Volume 4, Issue 3, March 2014 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

A Review Paper on ONDINE+

Vandana Sonawane, Prof. Srinu Dharavath

Computer Engineering

University Of Pune, Pune, India

Abstract— In this review paper of ONDINE system which allows the loading and the querying of a data warehouse opened on the Web, guided by an Ontological and Terminological Resource (OTR). The data warehouse, composed of data tables extracted from Web documents, has been built to supplement already existing local data sources. First, we present the main steps of our semiautomatic method to annotate data tables driven by an OTR. In this method is an XML/RDF data warehouse composed of XML documents representing data tables with their fuzzy RDF annotations, Flexible querying system which allows the local data sources and the data warehouse to be simultaneously and uniformly queried, using the OTR. Some system relies on SPARQL and allows approximate answers to be retrieved by comparing preferences expressed as fuzzy sets with fuzzy RDF annotations.

Keywords— Knowledge and data engineering tools and techniques, XML/XSL/RDF, "fuzzy," and probabilistic reasoning, representations, data structures, and transforms, knowledge modelling

I. INTRODUCTION

A huge amount of technical and scientific documents, available on the Web or the hidden Web (digital libraries,), include data tables. Those data tables can be seen as small relational databases even if they lack the explicit metadata associated with a database. They represent a very interesting potential external source for loading the data warehouse of a company dedicated to a given domain of application. They can be used to enrich local data sources. 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 as the one used to index the local data. We have designed a software called Ontology-based Data INtEgration (ONDINE), using the semantic Web framework and language recommendations (XML, RDF, OWL, SPARQL), to supplement existing local data sources with data tables which have been extracted from Web documents.

The ONDINE system relies on an Ontological and Terminological Resource (OTR) which is composed of two parts: on the one hand, a generic set of concepts dedicated to the data integration task and, on the other hand, a specific set of concepts and a terminology, dedicated to a given domain of application. ONDINE system is composed of two subsystems: 1) Web subsystem designed to load an XML/ RDF data warehouse with data tables which have been extracted from Web documents and semantically annotated using concepts from the OTR; 2) MIEL++ subsystem designed to query simultaneously and uniformly the local data sources and the XML/RDF data warehouse using the OTR in order to retrieve approximate answers in a homogeneous way. In the first step, relevant documents for the application domain described in the OTR are retrieved from the Web and filtered by a human expert. In the second step, data tables are semi automatically extracted from the documents. In third step, the extracted data tables are semantically annotated using the OTR. This step generates fuzzy annotations, represented in a fuzzy extension of RDF, which are associated with data tables represented in XML. In the fourth and last step, the end user has to validate the fuzzy RDF semantic annotations associated with data tables before loading them in the XML/RDF data warehouse. Let us notice that @Web subsystem does not pretend to annotate all data tables extracted from any Web documents, but to annotate accurately target data tables extracted from documents identified as relevant for a given domain. The human intervention at each of its step is therefore required to guarantee the accuracy of the approach. In this paper, we focus on the third step that is the semantic annotation method, of @Web subsystem. Its main originality is to produce fuzzy RDF annotations which allow: 1) the recognition and the representation of imprecise numerical data appearing in the cells of a data table; 2) the computation and explicit representation of the semantic distance between terms in the cells of a data table and terms of the OTR. MIEL++ subsystem allows the fuzzy RDF annotations to be queried using SPARQL2 which is recommended by W3C to query RDF data sources. This subsystem is an extension of the MIEL flexible querying system proposed in [1] and [2]. The main originalities of our new flexible querying subsystem are: 1) to retrieve not only exact answers compared with the selection criteria but also semantically close answers; 2) to compare the selection criteria expressed as fuzzy sets representing preferences with the fuzzy annotations of data tables. Some preliminary studies of this work have already been published in [3], [4], and [5]. This paper provides a synthetic overview of ONDINE system which relies on a new modeling of the OTR dedicated to the data integration task. The definition of this OTR, central in ONDINE system, was essential to consolidate the approach and ensure its sustainability and its future evolutions. @Web subsystem (previously presented in [3] and [4]) and MIEL++ subsystem (previously presented in [5]) have been revised to take into account this new OTR.

