# Semantic annotation of Web data applied to risk in food

Gaëlle Hignette, Patrice Buche, Olivier Couvert, Juliette Dibie-Barthelemy, David Doussot, Ollivier Haemmerlé, Eric Mettler, Lydie Soler

# Semantic annotation of Web data applied to risk in food

**Gaëlle Hignette*[1,2], Patrice Buche[1], Olivier Couvert[3], Juliette Dibie-Barthélemy[1,2], David Doussot[1,2], Ollivier Haemmerlé[4], Eric Mettler[5], Lydie Soler[1]**

[1]INRA MIA, Unité Mét@risk UR 1204,
16, rue Claude Bernard 75231 Paris Cédex 5 FRANCE
Tel +331 44 08 18 89, Fax 331 44 08 16 66
E-mail: gaelle.hignette@agroparistech.fr

[2]AgroParisTech, UFR Informatique, 16, rue Claude Bernard, 75 231 Paris Cedex 05, France
[3]ADRIA Développement, Creac'h Gwen, 29196 Quimper Cedex, France
[4]Département de Mathématiques-Informatique, UFR Sciences, Espaces et Sociétés, Université Toulouse le Mirail, 5, Allée Antonio Machado, F-31058 Toulouse Cedex 1, France
[5]Soredab (Groupe SOPARIND BONGRAIN), La Tremblaye, 78125 La Boissière-Ecole, France

* corresponding author

## Abstract

A preliminary step to risk in food assessment is the gathering of experimental data. In the framework of the Sym'Previus project (http://www.symprevius.org), a complete data integration system has been designed, grouping data provided by industrial partners and data extracted from papers published in the main scientific journals of the domain. Those data have been classified by means of a predefined vocabulary, called ontology. Our aim is to complement the database with data extracted from the Web. In the framework of the WebContent project (www.webcontent.fr), we have designed a semi-automatic acquisition tool, called @WEB, which retrieves scientific documents from the Web. During the @WEB process, data tables are extracted from the documents and then annotated with the ontology. We focus on the data tables as they contain, in general, a synthesis of data published in the documents. In this paper, we explain how the columns of the data tables are automatically annotated with data types of the ontology and how the relations represented by the table are recognized. We also give the results of our experimentation to assess the quality of such an annotation.

**Keywords :** Information extraction, database, ontology, predictive microbiology

# Introduction

A preliminary step to risk in food assessment is the gathering of experimental data, as stated by Tamplin, Baranyi and Paoli (2003) or Le Marc, Pin and Baranyi (2005). **In the field of food safety management, a lot of sources of information are available on the Web (see McMeekin & al. 2006 for a recent review). The generalisation of web publication for scientific laboratories and food authorities, combined with the excellent performance of the standard web crawlers like google, result in a huge amount of available information. Internet users can also access specialised databases containing experimental results. The difficulty is therefore to extract the pertinent quantitative information under a format that is rapidly and easily manageable. Such an extraction is time consuming if done manually and needs to be regularly repeated to remain accurate.**

In the framework of the Sym'Previus project (see Couvert *et al.* (2007) and http://www.symprevius.org), Buche, Dervin, Haemmerlé and Thomopoulos (2005) have designed a complete data integration system composed of data provided by industrial partners and data extracted from papers published in the main scientific journals of the domain. Those data have been classified by means of a predefined vocabulary, called ontology, representing the information that is relevant to food microbiology. In this sense, the database is comparable to the ComBase database described by Baranyi and Tamplin (2004). However, the data integration system that we propose has been designed in order to take into account an important characteristic of the data, their incompleteness. Data are relatively rare in the field of risk in food due to confidentiality and acquisition cost. Two solutions have been proposed to deal with that problem. The first solution relies on an extended querying system, called MIEL, which allows the user to retrieve the nearest data stored in the database corresponding to his/her selection criteria: the ontology is used in order to assess which data can be considered as "near" to the user's selection criteria. The second solution, which is under construction in the framework of the WebContent project (http://www.webcontent.fr), is detailed in this paper. It consists in searching data on the Web to complement the database. We have designed a semi-automatic acquisition tool, called @WEB (Annotating Tables from the WEB), which retrieves scientific documents from the Web and extracts data tables, which contain, in general, a synthesis of data published in the documents, and then annotates the tables using the ontology. The problem of finding data tables in documents has been widely addressed by the computer science research community (see for example the synthesis made by Zanibbi, Blostein, and Cordy (2004)). In this paper, we explain how the columns of the data tables are automatically annotated with the ontology. Once the tables are correctly annotated, they can be queried using the ontology in the same way as the existing database in MIEL presented by Buche, Dibie-Barthélemy, Haemmerlé and Hignette (2006).

