Ontology-based integration of heterogeneous, incomplete and imprecise data dedicated to a decision support system in food safety

Patrice Buche, Sandrine Contenot, Lydie Soler, Juliette Dibie-Barthelemy, David Doussot, Gaëlle Hignette, Liliana Ibanescu

Ontology-based integration of heterogeneous, incomplete and imprecise data dedicated to a decision support system for food safety

P. Buche, S. Contenot, and L. Soler
INRA,
Department of Applied Mathematics and Computer Science, Mét@risk Unit
E-mail: {patrice.buche, sandrine.contenot, lydie.soler}@ paris.inra.fr

J. Dibie-Barthélemy, D. Doussot, L. Ibanescu and G. Hignette
AgroParisTech,
Department Modélisation Mathématiques, Informatique et Physique
E-mail: {juliette.dibie, david.doussot, liliana.ibanescu, gaelle.hignette}@agroparistech.fr

16 rue Claude Bernard,
75231 PARIS Cedex 05, France.

ABSTRACT

This chapter presents an application in the field of food safety using an ontology-based data integration approach. An ontology is a vocabulary used to express the knowledge in a given domain of application. In this chapter, the ontology-based data integration approach permits to homogenise data sources which are heterogeneous in terms of structure and vocabulary. This approach is done in the framework of the Semantic Web, an international initiative which proposes annotating data sources using ontologies in order to manage them more efficiently. In this chapter, we explore three ways to integrate data according to a domain ontology: (1) a semantic annotation process to extend local data with Web data which have been semantically annotated according to a domain ontology, (2) a flexible querying system to query uniformly both local data and Web data and (3) an ontology alignment process to find correspondences between data from two sources indexed by distinct ontologies.

KEY-WORDS:

Ontologies, Flexible Query, Data Warehouse, Web-Based Database, Semantic matching

INTRODUCTION

The aim of the data integration systems is to put together a large amount of data coming from multiple and independent sources. One of the main problems of the data integration is the data heterogeneity. It can come from the structure of the data, the vocabulary used to index the data and the format of the data. These characteristics are in general specific to each source of data. Their harmonization is necessary to integrate the data. Another problem of the data integration is the data rarity. Although this problem can seem paradoxical, it can be explained by the fact that the numerous available data are not necessarily pertinent for a given application domain (in food safety for instance). A third problem may also occur in data integration: the imprecision of data. This imprecision can be intrinsic to the data (for instance an interval of pH values) or can correspond to a pertinence degree associated with the data according to the application domain.

We have developed a system, called CARAT (Chronic & Acute Risk Assessment), to estimate the exposure of a given population of consumers to chemical contaminants which relies on two distinct data sources. Both sources contain information about food products. The first one, called CONTA source, contains measures of the level of chemical contamination for food products. The second one, called CONSO source, stores household purchases of food products. Both sources have been indexed using their own domain ontology, the CONTA ontology and the CONSO ontology, an ontology representing a vocabulary used to express the knowledge in a given application domain. The CARAT system is composed of two sub-systems (see Figure 1): a decision support system that uses statistical methods to compute the exposure of a given population of consumers to chemical contaminants (Buche P, Soler L & Tressou J, 2006) and an ontology-based data integration system which feeds the decision support system with data about the chemical contamination and the consumption of food products. The data integration system is managed using a data warehouse approach: data sources provided by external partners are replicated locally and standardized using ETL technology.

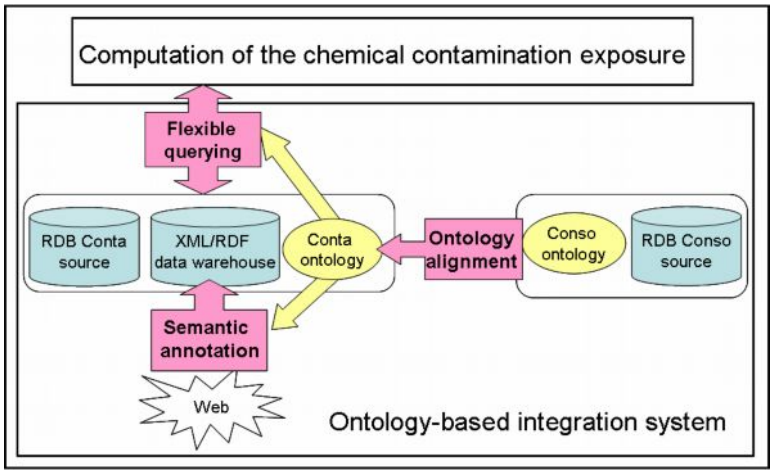


Figure 1 The CARAT system

In this chapter, we present the ontology-based data integration system which takes into account the three data integration problems presented above: data heterogeneity, data rarity and data imprecision. The ontology-based data integration system proposes three different ways to integrate data according to a domain ontology. The first one is a semantic annotation process

which allows a local database (the CONTA local database), indexed by a domain ontology, to be extended with data that have been extracted from the Web and semantically annotated according to this domain ontology. The second one, which is an original contribution of this chapter, is a querying system which allows the semantically annotated Web data to be integrated with the local data through a uniform flexible querying system relying on a domain ontology (the CONTA ontology). The third one is a ontology alignment method relying on rules which allow correspondences to be found between objects of a source ontology (the CONSO ontology) and objects of a target ontology (the CONTA ontology) according to their characteristics and associated values. Those three ways to integrate data have been designed using the Semantic web approach, an international initiative, which proposes annotating data sources using ontologies in order to manage them more efficiently.

In this chapter, we first present the ontology-based data integration system. We then provide some background on the topic. Third, current projects and future trends are presented. We conclude this chapter in the last section.

THE ONTOLOGY-BASED DATA INTEGRATION SYSTEM

This section describes the different construction steps of the ontology-based data integration system of the CARAT system. In the first section, we present the filling of its data sources. In the second section, we present its querying system. In the third section, we present the alignment between objects of its two data sources which are indexed by distinct ontologies.