II. REVIEW OF THE ONTOLOGICAL AND TERMINOLOGICAL RESOURCE

In [6], [7], [8], [9], [10] ontologies are associated with terminological and/or linguistic objects. In [6], Cimiano et al. 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 OWL3 ontology. Adapting LexInfo, [7] presents a model called Lexicon Model for Ontologies (lemon) 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 [8] is a model for the integration of conceptual, terminological and linguistic objects in ontologies. In [9] a meta model for ontological and terminological resources in OWL DL is presented, called an Ontological and Terminological Resource, extended afterward in [11] in order to be used for ontology-based information retrieval applied to automotive diagnosis.

The ontology we used in our previous works [3], [4], [5] was not designed to allow one to define the terminology and its variations (synonyms, multilingual, abbreviations,) denoting the concepts. We therefore propose to use an Ontological and Terminological Resource [9] allowing joint representation of an ontology and its associated terminology. According to [9], three factors influence the OTR structuring: the task to realize, the domain of interest and the application. The OTR used in ONDINE system has been designed for the data table integration (annotation and querying) task. In this paper, the domain of interest is food safety but the OTR structure we propose is generic enough to be applied to many other domains. For example, in this paper, experimental results in aeronautics will be also presented. The application is the construction of a data warehouse opened on the Web.

Since ONDINE system allows local data sources to be supplemented with data tables which have been extracted from Web documents, the domain specific part of the OTR was manually built by ontologists taking into account 1) the vocabulary used in the preexisting local databases in order to index the data and 2) the domain information available within the databases schema. Examples given in this paper concern the microbial risk domain. We present first, the conceptual component of the OTR and second, its terminological component, using the OWL2-DL model.

A) The Conceptual Component of the OTR:

The conceptual component is the ontology of the OTR. It is composed of two main parts: a generic part, commonly called core ontology, which contains the structuring concepts of the data table semantic annotation task, and a specific part, commonly called domain ontology, which contains the concepts specific to the domain of interest.

A data table is composed of columns, themselves composed of cells. In order to understand the structure of the core ontology, let us detail the data table semantic annotation task. A data table must be structured in a standardized way; otherwise preliminary transformations are applied on it using state-of-the-art tools like spreadsheets. The cells of a data table may contain terms or numerical values often followed by a measure unit. During the semantic annotation of a data table, cells content are semantically annotated in order to identify the symbolic concepts or quantities represented by its columns and finally the semantic n-ary relationships linking its columns. For instance, in Fig. 1, the cell content "E.coli" is associated with the symbolic concept Escherichia coli by our annotation method, the content of the three cells 4.9, 41.1, and 45.8 are associated with the quantity Temperature and the entire content of the second row of the data table is considered as an instance of the n-ary relation Growth Parameter Temperature which associates a given microorganism (like Escherichia coli) with its temperature growing conditions in a food product.

