web has emerged as one of the most promising solutions for large scale integration of distributed resources. This is made possible by a stack of World Wide Web Consortium (W3C) technologies such as the Resource Description Framework[1] (RDF), RDF Schema[2] (RDFS), Web Ontology Language[3] (OWL) and the SPARQL[4] Query Language. RDF forms the basis of the stack allows modeling information as a directed graph composed of triples that can be queried using SPARQL.

This entails that all data categories used in the interoperability specification should have URIs, and are ideally contained in an RDF resource, which will allow dereferencing.

Another consequence is that all (non-)standard models should be re-engineered if they are not available in XML-RDF/OWL already, and that all relevant ontologies should become networked.

3. Resources and Standards

There are a number of initiatives to make conceptual and linguistic classifications interoperable and exploitable in a uniform fashion. This has resulted in various (established/proposed/de facto) standards and best practices for encoding linguistic and terminological knowledge, both from the (computational) linguistic and the semantic web side.

The form and content in which knowledge resources come varies according to the format and content dimensions. According to the former, resources differ in their representation format and the level of formalization of this format. For instance, many linguistic resources such as text corpora, thesauri and dictionaries are encoded in XML[5], but an increasing number of linguistic resources are represented as populated RDF or OWL models, in order to be exploitable in semantic web applications. Another widely adopted format is the XML Metadata

[1] http://www.w3.org/RDF/
[2] http://www.w3.org/TR/rdf-schema/
[3] http://www.w3.org/TR/owl-ref/
[4] http://www.w3.org/TR/sparql11-query/
[5] http://www.w3.org/XML/

Interchange[6] (XMI).
The content side of knowledge resources covers the data categories that are used to capture standards and best practice information types. To name but a few, in the area of linguistic description the Lexical Markup Framework[7] (LMF) (Francopoulo et al., 2006) presents a linguistic description of lexical knowledge, whereas Lemon [8] (McCrae et al., 2012) is a model for sharing lexical information on the semantic web. The W3C Ontolex[9] interest group has developed a model for lexicons and the relation of lexical meaning with ontologies, and investigates the added value of using such a model in semantic web NLP applications. The Open Linguistics Working Group of the Open Knowledge Foundation[10] works towards a linked open data cloud of linguistic resources, which applies the linked data paradigm to linguistic knowledge. The NLP Interchange Format[11] (NIF) is an RDF/OWL-based format that aims to achieve interoperability between NLP tools, language resources and annotations.
As examples of domain-specific standards that are relevant to OpenMinTeD, formats such as the BioNLP format (Kim et al., 2011) and the BioC format promoted by BioCreative (Liu et al., 2013) are heavily used in the Life Sciences, promoting reusability of resources and interoperability of tools and Web services. A range of different tools, corpora and programming language implementations compliant with the BioC format have been recently implemented.
In the agricultural domain, the Agronomic Linked Data[12] (AgroLD) Project provides methods to aid data integration and knowledge management within the plant biology domain to improve information accessibility of heterogeneous data.
As illustrated, at present there are many converging developments in the form of (de facto) standardization of the representation of information elements required for interoperable text consumption and processing across domains. Given the existence of this variety of (standard) linguistic/terminological/ontological models, it is necessary to establish interoperability between their vocabularies in a principled way, in order to enable text mining tools to be brought together within the OpenMinTeD platform.

4. Models

We want our platform to be language agnostic and domain independent, in order to facilitate its use across domains and borders. For this purpose, we will adopt the Model Driven Architecture (MDA) (Miller et al., 2003) in the design and implementation of our data models. This is a development approach, strictly based on formal specifications of information structures and their semantics. MDA is promoted by the Object Management Group (OMG[13]) based on several modeling standards such as: Unified Modeling Language [14] (UML), Meta-Object Facility [15] (MOF), XML Metadata Interchange (XMI) and others.
When following the MDA approach, existing knowledge representation formalisms can be described and content can be instantiated in an integrated manner. Mappings between formalisms and integrating metamodels can then be used to transform or merge heterogeneous knowledge bases.
The Meta Object Facility (MOF) is an extensible model driven integration framework for defining, manipulating and integrating metadata and data in a platform and formalism independent manner. The Owl ontology metamodel as well as the UML profile are grounded in MOF, in that they are defined in terms of the MOF meta-metamodel. Basing ourselves on this will gives us a principled method for harmonizing, accessing and linking model elements from knowledge resources.
When harmonizing different knowledge bases the problem of classifying and linking concepts from heterogeneous vocabularies entails the adoption and linking of existing standards for the representation of multilingual linguistic, terminological and ontological information, in order to arrive at a practically motivated interoperability specification for TDM in OpenMinTeD. The re-use of existing (standard) data category semantics, data structures and linking strategies will ensure maximal consensus regarding standardization and best practice (Peters, 2013).
Linking data categories from different ontologies can be modelled in various ways. The most straightforward is the set of coarse grained lightweight thesaural mapping relations expressed by SKOS[16].
The second option is to define a mapping metamodel as in (Brockmans et al., 2006), and integrate it into the overall MOF picture. The advantage of this mapping meta-model is that it is formalism-independent. Each mapping between a source and target ontology has one or more mapping assertions that describe a semantic relation between a source ontology class and a target ontology class. In the mapping metamodel mappings are first-class (reified) objects that exist independently of the ontologies.
The difference is the granularity of the mapping relations that can be expressed. For now we consider the coarser set of SKOS relations, because this will make traversing the networked ontologies simpler. This is important because maintaining a network of related resource and standard-specific data categories rather than adopting a single data model for all integrated knowledge requires complex querying.
However, structural differences between ontologies involving permutations from for instance object properties to classes can be better handled by means of a separate mapping model (Scharffe et al, 2008). Even if ontologies share conceptually equivalent elements, they often express their content in different ways, because their information differs structurally.

