An Ontological and Terminological Resource for n-ary Relation Annotation in Web Data Tables

Rim Touhami, Patrice Buche, Juliette Dibie-Barthelemy, Liliana Ibanescu

An Ontological and Terminological Resource for n-ary Relation Annotation in Web Data Tables

Rim Touhami[1,3], Patrice Buche[1,2], Juliette Dibie-Barthélemy[3], and Liliana Ibănescu[3]

[1] INRA - UMR IATE, 2, place Pierre Viala, F-34060 Montpellier Cedex 2, France
[2] LIRMM, Montpellier, France
[3] INRA - Mét@risk & AgroParisTech, 16 rue Claude Bernard, F-75231 Paris Cedex 5, France
Patrice.Buche@supagro.inra.fr,
{rim.touhami,Juliette.Dibie,Liliana.Ibanescu,}@agroparistech.fr

Abstract. We propose, in this paper, a model for an Ontological and Terminological Resource (OTR) dedicated to the task of n-ary relations annotation in Web data tables. This task relies on the identification of the symbolic concepts and the quantities, defined in the OTR, which are represented in the tables' columns. We propose to guide the annotation by an OTR because it allows a separation between the terminological and conceptual components and allows dealing with abbreviations and synonyms which could denote the same concept in a multilingual context. The OTR is composed of a generic part to represent the structure of the ontology dedicated to the task of n-ary relations annotation in data tables for any application and of a specific part to represent a particular domain of interest. We present the model of our OTR and its use in an existing method for semantic annotation and querying of Web tables.

Keywords: Semantic integration, semantic data model, ontology engineering

1 Introduction

Today's Web is not only a set of semi-structured documents interconnected via hyper-links. A huge amount of technical and scientific documents, available on the Web or the hidden Web (digital libraries, ...), include data tables. They represent a very interesting potential external source for loading a data warehouse dedicated to a given domain of application. They can be used to enrich local data sources or to compare local data with external ones. In order to integrate data, a preliminary step consists in harmonizing external data with local ones, i.e. external data must be expressed with the same vocabulary, generally represented by an ontology, as the one used to index the local data. Ontology is a key notion in the Semantic Web and in data integration researches. According to [1], "Ontologies are part of the W3C standards stack for the Semantic Web, in which they are used to specify standard conceptual vocabularies in which to

exchange data among systems, provide services for answering queries, publish reusable knowledge bases, and offer services to facilitate interoperability across multiple, heterogeneous systems and databases".

In [2–6] ontologies are associated with terminological and/or linguistic objects. In [2] authors motivate why it is crucial to associate linguistic information (part-of-speech, inflection, decomposition, etc.) with ontology elements (concepts, relations, individuals, etc.) and they introduce *LexInfo*, an ontology-lexicon model, implemented as an OWL[4] ontology. Adapting *LexInfo*, [3] presents a model called *lemon* (Lexicon Model for Ontologies) that supports the sharing of terminological and lexicon resources on the Semantic Web as well as their linking to the existing semantic representations provided by ontologies. The *CTL* model from [4] is a model for the integration of conceptual, terminological and linguistic objects in ontologies. In [5] a meta-model for ontological and terminological resources in OWL DL is presented, called an *Ontological and Terminological Resource (OTR)*, extended afterward in [7] in order to be used for ontology based information retrieval applied to automotic diagnosis.

In the same trend, we present in this paper an Ontological and Terminological Resource (OTR) dedicated to the task of data tables integration. An Ontological and Terminological Resource (OTR) [6, 5] is a model allowing joint representation of an ontology and its associated terminology. According to [5], the OTR structuring can be guided by three factors: the task to realize, the domain of interest and the application. In this paper, the domain of interest is the food safety but the OTR structure we propose is generic enough to be applied to many other domains. The application is the construction of a data warehouse opened on the Web. We are interested in loading our data warehouse with data coming from external sources such as scientific papers, international reports or Web pages and in its querying.

In previous works [8, 9], we proposed a data tables semantic annotation method guided by an ontology, but we did not especially pay attention to the ontology modelling and only use a preliminary version built from scratch by domain ontologists. Nevertheless, since our ontology is at the heart of our method, it appears that its modelling is essential to the sustainability of our approach and more generally to the data tables semantic annotation task. Like in [10, 11], we are addressing the situation when data tables consist of a header row that represents semantic relationships between concepts which may be symbolic concepts or quantities associated with units. [10] proposes a method to discover semantic relations between concepts. Our purpose is different: the semantic annotation of a data table consists in (i) recognizing the semantic relations defined in the OTR and represented in the data table; (ii) instantiating each recognized relation in each row of the data table, that is identifying their values in each row. Our final objective is to integrate in the same 'schema', Web data tables. The work of [11] can be considered as a sub-task of ours as they focus on the recognition of quantities in columns of the tables.

[4] http://www.w3.org/TR/2004/REC-owl-features-20040210/

The model of an OTR proposed in this paper is dedicated to the task of n-ary relations annotation in Web data tables. In the OTR, a clear separation is done between conceptual aspects and terminological ones. The conceptual part represents the semantic expressed by concepts while the terminological part allows one to define the terminology and its variations (multilingual, synonyms, abbreviations) denoting the concepts. The terminological part of the OTR allows one to improve the semantic annotation of data tables in a multilingual context thanks to the synonyms and abbreviations management. Moreover, this clear distinction between conceptual and terminological aspects and the management of unit conversions allow one to improve the querying of the data warehouse. As a matter of fact, since the data are annotated with concepts of the OTR, their querying can also be performed thanks to these concepts without worrying about the terminological variations and unit conversions (see [12, 13] for preliminary works).

The structure of this paper is as follows. We first present the OTR in Section 2, with its conceptual and terminological parts. Then, in Section 3 a semantic annotation method of data tables guided by this OTR is presented. We finally conclude and present our future work in Section 4.