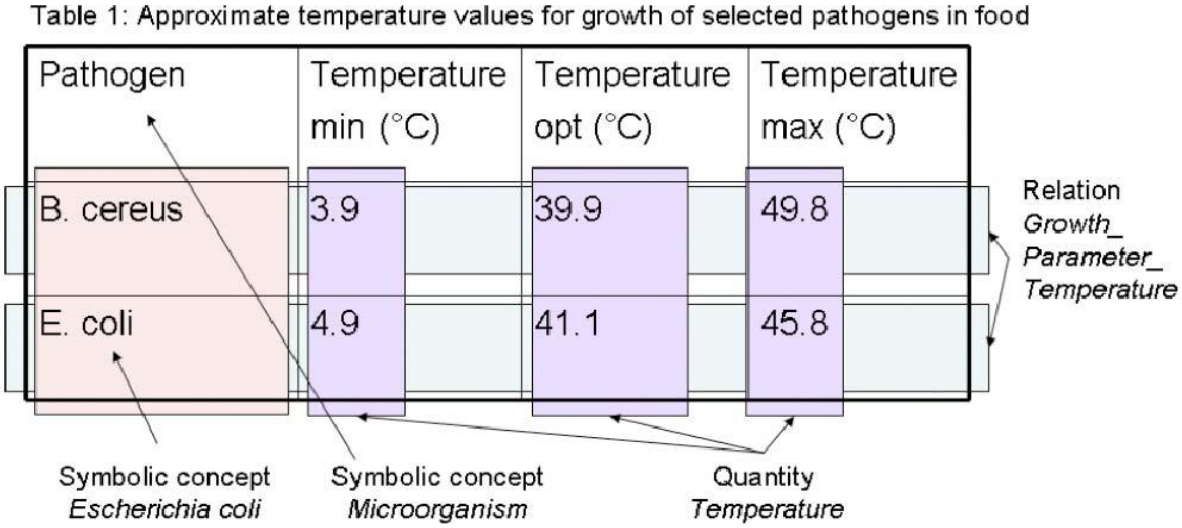

Fig.1 Annotation of a table according to concepts defined in the OTR

The core ontology is therefore composed of three kinds of generic concepts: 1) simple concepts which contain the symbolic concepts and the quantities, 2) unit concepts which contain the units used to characterize the quantities, and 3) relations which allow n-ary relationships to be represented between simple concepts. The concepts belonging to the domain ontology, called specific concepts, appear in the OTR as sub concepts of the generic concepts. Fig. 2 presents an excerpt of the conceptual component of the OTR in microbial risk domain. In OWL, all concepts are represented by

classes which are pair wise disjoints and are hierarchically organized by the sub Class Of relationship. The nodes represent the OWL classes, the solid arrows the "is-a" relationship between classes and the dashed arrows properties between classes. For instance, the property has Unit links a quantity (e.g., a Temperature) with its units of measurement (e.g., Celsius_ Degree and Fahrenheit Degree). We detail below the three kinds of generic concepts and their sub specific concepts in microbial risk domain.

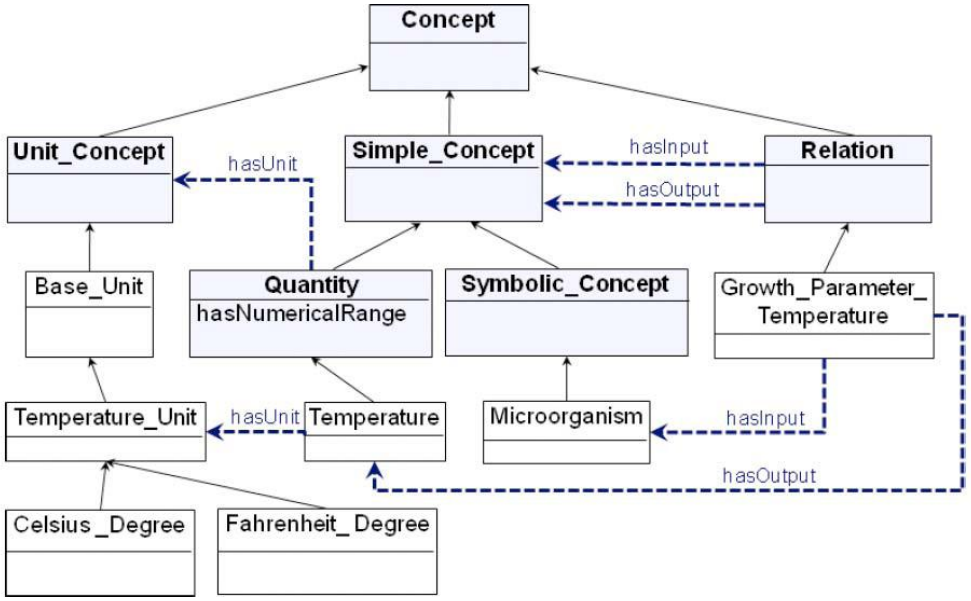

Fig.2**.** An excerpt of the conceptual component of the OTR in microbial risk domain

i) Unit Concepts: Unit concepts allow the meaning of units to be represented. Our classification relies on the international system of units which decomposes the units into base units and derived units. There exist several ontologies dedicated to quantities and associated units (OM,7 QUDT,8 QUOMOS, OBOE,9 . . .). We learn from these ontologies to build ours, but they cannot contain all the required specific units for a given domain. For instance, in microbial risk domain, the ontologist has added some units such as ppm10 or CFU=g.11 Fig. 3 presents an excerpt of the unit concepts in microbial risk domain.