Filling the data warehouse

There are two types of data available in the CARAT system: contamination data and consumption data. Both types of data concern food products but their content and their treatment are not the same. The contamination data are measures of level of chemical contamination for food products whereas the consumption data are about household purchases of food products during a year.

The contamination data are stored in a relational database, called CONTA local database, which has been defined and filled by our research team from different sources. It is indexed by the CONTA ontology. The consumption data are stored in a relational database, called CONSO database, which is filled from the TNS WORLD PANEL source, a private source of household purchases. It is indexed by the CONSO ontology. Both databases are filled using ETL (Extract, Transform, Load) technology.

In this section, we make a focus on two original characteristics of the contamination data which must be taken into account during their storage: their imprecision and their rarity. On the one hand, we propose to use the fuzzy set theory in order to represent imprecise data. On the other hand, we propose to search and annotate data from the Web using the CONTA ontology in order to extend the CONTA local database. We first present the structure of the CONTA ontology. We then present the fuzzy set theory used to treat the imprecise data. Finally, we detail our semantic annotation process which allows the CONTA local database to be enriched with Web data.

The structure of the CONTA ontology

The CONTA ontology is composed of datatypes -numeric types and symbolic types- and of relations that allow one to link datatypes.

Numeric types are used to define the numeric data. A numeric type is described by the name of the type, the units in which data of this type is usually expressed, and the interval of possible values for this type. For example, the type *Contamination Level* can be expressed in the units µg/g, µg/kg, ng/g, µg/l and has a range of [0, 1000].

Symbolic types are used when the data of interest are represented as a string. A symbolic type is described by the name of the type and the type hierarchy (which is the set of possible values for the type, partially ordered by the subsumption relation). For example, *FoodProduct* and *Contaminant* are symbolic types.

Relations are used to represent semantic links between datatypes. A relation is described by the name of the relation and its signature. For example, the relation *ContaminationRange* represents the average level of contamination of a food product by a contaminant for a given number of samples. This relation has for domain the symbolic types *FoodProduct*, *Contaminant* and *Samples Total Number* and for range the numeric type *Contamination Level*. Figure 2 shows the structure of an excerpt of the CONTA ontology.

This ontology has been expressed in OWL distinguishing two types of knowledge: (i) generic knowledge which define the structure of the ontology: for instance, the class *numericalType* (resp. the class *Relation*) which is the superclass of all numerical types (resp. relations); (ii) domain-dependant knowledge: for instance, the class *ContaminationRange* is a subclass of the class *Relation* and the class *ContaminationLevel* is a subclass of the class *numericalType*.

Example 1
Figure 2 gives an excerpt of the CONTA ontology.

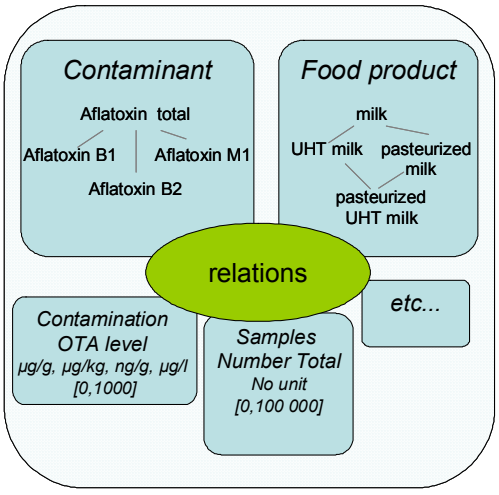


Figure 2 A simplified excerpt of the CONTA ontology

The fuzzy set theory

We propose to use the fuzzy set theory to represent imprecise data. In this chapter, we use the representation of fuzzy sets proposed in (Zadeh, 1965).

Definition A fuzzy set f on a definition domain D(f) is defined by a membership function μ from D(f) to [0,1] that associates the degree to which x belongs to f with each element x of D(f). We call support of f the subset of D(f) such that support(f)= $\{a \in D(f) \mid \mu_f(a) > 0\}$. We call kernel of f the subset of D(f) such that kernel(f) = $\{a \in D(f) \mid \mu_f(a) = 1\}$.

We distinguish two kinds of fuzzy sets: (i) discrete fuzzy sets and (ii) continuous fuzzy sets.

Definition A discrete fuzzy set f is a fuzzy set associated with a symbolic type of the ontology. Its definition domain is the type hierarchy.

Definition A continuous fuzzy set f is a trapezoidal fuzzy set associated with a numeric type of the ontology. A trapezoidal fuzzy set is defined by its four characteristic points which correspond to min(support(f)), min(kernel(f)), max(kernel(f)) and max(support(f)). Its definition domain is the interval of possible values of the type.

The fuzzy set formalism can be used in three different ways as defined in (Dubois, D., & Prade, H., 1988): (i) in the database, in order to represent imprecise data as an ordered disjunction of exclusive possible values, (ii) in the database as a result of the annotation process, in order to represent the similarity between a value from the web and values from the ontology or (iii) in the queries, in order to represent fuzzy selection criteria which express the preferences of the end-user.

Example 2

The fuzzy set *ContaminationLevel_FS* of Figure 3 is a continuous fuzzy set denoted [4,5,6,7]. It represents the possible values of a level of contamination. The fuzzy set *FoodProduct_Similarity* is a discrete fuzzy set denoted (0.66/rice + 0.5/rice flour). It represents the set of terms of the ontology that are similar with different degrees to the term Basmati rice found in a document retrieved from the Web. The fuzzy set *FoodProduct_Preferences* is a discrete one denoted (1/rice + 0.5/cereal). Used in a query, it means that the end-user is interested by information about rice but also with a lowest interest about cereal.