The ontology used in our data integration system is composed of data types meaningful in the field of risk in food, and of relations that allow one to link those data types.

Data types are described in the ontology in two different ways depending on whether their associated values are symbolic (for example *Food Product*, where values are product names) or numeric (for example *Temperature*). Our ontology contains 3 symbolic types and 18 numeric types. Symbolic types are described by a type name (*Food Product, Microorganism* or *Response*) and a taxonomy of possible values (a taxonomy of food products, a taxonomy of microorganisms and, for the type *Response*, the possible responses of a microorganism to a treatment: growth, absence of growth or death). The possible values of a symbolic type defined in the type taxonomy are called terms. Numeric types are described by a type name (for example *Time*, *Temperature* or *Colony count*), the set of units in which the type can be expressed (for example, °C or °F for *Temperature*, but no unit for *pH* or $a_w$), and eventually a numerical range (for example, [0,14] for *pH* or [0,1] for $a_w$). See Table 1 for a description of the numeric types of the ontology.

Relations are used to describe the meaning of different datatypes grouped together: for example, linking the type *pH* with the type *Food product* within the relation *Product parameter - pH* allows one to measure the pH of a food product, while linking the type *pH* with the type *Microorganism* in the relation *growth parameter - pH* allows one to measure the pH at which the microorganism is able to grow. The relations are described in the ontology by their name and their signature: the signature of a relation is composed of the result type (the measure that is the object of the experiment) and the access types (the factors that influence the result type measure). For example, in the relation *Product parameter – pH*, the access type is *Food product* and the result type is *pH;* in the relation *Growth kinetics*, the result type is *Microorganism concentration* and the access types are *Microorganism*, *Food product*, *Temperature* and *Time* (a typical experiment would be setting the microorganism, food product and temperature, and measuring the microorganism concentration as a factor of time. Additional environmental factors might be controlled in the experiment, but the ontology models what information is most commonly available; other factors might be modelled as part of other relations, for example *Product parameter – $a_w$...*).

The ontology we currently use in our annotation process is very small, but permits the representation of quite a lot of information. Of course, it will be possible to extend the ontology to represent some more information: this extension has to be done by manually adding data types or relations that lack in the ontology to represent other kind of data that we want to integrate into the data integration system.

# Materials and methods

Our annotation algorithm is divided in three steps described in Figure 1. First, we distinguish between columns containing numeric data and columns containing symbolic data. Then we annotate the columns, using a different method according to whether the column is symbolic or numeric. The final step of our algorithm is the recognition of the relations represented by the table.

/* IL FAUDRAIT AJOUTER UNE FIGURE QUI REPREND L'UN DES TABLEAUX EXEMPLES DE LA SUITE DU PAPIER POUR DONNER UNE VISION GLOBALE DU RESULTAT DE L'ANNOTATION : 1) LE NOM DE LA RELATION SEMANTIQUE A RETROUVER, 2) LE NOM DES TYPES DE L'ONTO CORRESPONDANT AUX COLONNES, 3) LES NOMS DE PRODUITS, MICROORG DE L'ONTO SIMILAIRES AU CONTENU DES CELLULES DE TYPE SYMBOLIQUE */

### Distinction between numeric and symbolic columns

The distinction between numeric and symbolic columns is not as simple as it seems: symbolic columns may contain numbers (for example, the strain of a microorganism) and numeric columns often contain character strings such as units, etc. We thus propose a method that uses the units defined in the ontology in order to classify the columns.