2 Modelling of the Ontological and Terminological Resource (OTR)

Since the modelling of the OTR is dedicated to the task of n-ary relations annotation in Web data tables, we present in Figure 1 an example of a semantic annotation of a Web data table extracted from a scientific paper in food science.

Cells of a data table contain terms (e.g. *MFC film A*) denoting symbolic concepts (e.g. *Packaging Material*) or numerical values (e.g. 3), often followed by a unit of measurement (e.g. ml m-2 day-1). Usually a data table represents semantic relationships between concepts which may be symbolic concepts or quantities characterized by units. The semantic annotation of a data table consists in recognizing the relations represented by the data table, which suppose to recognize symbolic concepts but also quantities and units. In the data table from Figure 1, the semantic relation *O2Permeability_Relation* which represents oxygen permeability for a food packaging material given its thickness, temperature and humidity has been partially recognized: the symbolic concept *Packaging* has been recognized in the first column, the quantity *Thickness* in the third column and the quantity *O2Permeability* in the last column, but the quantities *Temperature* and *Relative_Humidity* have not been recognized.

The OTR used for the semantic annotation of data tables should contain symbolic concepts, quantities and associated units, semantic relations linking symbolic concepts and quantities. Figure 2 presents an excerpt of our OTR in food science domain. An OTR is composed of a conceptual component, the ontology, and a terminological component, the terminology. We first present, in Subsection 2.1, the conceptual component of the OTR and, in Subsection 2.2, its

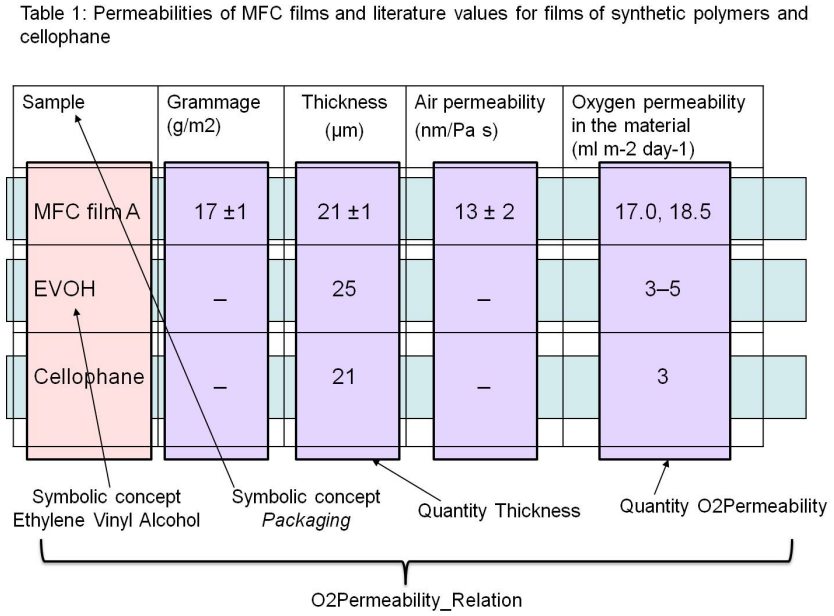

Fig. 1. Example of a web data table

terminological component. We modelled the conceptual and the terminological component of our OTR using the OWL2-DL[5] model.

2.1 Conceptual component of the OTR

The conceptual component of the OTR is composed of two main parts: on the one hand, a generic part, commonly called core ontology, which allows the representation of the structure of the ontology and is dedicated to the n-ary relations annotation task in data tables and, on the other hand, a specific part, commonly called domain ontology, which depends on the domain of interest. Our OTR is generic because it allows n-ary relation to be instantiated in data tables for any application. Additionally, its specific part allows the representation of a particular domain of interest.

Figure 3 presents the generic part of our OTR which does not depend on a domain of interest. There are three categories of generic concept: *Dimension*, *T_Concept* and *UM_Concept*. The generic concept *Dimension* represents dimensions that allow quantities and unit concepts (e.g *Temperature*, *Length*, *Time*, ...) to be classified. The generic concept *T_Concept* contains concepts to be recognized in data tables (in their cells, columns and rows) and are of three kinds: *Relation*, *Simple_Concept* or *Unit_Concept*. It is called *T_Concept* for *Terminological Concept*, because as detailed in Subsection 2.2, it contains concepts

[5] http://www.w3.org/TR/owl2-overview/

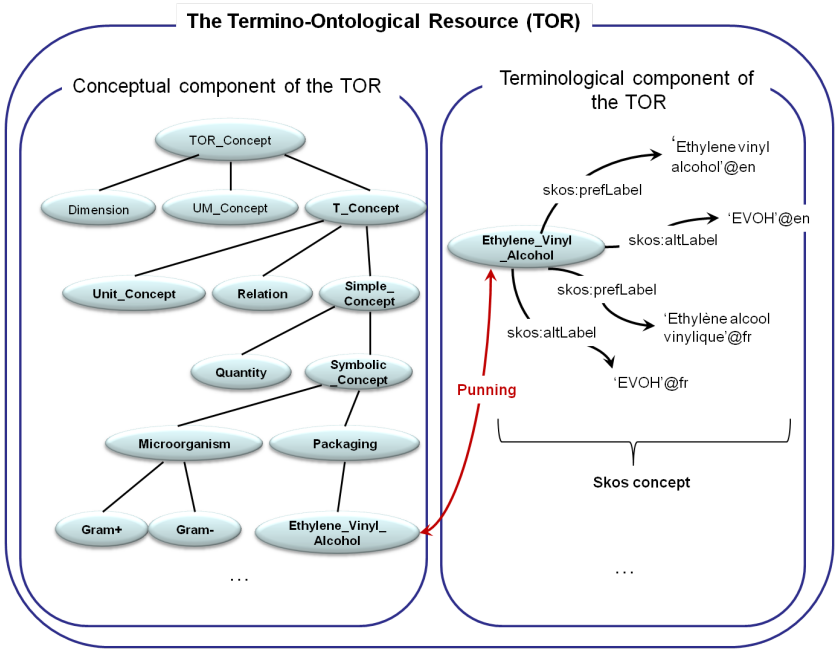

Fig. 2. An excerpt of the OTR in food science domain.

having one or several terms associated with the terminological component. The generic concept *UM_Concept* contains concepts which are used to manage units of measurement, especially conversions between units of measurement.

