# A framework for mapping refinement specification

Fayçal Hamdi, Chantal Reynaud, Brigitte Safar

# A Framework for mapping refinements specification

**F. Hamdi**[1], **C. Reynaud**[1], **B. Safar**[1]

**Abstract.** Ontology alignment is an important task for information integration. Current ontology matchers are not efficient for all application domains or ontologies. Very often the quality of the results can be improved by considering the specificities of the ontologies domain. In this paper, we propose an environment, called TaxoMap Framework, based on TaxoMap, an alignment tool, which helps an expert to specify treatments based on alignment results. The aim is to refine these results or to merge, restructure or enrich ontologies. We apply our approach to mapping refinement in the topographic field within the ANR (The French National Research Agency) project, GéOnto.

## 1 INTRODUCTION

The explosion of the number of data sources available in the web increases the need for techniques which allowed the integration of these sources. The ontologies are considered as an essential element in each integration system because it defines the concepts related to any domains. The task of ontology alignment is particularly important in such systems because it makes different resources interoperate, even if they are described by various and heterogeneous ontologies. The current alignment tools [4] are not efficient in all domains and in all ontologies. They are very good in some cases, worse in others. The quality of their results is not always guaranteed and could often be improved if the alignment process takes more into account the specificities of the aligned ontologies.

Taking into account these specific aspects can be done in different ways: (1) during the alignment process itself or (2) by refining the results generated by the alignment, considered as preliminaries. In the first case, the adaptation of the handled ontologies is possible by the modification of the alignment process parameters or by the definition of a particular combination of the alignment systems. No differentiation is thus made in the way the different elements of the ontologies are treated. Inversely, the refinement of mappings (the alignment results) extends the alignment process, applied in the same way to all ontologies, and completes it. This second solution allows a finer adaptation of the alignment to the specificities of the handled ontologies. It allows also performing differentiated refinements according to the generated results. We retained this solution and we extend it to consider, not only the improvement of the quality of an alignment but also other tasks such as to merge, restructure or enrich ontologies. All these tasks are based on mappings and are specific to the characteristics of the treated ontologies, for example, how these ontologies are structured or how the labels of their concepts are built. They must be also made in interaction with the expert.

Currently, there is no tool which allows to specify easily particulars treatments to be applied to an alignment, this is why we propose the environment TaxoMap Framework, which allows these specifications based on the alignment tool TaxoMap [14] [7].

Our contributions, in this paper, focus on the conception of this environment, on the definition of a first set of primitives to support the specification of treatments, and on the presentation of a use of the environment for the mapping refinements in the topographic field.

The paper is organized as follows. In the next section, we present the context of this work, in particular the ontology alignment tool TaxoMap and the objectives aimed by the conception of TaxoMap Framework. In Section 3 we presents the approach adopted in TaxoMap Framework and in Section 4 we describe the use of this approach for the mapping refinements applied in the topographic field within the ANR project GéOnto [5]. In Section 5 we present some related works. Finally we conclude and give some perspectives in Section 6.

## 2 CONTEXT

TaxoMap Framework is based on the alignment tool TaxoMap [14] [7]. We describe the tool in Section 2.1 and the objectives of the approach in Section 2.2.

## 2.1 TAXOMAP

TaxoMap has been designed to align ontologies $O = (C, H)$ ($C$ a set of concepts and $H$ a subsumption hierarchy). The alignment process is an oriented process which tries to connect the concepts of the source ontology $O_S$ to the concepts of the target ontology $O_C$. The correspondences found are equivalence relations (*isEq*), subsumption relations (*isA*) or proximity relations (*isClose*).

To identify these correspondences, TaxoMap implements techniques which exploit the labels of the concepts and the subsumption links that connect the concepts in the hierarchy [6]. These techniques are based first, on the use of morpho-syntactic analysis tool, *TreeTagger* [15], a tool for tagging text, and second, on a similarity measure which compares the tri-grams of the labels of the concepts [12].