Let *col* be a column of the table we want to annotate. We search *col* for all occurrences of numbers (in decimal or scientific format) and for all occurrences of units of numeric types described in the ontology. We also search *col* for all words, which are defined as alphabetic character sequences that are neither units nor "no result indicators" (the "no result indicators" are character sequences that indicate that the cell contains no result, such as "not specified", "not available", "no result" etc.).

Let *c* be a cell of the column *col*. We apply the following classification rules:

- if *c* contains a number immediately followed by a unit, or a number in scientific format, then *c* is numeric;
- else, if *c* contains more numbers and units than words, then *c* is numeric;
- else, if *c* contains more words than numbers and units, then *c* is symbolic;

- else (number of words equal to number of units and numbers) the status of *c* is considered as unknown.

Once all cells of the column *col* have been classified using the above rules, we count the number of cells in *col* that were classified as numeric or as symbolic (the cells classified as unknown are not taken into account). The column *col* is classified as symbolic if there are more cells classified as symbolic than numeric. Else, the column is classified as numeric (we have experimentally shown that when numbers of symbolic and numeric cells are equal, it usually corresponds to a high rate of absent data, which is more frequent in numeric columns).

**Symbolic column annotation**

Once a column has been recognised as symbolic, we annotate each cell in the column with the terms from the taxonomies of each symbolic type in the ontology. For that, we use a similarity measure between a term from the Web (found in the cell of a symbolic column), and a term from the ontology. All terms are transformed into weighted vectors: the coordinate axis of the vectors represent all possible words (i.e. all words in the ontology plus the words in the terms to compare with the ontology), the coordinate values represent the weight of those words in the term. Table 2 presents an example of such a vector representation of terms. For terms from the ontology, each word is manually weighted according to its importance in the meaning of the term. A weight of 1 means that the word is essential to the meaning of the term ; a weight of 0.2 means that the word is secondary to the meaning of the term. For terms from the Web, each word has a weight of 1, as the meaning of the term is not known *a priori*. Terms are lemmatised, i.e. grammatical forms of plural or conjugaison are removed, so that “carrot cuts” and “cut carrots” will be considered as the same. Words consisting of only one letter or terms that belong to a defined “stopword list” are not taken into account (the stopwords are words that are very common and bear no real semantics, such as articles and conjunctions).

The similarity between a term from the Web and a term from the ontology is computed as the cosine similarity measure between the two weighted vectors, which is one of the different similarity measures described by Lin (98). Let *w* be a term from the Web, represented as the weighted vector $w = (w_1, \ldots, w_n)$ and *o* a term from the ontology, represented as the weighted vector $o = (o_1, \ldots, o_n)$. The similarity between *w* and *o* is computed as:

$$sim(w,o) = \frac{\sum_{k=1}^{n} w_k \times o_k}{\sqrt{\sum_{k=1}^{n} {w_k}^2 \times \sum_{k=1}^{n} {o_k}^2}} \qquad \textbf{(1)}$$

For example, using the terms given in Table 2, we compute the following similarities:

$$sim(ground\,meat,\ fresh\,meat) = \frac{1\times0+1\times1+0\times0.2+0\times0}{\sqrt{(1^2+1^2)\times(1^2+0.2^2)}} \approx 0.57$$

$$sim(ground\,meat, ground\,beef\,) = \frac{1\times0.2+1\times0+0\times0+0\times1}{\sqrt{(1^2+1^2)\times(0.2^2+1^2)}} \approx 0.11$$

For each cell in the symbolic column, we compute the similarity measure with each term from the taxonomies of symbolic types of the ontology. Then, for each cell, we compute the sum of such similarities for each symbolic type. A cell is considered as having the type which has the best sum of similarities, provided that this sum of similarity is sufficiently higher than the second best sum of similarities. This notion of "sufficiently higher" is computed using the proportional advantage: let *best* be the type with the best sum of similarities for the cell *c*, and *secondBest* be the type with the second best sum of similarities for the cell *c*; let *Taxo(type)* be the set of terms in the taxonomy of a symbolic type *type* and *Term(c)* the term that is contained in the cell *c*. Then the proportional advantage of the type *best* for the cell *c* is computed as:

$$adv(best,c) = \frac{\sum_{t_1 \in Taxo(best)} sim(Term(c),t_1) - \sum_{t_2 \in Taxo(secondBest)} sim(Term(c),t_2)}{\sum_{t_1 \in Taxo(best)} sim(Term(c),t_1)} \qquad \textbf{(2)}$$

The type *best* is then considered as the type of the cell *c* if its proportional advantage for the cell is higher than a specified threshold. If the proportional advantage of *best* for the cell *c* is lower than the specified threshold, then the cell *c* is considered as of unknown type.