The specific part of the OTR allows the representation of all concepts which are specific to a domain of interest. They appear in the OTR as sub concepts of the generic concepts. In OWL, all the concepts are represented by OWL classes, which are hierarchically organized by the *subClassOf* relationship and are pairwise disjoints.

We detail below the three kinds of the generic concept *T_Concept* with an example of specific sub concepts in food science domain and we present the management of conversions between units of measurement.

Presentation of the generic concept *Simple_Concept*: Simple concepts include symbolic concepts (*Symbolic_Concept*) and quantities (*Quantity*).

1. *Symbolic_Concept*: A symbolic concept is characterized by its label (i.e. a term composed of one or more words), defined in the terminological part of the OTR, and by its hierarchy of possible values.

   Example 1. Figure 4 presents an excerpt of the symbolic concepts hierarchy in food science domain. The specific symbolic concepts are sub-concepts of the generic concept *Symbolic_Concept*. For example, *Food_Product* and *Cereal*

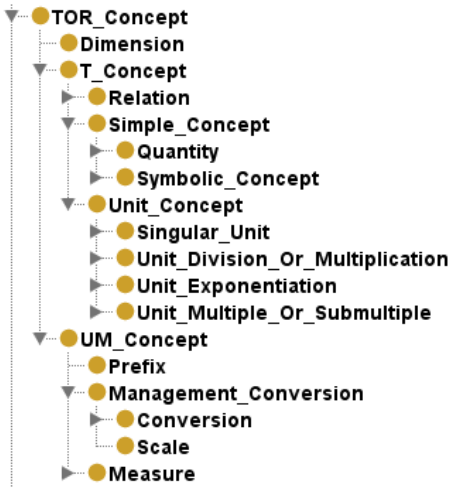

Fig. 3. The generic part of the OTR.

are two specific symbolic concepts, *Cereal* is a kind of *Food_Product*. The food science domain OTR contains 4 distinct hierarchies of specific symbolic concepts:

- *Food_Product* which has more than 500 sub concepts,
- *Microorganism* which has more than 150 sub concepts,
- *Packaging* which has more than 150 sub concepts, and
- *Response* which has three sub concepts: *growth*, *absence of growth* and *death*, which represent possible responses of a micro-organism to a treatment.

Let us notice that we could not reuse pre-existing terminologies for food products as AGROVOC[6] (from FAO - Food and Agriculture Organisation of the United Nations) or Gems-Food[7] (from WHO - World Health Organisation), because those terminologies are not specific enough compared to the one founded in our corpus in food science (respectively only 20% and 34% of common words).

2. *Quantity*: A quantity is characterized by its label, defined in the terminological part of the OTR, a set of units, which are sub concepts of the unit concept *Unit_Concept*, a dimension, which is sub concept of the dimension concept *Dimension*, and eventually a numerical range. An OWL object property *hasUnitConcept* associates a quantity with a set of unit concepts: it has for domain the generic concept *Quantity* and for range the generic concept *Unit_Concept*. An OWL object property *hasDimension* associates a quantity with a dimension: it has for domain the generic concept *Quantity* and for range the generic concept *Dimension*. We use the numerical restrictions

[6] http://aims.fao.org/website/AGROVOC-Thesaurus
[7] http://www.who.int/foodsafety/chem/gems/en/

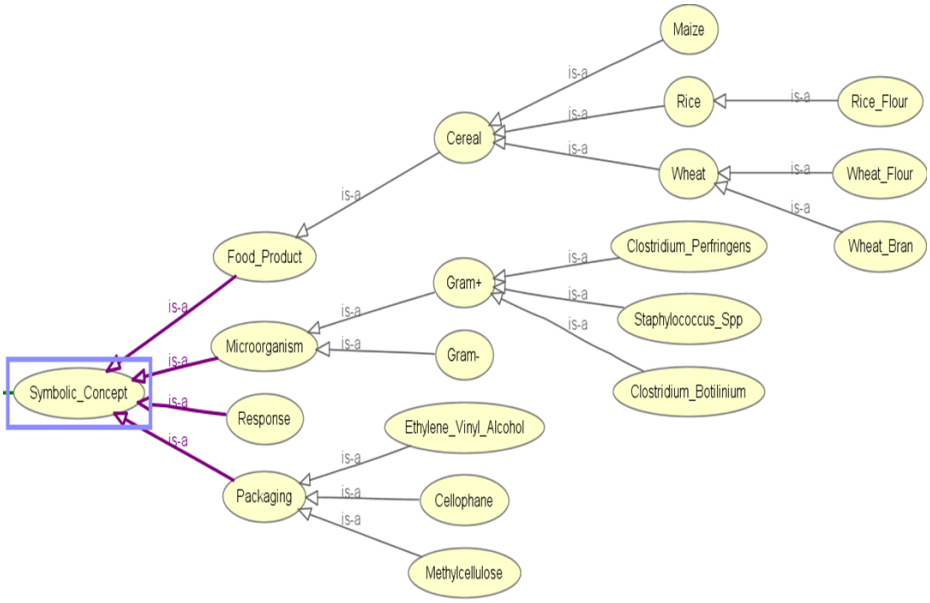

Fig. 4. An excerpt of the symbolic concepts hierarchy in food science domain.

of OWL2 (e.g. *minInclusive* and *maxInclusive*) to represent the maximal and minimal values associated with a quantity.

Example 2. Figure 5 presents an excerpt of quantities in food science domain. The specific quantities, such as *PH*, *Permeability* or *Relative_Humidity*, are sub-concepts of the generic concept *Quantity*. The food science domain OTR contains 22 quantities. Figure 6 shows that the specific quantity *Relative_Humidity* can be expressed using the unit *Percent* or the unit *One*, which indicates dimensionless quantity, and it is restricted to the numerical range [0, 100].