Once classified by *TreeTagger*, the words of the labels are divided into two classes, *full words* and *complementary words*, according to their category and their position in the label. This repartition between *full* and *complementary words* is then used to give more weight to the full words in the calculation of similarity between concepts.

[1] LRI – University of Paris-Sud 11, CNRS & INRIA Saclay Île-de-France
Parc Orsay Université – 4 rue Jacques Monod – 91893 Orsay (France)
{Faycal.Hamdi, Chantal.Reynaud, Brigitte.Safar}@lri.fr

## 2.2 TAXOMAP FRAMEWORK OBJECTIVES

Many ontology alignment tools have been developed in these last years but as shown in the results of OAEI campaign (Ontology Alignment Evaluation Initiative) [9] organized every year since 2004 [3][1], no tool reaches 100% of precision and recall, even if the results obtained by some of these tools are very good. These results concern also TaxoMap. We observed TaxoMap trough its results in this competition in the two last years [7][6] and also through our participation in the ANR project GéOnto [5]. The aim of this project is the construction and the enrichment of a topographic ontology from documents of different geographic domain. The enrichment is based on the results of the alignment of this ontology with other ontologies of the same domain. Tests performed on taxonomies provided by the COGIT-IGN (project partner), have shown that TaxoMap gives in this context very good results (precision 92.3%) but that they could still be improved.

A study showed that the improvements desired by the experts are often specific to the aligned ontologies. To not make TaxoMap a tool used only to align the topographical taxonomies (and thus the quality of results would not be guaranteed when it align other ontologies), we propose to the experts of the project an environment allowing them to specify and perform different treatments. This environment will be used to improve the quality of an alignment provided by TaxoMap, but also for any other treatment based on the results of an alignment between ontologies, such treatments of merging, restructuring or enriching ontologies.

## 3 TAXOMAP FRAMEWORK APPROACH

The approach TaxoMap Framework has been designed to meet the objectives described in section 2.2. We describe the approach and a diagram representing the architecture of this environment respectively in Section 3.1 and 3.2. This environment allows the specification of treatment from predefined primitives. These are presented in Section 3.3.

## 3.1 PRESENTATION OF THE APPROACH

An important feature of the approach is to allow a declarative specification of treatments based on particular alignment results and concerning particular ontologies, using a set of generic and predefined primitives.

Treatments which can be specified depend on the characteristics of the concerned ontologies and the aimed task (mapping refinements, ontology merging, restructuring, enriching). These treatments are thus associated to independent specifications modules, one for each task, having each their own set of primitives of specification. The approach is extensible and a priori applicable to any treatment based on the alignment results.

This approach should help to refine the alignment results produced by TaxoMap. It must be possible, for example, to specify that the subsumption mapping "*isA*" generated between "Way and coastal path" and "Path", as shown in Figure 1 must be replaced by a mapping of the same type but between "Way and coastal path" and "Way". Indeed, "Path" is defined in $O_C$ as a kind of "Way" and the term "Way" itself is used in the label "Way and coastal path". The expert would thus prefer to establish a mapping directly between "Way and coastal path" and "Way".

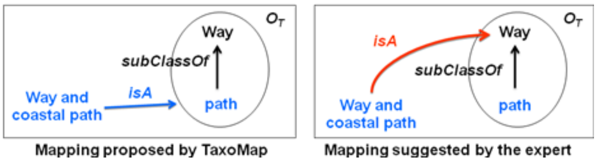


**Figure1**. Example of treatment to specify

The specification of treatments should be expressed as generic as possible. Thus, the treatment shown in Figure1 should not refer directly to the concepts denoted by "Way", "Path" and "Way and coastal path". To help the expert to clarify the conditions for applying the treatment he wishes to implement, we propose a set of predefined generic primitives. These primitive allow a representation of various conditions which can be tested on the mapping concepts identified by TaxoMap. By analyzing alignment results and by leaning on proposed primitives, the expert will be able to identify a "group" of mappings requiring the same refinement and to specify the appropriate treatment to apply to each identified group. The specification will be so declared in a generic way then instantiated on the alignment results and the concerned ontologies to perform the corresponding treatments.