***Example:*** We consider the first column of Table 3. The first cell of the column contains the term «Canned foods "Neutral"» which has common words with several terms from the *Food product* taxonomy: «Baby foods» (similarity of 0.258), «Deep frozen foods» (similarity of 0.236), «Hospital food» (similarity of 0.516), «Food products» (similarity of 0.516) and «Rice baby food» (similarity of 0.192). The sum of similarities of the type *Food product* for the cell is then 1.718, while the other symbolic types (i.e. *Microorganism* and *Response*) have sums of similarities of 0. The cell is thus considered as having the type *Food product*. The second cell in the column contains the term «Canned foods "Acid"». This term has the same similarity measures with the terms from the *Food product* taxonomy as the term in the first cell of the column, but it also has common words with some terms from the *Microorganism* taxonomy: «Lactic acid bacteria» (similarity of 0.333), «Lactic acid microorganisms» (similarity of 0.333) and «Acidophilic lactic acid microorganisms» (similarity of 0.289). The sum of similarities of the type *Food Product* for the cell is 1.718 and the sum of similarities of the type *Microorganism* for the cell is 0.955. The proportional advantage of *Food Product* for the cell is then (1.718-0.955)/1.718 = 44.4%. If this is higher than the specified threshold, then the cell is considered as having the type *Food product*.

When each cell of the column is assigned a type, we compute the score of a symbolic type *type* for the column *col* according to the column contents, noted $score_{contents}(type, col)$, as the proportion of cells in that column that were considered as having this type.

We also compute the score of a symbolic type *type* for the column *col* according to the column title, noted $score_{title}(type, col)$, as the cosine similarity measure between the column title and the type name.

Then the final score of a symbolic type *type* for the column *col* is computed as follows:

$$score_{final}(type, col)=1-(1- score_{contents}(type, col))(1- score_{title}(type, col)) \qquad \textbf{(3)}$$

The type of the column is then the type that has the best final score for this column, provided that this score is sufficiently higher than the second best score according to the proportional advantage measure: the proportional advantage is computed in the same way as described in equation **(2)**, replacing the sum of similarities  with the final score of the type for the column. The type with the best final score is then considered as the type of the column if its proportional advantage is higher than a specified threshold. If the proportional advantage is lower than the specified threshold, then the column is considered as of unknown type.

***Example:*** We consider the first column of Table 3. Assuming that the threshold of proportional advantage to adopt a type for a cell is lower than 44,4% (see preceding example), the score of the type *Food product* for the column according to the column contents is 1 (two cells over two are classified as *Food product*). The scores of the types *Microorganism* and *Response* for the column according to the column contents are both 0. The title of the column is the term «Food»: the score of the type *Food product* according to the column title is thus 0.577, while the scores of *Microorganism* and *Response* according to the column title are both 0. The final score of the type *Food product* for the column is computed as 1-(1-1)*(1-0.577) = 1, while the final score of the two other symbolic types is 0. The column is then considered as having the type *Food product*.

**Numeric column annotation**

When a column has been recognised as numeric, we look at all the units that are presented in this column. Let *num* be a function that associates to a unit *u* the number *num(u)* of numeric types in the ontology that can be expressed with this unit. Let *units* be a function that associates to a numeric type *type* and a column *col* the set *units(type,col)* of all units that are present in the column *col* and that can be used to represent data of the type *type*. Then the score of the numeric type *type* for the column *col* according to the units presented in the column is:

$$score_{units}(type,col) = \max_{u \in units(type,col)} \frac{1}{num(u)} \qquad \textbf{(4)}$$

***Example:*** We consider the second column of Table 4. The only unit which is present in the column is %. There are five numerical types that can be expressed with this : *NaCl, N2, CO2, O2* and *Samples Positive*. As there is only one unit in the column, $\text{score}_{\text{units}}$(type, column)=0.2 for those five numerical types.