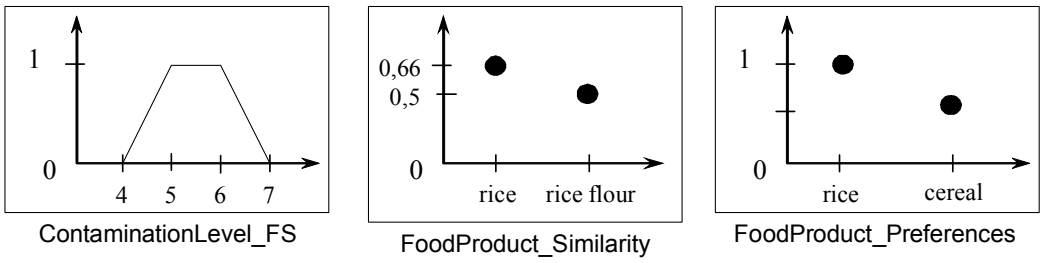


Figure 3 Examples of fuzzy sets

The semantic annotation process

In order to deal with the data rarity problem of the CONTA local database, we propose to extend the local database with data extracted from the Web. We have designed for that purpose a semi-automatic acquisition tool, called @WEB (Annotating Tables from the WEB). This tool relies on three steps as described in Figure 4. In the first step, relevant documents for the application

domain are retrieved using the domain ontology thanks to crawlers and RSS feeds. We focus on documents which contain data tables. This may be seen as a restriction of our approach. But, in a lot of application domains, especially in the scientific field, data tables are often a source of relevant, reliable and synthetic data. Moreover, their tabular structure is obviously easier to automatically parse than natural language. In the second step, the Web documents in html or most usually in pdf are translated into a generic XML format, which allows the representation of data tables in a classical and generic way -- a table is a set of lines, each line being a set of cells. In the third step, the tables are semantically annotated according to the domain ontology.

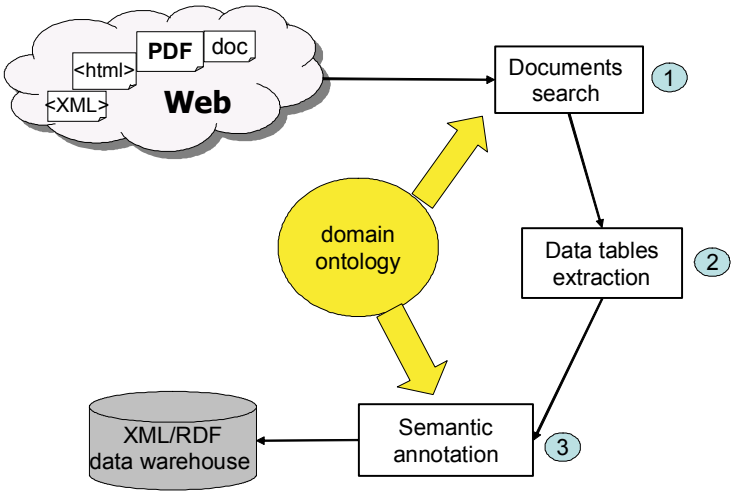


Figure 4 @WEB architecture

The semantic annotation process of a table extracted from the web consists in identifying which semantic relations from the domain ontology are represented in the table. The different steps of our semantic annotation process are detailed in Hignette & al. (2007).

The semantic annotation process generates RDF descriptions which represent the semantic relations of the ontology recognized in each row of the Web data table. Some of these RDF descriptions include values expressed as fuzzy sets. The fuzzy values used to annotate Web data tables may express similarity or imprecision. A fuzzy set having a semantic of similarity is associated with each cell belonging to a symbolic column. It represents the ordered list of the most similar values of the ontology associated with the value present in the cell. A fuzzy set having a semantic of imprecision may be associated with cells belonging to numerical columns. It represents an ordered disjunction of exclusive possible values.

Example 3
Table 1 presents an example of a Web data table in which the semantic relation *ContaminationRange* of domain the symbolic types *FoodProduct*, *Contaminant* and *Samples Total Number* and of range the numeric type *Contamination Level* has been identified.

Food	Contaminant	Contamination Level (ng/g)
Basmati rice	OTA	1.65-1.95
Chili powder	OTA	2.34-4.91
Grape raisins	OTA	0.93-1.20

Table 1 A Web data table

Figure 5 presents a part of the RDF descriptions corresponding to the recognition of the relation *ContaminationRange* in the first row of Table 1. The first row (having the URI *uriRow1* in the XML document) is an instance of the *ContaminationRange* relation recognized with a pertinence score of 0.75. This pertinence score is computed by the annotation process as the proportion of recognized types of the relation. It expresses the degree of "certainty" associated with the relation recognition. A part of the domain of the relation presented in the example (typed by the OWL class *AssociatedKey*) is an instance of the symbolic type *FoodProduct (food1)* and is annotated by a discrete fuzzy set (*DFS1*) which has a semantic of similarity. It indicates the list of closest values of the ontology (*Rice* and *Rice Flour*) compared to the value *Basmati Rice*. The range of the relation (typed by the OWL class *AssociatedResult*) is an instance of the numeric type *ContaminationLevel* and is annotated by a continuous fuzzy set (*CFS1*) which has a semantic of imprecision. It indicates the possible contamination limits ([1.65, 1.95]) represented as the support and the kernel of the fuzzy set.