The approach should also allow other treatments such as the restructuration of an ontology *O'* built from $O_S$ and $O_C$, and the alignments generated by TaxoMap between these two ontologies. Thus, it should help to explain a processing deciding, for example, what mappings "*isA*" must be transformed into "*subClassOf*" relations and accompanying this transformation by importing in *O'* dependant concepts.

## 3.2 ARCHITECTURE OF TAXOMAP FRAMEWORK

Figure 2 presents the environment of specification implementing the approach TaxoMap Framework. This environment has three modules: "controller", "knowledge" and "treatment".

The "knowledge" module includes all knowledge on which treatments may be specified. It thus includes the ontologies aligned by TaxoMap $O_S$ and $O_C$ and the generated alignment (Mappings database). According to the treatment performed, we can also find ontology $O_F$ resulting from the fusion of $O_S$ and $O_C$ performed by exploiting the mappings database or by exploiting an ontology $O'_F$ corresponding to a restructured or enriched version of $O_F$.

The "treatment" module includes the alignment tool TaxoMap and all modules associated to the different tasks to perform. TaxoMap execute sequentially 9 techniques, which may or may not be chosen during a particular session. In addition, the execution order of these techniques is customizable. The modules associated to the tasks allow to specify particular treatments that an expert wishes to implement on particular ontologies or alignments, and also to execute these treatments. Additional modules can easily be added if they combine the appropriate primitives (which may be taken from primitives proposed in other modules).

The "controller" module can manage all possible treatments using this environment, i.e. the specification of treatments and their execution, the access to relevant data and the storage of the obtained results.

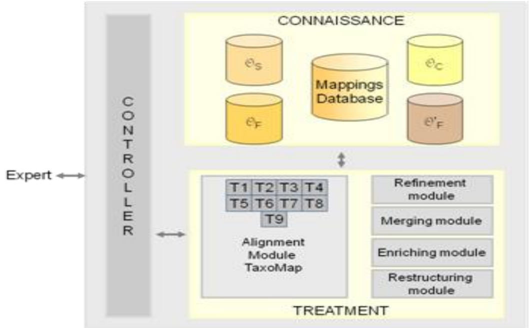


**Figure 2.** Architecture of TaxoMap Framework

## 3.3 PRIMITIVES OF TAXOMAP FRAMEWORK

The specifications of the treatments to implement on the alignment results should be easy. All the elements relevant to their specification are so given to the expert in the form of primitives. These sets of primitive differ according to the aimed task (mapping refinements, ontology merging, restructuring, enriching).

The specification of a treatment has two parts: a “condition” part which must be satisfied to make the execution of the treatment possible, and an “action” part which expresses the process to achieve when the “condition” is satisfied.

The **condition part** is expressed through a set of primitives, identified as necessary to translate the specifications of the treatments proposed by experts. These primitives are designed to test (1) the technique used to identify the considered mapping, (2) the structural constraints on mapped elements, for example, the fact that they are related by a subsumption relation to concepts verifying or not some properties, or (3) the terminological constraints, for example, the fact that the labels of a concept are included in the labels of other concepts. These conditions are represented using three kinds of predicates:

**The predicates relating to the type of techniques applied in the identification of a mapping by TaxoMap**. By testing the existence in the mappings database of a particular relation generated by a given technique, these predicates test implicitly the conditions for the application of this technique. The expert neither needs to know precisely the techniques used nor to re-specify them. For example the primitive ”*isAStrictInclusion(X, Y)*” tests the existence of a mapping “*isA*” generated between two concepts X and Y using the technique $t_2$. It validates implicitly at the same time the conditions for the application of $t_2$, i.e. (1) one of the labels of Y is included in one of the labels of X, (2) all the words of the labels of Y are classified as *full word* by *TreeTagger*, and (3) the concept Y is the concept of $O_C$ having the highest similarity value with the concept X.