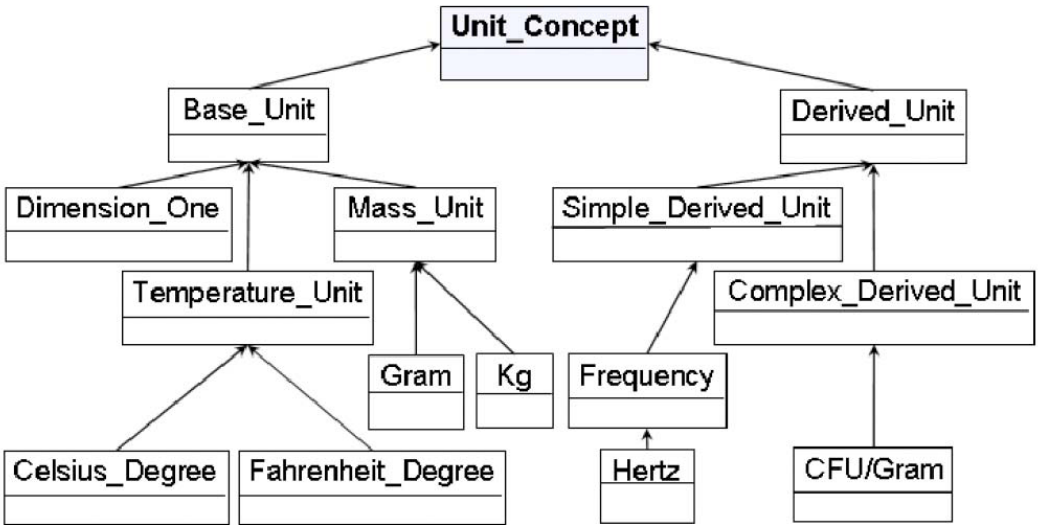

Fig. 3 An excerpt of the unit concepts in microbial risk domain

ii)Simple Concept: Symbolic concepts allow the meaning of terms to be represented. Symbolic concepts are hierarchically organized by the "is-a" relationship. Fig. 4 presents an excerpt of the specific symbolic concepts in microbial risk domain. The microbial risk domain OTR contains three distinct sub hierarchies of specific symbolic concepts: the specific symbolic concept Food_Product with more than 400 subconcepts, the specific symbolic concept Microorganism with more than 150 subconcepts and the specific symbolic concept Response with three subconcepts: growth, absence of growth, and death, which represent the possible responses of a microorganism to a treatment. These subhierarchies have been defined by ontologists. We could not reuse pre existing terminologies for food products such as AGROVOC12 (from FAO—Food and Agriculture Organisation of the United Nations) or Gems-Food13 (from WHO—World Health Organisation) because those terminologies are not specific enough compared with the one built from our corpus in microbial risk (only 20 and 34 percent of common words, respectively).

A quantity is described by a set of units, which are sub concepts of the unit concept, and eventually a numerical range. Quantities allow the meaning of numerical values to be represented. Two properties has Unit and has Numerical Range, belonging to the core ontology, link, respectively, quantities to their associated units and numerical range. The OWL object property has Unit allows a quantity to be described by one or several unit concepts. OWL2-DL data type restrictions using facet spaces allow the numerical range of a quantity to be represented in the OWL data type property has Numerical Range. Fig. 5 presents an excerpt of the quantities in microbial risk domain. Eighteen specific quantities have been defined for the microbial risk domain. The specific quantity Temperature can be expressed using the unit _C (represented by the concept Celsius_Degree) or _F(represented by the concept Fahrenheit_Degree) and has no numerical range.