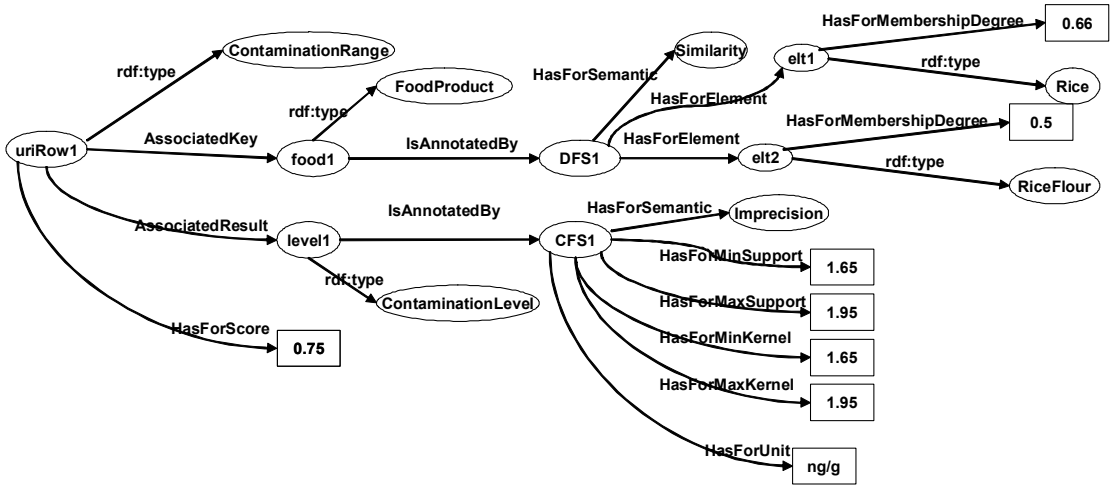


Figure 5 RDF annotations of the first row of the Web data table presented in Table 1

The output of the @WEB system is an XML/RDF data warehouse composed of a set of XML documents which represent data tables and contain their RDF annotations.

The flexible querying system

In order to deal with the data heterogeneity in the CONTA sources, we propose to the end-user a unified querying system, called MIEL++, which permits to query simultaneously the CONTA local relational database and the CONTA XML/RDF data warehouse in a transparent way. The MIEL++ system is a flexible querying system relying on a given domain ontology, the CONTA ontology. It allows the end-user to retrieve the nearest data stored in both sources corresponding

to his/her selection criteria: the CONTA ontology -more precisely the type hierarchies- is used in order to assess which data can be considered as "near" to the user's selection criteria.

Figure 6 gives an overview of the MIEL++ architecture. When a query is asked to the MIEL++ system, that query is asked through a single graphical user interface, which relies on the domain ontology. The query is translated by each subsystem's wrapper into a query expressed in the query language of the subsystem: an SQL query in the relational subsystem (see Buche & al. 2005 for more details about the SQL subsystem), a SPARQL query in the XML/RDF subsystem (SPARQL is a query language recommended by the W3C to query RDF data sources http://www.w3.org/TR/rdf-sparql-query/). Finally, the global answer to the query is the union of the local results of the two subsystems, which are ordered according to their relevance to the query selection criteria.

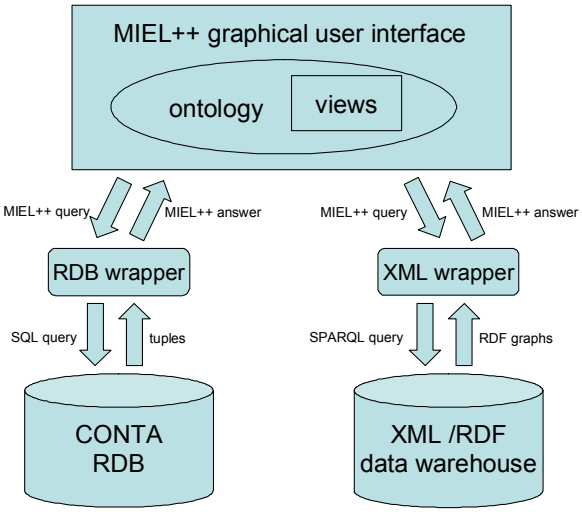


Figure 6 MIEL++ architecture

In this section, we are interested in the XML/RDF subsystem which allows the end-user to query RDF annotations of Web data tables represented in XML documents. A MIEL++ query is asked in a view which corresponds to a given relation of the ontology. A view is characterized by its set of queryable attributes and by its actual definition. Each queryable attribute corresponds to a type of the relation represented by the view. The concept of view must be understood with the meaning of the relational database model: it permits to hide the complexity of the querying in a given subsystem to the end-user. The end-user uses a view to build his query. In the XML/RDF subsystem, a view is defined by means of a SPARQL generic query where the SELECT clause contains the queryable attributes and the WHERE clause contains the definition of the view.

Example 4
Let us consider the view V associated with the relation ContaminationRange of domain FoodProduct, Contaminant, SamplesNumberTotal and of range ContaminationLevel. The SPARQL query associated with V is presented in Figure 7, the where part of the query being shown graphically for readability reasons.

V = {FoodProduct, Contaminant, SamplesNumberTotal, ContaminationLevel | ContaminationRange (FoodProduct, Contaminant, SamplesNumberTotal, ContaminationLevel)}