TaxoMap including several alignment techniques and thus, several predicates will be defined. More formally, let:
$R_M$ = {*isEq*, *isA*, *isClose*}, the set of correspondence relations used by TaxoMap,
$T = \{t_1, t_2, t_3, t_4, t_5, t_6, t_7, t_8, t_9\}$, the set of techniques.
$T_M$, the table storing generated mappings in the form of 4-tuple (x, y, r, t) where $x \in C_S$, $y \in C_C$, $r \in R_M$, $t \in T$.
The pairs of variables *(X, Y)* which can instantiate these primitives will take their values in the set $\{(x, y) \mid (x, y, r, t) \in T_M\}$.
The primitives necessary for the task of refinement are (in this paper we present just three primitives) :
- *isEquivalence (X,Y)* is true iff $\exists$ (X, Y,isEq, $t_1$) $\in T_M$
- *isAStrictInclusion (X,Y)* is true iff $\exists$ (X,Y,isA,$t_2$)$\in T_M$
- *isCloseStrictInclusion(X,Y)* is true iff $\exists$ (X,Y,isClose, $t_3$) $\in T_M$

These primitive will be presented to the expert via an interface, with comments clarifying their conditions of validation as well as examples and counter-examples of use.

**The predicates expressing structural relations between concepts X and Y of the same ontology *O* = (*C*,*H*).** Note that the instances of variables in these predicates will be constrained, either directly because they instantiate the previous predicates, i.e. concerning the type of the applied techniques, or indirectly by having to be in relation with other instances.
- *isSubClassOf(X,Y,O)* is true $\Leftrightarrow$ subClassOf(X,Y) $\in H$
- *isSuperClassOf(X,Y,O)* is true $\Leftrightarrow$ subClassOf(Y,X) $\in H$

**The predicates expressing terminological relations between the labels of the concepts**:
- *strictInclusionLabel*(X,Y) is defined as follows:
  For each label $L_1$ of X
    For each label L2 of Y
      If $L_1 \subseteq$ *FullWords*($L_2$,$L_1$) then return true
  End

Where X and Y $\in C_S \cup C_C$ and *FullWords*($L_2$,$L_1$) is a function, which calculates all the terms of $L_2$ considered as full words in its comparison with $L_1$.
- *conceptsDifferent*(X,Y) is true $\Leftrightarrow$ *ID*(X) $\neq$ *ID*(Y) with *ID*(X) is the identifier of the concept X.
- *inclusionInLabel*(X,Y) is true iff $\exists$ a label $L_1$ of Y / X $\subset L_1$, where X $\in$ {“and”, “or”} and Y $\in C_S \cup C_C$.

The **action part** describes the procedures to be performed. We identified an initial set of actions. They are represented using the following three procedures:
- *Add_Mapping*(X, Y, R) has the effect of adding a tuple to the table $T_M$ which becomes $T_M \cup$ {(X, Y, R, t)} where R and t are fixed in the treatment condition, by instantiating the predicate identifying the technical question.
- *Delete_Mapping*(X, _, Y) has the effect of removing a tuple from the table $T_M$ which becomes $T_M$ - {(X, Y, _ , _)}
- *Add_Relation*(X, Y, R) corresponds to the addition of a relation R between X and Y with R $\in R_M \cup$ {subClassOf}.

In the Framework, the expert will be able to select all primitives through an appropriate GUI. Note that the approach is based on the use of TaxoMap as alignment tool, but it could be based on another tool if the primitives associated with this tool have been defined. In principle, the method is reproducible. Other predicates that express structural or terminological relations between concepts will be probably introduced for the treatment of the enriching or the restructuring tasks.

