TOK: A meta-model and ontology for heterogenous terminological, linguistic and ontological knowledge resources

Nizar Ghoula, Gilles Falquet and Jacques Guyot
ISI lab, Centre Universitaire d'Informatique
Information System Department, University of Geneva
Geneva, Switzerland
{*Nizar.Ghoula, Gilles.Falquet, Jacques.Guyot*}*@unige.ch*

***Abstract*—Documents are very rich resources containing knowledge describing a specific domain. That's why their treatment is a common task, which is based on the use of terminological and ontological resources. Various types of ontologies, thesauri, and a large list of resources are commonly used in the process of knowledge extraction. The modeling and reuse of these resources is intended to support knowledge management. In this paper, we propose a methodology and a model for ontological and terminological resource management. Our aim is to build a resources repository that offers operations for loading, storing, indexing, translating, generating and matching different resources. In this contribution we propose an ontology as a model of these resources and we explain how can we represent, annotate and load new resources into our repository.**

***Keywords*-Ontology of Resources; Multilingual; Terminology; Alignment; Resources Repository.**

I. Introduction

Many tasks related to documents, such as indexing, retrieving, annotation, or translation are based on linguistic, terminological and ontological knowledge. This knowledge currently exists in resources of different types such as terminologies, glossaries, ontologies, multilingual dictionaries or parallel text corpuses. These resources are represented using various formalisms and languages (predicate logic, description logic, semantic networks, conceptual graphs, documents, etc.). Producing or finding linguistic, terminological and ontological knowledge resources is not a simple task, it is generally difficult to find the right resources for the right usage. Some ontology repositories have been created to offer a more effective indexing for these resources than common search engines. For example, Swoogle[1] indexes approximately 10 000 ontologies; the DAML site [2] provides search based on ontology components (classes, properties, ...) or metadata (URI, funding source, ...); BioPortal[3] has similar searching and browsing tools [1].

However, users need more than ontologies to perform knowledge engineering tasks, then it is important to have repositories offering access to more diverse resources in different formats. Available resources generally do not fit exactly with the user needs. Thus the user must be provided with tools to derive new resources from existing ones. This derivation may involve operations such as selecting a part of a resource, composing it with another one, translating it to another language or representing it in a different formalism.

In this contribution, we present our approach called TOK (*Terminological, Linguistic and Ontological Knowledge Resource Management*). TOK is based on the principles of semantic web, metadata and ontologies to facilitate the representation, storage and alignment of heterogeneous and multilingual resources. In the first part of this paper we will identify the kinds of heterogenous resources that we may process and discuss some of the proposed resource's models. Afterwards we will describe our approach of resources representation structure, which is ensured by different levels. In the third part we will introduce our ontology of resources, which is an implementation of our general model of heterogenous resources. The final section describes the storage of these resources in the repository and introduces the next part of our work, which will be focused on the identification and the definition of operations that can be used in the processing and treatment of the stored resources.

II. Knowledge resources

A central point of our approach is to build a repository of knowledge resources. This repository is a collection of heterogenous resources that are based on different formalisms or models. In this paper we will focus on the ontology that describes this repository and on the way we manage the resource representation by means a resource representation ontology.

A. Resources Identification

In the TOK resource model we distinguish two may categories of resources: autonomous and enrichment resources

Autonomous resources: these are resources whose existence is independent of any other resource, like ontologies, thesauri, terminologies, corpuses, documents ....

[1]http://swoogle.umbc.edu
[2]http://www.daml.org/ontologies
[3]http://bioportal.bioontology.org

Enrichment Resources: These resources interconnect two or more resources, they result from the application of a process (automatic or human) on these resources.

- *annotations*; they are intended to enrich an initial resource by associating to its content some conceptual elements of another type to enable it to be usable, accessible and recognized by a set of actors or agents.
- *alignments*; they are mappings between two resources of the same type [2]. The alignment process is based on finding similar entities in different resources while preserving their independence and integrity.

B. Resources representation models

There are many models and languages for knowledge representation, but they usually focus on one or two aspects only: ontological, terminological, lexical, textual, documentary, etc. It is more difficult to find models representing various aspects of knowledge or resources of different kinds.

A model of the multilingual aspect in ontology has been proposed by [3], its development is an association between a meta-model of ontologies (describing classes and properties) and a linguistic model. Another model has been developed to centralize the management of linguistic resources a linguistic platform called Intuition [4] and unify the management of these resources in multilingual environment. This model is characterized by its exploration of the structure of linguistic forms. The application of this model allows to represent ontological entities and identify lexical units by taking into account the syntactic and semantic multilingual relations [5]. Every representation starts with the conceptual entity in an ontology and subsequently describes the corresponding token. This model cannot represent pure linguistic resources.