Presentation of the generic concept *Unit_concept*: A unit concept represents a unit of measurement. It is characterized by its label, defined in the terminological part of the OTR, a dimension and eventually by conversions. Our classification relies on the International System of Units[8]. There exist several ontologies dedicated to quantities and associated units (OM[9], OBOE[10], QUDT[11], QUOMOS, ...). We learned from these ontologies how to build ours, but they cannot contain all the required specific units for a given domain. For instance,

[8] http://www.bipm.org/en/si/
[9] http://www.wurvoc.org/vocabularies/om-1.8/
[10] http://marinemetadata.org/references/oboeontology
[11] http://www.qudt.org/

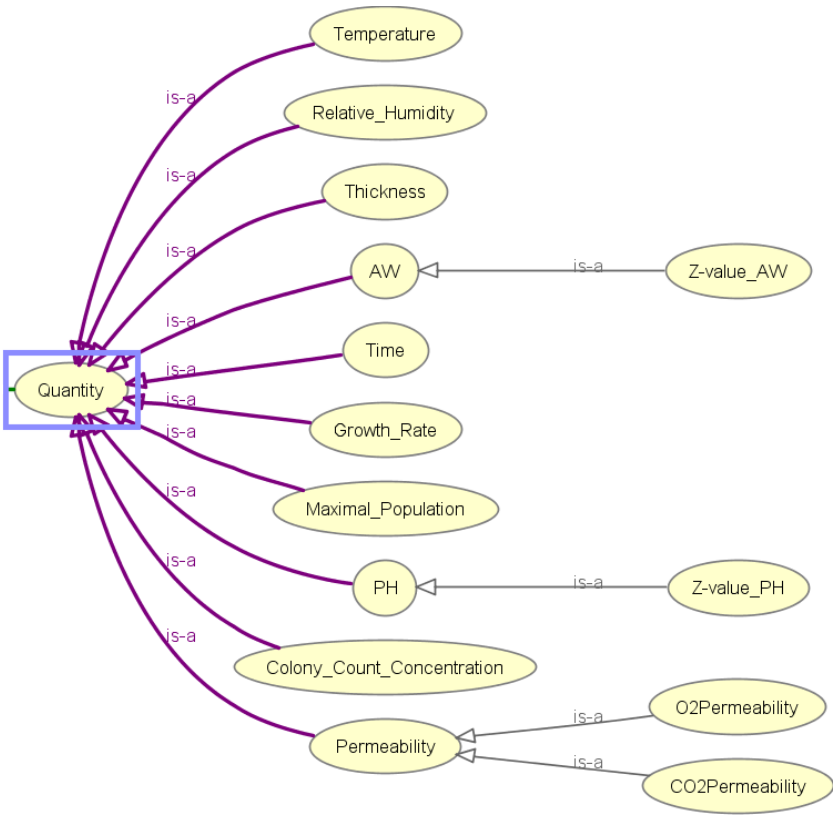

Fig. 5. An excerpt of the quantities in food science domain.

in food science domain, the ontologist has added some units such as ppm[12] or CFU/g[13].

Example 3. Figure 7 presents an excerpt of the unit concepts hierarchy in food science domain. Specific concepts, such as *Day* (d), *Square_Metre* (m^2), *Micrometre* (µm) or *Cubic_Centimetre_By_25_Micrometre_Per_Square_Metre_Per_Day_Per_Atmosphere* ($cm^3 25\ \mu m/m^2/d/atm$) appear as sub concepts of the generic concepts *Singular_Unit*, *Unit_Exponentiation*, *Unit_Multiple_Or_Submultiple* and *Unit_Division_Or_Multiplication*. The concept *Measure* is used to represent components of units of measurement which are written in the form of a constant multiplied by a unit (e.g 25 µm). The concept *Prefix* is used to represent constant values defined in the International System of Units (e.g *Micro*(µ)).

[12] parts per million. ppm is a unit of concentration often used when measuring levels of pollutants in air, water, body fluids, etc.

[13] colony-forming units per gram. Colony-forming units (CFU) is a measure of viable bacterial or fungal numbers in microbiology

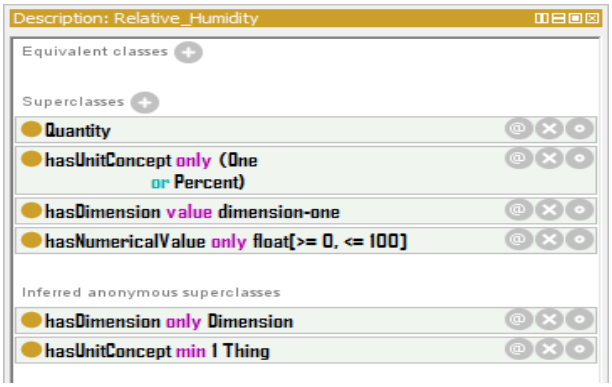

Fig. 6. The specific quantity *Relative_Humidity*.

Presentation of the generic concept *Relation*: The concept *Relation* allows a n-ary relationship between simple concepts to be represented. A relation is characterized by its label, defined in the terminological part of the OTR, and by its signature (i.e. the set of simple concepts which are linked by the relation). The signature of a relation is defined by a domain and a range. The range is limited to only one simple concept, called *result concept*, while the domain contains one or several simple concepts, called *access concepts*. The restriction of the range to only one result concept is justified by the fact that, in a data table, a relation often represents a semantic n-ary relationship between simple concepts with only one result, such as an experimental result with several measured parameters. If a data table contains several result columns, it is then represented by as many relations as it has results. As suggested in [14], a n-ary relation is represented in OWL by a class associated with the access concepts of its signature via the OWL object property *AccessConcept* and the result concept of its signature via the OWL functional object property *ResultConcept*.