SELECT ?foodproduct ?contaminant ?samplesnumbertotal ?contaminationlevel
WHERE

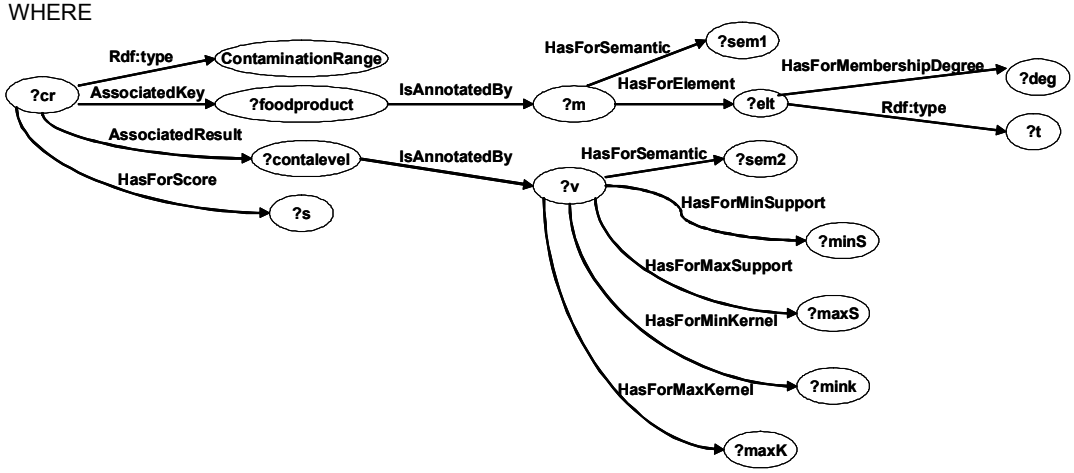


Figure 7 The SPARQL query associated with a view

A MIEL++ query is an instance of a given view specified by the end-user, by choosing, among the set of queryable attributes of the view, which are the conjunctive selection attributes and their corresponding searched values, which are the projection attributes and which is the minimal threshold on the pertinence score associated with the relation represented by the view. In a MIEL++ query, the end-user can express preferences in his/her selection criteria. These preferences are expressed by fuzzy sets as presented in Subsection "The fuzzy set theory". Since fuzzy sets are not supported in a standard SPARQL query, we propose to "defuzzify" the MIEL++ query before translating it into SPARQL. This defuzzification is performed in two steps.

When the fuzzy value of a selection criterion has a hierarchized symbolic definition domain, it is represented by a fuzzy set defined on a subset of its definition domain. Such a fuzzy set defines degrees implicitly on the whole definition domain of the selection attribute. In order to take those implicit degrees into account, we propose to perform a closure of the fuzzy set as defined in Thomopoulos & al. (2006). Intuitively, the closure propagates the degrees to more specific values of the hierarchy. Then, for each selection criterion represented by a fuzzy set, we can perform the defuzzification of the fuzzy set which consists in deleting the degrees in the case of a DFS and in only keeping the interval which corresponds to the support in the case of a CFS.

Example 5
Let us consider the following MIEL++ query Q where FoodProduct, Contaminant, ContaminationLevel are the projection attributes and where FoodProduct and ContaminationLevel are the selection attributes. Figure 8 presents (i) on the left, the closure and the defuzzification of the fuzzy value FoodProduct_Preferences={1.0/rice + 0.5/cereal} associated with the selection criterion FoodProduct according to the type hierarchy of the type FoodProduct of Figure 2 and (ii) on the right, the defuzzification of the fuzzy value Conta_Preferences={0.5, 0.75, 1.7, 1.8} associated with the selection criterion ContaminationLevel.

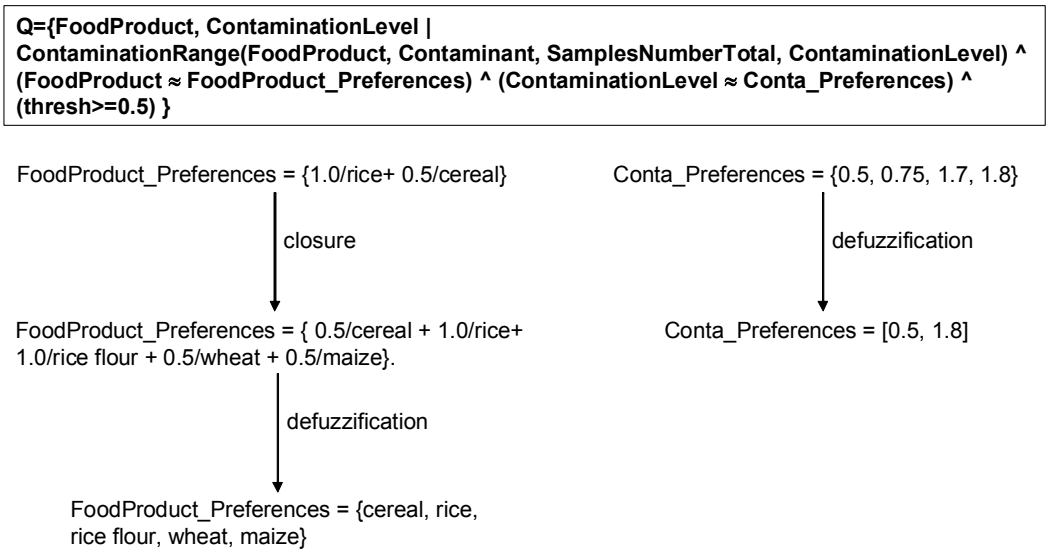


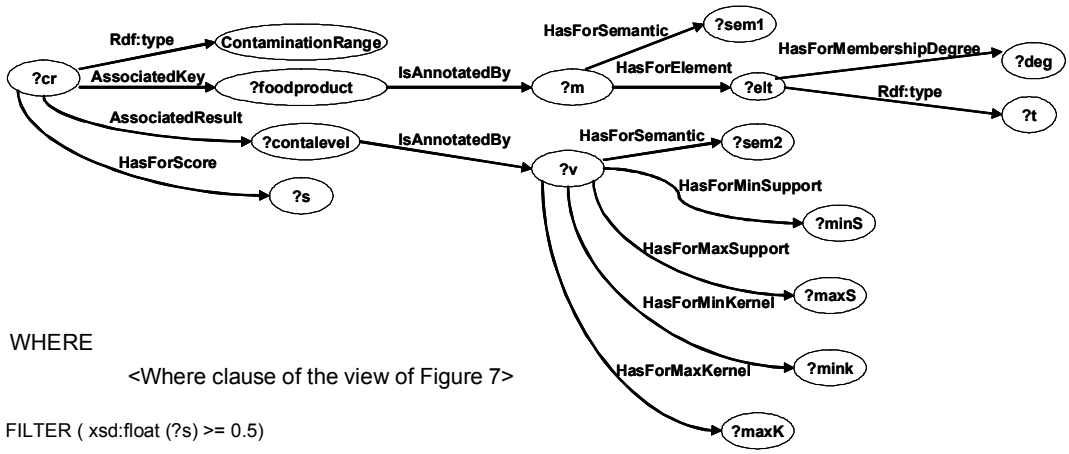
Figure 8 Defuzzification of a MIEL++ query

The defuzzified MIEL++ query can now be translated into a SPARQL query where the CONSTRUCT clause allows the answers of the query to be built according to the projection attributes of the MIEL++ query and the SELECT clause contains the selection criteria and the threshold of the MIEL++ query. All the selection criteria are represented into the FILTER clause of the SELECT clause.