# 4 APPLICATION TO THE MAPPING REFINEMENTS

The mapping refinements module is the first module of TaxoMap Framework, realized within the ANR project, GéoOnto [5]. We describe the application setting and then we present the specifications of the treatment of mapping refinements required by the experts of the COGIT-IGN (project partner).

## 4.1 APPLICATION DOMAIN

One of the goals of the GéoOnto project is to build an ontology of topographic concepts, as complete as possible, by enriching an initial taxonomy of terms. Topo-Cogit is an ontology already achieved by the COGIT. The enrichment is carried out, by the alignment of this ontology with other ontologies of the same domain. Thus, within the project, other partners are developing an ontology based on the topographic specifications of the IGN databases and on travel books of the library of Pau [8, 10]. The enrichment process should be automated and reused in the future on other ontologies domain. As this process is based on the results of alignments, they must be as accurate as possible, to minimize the contributions of experts.

The first tests have been performed using a second ontology, Carto-Cogit built manually from the specification of the IGN database. In these tests, the objective is to align the 495 concepts of the source ontology Carto-Cogit with the 600 concepts of the target ontology to enrich, Topo-Cogit. During these tests, 326 mappings have been identified by TaxoMap and presented to experts following the techniques used to obtain them. 25 mappings (precision 92.33%) have been deemed as invalid. For other mappings, the expert proposed alternative mappings. The treatments of mapping refinements proposed below intended to obtain these alternative mappings.

## 4.2 THE MAPPING REFINEMENTS SPECIFICATION

We present in this section two expected changes and in each change the specification of treatment such as they can be expressed in the environment TaxoMap Framework.

- **Case 1:** The first improvement is presented in Section 3.1 (see Figure 1). Generally, it concerns mappings connecting by a subsumption relation "*isA*" a concept $c_S$ of the source ontology $O_S$ to a concept $c_{TMax}$ of the target ontology $O_C$, such as one of the labels of $c_{TMax}$ is included in the labels of $c_S$. If one of labels of the concept $A$ that subsumes $c_{TMax}$ in $C_C$ is also included in the label $c_S$, (see Figure 3), the expert prefers to attach $c_S$ with $A$, the most general concept of $O_C$.

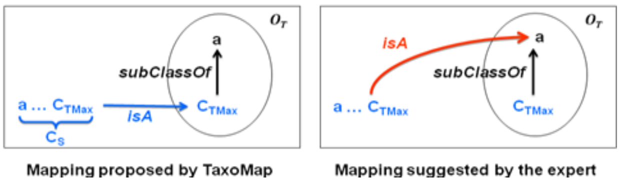


**Figure 3.** Illustration of the treatment1

The specification of the correspondent treatment is:
Conditions of application:

∃X ∃Y *isAStricInclusion* (X,Y)
∧ ∃Z *isSubClassOf* (Y,Z,$O_C$)
∧ *strictInclusionLabel* (Z,X))

Actions: *Delete_Mapping* (X,_,Y)
∧ *Add_Mapping* (X,Z, isA)

The application of this treatment on the example presented in Figure 1, allows first to select from the mappings database the mapping (*ID*("Way and coastal path"), *ID*("Path"), isA, $t_2$) satisfying the primitive *isAStricInclusion*(*ID*("Way and coastal path"), *ID*("Path")). The variables X and Y are instantiated respectively by X/*ID*("Way and coastal path") and Y/*ID*("Path"). The use of the structural predicate *isSubClassOf*(*ID*("Path"), Z, $O_T$) allows the instantiation of the variable Z, Z/ID("Way") and the verification of the terminological predicate *strictInclusionLabel*(*ID*("Way"), *ID*("Way and coastal path")).
The mapping (ID ("Way and coastal path"), ID ("Path"), isA, $t_2$) is removed from the database and replaced by the mapping (ID ("Way and coastal path"), ID ("Way"), isA, $t_2$).