[6] http://www.omg.org/spec/XMI/
[7] http://www.lexicalmarkupframework.org/
[8] http://lemon-model.net/
[9] http://www.w3.org/community/ontolex/
[10] http://okfn.org/
[11] http://nlp2rdf.org/nif-1-0
[12] http://volvestre.cirad.fr:8080/agrold/index.jsp
[13] http://omg.org/
[14] http://www.uml.org/
[15] http://www.omg.org/mof/
[16] https://www.w3.org/2004/02/skos/

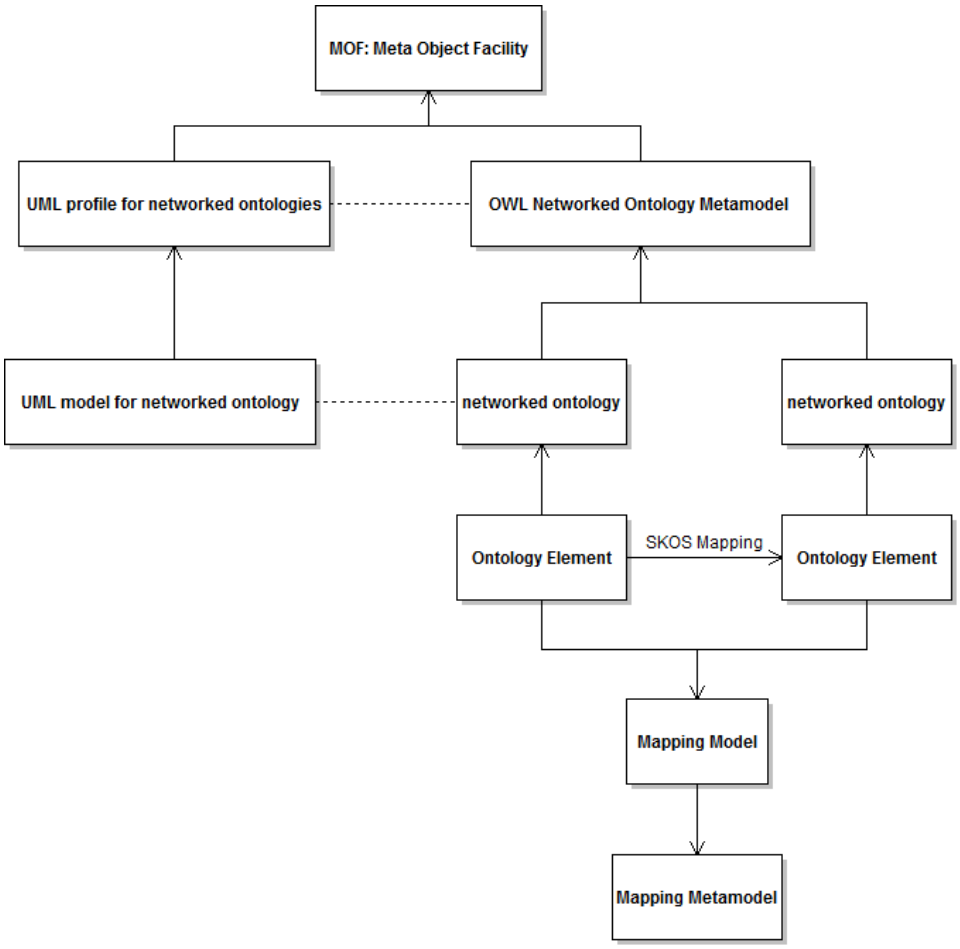

Figure 1: MOF-based Model Structure

For instance, the following more or less equivalent sets of data categories from various sources.

1. Token; pos='noun´; lemma='kidney´
2. Noun; lemma='kidney' ´
3. Noun;Token.root='kidney'

This example above shows that the features 'root' in 3 and 'lemma' in 1 and 2 are equivalent. Their transformation can be expressed by means of a simple identity relation (SKOS:exactmatch). The concept Noun in 2. is equivalent to the concept Token with the value 'noun' of the feature 'pos' in 1. This requires a complex transposition. This is a typical example of a "class to class-plus-attribute" transformation pattern, which is one of a series of structural transformations observed and collected by (Scharffe et al, 2008)[17], which regulate regularly observed structural transformations between different configurations. Reified mappings can reference these transformations. Figure 1 illustrates the overall architecture with the two mapping modeling options.