Example 6

The defuzzified MIEL++ query of example 5 can be translated into the SPARQL query of Figure 9 in which, for readability reasons, we do not detail the where part of the query already given in Figure 7.

```
CONSTRUCT
```




```
WHERE
        <Where clause of the view of Figure 7>
FILTER ( xsd:float (?s) >= 0.5)
FILTER ( ?t = 'cereal' || ?t = 'rice' || ?t = 'rice flour' || ?t = 'wheat' || ?t = 'maize')
FILTER ( ( xsd:float (?minS) >= 0.5 && xsd:float (?minS) <= 1.8) || ( xsd:float (?maxS) >= 0.5 && xsd:float (?maxS) <= 1.8  ||
( xsd:float (?maxS) <= 1.8 && xsd:float (?minS) >= 0.5 ) )
```

Figure 9 The SPARQL query associated with the defuzzified MIEL++ query of example 5

An answer to a MIEL++ query must (1) satisfy the minimal acceptable pertinence score associated with the relation represented by the query; (2) satisfy all the selection criteria of the

query and (3) associate a constant value with each projection attribute of the query. The answer to a MIEL++ query in the XML/RDF subsystem is computed in two steps. First the corresponding SPARQL query is generated and executed into the XML/RDF data warehouse. Then, the answer to this SPARL query must be "refuzzified" in order to be able to measure how it satisfies the selection criteria.

To measure the satisfaction of a selection criterium, we have to consider the two semantics - imprecision and similarity- associated with fuzzy values of the XML/RDF data warehouse. On the one hand, two classical measures (Dubois & Prade, 1988) have been proposed to compare a fuzzy set representing preferences to a fuzzy set having a semantic of imprecision: a possibility degree of matching denoted Π and a necessity degree of matching denoted *N*. On the other hand, we propose to use the adequation degree as proposed in (Baziz & al., 2006) to compare a fuzzy set representing preferences to a fuzzy set having a semantic of similarity. The comparison results of fuzzy sets having the same semantic (similarity or imprecision) are aggregated using the min operator (which is classically used to interpret the conjunction).

Therefore, an answer is a set of tuples composed of the pertinence score *ps* associated with the relation, three comparison scores associated with the selection criteria in the data warehouse: a global adequation score *ad* associated with the comparison results having a semantic of similarity and two global matching scores Π and *N* associated with the comparison results having a semantic of imprecision, and, the values associated with each projection attribute. Based on those scores, we propose to define a total order on the answers which gives greater importance to the most pertinent answers compared with the ontology. Thus, the answers are successively sorted according to firstly ps, then *ad* and thirdly a total order defined on Π and *N* where *N* is considered as of greater importance than Π.

Example 7
The answer to the query of Example 6 compared with the first row of the table presented in Table 1Table 1 A Web data table and annotated in Figure 4 is given below:
Result = {ps= 0.75, ad=0.66, N= 0.0 , Π = 1.0, FoodProduct=(0.66/ Rice + 0.5/Rice Flour), ContaminationLevel=[1.65,1.65,1.95,1.95]}.

The Ontology alignment

As already said, the CARAT system is composed of contamination data indexed by the CONTA ontology and consumption data indexed by the CONSO ontology. Both types of data concern food products: the contamination data are measures of chemical contamination for food products and the consumption data are about household purchases of food products. Therefore, the decision support system of the CARAT system needs correspondences to be found between food products of the CONTA ontology and food products of the CONSO in order to estimate the exposure of a given population of consumers to chemical contaminants.

Since the CONSO ontology is updated every year by the company which provides the TNS WORLD PANEL data and, on the contrary, the CONTA ontology remains stable, the CONSO ontology is considered as the source ontology in the alignment process and the CONTA ontology as the target ontology. A simple mapping between food product names of the CONSO ontology

and food product names of the CONTA ontology is not efficient because only a little set of names have words in common. Therefore we have used an additional knowledge to map food products: the food product description available in both ontologies. For this purpose, the content of the CONTA ontology presented in Subsection "The structure of the CONTA ontology" is extended with an international food description vocabulary called Langual (Ireland & Moller, 2000). Langual is composed of predefined characteristics and of predefined associated values partially order by the subsumption relation. Figure 10 gives an excerpt of the extended version of the CONTA ontology expressed in RDFS: the symbolic type Food product presented in Figure 2 is extended with the Langual vocabulary (colored in grey) which permits to describe the food product.

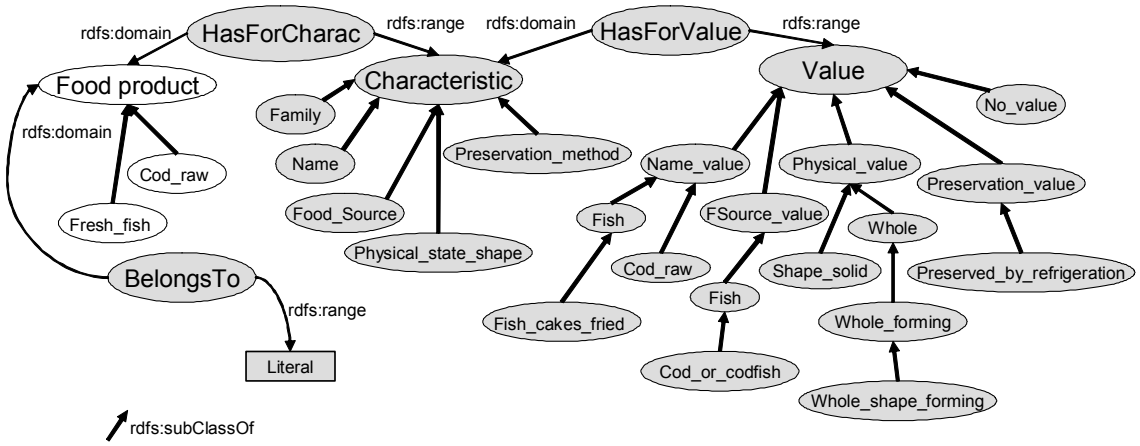


Figure 10 An excerpt of the extended version of the CONTA ontology

The CONSO ontology has the same structure. It is restricted to the symbolic type FoodProduct and associated vocabulary used to describe the food product. This food description vocabulary is extracted from the TNS WORLD PANEL data source. As for the Langual vocabulary, it is composed of a food characterization list and predefined associated values.