In the context of mapping linguistic and ontological resources, [6] have proposed an approach to integrate and merge Wikipedia and WordNet to enrich an ontology (YAGO[4]). The ontology is extracted from these two resources by adding new facts[5] extracted from Wikipedia as individuals, classes from the conceptual categories in Wikipedia and each "synset" of WordNet. The extraction process is based on the identification of a number of relationships such as: *Type, SubClassOf, Means, Context*. The model of the resulting ontology is extensible based on the resource use and dedicated to the representation of facts. This approach shows that the combination of multiple resources makes possible building or extending existing resources. If they have used a more generic model capable of representing resources, their methodology become less complex and less costly than the proposed one.

For the integration of heterogeneous resources, [7] have proposed a model of terminologies and ontologies. This model provides a more general formalism and states new constructors that provide additional expressivity to a terminology. This representation is based on the differentiation of resource entities, and remains faithful to the representation of each resource model without using common abstract entities, for example instead of considering a term or a concept as an abstract entity these classes have different representations depending on the resource which creates redundancy in the instances. Otherwise we have not access to the whole model, nor the approach of heterogenous resources treatment and how the model is used to represent resources and ensure their interoperability.

[4]Yet Another Great Ontology

[5]relative to all existing data in a knowledge base

III. Structure of resources representation

Our approach is based on a model (*TOK*) having three levels: resources, metadata and content related to resources, and an ontology called *TOK_Onto* defining content representation models, and common metadata of resources.

Table I
LEVELS OF TOK MODEL

level	function
Ontology (*TOK_Onto*)	defining a representation language reasoning, querying (access)
Knowledge base	representing each resource metadata representation content representation
Resource	storage of each resources

At each one of these two levels, resources and entities are represented by instances of a global representation model (in our case a general ontology called *TOK_Onto*). The way of representing the content of a resource depends on each model. Resources representation models share the use of the resource's entities and define the content of the resource's representation.

At the knowledge base level, we create instances of the resource and its entities. Depending on treatments that we need to apply, this representation may use a specific kind of model, for example if we need to build an alignment between two ontologies driven by the hierarchy and the linguistic labels, we do not need to represent axioms or blank nodes, so we use an hierarchy model to represent these resources. The abstraction degree of the resource representation can vary from simple (it sees the resource as a whole) to the most detailed (we represent each entity of the resource and all its properties).

To create a repository we need to store and build resources, which is the third level of our modeling approach. At this level, resources are stored using a same physical model that implements each representation model.

A. Metadata

The ontology level defines classes of resources and their properties. A resource is represented as an instance of a class

of *TOK_Onto*. Depending on the resource type, its formalism and the degree of the desired accuracy of representation an instance of a class is created to represent the resource.

B. Contents

Since there exist many different (and incompatible) ways to express knowledge in resources (from formal first order or description logic to semi-formal models and natural languages), it is impossible to devise a single representation model for the content of resources. Moreover, the same resource may be involved in processes that can only handle resources represented in some specific form. For instance, an ontology alignment algorithm may only accept OWL ontologies, while another one requires ontologies in a WordNet-like format. The same condition is true for other processes like automated text annotation, multilingual text alignment, word sense disambiguation, etc.

To address these issues, we propose to support multiple representations for the content of a resource. Once a resource is present in the repository it is possible to create as many content representations as needed, depending on the tasks it is used for. For instance, the content of an ontology expressed in description logics may be viewed as a simple hierarchy of concepts, or as a semantic network, even as a mere list of terms. Conversely, a simple glossary of terms can be seen as a (flat) lexical ontology.

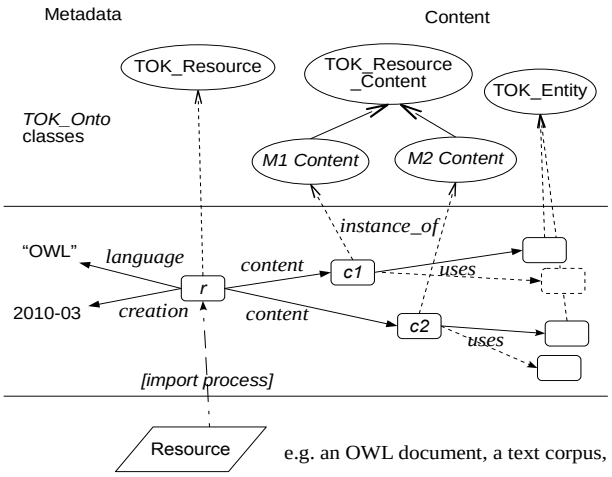

Figure 1. Representation of a resource with its metadata and different representations of its content (*c1* and *c2*). Representation elements are instances of *TOK_Onto* classes.