Example 4. Figure 8 presents the specific relation *O2Permeability_Relation* which has for access concepts the specific symbolic concepts : *Packaging*, *Relative_Humidity*, *Temperature* and *Thickness* and for result concept the specific quantity *O2Permeability*. It represents oxygen permeability for a packaging material given its thickness, temperature and humidity. The food science domain OTR contains 16 relations.

Management of conversions between units of measurement : As pointed out in Section 1 the modelling of our OTR, dedicated to the task of data tables integration, has been guided by the construction of a data warehouse opened on the Web. In order to load and query the data warehouse and to be able to use data in decision models, we will have to convert automatically numerical data. We define the generic concept *Conversion*, sub concept of the generic concept

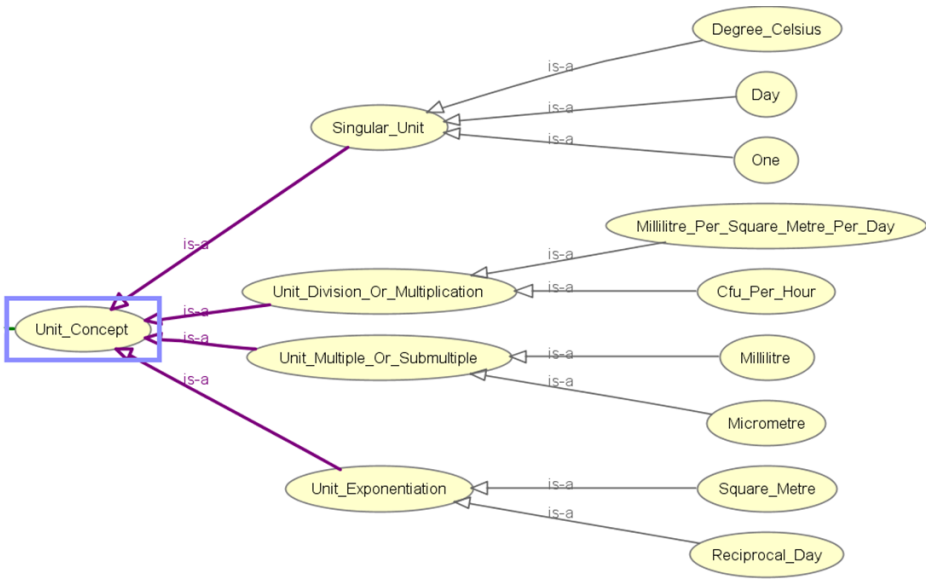

Fig. 7. An excerpt of the unit concepts hierarchy in food science domain.

UM_Conversion (see Figure 3), which is associated with units of measurement through the property *hasConversion*.

In this paper, we consider conversions between units of measurement which can be modelled by the following equation: $v_t = (v_s + o) * s$, where v_t is the value expressed in the target unit, v_s is a value expressed in a source unit, o is the offset, and s is the scale. A lot of conversions between units of measurement can be done using a conversion factor (the scale) as those published by the US National Institute of Standards and Technology[14]. Conversions between units of measurement for temperatures require to introduce an additional offset (see for instance http://en.wikipedia.org/wiki/Fahrenheit).

Let us illustrate the management of conversions between units of measurement through one example.

Example 5. To convert a temperature value expressed in Fahrenheit into Celsius, we use the following formula: $v_{\circ \mathrm{C}} = (v_{\circ \mathrm{F}} - 32) \times \frac{5}{9}$. To do this, we define the class *FahrenheitToCelsius*, detailled in Figure 9, as a subclass of the class *Conversion*, where the class *Degree_Fahrenheit* is a subclass of the generic concept *Singular_unit*.

2.2 Terminological component of the OTR

The terminological component represents the terminology of the OTR: it contains the set of terms of the domain of interest. As mentioned in Section 2.1,

[14] http://ts.nist.gov/WeightsAndMeasures/Publications/appxc.cfm

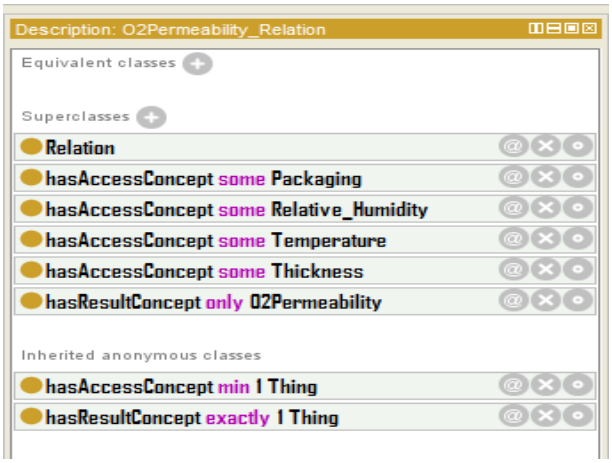

Fig. 8. The specific relation *O2Permeability_Relation*

Fig. 9. An example of conversion for temperature

at least one term of the terminological component is associated with each sub concept of the generic concept *T_Concept*; for exemple the term *Ethylene vinyl alcohol* is associated with the concept *Ethylene_Vinyl_Alcohol*. As a matter of fact, each sub concept of *T_Concept* is characterized by a label, i.e. a sequence of words defined in a given language. More precisely, it is characterized by a preferred label in a given language, but it may also be characterized by alternate labels, which correspond to synonyms or abbreviations, this in different languages. Those labels associated with a given concept are used in the semantic annotation of data tables: they are compared with the terms present in the data tables (in their cells, columns' titles, table title) in order to be able to recognize the concepts of the OTR (more precisely the sub concepts of *T_Concept*) that the data tables represent.

We propose to associate labels with each sub concept of *T_Concept* using the labeling properties of SKOS[15] (Simple Knowledge Organization Scheme) which is a W3C recommendation and is based on RDF language. Thanks to the metamodelling of OWL2-DL, each sub concept of *T_Concept* is defined at the same time as an OWL class and as an instance of the class *OWL SKOS : Concept* (see example in Figure 2). More precisely, the same identifier (URI) is associated with its OWL class representation and its individual representation, using the punning[16] metamodelling capabilities available in OWL2-DL. Therefore, each sub concept of *T_Concept* is defined, on the one hand, as an OWL class in order to be instantiated in rows of a data table and, on the other hand, as an instance in order to allow one to compare its associated labels with the terms present in the data tables.