[17]http://ontologydesignpatterns.org/wiki/Category:AlignmentOP

5. Schema Selection

Now we have a modelling framework, we can populate it by selecting select resources for inclusion. This process of resource (schema) aggregation involves a schema selection methodology that should adhere to the following methodological requirements:

1. The process is extendable and bottom up in the sense that it allows an incremental inclusion of resource schemas. From this follows that its content will not be exhaustive but sufficiently populated for the interoperability task at hand. Where necessary, linking relations need to be defined between vocabulary elements. For this purpose the use of the SKOS linking vocabulary (section 4) is required.
2. The extension is driven by the OpenMinTeD use cases, which describe the interaction of users with the OpenMinTed platform within selected application domains, and determine which additional resources should be taken into account. Also, in this stage SKOS linking relations will establish the interoperability between schema elements.
3. The schemas/vocabularies that are selected from the start as representative vocabularies need to

be representative and widely used in concrete applications. In other words, they must be popular resources or de facto standards for capturing linguistic and terminological standards. Obvious candidates for inclusion are Universal Dependencies [18], OLIA [19], SKOS, TBX [20] and OBO[21], and linguistic reference vocabularies such as NIF[22], OntoLex[23] and Lemon[24]. Some of these resources are already linked within the LLOD cloud[25].

4. Ideally the vocabularies should maximally reflect standardisation in terms of both content representation and data category linking. Where application-specific schema elements need to be integrated, user friendly link facilities should be provided.

6. Conclusion

In this paper we presented a principled modelling configuration, which, together with a descriptively adequate mapping facility, will allow us to incrementally build a network of resource data category vocabularies for TDM. In its RDF format this network allows flexible traversal in SPARQL, enables the detection and definition of interoperability at the level of data category semantics, and guarantees the preservation of resource specific and standard data categories without relying on a single common data model for capturing knowledge.

7. Acknowledgements

Work was funded by the OpenMinTed project (Open Mining INfrastructure for TExt and Data); H2020 654021). It reflects only the author's views and the EU is not liable for any use that may be made of the information contained therein.

8. Bibliographical References

Bechhofer and Miles, A. (2009), SKOS Simple Knowledge Organization System Reference. W3C recommendation, W3C http://www.w3.org/TR/2009/REC-skos-reference-20090818/

Bizer, C., Heath, T., Berners-Lee, T. (2009). Linked Data·The Story So Far. International Journal on Semantic Web and Information Systems 5 (3): 1¶22. doi:10.4018/jswis.2009081901. ISSN 1552-6283

Miller, J. and Mukerji.J. (2003). MDA Guide Version Technical report, Object Management Group (OMG).

Brockmans, S., Haase, P., Stuckenschmidt, H. (2006). Formalism-Independent Specification of Ontology Mappings - A Metamodeling Approach, In: Robert Meersman, R., Tari, Z. et al. (eds), OTM 2006 Conferences, Springer Verlag, Montpellier, France (2006)

Francopoulo, G., George, M., Calzolari, N. Monachini, M., Bel, N., Pet, M., Soria, C. (2006). LMF for multilingual, specialized lexicons. In: LREC, Genova, Italy

Kim, JD., Pyysalo, S., Ohta, T., Bossy, R., Nguyen, N.,Tsujii, J. (2011)). Overview of the BioNLP Shared Task 2011, Biomedical Natural Language Processing Shared Task Workshop, ACL, Portland, Oregon, USA

Liu, W., Comeau, D. C., Dougan, R.D., Islamaj, R. and Wilbur, W. J. (2013) Extending BioC Implementation to More Languages, in Proceedings of the Fourth BioCreative Challenge Evaluation Workshop vol. 1., pp. 31-37.

McCrae, J., Aguado-de Cea, G., Buitelaar, P., Cimiano, P., Declerck, T., Gomez-Perez, A., et al. (2012). Interchanging lexical resources on the Semantic Web. Language Resources and Evaluation, 46(4), 701-719.

Peters, W. (2013), Establishing Interoperability between Linguistic and Terminological Ontologies, In: Oltramari, A.; Vossen, P.; Qin, L.; Hovy, E. (Eds.), New Trends of Research in Ontologies and Lexical Resources, Springer 2013.

Scharffe, F. Euzenat, J. and Fensel, D. (2008). Towards design patterns for ontology alignment. In R.L. Wainwright and H. Haddad (eds.): Proceedings of the 2008 ACM Symposium on Applied Computing (SAC), Fortaleza, Ceara, Brazil.

[18] http://universaldependencies.org
[19] http://acoli.cs.uni-frankfurt.de/resources/olia/
[20] http://www.ttt.org/oscarStandards/tbx/
[21] http://oboformat.googlecode.com/svn/trunk/doc/GO.format.obo-1_2.html#S.2
[22] http://persistence.uni-leipzig.org/nlp2rdf/
[23] https://www.w3.org/community/ontolex/
[24] http://lemon-model.net/
[25] http://www.linguistic-lod.org/