IV. The Resource Ontology: *TOK_Onto*

The resource study within the criteria we have presented have allowed us to build a resource classification and to develop the first layer of our general ontology *TOK_Onto*[6]. This ontology is intended to represent heterogeneous knowledge resources. *TOK_Onto* is an OWL ontology with the degree of expressiveness *SRIQ(D)*, it has been developed with the Protégé ontology editor. *TOK_Onto* contains 192 concepts (named and unnamed), 116 properties, 450 axioms, and 2000 annotations (comments and labels).

In this section we briefly describe the design choices we made in *TOK_Onto* to represent resources, their content (with multiple models).

A. Resources

As described in the previous section, the *TOK_Onto* ontology must describe both the metadata associated to each resource and the resource content. It must also describe the relationships between enrichment resources (annotations and alignments) and the resource they enrich.

The *TOK_Resource* class is used to model resources in an abstract way, it has subclasses that represent the different types of resources that are supported by the repository (ontology, thesaurus, corpus, etc.). All resources added in the repository are instances of this class and have specific metadata represented by object and data properties (name, URL, creation_date, formalism, language, etc.) related to the class *TOK_Resource*.

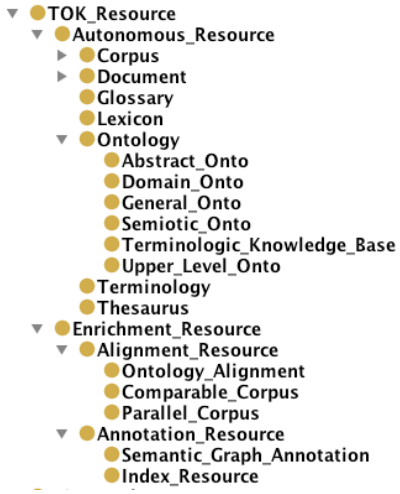

Figure 2. Partial view of the resources classification in *TOK_Onto*.

If a resource is aligned with another one, it is connected to the corresponding alignment resource through an has_alignment property. Similarly, it can be connected to annotation resources through the has_annotation property.

The UML class diagram of Figure 3 shows the properties with their domain, ranges and cardinality constraints.

B. Multi-model resource contents

In order to handle multiple content representations, the TOK ontology contains a model for each supported representation. Each resource can be associated to one

[6]Accessible online at: http://cui.unige.ch/isi/onto/tok/OWL_Doc/

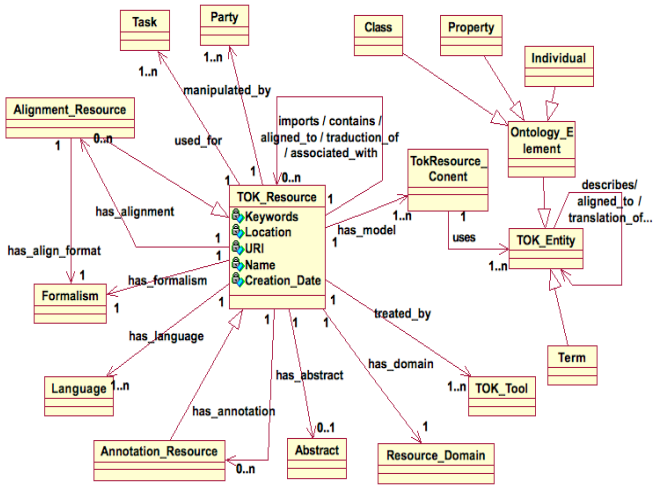

Figure 3. Partial view of the TOK model of heterogenous terminological, linguistic and ontological knowledge resources.

or more content that belong to a content model. The *TOK_Resource_Content* class is the abstract superclass of all content models. Typical content models are: concept hierarchies (a model for simple concept taxonomies); WordNet-like lexical ontologies; description logic ontologies; bitext (corpuses of texts associated to their translation); translation memory (text segments aligned with their translation). Each model is formally defined by a set of ontological axioms that indicate the type of entities that belong to the model and how they are interconnected. Entities used in model definition are subclasses of the *TOK_Entity* abstract class, as shown in Figure 4 The actual representation of a resource content according to a content model is thus an instance of the content model, i.e. a set of instances of the *TOK_Onto* classes and properties that make up the model. The representation of a resource together with its metadata and contents is depicted in Figure 1.