Example 6. In food science OTR, the symbolic concept *Ethylene_Vinyl_ Alcohol* was defined both as an OWL class in order to be able to instantiate it in a data table, and as an instance of the class *OWL SKOS : Concept* allowing to represent its terminological characteristics by using the labeling properties *prefLabel* and *altLabel* of SKOS. The concept *Ethylene_Vinyl_Alcohol* is then defined as follows:

```
<owl:Class rdf:ID="Ethylene Vinyl Alcohol">
  <rdfs:subClassOf rdf:resource="#Packaging"/>
  <rdf:type rdf:resource="http://www.w3.org/2004/02/skos/core#Concept"/>
  <skos:prefLabel xml:lang="en">Ethylene vinyl alcohol</skos:prefLabel>
  <skos:altLabel xml:lang="en">EVOH</skos:altLabel>
  <skos:prefLabel xml:lang="fr">Ethylï¿½ne alcool vinylique</skos:prefLabel>
  <skos:altLabel xml:lang="fr">EVOH</skos:altLabel>
</owl:Class>
```

3 Using the OTR to annotate and query data tables

We propose to illustrate the relevance of our modelling choices made in our OTR, by using it in the semantic annotation method of Web data tables proposed in [9]. We briefly present, in this section, the main steps of this method (see Figure 10) and its adaptation to our OTR through an example: the annotation of the data table from Figure 1.

Distinction between symbolic and numerical columns. The first step of the semantic annotation method is to distinguish between symbolic and numerical columns, by counting occurrences of numerical values and terms found in each column and by using some of the knowledge described in the OTR (e.g terms denoted unit concepts are accepted in numerical columns).

Example 7. In Table 1 from Figure 1 the first column is identified as a symbolic column: it contains only terms. The other columns are identified as numerical ones: they contain only numerical values or ranges of numerical values (e.g 21 ± 1).

[15] http://www.w3.org/TR/skos-reference/
[16] http://www.w3.org/TR/owl2-new-features/#F12:_Punning

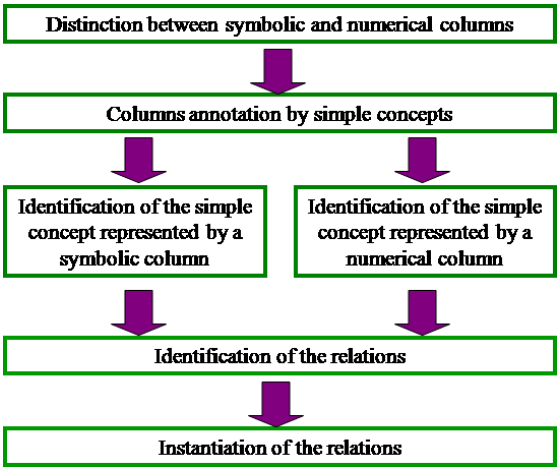

Fig. 10. The main steps of the semantic annotation method of a table driven by an OTR.

Columns annotation by simple concepts. Once a column has been classified as a symbolic column or as a numerical one, this step identifies which simple concept of the domain OTR corresponds to the column. In order to annotate a column *col* by a simple concept *c*, two scores are combined: the score of the simple concept *c* for the column *col* according to the column title, and the score of the simple concept *c* for the column *col* according to the column content. Only the simple concepts of the OTR which appear in the signatures of the relations of the OTR are considered. As a matter of fact, the main objective of the semantic annotation method is to identify which relations of the OTR are represented in a Web data table: those simple concepts are called *simple target concepts*.

Example 8. The domain of food science is composed of four symbolic target concepts : *Food_Product*, *Microorganism*, *Packaging* and *Response* (see Figure 4), and it is composed of 22 target quantities of which an excerpt is presented in Figure 5.

Identification of the simple concept represented by a symbolic column. The annotation of a symbolic column by a symbolic target concept relies on a comparison between the terms present in each cell of the column and the list of preferred and alternative labels associated with the concepts which belong to the hierarchy of each symbolic target concept of the OTR. We use the cosine similarity measure [15] to compare terms which have been previously transformed into a vector of lemmatized words using WordNet.

Example 9. Let us consider the first column of Table 1 which was identified as a symbolic column. The following steps allows to annotate this first column with the symbolic target concept *Packaging*.

– The second cell of the first column which contains the term *EVOH* is annotated with the symbolic target concept *Packaging* because this term is among the labels denoting the symbolic concept *Ethylene_Vinyl_Alcohol*, which is a sub concept of the symbolic target concept *Packaging* (see Example 6 and Figure 2). The score of a symbolic target concept $TargetConcept$ for a cell *cell* is computed as the maximum for all the cosine similarity measures between the terms t_i denoting $TargetConcept$ or one of its sub concepts in the OTR and the term contained into the cell:

$$score_{cell}(TargetConcept) = max_i sim(t_i, content(cell)).$$

Then the score of the symbolic target concept *Packaging* for the second cell of the first colums is 1: $score_{cell_{21}}(Packaging) = max_i sim(t_i, EVOH) = 1$.
– The scores of the symbolic target concept *Packaging* for the other cells (i.e *MFC film A* and *Cellophane*) of this column are also computed and equal to 1.
– The score of a symbolic target concept $TargetConcept$ for a column *col* according to the column content is

$$score_{ContentCol}(TargetConcept) = \frac{\#(\text{cells of } col \text{ annotated by } TargetConcept)}{\#(\text{cells of } col)}.$$