We also compute the score of a numeric type *type* for the column *col* according to the column title, noted $score_{title}(type, col)$, as the cosine similarity measure between the column title and the type name.

Then the final score of a numeric type *type* for the column *col* is computed as follows:

- if the numeric contents of the column are not compatible with the value range defined in the ontology for the numeric type *type* , then $score_{final}(type, col)=0$ (for example, a column with no unit containing the numeric value 16 can neither be of type $a_w$ nor of type pH);
- else (if all numbers in the column are inside the value range of the type *type*), the final score of the type *type* for the column *col* is

$$score_{final}(type, col)=1-(1- score_{units}(type, col))(1- score_{title}(type, col)) \qquad \textbf{(5)}$$

The type of the column is then the type that has the best final score for this column, provided that its proportional advantage (computed in the same way as in equation **(2)**, replacing the sum of similarities with the final score of the type for the column) is better than a given threshold, otherwise the column is considered as unknown.

***Example:*** We consider the second column of Table 4. As seen in the preceding example, $\text{score}_{\text{units}}$(type, column)=0.2 for the five numeric types *Samples Positive*, *NaCl*, *N2*, *CO2* and *O2*. The title of the column is the term «Positive for Campylobacter»: the score of the type *Samples Positive* according to the column title is 0.5, while the scores of the other types according to the column title are all 0. The final score of the type *Samples Positive* for the column is then computed as 1-(1-0.2)*(1-0.5) = 0.6, while the final score of the four other numerical types is 0.2. The column is then considered as having the type *Samples Positive*.

## Finding the semantic relations represented by the table

Once the types of all columns of a table have been recognized, we look for the relation(s) of the ontology that are represented in the table. As for the column types recognition, the final score of a relation for the table is the combination of two scores: the score of the relation for the table according to the table title, and the score of the relation for the table according to the table signature (the set of its recognized columns).

The score of a relation for the table according to the table title is computed as the cosine similarity measure between the table title and the relation name.

The score of a relation *rel* for the table *tab* according to the table signature is computed as follows:

- if the result type of the relation *rel* was not recognized as a type of a column of the table, then $score_{signature}(rel, tab) = 0$
- else, the score of the relation for the table is the proportion of types in its signature that were recognized in the table columns. Let $Sign_{rel}$ be the set of types in the signature of relation *rel* (i.e. the access types and the result type), $Sign_{tab}$ the set of types that were recognized for the table columns and *card* the function that associates to a set the number of items in this set, then

$$score_{signature}(rel, tab) = \frac{card(Sign_{rel} \cap Sign_{tab})}{card(Sign_{rel})} \quad \textbf{(6)}$$

Then the final score of a relation *rel* for the table *tab* is computed as:

$$score_{final}(rel, tab) = 1-(1-score_{title}(rel, tab))(1-score_{signature}(rel, tab)) \quad (7)$$

When the scores of all relations of the ontology have been computed for the table, we choose the relation(s) with which the table is annotated. A table can represent several relations at a time: this is mainly due to our modelling of relations, which only have one result type. For example, if a table gives the pH and the water activity of a food product, we will consider it as two separate relations: *food pH* and *food water activity*.

Two relations are called concurrent if they have the same result type. A relation *rel* with a non-zero final score for the table is kept or not for the annotation of the table according to the status of its concurrent relations:

- if the relation *rel* has no concurrent relation, then *rel* is used to annotate the table;
- if the relation *rel* has a concurrent relation *rel2* with a better final score for the table, then *rel* is excluded from the annotation of the table;
- if the relation *rel* has concurrent relations, but all those concurrent relations have final scores for the table that are lower or equal to the final score of *rel* for the table, then *rel* is used to annotate the table.