Our ontology alignment process consists in considering the ontology alignment problem as a rule application problem in which the food descriptions of the CONSO ontology and the food descriptions of the CONTA ontology are put together into a fact base and in which rules are defined from the food descriptions of the CONTA ontology. The different steps of our ontology alignment process is detailed in Buche & al. (2008). At the end of our ontology alignment process, food products of the CONSO ontology are annotated by sets of food products of the CONTA ontology that are candidates for the alignment.

BACKGROUND

The contribution of our system can be evaluated as an application in the field of food safety or as methods in semantic annotation, flexible querying and ontology alignment.

In the field of food safety, a lot of sources of information are available on the Web (see McMeekin & al., 2006 for a recent review). Many efforts have been done to standardize and to classify the food product names used to index the data in those sources at an international level

(see Ireland & Moller, 2000 for a review). Recent works have also proposed to build ontologies using Semantic Web languages (Soergel & al., 2004). But, for the best of our knowledge, we think that our approach, which permits to semi-automatically integrate data extracted from heterogeneous sources using semantic annotation and ontology alignment methods, is an original contribution.

We have chosen to build an unsupervised annotation system which recognizes predefined relations in tables: first, the ontology can be easily built from explicit metadata associated with relational local databases, which correspond to the relational schemas of the databases and their attributes with their associated domains. This approach has been experimentally tested on three different domains (microbial risk in food, chemical risk in food and aeronautics): three OWL ontologies have been created within a couple of hours thanks to preexisting information retrieved from local databases and a very simple tool which translates automatically csv files containing the metadata into an OWL ontology; second, the structure of data tables is highly variable (even tables in the same paper don't have the same structure) and terms appear in tables with no linguistic context, that invalidates the annotation techniques that learn wrappers based on structure and/or textual context such as Lixto (Baumgartner & al., 2001) or BWI (Freitag & Kushmerick, 2000). Our approach can be compared to the construction of frames from tables described in Pivk & al. (2004) but they use a generic ontology and create new relations according to the table signature, whereas we want to recognize predefined relations in an ontology specific to the target domain.

In the framework of XML database flexible querying, different approaches have been proposed to extend either XPATH or SPARQL. (Campi & al., 2006) proposes FUZZYXPATH, a fuzzy extension of XPATH to query XML documents. Extensions are of two kinds : (i) the 'deep-similar' function permits a relaxed comparison in term of structure between the query tree and the data tree; (ii) the 'close' and 'similar' predicates extend the equality comparison to a similarity comparison between the content of a node and a given value expressed in the query. (Hutardo & al., 2006) proposes an extension of the SPARQL 'Optional' clause (called Relax). This clause permits to compute a set of generalizations of the RDF triplets involved in the SPARQL query using especially declarations done in the RDF Schema. (Corbi & al., 2004) also proposes the same kind of extension of the SPARQL query using a distance function applied to the classes and properties of the RDF Schema. The originality of our approach in flexible querying is that we propose a complete and integrated solution which permits (1) to annotate data tables with the vocabulary defined in an OWL ontology, (2) to execute a flexible query of the annotated tables using the same vocabulary and taking into account the pertinence degrees generated by the annotation system.

Finally, the ontology alignment problem has been widely investigated in the literature (Euzenat & Shvaiko, 2007; Noy, 2004; Kalfoglou & Schorlemmer, 2005; Castano & al., 2007). Our originality is to treat that problem as a rule application problem where a source ontology, considered as a fact base, is aligned with a target one, considered as a rule base.

FUTURE RESEARCH DIRECTIONS

The domain ontology is the central element of our data integration system. In the future, we want to carry on our work on data integration based on ontology.

First, we intend enhancing the performance of the annotation system using machine learning techniques (see Doan & al., 2003) on the knowledge of the ontology but without manual training on a subset of the corpus. By example, a new classifier for symbolic types can be added to the existing one and trained using the domain of values associated with the symbolic type in the ontology. Second, we want to integrate the user's opinion on the query result in order to improve the underlying semantic annotation process and consequently to enrich the ontology. Third, since our flexible querying system allows the user to query uniformly several sources indexed by the same ontology, we want to extend our system in order to be able to query several sources relying on distinct ontologies which have been previously aligned. Fourth, one important feature which must be added to @Web is to be able to detect that data included in tables retrieved from different documents of the Web are redundant. We want to use reference reconciliation methods (see Sais & al, 2007) to deal with this problem.

CONCLUSION

In this chapter, we have presented an ontology-based data integration system in the field of food safety. This system allows data of different nature (contamination data and consumption data) and of different sources (filled manually, coming from existing databases or extracted from the Web) to feed together a decision support system to compute the exposure of a given population of consumers to chemical contaminants.

The essential point to retain from this chapter is that the ontology is the core of our data integration system. We have proposed three original processes to integrate data according to a domain ontology. First, the semantic annotation process proposes an unsupervised aggregation approach from cells to relations to annotate Web data tables according to a domain ontology. Second, the querying process relies on a flexible querying system which takes into account the pertinence degrees generated by the semantic annotation process. Third, the ontology alignment process proposes to find correspondences between objects of a source ontology and objects of a target ontology by means of rules which exploit the characteristics and their values associated with each objects of both ontologies.