Then the score of the symbolic target concept *Packaging* according to the content of the first column is $score_{ContentCol_1}(Packaging) = \frac{3}{3} = 1$.
– In the same way, the scores of the others target concepts of the food science OTR for the first column according to its are computed and equal to 0.
– Furthermore, the score of a symbolic target concept $TargetConcept$ for a column *col* according to the column title is

$$score_{TitleCol}(TargetConcept) = max_i sim(t_i, col(title))$$

where terms t_i denote $TargetConcept$ and $col(title)$ is the content of the title of the column. $score_{TitleCol_1}(Packaging) = max_i sim(t_i, Sample) = 0$ because the symbolic target concept *Packaging* is not denoted by a label syntactically close to the term *Sample*.
– The final score for a column *col* is defined by

$$score_{col}(TargetConcept) =$$

$$= 1-(1-score_{TitleCol}(TargetConcept))(1-score_{ContentCol}(TargetConcept)).$$

Therefore the final score of the symbolic target concept *Packaging* for the first column of Table 1 is : $score_{col_1}(Packaging) = 1-(1-score_{TitleCont_1}(Packaging))(1-score_{ContentCol_1}(Packaging)) = 1-(1-0)(1-1) = 1$. Since all the others symbolic target concepts have a null final score for the first column, the first column of Table 1 is annotated by the symbolic target concept *Packaging*.

Identification of the simple concept represented by a numerical column. The annotation of a numerical column by a target quantity (called $TargetQ$ in the following) relies on the units present in the column and its numerical values, which must be compatible with the numerical range of the target quantity.

Example 10. Let us consider the last column of Table 1 which was identified as a numerical column. The following steps allows to annotate this last column with the target quantity *O2Permeability*:

- First the annotation method identifies in the column the unit concept *Millilitre_Per_Square_Metre_Per_Day* because the label *ml m-2 day-1* is an alternative label for this concept. In food science OTR, this unit concept is only associated with the quantity *O2Permeability*.As the score for a unit *unit* is defined by:

$$score(unit) = \frac{1}{\#\{TargetQ | unit \in hasUnitConcept(TargetQ)\}}.$$

  then, $score(Millilitre\_Per\_Square\_Metre\_Per\_Day) = \frac{1}{1} = 1$.
- As *ml m-2 day-1* is the only unit in the last column and the target quantity *O2Permeability* has no numerical range defined in the OTR, then the score of the target quantity *O2Permeability* for this column according to its content is : $score\_contentCol_5$(O2Permeability)=1.
- In the same way, the scores of the other target quantities of the food science OTR for the column according to the column content are computed and equal to 0.
- Furthermore, the score of a target quantity $TargetQ$ for a column *col* according to the column title is

$$score_{TitleCol}(TargetQ) = max_i sim(t_i, col(title))$$

  where terms t_i denote $TargetQ$ and $col(title)$ is the content of the title of the column.
  $score_{TitleCol_5}(O2Permeability) = max_i sim(t_i, Oxygen\ permeability\ in\ the\ material) = sim(Oxygen\ permeability, Oxygen\ permeability\ in\ the\ material) = 0.816$
  because the target quantity *O2Permeability* is, in particular, denoted by the english preferred label *Oxygen permeability*.
  Besides, the score of the target quantity *CO2Permeability* for the column according to the column title is also computed as follows: $score_{TitleCol_5}(CO2Permeability) = sim(Carbon\ Dioxide\ permeability, Oxygen\ permeability\ in\ the\ material) = 0.408$.
- The final score for a column *col* is $score_{col}(TargetQ) =$

$$= 1 - (1 - score_{TitleCol}(TargetQ))(1 - score_{ContentCol}(TargetQ)).$$

  Therefore, the final scores of the target quantities *O2Permeability* and *CO2-Permeability* for the last column of Table 1 are:

$score_{col_5}(O2Permeability) = 1 - (1 - 0.816)(1 - 1) = 1$,
$score_{col_5}(CO2Permeability) = 1 - (1 - 0.408)(1 - 0) = 0.408$
Since all the others target quantities have a null final score for the last column, the last column of Table 1 is annotated by the target quantity *O2Permeability* which has the best score.

Using the same method, we also determine that the third column of Table 1 is annotated by the target quantity *Thickness*. Furthermore, since no target quantity from the OTR has been identified to annotate the second and the fourth column of Table 1, they are annotated by the generic concept *Quantity*.

Identification of the relations. Once all the columns of a data table have been annotated by concepts of the domain OTR, the fourth step of the annotation method consists in identifying which relations of the OTR are represented in the data table. In order to annotate a data table by a relation, two scores are combined: the score of the relation for the data table according to the data table title and the score of the relation for the data table according to the data table content. This second score depends on the proportion of simple concepts in the relation's signature which were represented by columns of the data table, the result concept recognition being required. Let us notice that a data table can be annotated by several relations.

Example 11. According to Examples 9 and 10, the first column of Table 1 has been annotated by the symbolic target concept *Packaging*, the third column by the target quantity *Thickness*, the last column by the target quantity *O2Permeability* and the second and fourth columns by the generic concept *Quantity*. The data table can be annotated by the relation *O2Permeability_Relation* of the OTR, which has the target quantity *O2Permeability* as result concept. The score of a relation *Rel* according to its signature is:

$$score_{signature}(Rel) = \frac{\#(\text{recognized concept in } Rel \text{ signature})}{\#(\text{concepts in } Rel \text{ signature})}.$$

The score of the relation *O2Permeability_Relation* for Table 1 according to its signature (see Example 4) is: $score_{signature}(O2Permeability\_Relation) = \frac{3}{5} = 0.6$.

The score of a relation *Rel* for the data table *table* according to the data table title is computed as the maximum cosine similarity measure between the terms t_i denoting *Rel* in the OTR and the data table title.

$$score_{TitleTable}(Rel) = max_i sim(t_i, table(title)).$$

As the title of Table 1 is *Permeabilities of MFC films and literature values for films of synthetic polymers and cellophane*, the score of relation *O2Permeability_Relation* is: $score_{TitleTable_1}(O2Permeability\_Relation) = 0.35$.