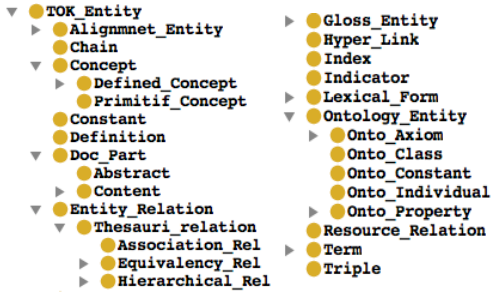

Figure 4. Types of entities used in the resource models.

Example: Let us consider a simplified version of the WordNet [8] lexical ontology model. In this model each term is connected to a meaning (a concept) and to a lexical form (a character string). In *TOK_Onto* this model will be represented by a class (WN_Like). The fact that the model makes use of entities such as *Concept*, *Term*, *Lexical Form*, *Sentence*, and specific relational entities (*Hyperonymy*, *Meronymy*, ...) will be formally represented in the ontology by axioms such as :

- WN_Like $\sqsubseteq$ uses **only** (Concept **or** Term **or** LexicalForm **or** Sentence **or** Par_of_speech **or** WNRelation)
- Term_Meaning $\sqsubseteq$ WNRelation
- Term_Meaning $\sqsubseteq$ src **only** Term
- Term_Form $\sqsubseteq$ WNRelation
- Hypernym $\sqsubseteq$ WNRelation
- ...

Table II shows some example of models that have been implemented in the current version of *TOK_Onto*.

Table II
EXAMPLES OF RESOURCE CONTENT MODELS AND THEIR PRINCIPAL COMPONENTS

Model	Components
Concept hierarchy	Concept, ISA_Relation, ...
WordNet Like	Concept, Term, Lexical_Form, Hypernym_Relation, Meronym_Relation, Term_Form_Relation, Term_Meaning_Relation, ...
Description Logics	Class, Property, Property_Restriction, Union, ..., Axiom, SubClass_Relation, Equiv_Relation, etc.
Graph ontology	Class, Taxonomic_Relation, Relation, Relation_Label, etc.
Class Diagram	Class, Association, ...
Translation memory	Text_Segment, Language, Translation_Relation, Language_Relation
Ontology Alignment	Concept, Correspondence_Relation, ...

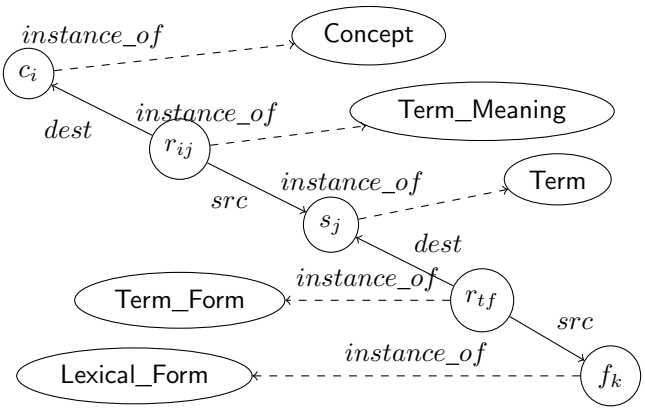

Figure 5. Part of an instantiation of the WN_Like model

When a resource is imported in the repository, if the user wants it to be represented as a WordNet-like ontology, the import procedure will analyze the resource and map its content to a WordNet-like content. The resulting representation

will be a set of *TOK_Entity* instances, as shown on figure 5. This operation may be straightforward if the resource is already expressed in an ontological language. It may become more complex or even impossible for other types of resources.

V. Storing resource representations

We have built a storage space to represent resources, their entities and relations between them. This storage space is the basic step for our further work such as the detection of alignments between concepts or entities, the translation and alignment of multilingual lexical units, the selection of resources and their exportation to a specific formalism as well as other operations.

We have implemented the storage space on top of a standard relational database system. Databases technology usage is justified by the important size of the treated resources. We wanted to exploit the performance of this type of storage with its effective and simple query language. Instead of mapping *TOK_Onto* directly to the database schema, we have build a generic node-link model that resembles RDF triples, supplemented with specific attributes attached to nodes and links [7].