REFERENCES

Baumgartner, R., Flesca, S., & Gottlob, G. (2001). Visual Web Information Extraction with Lixto. *Proceedings of the 27th International Conference on Very Large Data Bases*, (pp. 119-128).

Baziz M; Boughanem, M., Prade, H., & Pasi, G., (2006). A fuzzy logic approach to information retrieval using a ontology-based representation of documents in E. Sanchez (Ed.), *Fuzzy logic and the Semantic* (pp. 363-377). Web Elsevier, 18.

Buche, P., Dervin, C., Haemmerlé, O., & Thomopoulos, R. (2005) Fuzzy querying of incomplete, imprecise, and heterogeneously structured data in the relational model using ontologies and rules. IEEE T. Fuzzy Systems, 13(3), 373-383.

Buche, P., Dibie-Barthélemy, J., & Ibanescu, L. (2008) Ontology Mapping Using Fuzzy Conceptual Graphs and Rules. *International Conference on Conceptual Structures*, ICCS Supplement, (pp. 17-24).

Buche P., Soler L. & Tressou J. (2006) Le logiciel CARAT. In : Bertail P., Feinberg M., Tressou J., Verger P. (Eds.), *Analyse des Risques alimentaires*, (pp 305-333). Lavoisier Tech&Doc.

Campi, A., Damiani, E., Guinea, S., Marrara, S., Pasi, G., & Spoletini, P. (2006). A Fuzzy Extension for the Xpath Query Language. *Flexible Query Answering Systems*, Lecture Notes in Computer Science: Vol. 4027 (pp. 210-221).

Castano, S., Ferrara, A., Montanelli, S., Hess, G. N., & Bruno, S., (2007). BOEMIE (Bootstrapping Ontology Evolution with Multimedia Information Extraction). *State of the Art on Ontology Coordination and Matching*, FP-027538 Delivrable 4.4.

Corbi, O., Dieng-Kuntz, R. & Faron-Zucker, C. (2004). Querying the Semantic Web with Corese Search Engine. *Proceedings of the 16th European Conference on Artificial Intelligence*, subconference PAIS'2004, IOS Press, (pp. 705-709).

Doan, A., Domingos, P. & Halevy, A.Y. (2003) Learning to match the schemas of data sources: A multistrategy approach. *Machine Learning* 50(3), 279–301

Dubois, D., & Prade, H., (1988). *Possibility theory: an approach to computerized processing of uncertainty*, New York: Plenum Press.

Euzenat, J. & Shvaiko, P., (2007). Ontology Matching. Berlin: Springer.

Freitag, D., & Kushmerick, N. (2000). Boosted Wrapper Induction. *Proceedings of the 17th National Conference on Artificial Intelligence and 20th Conference on Innovative Applications of Artificial Intelligence* , (pp. 577-583).

Hignette, G., Buche, P., Dibie-Barthélemy, J. & Haemmerlé, O. (2007). An Ontology-Driven Annotation of Data Tables. *Web Information Systems Engineering 2007 Workshops*, (pp. 29-40)

Hutardo, C. A., Poulovassilis, A., & Wood, P. T., (2006). A Relaxed Approach to RDF Querying. *Proceedings of the 5th International. Semantic Web Conference*, Lecture Notes in Computer Science : Vol. 4273 (pp. 314-328).

Ireland, J. D. & Moller, A. (2000). Review of international food classification and description. *Journal of food composition and analysis*, 13, 529-538.

Kalfoglou, Y., & Schorlemmer, M., (2005). Ontology Mapping: The State of the Art. *The Knowledge Engineering Review*, 18(1), 1-31.

Mc Meekin, T.A., Baranyi, J., Bowman, J., Dalgaard, P., Kirk, M., Ross, T., Schmid, S., & Zwietering, M. H. (2006). Information systems in food safety management. *International Journal of Food Microbiology,* 112(3), 181-194.

Noy, N. F., (2004). Semantic Integration: A Survey Of Ontology-Based Approaches. *ACM SIGMOD Record*, 33(4), 65-70.

Pivk, A., Cimiano, P., & Sure, Y. (2004). From Tables to Frames. *International Semantic Web Conference*, Lecture Notes in Computer Science : Vol. 3298 (pp. 116-181)

Saïs, F., Pernelle, N., Rousset, M. C. (2007) L2R: A Logical Method for Reference Reconciliation. AAAI 2007 (pp. 329-334).

Soergel, D., Lauser, B., Liang, A., Fisseha, F., Keizer, J., & Katz, S. (2004). Reengineering Thesauri for New Applications: The AGROVOC Example. *Journal of Digital Information*, 4(4).

Thomopoulos, R., Buche, B. & Haemmerle, O. (2006). Fuzzy Sets Defined on a Hierarchical Domain. *IEEE Transactions on Knowledge and Data Engineering* 18(10), 1397-1410.

Van Rijsbergen, C.J., (1979). *Information Retrieval, second edition*, Department of computer science, university of Glasgow.

Zadeh, L. A., (1965). Fuzzy Sets. *Information and Control*, 8, 338-353.

KEY TERMS & DEFINITIONS

Flexible Querying: methods for querying a database which enhance standard querying expressiveness in various ways such as the expression of user's preferences in order to facilitate the extraction of relevant data.

Fuzzy Set: a mapping from a universe of discourse – definition domain of the fuzzy set – into the interval [0,1]. The concept of fuzzy set extends the notion of Boolean membership to a set to the notion of degree of membership.

MIEL++ query: a conjunctive query where the selection value associated with a queried attribute is expressed by a fuzzy set representing preferences.

Ontology: the vocabulary used to express the knowledge specific to the application domain.

Semantic annotation: process for identifying which semantic relations of a given domain ontology are represented in a Web data tables.