The final score of a relation *Rel* for a data table *table* is

$$score_{table}(Rel) = 1 - (1 - score_{TitleTable}(Rel))(1 - score_{signature}(Rel)).$$

Therefore, the final score of the relation *O2Permeability_Relation* for the data table Table 1 is: $score_{Table1}(O2Permeability\_Relation) = 1 - (1 - 0.35)(1 - 0.6) = 0.74$.
Since no other relation of the OTR has the target quantity *O2Permeability* as result concept, Table 1 is annotated by the relation *O2Permeability_Relation*.

Instantiation of the relations: The fifth and last step of the annotation method (see Figure 10) is the instantiation of each identified relation for each row of the considered data table. The instantiation of a relation relies on the instantiation of the symbolic target concepts and the target quantities which belong to its signature and were represented by columns of the data table (see [9] for more detail).

Example 12. The instantiation of the relation *O2Permeability_Relation* for the second row of Table 1 is represented by the set of pairs {(original value, recognized simple target concept : (annotation values[17]))} :
{(EVOH, Packaging: (Ethylene Vinyl Alcohol)), (25 μm, Thickness: (value: 25, unit concept: Micrometre)), (3-5 ml m-2 day-1, O2Permeability : (interval of values: [3, 5], unit concept: Millilitre_Per_Square_Metre_Per_Day)) }.

4 Conclusion

We have proposed, in this paper, a model for an Ontological and Terminological Resource (OTR) dedicated to the task of n-ary relations annotation in Web data tables. In this OTR, a clear separation is made between the conceptual and the terminological components using the latest W3C recommendations (OWL2/DL and SKOS). In this OTR, a special effort has been made to distinguish the generic part (core ontology) dedicated to the n-ary relation annotation task for any application from the specific part dedicated to a given application domain. We have demonstrated the relevance of this model by applying it in a semantic annotation method of Web data tables proposed in [9]. Consequently, the OTR model can be reused for any application domain, redefining only its specific part, to annotate n-ary relations from Web data tables. As a matter of fact, since the data are annotated with concepts of the OTR, their querying can also be performed thanks to these concepts without worrying about the terminological variations and unit conversions.

As a short term perspective, we want to propose a method of evolution and enrichment of our OTR to improve the quality of the annotation of web data tables. This method should be able to take into account different types of changes: changes explicitly required by ontologists, changes due to an alignment with external ontologies, changes required after analyzing of the OTR to fulfil ontology quality assurance criteria and changes required after manual validation of new annotations. Another exciting perspective will be to extend our model to

[17] see [9] for more detail

be able to annotate n-ary relations not only in data tables extracted from Web documents but also using the information available in the plain text of those documents.

References

1. Gruber, T.: Ontology. In Liu, L., Özsu, M.T., eds.: Encyclopedia of Database Systems. Springer US (2009) 1963–1965
2. Cimiano, P., Buitelaar, P., McCrae, J., Sintek, M.: Lexinfo: A declarative model for the lexicon-ontology interface. J. Web Sem. **9**(1) (2011) 29–51
3. McCrae, J., Spohr, D., Cimiano, P.: Linking lexical resources and ontologies on the semantic web with lemon. In Antoniou, G., Grobelnik, M., Simperl, E.P.B., Parsia, B., Plexousakis, D., Leenheer, P.D., Pan, J.Z., eds.: ESWC (1). Volume 6643 of Lecture Notes in Computer Science., Springer (2011) 245–259
4. Declerck, T., Lendvai, P.: Towards a standardized linguistic annotation of the textual content of labels in knowledge representation systems. In: LREC, European Language Resources Association (2010)
5. Reymonet, A., Thomas, J., Aussenac-Gilles, N.: Modelling ontological and terminological resources in OWL DL. In: OntoLex 2007, ISWC Workshop. (2007)
6. Roche, C., Calberg-Challot, M., Damas, L., Rouard, P.: Ontoterminology - a new paradigm for terminology. In Dietz, J.L.G., ed.: KEOD, INSTICC Press (2009) 321–326
7. Reymonet, A., Thomas, J., Aussenac-Gilles, N.: Ontology based information retrieval: an application to automotive diagnosis. In: International Workshop on Principles of Diagnosis (DX 2009). (2009) 9–14
8. Hignette, G., Buche, P., Dibie-Barthélemy, J., Haemmerlé, O.: An ontology-driven annotation of data tables. In: WISE Workshops. Web Data Integration and Management for Life Sciences. Volume 4832 of LNCS. (2007) 29–40
9. Hignette, G., Buche, P., Dibie-Barthélemy, J., Haemmerlé, O.: Fuzzy annotation of web data tables driven by a domain ontology. In: ESWC. Volume 5554 of Lecture Notes in Computer Science. (2009) 638–653
10. Lynn, S., Embley, D.W.: Semantically conceptualizing and annotating tables. In: Proceedings of the 3rd Asian Semantic Web Conference on The Semantic Web. ASWC '08, Berlin, Heidelberg, Springer-Verlag (2008) 345–359
11. van Assem, M., Rijgersberg, H., Wigham, M., Top, J.: Converting and annotating quantitative data tables. In: Proceedings of the 9th international semantic web conference on The semantic web - Volume Part I. ISWC'10, Berlin, Heidelberg, Springer-Verlag (2010) 16–31
12. Buche, P., Haemmerlé, O.: Towards a unified querying system of both structured and semi-structured imprecise data using fuzzy view. In: ICCS 2000. Volume 1867 of Lecture Notes in Computer Science. (2000) 207–220
13. Buche, P., Dibie-Barthélemy, J., Chebil, H.: Flexible sparql querying of web data tables driven by an ontology. In: FQAS. Volume 5822 of Lecture Notes in Computer Science. (2009) 345–357
14. Noy, N., Rector, A., Hayes, P., Welty, C.: Defining n-ary relations on the semantic web. W3C working group note http://www.w3.org/TR/swbp-n-aryRelations.
15. van Rijsbergen, C.J.: Information Retrieval. Butterworth (1979)

Acknowledgments

Financial support from the French National Research Agency (ANR) for the project Map'OPT is gratefully acknowledged.