The repository is currently in use. We gradually add new resources. We've added AGROVOC[8] in 17 languages, WordNet in English, German, Catalan, Spanish, etc., UNL[9] in French, Arabic, Japanese, CityGML[10], the URBAMET thesaurus [11], etc. For the moment, the resources repository contains approximately 950 000 lexical forms in 24 languages, 173 000 concepts and 335 000 sentence.

To import these resources we have developed several tools to map different resource formats and languages (OWL/XML, WordNet, AGROVOC, XML Schema, interlinked HTML pages, ...) to supported models such as *WordNet_like Lexical Ontology*, *Graph Ontology*, or *Translation Memory*.

We have generated new resources with the imported resources . For instance, we have implemented a simple syntactic ontology alignment algorithm on the WordNet_Like model. Then we have aligned the English WordNet with other resources, such as AGROVOC, URBAMET, UNL, that have lexical forms in multiple languages. This enabled us import these lexical forms into the English Wordnet and associate them to the corresponding concepts, thus obtaining an enriched WordNet.

[7] We intend to turn to an RDF store in the near future

[8] AGROVOC is a multilingual structured vocabulary developed by the FAO terminology covering all domains related to agriculture, fisheries, food and related fields (environment, for example).

[9] Universal Networking Language, is an artificial language that can be used as a pivot language for translation systems or as a language of knowledge representation in applications of information retrieval.

[10] CityGML is a common information model for representing 3D objects in urban environments.

[11] URBAMET is a is a bibliographic database in the urban domain, its thesaurus has approximately 4000 entries in 3 languages

VI. Conclusion

This work is focused on linguistic, terminological, and ontological knowledge resources. We have proposed a representation model of these resources and we explained its structure and use. This model covers the (I) treatment of a broad spectrum of resources represented in different formalisms, (II) definition of a transformation process of resources, (III) and ensures the prospect of modeling tasks of semantic processing on resources.

The main objective of our approach is to generate new resources from the composition of existing ones in the repository and instantiated in the ontology. Thus, extending the knowledge in the repository is done every time we load a new resource. Based on the developed storage repository, we will be able to generate, integrate, and use knowledge and produce new resources in various formalisms. This toolbox is based on the data repository, the ontology *TOK_Onto* and an implementation of all operators describing tasks of knowledge processing.

The next stage of this work is to define rules and axioms that can model each knowledge management task, the corresponding representation operators or the combination of operators to perform it.

References

[1] N. Noy, N. Griffith, and M. Musen, "Collecting community-based mappings in an ontology repository," in *7th International Semantic Web Conference (ISWC2008)*, October 2008.

[2] J. Euzenat and P. Shvaiko, *Ontology matching*. Heidelberg (DE): Springer-Verlag, 2007.

[3] E. Montiel-Ponsoda, G. Aguado de Cea, A. Gómez-Pérez, and W. Peters, "Modelling multilinguality in ontologies," in *Coling 2008: Companion volume: Posters*. Manchester, UK: Coling 2008 Organizing Committee, August 2008, pp. 67–70.

[4] F. Cailliau, "Un modle pour unifier la gestion de ressources linguistiques en contexte multilingue," in *Verbum ex machina: actes de la 13e Confrence sur le Traitement Automatique des Langues Naturelles (TALN 2006) : Leuven.*, P. Mertens, Ed. Presses univ. de Louvain, 2006, 2006, pp. 454–461.

[5] G. Falquet, C.-L. M. Jiang, and J. Guyot, "Un modèle et une algèbre pour les systèmes de gestion dontologies," in *EGC*, 2008, pp. 697–702.

[6] F. Suchanek, G. Kasneci, and G. Weikum, "YAGO: A core of semantic knowledge - unifying WordNet and Wikipedia," in *16th International World Wide Web Conference (WWW 2007)*, C. L. Williamson, M. E. Zurko, and P. J. Patel-Schneider, Peter F. Shenoy, Eds. Banff, Canada: ACM, 2007, pp. 697–706.

[7] P.-Y. Vandenbussche and J. Charlet, "Méta-modèle général de description de ressources terminologiques et ontologiques," in *Actes d'IC*, F. L. Gandon, Ed. PUG, 2009, pp. 193–204.

[8] C. Fellbaum, Ed., *WordNet: An Electronic Lexical Database*, ser. Language, Speech, and Communication. Cambridge, Mass.: MIT Press, 1998.