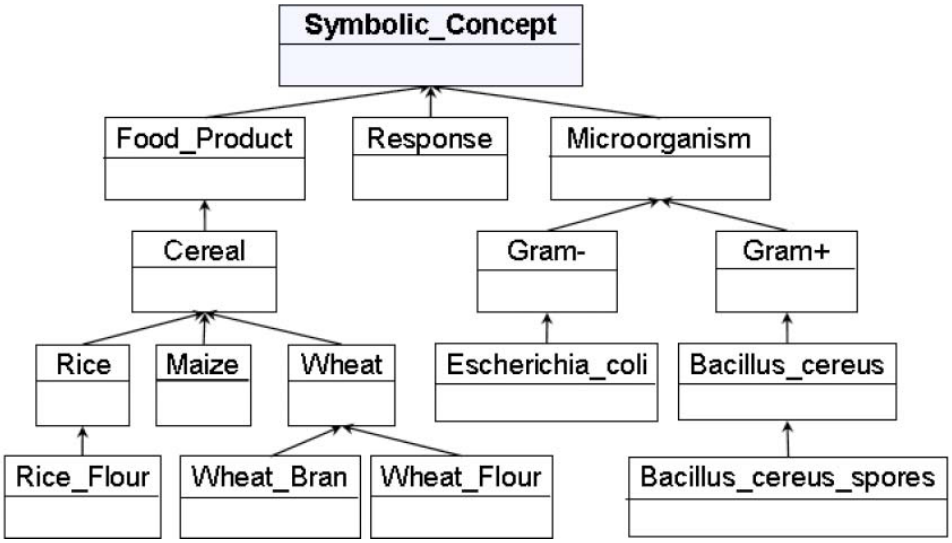

Fig.4 An excerpt of the symbolic concepts in microbial risk domain.

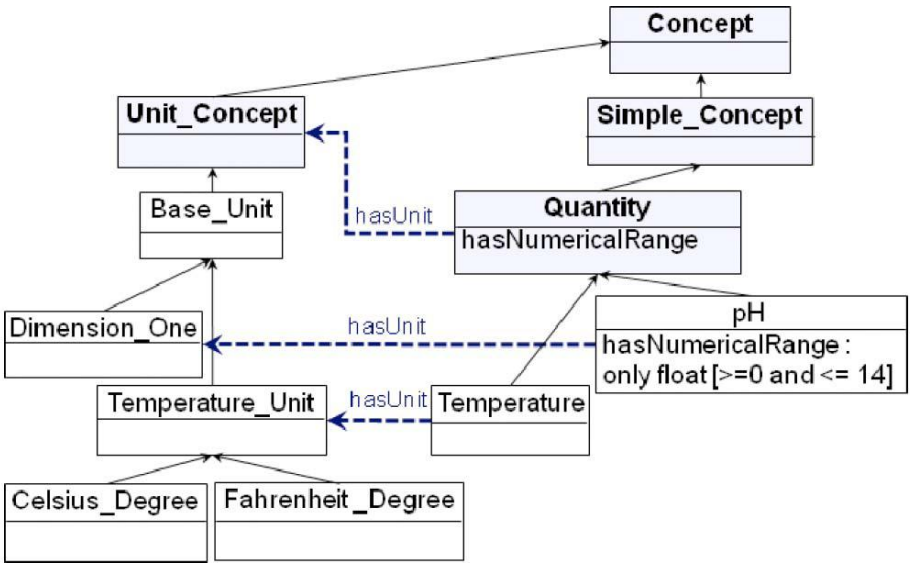

Fig.5 An excerpt of the quantities associated with unit concepts in microbial risk domain

iii) Relations: Relations allow the meaning of n-ary relationships between simple concepts to be represented. A relation is defined by its signature, which is composed of several input simple concepts and one output simple concept. The input simple concepts represent the domain of the relation. A relation may have several input simple concepts. The output simple concepts represent the range of the relation. The restriction of the range to only one output simple 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 which may have several entry factors. If a data table contains several result columns, it is then represented by as many relations as it has results. Two properties, belonging to the core ontology, called hasInput and hasOutput, link a relation to its domain and range. Since a relation represents in the OTR a n-ary relationship, we learned from [12] which suggests to decompose a n-ary relationship into n binary relationships. Consequently, a n-ary relation is represented in OWL by a class associated with the simple concepts of its signature via the OWL object property hasInput or the OWL functional object property hasOutput. In Fig. 2, for instance, the specific relation Growth_Paramater_Temperature has for input the specific symbolic concept Microorganism and for output the specific quantity Temperature. The microbial risk domain OTR contains 16 relations.

B). The Terminological Component of the OTR:

The terminological component represents the terminology of the OTR: it contains the terms set of the domain of interest. A term is defined as a sequence of words, in a language, and has a label. Terms are divided according to their source language. A term denotes a concept; it must denote at least one concept and it can denote several concepts. The OWL object property denotes, belonging to the core ontology, allows a term to denote a concept. The OWL functional data properties Label and Language, belonging to the core ontology, allow a term to be associated with its label and its language, which are represented as a string. Fig. 6 presents an excerpt of the OTR in microbial risk domain. The specific terms T_Produit_Alimentaire and T_Food_Product both denote the specific symbolic concept Food_Product.