- **Case 2:** This case concerns the mapping connecting by a relation of proximity "*isClose*" a concepts $c_S$ of the source ontology $O_S$ to a concept $c_{TMax}$ of the target ontology $O_C$, such as one of the labels of $c_S$ is included in $c_{TMax}$ labels. If another label of a concept in $O_C$ contains also $c_S$ labels and if this concept has the same father $p$ in $O_C$ that $c_{TMax}$, the expert prefers to connect to $c_S$ to $p$. An illustration is given in Figure 4.

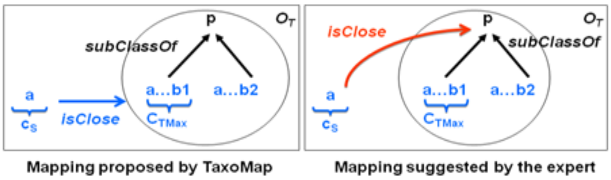


**Figure 4.** Illustration of the treatment2

The specification of the treatment associated to the case 2 is the following:
Conditions of application:

∃X ∃Y (*isCloseStrictInclusion* (X,Y)
∧ ∃Z ∃P (*isSuperClassOf* (P,Y,$O_C$)
∧ *isSuperClassOf* (P,Z,$O_C$ )
∧ *conceptsDifferent*(Y,P)
∧ *strictInclusionLabel* (X,Z) ))

Actions: *Delete_Mapping* (X,_,Y)
∧ *Add_Mapping* (X,P, isClose)

# 5 RELATED WORKS

Many alignment tools existing today generate good results in certain cases and less good in others. This observation should direct research to treat several problems [16] such as: the choice of the most adapted tool, the combination of the alignment techniques and the problem of the regulation of the parameters (thresholds, coefficient of formulas, etc.) used in the alignment tools.

Our works are issued from the same observation but has been developed in a different direction, the alignment refinement, extended in second time to the assistant of the specification of the treatments based on mappings. They can be then compared with those developed in the alignment system COMA++ [2]. This system aims to build powerful alignment tool by the combination of existing tools then to refine the obtained alignment results considered as preliminary. The refinement process is totally automatic.

It consists of the reapplication of the COMA++ alignment process on groups of elements whose proximity has been established by a first treatment applied to ontologies. The refinement of the alignment can also be seen as an adaptation of the alignment solutions to the context of an application. Thus, the system eTunes [11] adapts an alignment by looking automatically to the most adapted values for the parameters of the alignment system. Finally, PROMPT-Suite a system which integrate the ontology merging tool IPROMPT [13], and other tools for the management of multiple ontologies such as the alignment tool Anchor-PROMPT, management of versions, comparison, translation, within the same environment. These tools are interactive and semi-automatic. For example, in the fusion process the system makes suggestions. The expert can hold one of them or specify an operation to perform. The system executes then the operation, calculates the resulting changes, made other suggestions and detects any inconsistencies.

All systems combining several alignment systems are very modular. The possibility of defining the strategy of combination or of adapting automatically parameters makes them adaptable to a new field of application. This modularity and adaptability are strong points which also characterize our approach. The treatments which can be specified in TaxoMap Framework are indeed modular and conceived to integrate the very particular characteristics of the treated ontologies. It goes beyond the possibilities of the tools previously mentioned.

On the other hand, TaxoMap Framework differs from existing tools (such COMA++, eTunes or PROMPT-Suite) by considering that the performance of an alignment tool implementing general alignment algorithms are necessarily limited (even if the values of parameters are optimal). Some improvements can be obtained only after taking into account the particularities of the aligned ontology, which involves various improvements as ontologies. The definition of such improvements needs to be familiar with aligned ontologies. This process cannot thus be automatic, only an expert of the domain is able to make it. As in PROMPT-Suite, we offer an interactive environment to help the expert to carry out this task, but we do it differently. We allow him to define particular generic treatments. In PROMPT-Suite, this is not possible. The treatments are all pre-defined.

# 6 CONCLUSION AND FUTURE WORK

TaxoMap Framework is an environment for the specification of treatments based on the alignment results generated by TaxoMap. We presented the implemented approach in this system, its architecture, then a first set of pre-defined primitive allowing an expert of the domain to specify easily the treatments that would apply to an alignment. We presented the module for assistance to the mapping refinements that we conceived basing on the results of experiments realized within the ANR project, GéOnto.

The conception of TaxoMap Framework is adapted to the specification of other treatments such as merging, restructuring and enriching of ontologies based on alignment. These modules are being implemented as well as the graphical interface that will allow to experts to easily select the appropriate primitive.

# ACKNOWLEDGMENTS

This research is financed by The French National Research Agency through the GEONTO project (ANR-07-MDCO-005, http://geonto.lri.fr/).

# REFERENCES

[1] C. Caraciolo, J. Euzenat, L. Hollink, R. Ichise, A. Isaac, V. Malaisé, C. Meilike, J. Pane, P. Shvaiko, H. Stuckenschmidt, O. Svab, V. Svatek, *Results of the ontology alignment initiative 2008*, in P. Shvaiko, J. Euzenat, F. Giunchiglia, H. Stuckenschmidt (Eds), Proceedings 3rd ISWC workshop on Ontology Matching, Karlsruhe (DE), pp 73-119, 2008.

[2] H.-H. Do, E. Rahm, *Matching large schemas: Approaches and Evaluation*, Information Systems 32 (2007), 857-885.

[3] J. Euzenat, A. Isaac, C. Meilike, P. Shvaiko, H. Stuckenschmidt, O. Svab, V. Svatek, W. van Hage, M. Yatskevich, *Results of the ontology alignment initiative 2007*. In Ontology Matching, 2007.

[4] J. Euzenat, P. Shvaiko, *Ontology Matching*, Springer-Verlag, Heidelberg (DE), 2007.

[5] GéOnto : http://geonto.lri.fr/

[6] F. Hamdi, B. Safar, N. Niraula, C. Reynaud, TaxoMap in the OAEI 2009 alignment contest. – Ontology Matching 2009.

[7] F. Hamdi, H. Zargayouna, B. Safar, C. Reynaud, TaxoMap in the OAEI 2008 alignment contest. – Ontology Matching 2008.

[8] M. Kamel, N. Aussenac-Gilles, *Ontology Learning by Analysing XML Document Structure and Content*, Knowledge Engineering and Ontology Development (KEOD), Madère, Portugal, 2009.

[9] http://www.ontologymatching.org/

[10] E. Kergosien, M. Kamel, C. Sallaberry, M.-N. Bessagnet, N. Aussenac, M. Gaio, *Construction automatique d'ontologie et enrichissement à partir de ressources externes*, JFO'2009, Poitiers.

[11] Y. Lee, M. Sayyadian, A. Doan, A.S. Rosenthal, *eTuner: tuning schema matching software using synthetic scenarios*, The VLDB Journal (2007) 16:97-122.

[12] D. Lin, *An Information-Theoretic definition of Similarity*, in proc of ICML-98, Madison, pp. 296-304, 1998.

[13] N. F. Noy, M.A. Musen, *The PROMPT Suite: Interactive Tools For Ontology Merging And Mapping*, IJHCS, 59(6), pp. 983-1024, 2003.

[14] C. Reynaud, B. Safar, *Techniques structurelles d'alignement pour portails Web*, Revue RNTI W-3, Fouille du Web, Cépaduès, 2007.

[15] H. Schmid, *Probabilistic Part-of-Speech Tagging Using Decision Trees*, Int. Conf. on New Methods in Language Processing, 1994.

[16] P. Shvaiko, J. Euzenat, *Ten Challenges for Ontology Matching*, in proc. Of the 7th International Conference on Ontologies, DataBases, and Applications of Semantics (ODBASE), 2008.