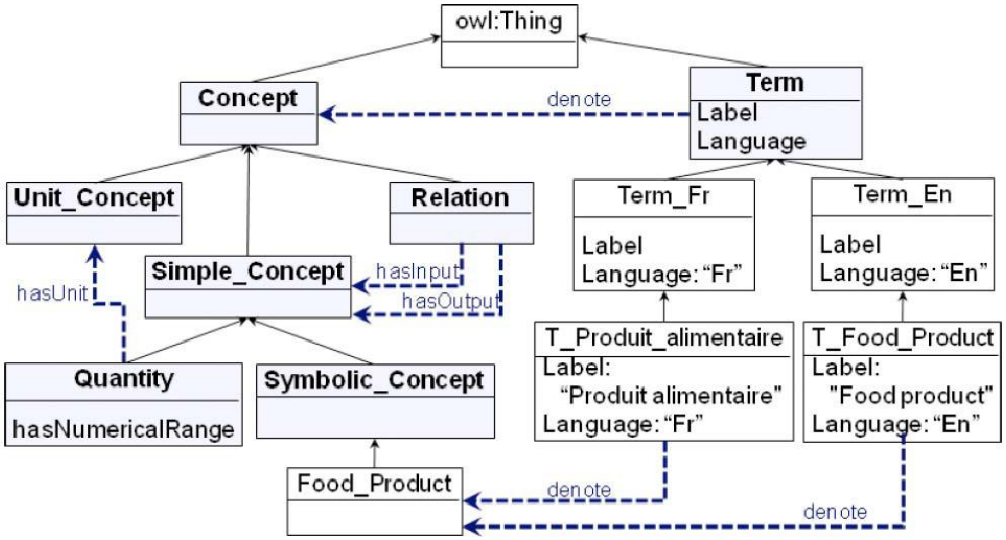

Fig.6 An excerpt of the OTR in microbial risk domain.

The OTR presented above is at the heart of the ONDINE system which allows local data sources to be supplemented with annotated Web data tables. We present in the next two sections the semantic annotation method of @Web subsystem which allow data tables, extracted from Web documents, to be annotated thanks to the OTR, before being added to the XML/RDF data warehouse. The semantic annotation of a data table is composed of two steps: 1) identifying which relations defined in the OTR are represented in the data table, 2) instantiating the identified relations, which consists in associating a set of fuzzy RDF annotation graphs with each row of the data table.

III. COMPARATIVE STUDY OF DIFFERENT METHODS

Paper	Technique	Drawback	Advantage
P. Buche and O. Haemmerle,"Towards a Unified Querying System of Both Structured and Semi-Structured Imprecise Data Using Fuzzy Views,"	UQS Querying Language, CFQ Engine,	1). A drawback to this simple mechanism is its important lack of semantics.	1).The CFQ engine enlarges queries when they are too restrictive because of the actual content of the database.
P. Cimiano, P. Buitelaar, J. McCrae, and M. Sintek, "Lexinfo: A Declarative Model for the Lexicon-Ontology Interface,"	Separation Between Lexical and Ontological Layer , More Flexible Coupling Between Ontological and the Linguistic System	1). disadvantage is that it is undecidable and thus not relevant for practical implementations.	The advantage of modeling The di_erent subcategorization frames as subclasses (as we have done) is that this allows us to formulate additional axioms,
A. Reymonet, J. Thomas, and N. Aussenac-Gilles, "Modelling Ontological and Terminological Resources in OWL DL,"	Complementarity of Ontology and Terminology, OWL Standard	1).A drawback of this approach lies in the scalability of the meta-model which creates up to twice as many instances as the number of found term occurrences.	Due to the term reification, the proposed OTR model has the advantage to enable a direct manipulation of terms, totally free from the way concepts are

			represented.
C. Roche, M. Calberg-Challot, L. Damas, and P. Rouard, "Ontoterminology - A New Paradigm for Terminology,"	Ontoterminology	The Web Ontology Language [OWL], combines the disadvantages of these two approaches.	if term definitions written in natural language are useful, they are not always consensual unlike domain conceptualisation.
A. Reymonet, J. Thomas, and N. Aussenac-Gilles, "Ontology Based Information Retrieval: An Application to Automotive Diagnosis,"	Ontological and Terminological Resource (OTR) , Complementarity of Ontology and Terminology	A drawback of this approach lies in the scalability of the meta-model which creates up to twice as many instances as the number of found term occurrences.	Due to the term reification, the proposed OTR model has the advantage to enable a direct manipulation of terms, totally free from the way concepts are represented.

IV. CONCLUSIONS

We have presented in this paper a complete system, called ONDINE, built, using the recommendations of the W3C, on a generic OTR expressed in OWL.ONDINE system allows XML data tables, which have been extracted from Web documents, to be annotated with fuzzy RDF descriptions and to be flexibly queried using SPARQL. Fuzzy RDF annotations are used to represent (1) imprecise values associated with a quantity expressed in one or several numerical columns, (2) the set of most similar symbolic concepts of the O TR which are automatically associated with the content of a cell belonging to a symbolic column, (3) a degree of certainty associated with each n-ary relation recognized in a data table. ONDINE system has been implemented through the development of @Web software on the one hand and the development of MIEL++ software on the other hand. We then present our flexible querying system which allows the local data sources and the data warehouse to be simultaneously and uniformly queried, using the OTR. This system relies on SPARQL and allows approximate answers to be retrieved by comparing preferences expressed as fuzzy sets with fuzzy RDF annotations.

REFERENCES

[1] P. Buche and O. Haemmerle´, "Towards a Unified Querying System of Both Structured and Semi-Structured Imprecise Data Using Fuzzy Views," Proc. Linguistic on Conceptual Structures: Logical Linguistic, and Computational Issues (ICCS), pp. 207-220, 2000.

[2] P. Buche, C. Dervin, O. Haemmerle´, and R. Thomopoulos, "Fuzzy Querying of Incomplete, Imprecise, and Heterogeneously Structured Data in the Relational Model Using Ontologies and Rules," IEEE Trans. Fuzzy Systems, vol. 13, no. 3, pp. 373-383, June 2005.

[3] G. Hignette, P. Buche, J. Dibie-Barthe´lemy, and O. Haemmerle´, "An Ontology-Driven Annotation of Data Tables," Proc. WISE Workshops Web Data Integration and Management for Life Sciences, pp. 29-40, 2007.

[4] G. Hignette, P. Buche, J. Dibie-Barthe´lemy, and O. Haemmerle´, "Fuzzy Annotation of Web Data Tables Driven by a Domain Ontology," Proc. Sixth European Semantic Web Conf. The Semantic Web: Research and Applications (ESWC), pp. 638-653, 2009.

[5] P. Buche, J. Dibie-Barthe´lemy, and H. Chebil, "Flexible Sparql Querying of Web Data Tables Driven by an Ontology," Proc. Eight Int'l Conf. Flexible Query Answering Systems (FQAS), pp. 345-357, 2009.

[6] P. Cimiano, P. Buitelaar, J. McCrae, and M. Sintek, "Lexinfo: A Declarative Model for the Lexicon-Ontology Interface," J. Web Semantics, vol. 9, no. 1, pp. 29-51, 2011.

[7] J. McCrae, D. Spohr, and P. Cimiano, "Linking Lexical Resources and Ontologies on the Semantic Web with Lemon," Proc. Eight Extended Semantic Web Conf. The Semantic Web: Research and Applications (ESWC), pp. 245-259, 2011.

[8] T. Declerck and P. Lendvai, "Towards a Standardized Linguistic Annotation of the Textual Content of Labels in Knowledge Representation Systems," Proc. Seventh Int'l Conf. Language Resources and Evaluation (LREC '10), 2010.

[9] A. Reymonet, J. Thomas, and N. Aussenac-Gilles, "Modelling Ontological and Terminological Resources in OWL DL," Proc. OntoLex 2007 - Workshop associated with ISWC '07, Sixth Int'l Semantic Web Conf. (ISWC '07), 2007.

[10] C. Roche, M. Calberg-Challot, L. Damas, and P. Rouard, "Ontoterminology - A New Paradigm for Terminology," Proc. Int'l Conf. Knowledge Eng. and Ontology Development (KEOD), pp. 321-326. 2009.

[11] A. Reymonet, J. Thomas, and N. Aussenac-Gilles, "Ontology Based Information Retrieval: An Application to Automotive Diagnosis," Proc. Int'l Workshop Principles of Diagnosis, pp. 9-14, 2009

[12] N. Noy, A. Rector, P. Hayes, and C. Welty, "Defining Nary Relations on the Semantic Web W3C Working Group Note," http://www.w3.org/TR/swbp-n-aryRelations, 2012.