***Example:*** We consider the example presented in Table 5. The first column is of unknown type, while the second has been recognised as of type *pH*. The only relations of our ontology having *pH* as result type are *Growth parameter – pH* (access type: *Microorganism*) and *Product property – pH* (access type: *Food product*). For both these relations, only one over the two types of the signature is recognised: the scores of these relations for the table according to the column types are both 0.5. The table title contains the word “growth” which is in common with the name of the relation *Growth parameter – pH* (score of the relation according to the table title: 0.218), while the table title has no common word with the relation *Product property – pH* (score of the relation according to the table title: 0) . The final score of the relation *Growth parameter – pH* for the table is computed as: 1-(1-0.5)*(1-0.218) = 0.609 while the final score of the relation *Product property – pH* for the table is computed as: 1-(1-0.5)*(1-0) = 0.5. The table is then annotated with the relation *Growth parameter – pH.*

## Experimental approach

Our annotation algorithm was tested on 60 tables extracted from publications on food microbiology. The tables were manually annotated to give a type to each of the 349 columns belonging to those tables: the columns were first separated between numeric and symbolic, then the symbolic columns were annotated with the types *Microorganism*, *Food Product*, *Response* or "other" if the column contained other precisions that did not match any of the symbolic types of our ontology. The numeric columns were annotated with the 18 numeric types of our ontology. The tables were then manually annotated with the relations in the ontology corresponding to the meaning of the data represented in the table.

We ran our annotation algorithm on the 60 tables, comparing the computed column types and the computed relations with the ones that had been manually chosen. The thresholds of proportional advantage for recognizing the symbolic cell type, the symbolic column type and the numeric column type were all set to 10%.

The quality of our method to distinguish between symbolic and numeric columns was assessed against a "naive" classifier: in that classifier, the units defined in the ontology, as well as the list of "no result indicators" are not used. In the naive classifier, a cell is considered as numeric if and only if it contains a number, and a column is numeric if at least half of its cells are numeric (else the column is symbolic). The quality of the rest of the annotations is assessed using two common measures: precision and recall. Precision is the ratio of correct computed annotations over the total number of computed annotations (correct and wrong). Recall is the ratio of correct computed annotations over the number of manual annotations.

# Results and discussion

The results of the distinction between numeric and symbolic columns are given in Table 6. Our method gives much better results than the naive classifier because it is able to consider as non-numeric a cell that contains numbers (for example a microorganism with a strain number). It is also able to deal with unknown data: the “no result indicators” are not considered as words, so a cell containing only a “no result indicator” is considered as unknown, whereas the naive classifier considers it as symbolic.

Table 7 shows the results of the annotation of 81 symbolic columns that were correctly recognized as symbolic in the first step of our algorithm. Our method gives a good overall precision (89%) and a lower overall recall (81%). This is due to the fact that the column is considered as unknown whenever there is a doubt on its type: such an annotation is not considered as a real annotation (this leads to a good precision, as it is not added to the wrong annotations, but to a lower recall, as it is not added to the correct annotations).

The annotation of numeric columns gives even better results, with 99.6% precision and 93.9% recall, which is mainly due to a lesser extent of variations in column titles (for example, Temperature is always called Temperature) and to the use of some very indicative units (for example, cfu will only denote a microorganism concentration). Such annotation results can be considered as very good as they are obtained via a fully-automatic method.

For the relations, we obtained a 69% precision and 95% recall. Nevertheless, it is possible to get a better precision by using a threshold on the final score of the relations: the relations are kept for the annotations only if their final score for the table is higher than the given threshold.

Figure 2 shows the evolution of precision and recall according to the value of the threshold. Using a threshold of 0.5 permits a switch of precision and recall: we get a much better precision (96%) at the cost of a lower recall (76%). The switch of precision and recall at a threshold of 0.5 is due to the existence of several relations having only one access type and the same result type (for example *Growth parameter-pH* and *Product property-pH*, or *Growth parameter-*$a_w$ and *Product property-*$a_w$): when only the result type is recognized and the table title gives no indication, the score is 0.5. If the threshold is lower than 0.5, both concurrent relations are kept (one is correct, the other one is false: thus a low precision). If the threshold is higher than 0.5, none of the relations is used to annotate the table (no false annotation, thus a higher precision, but no correct annotation either, thus a lower recall).

The choice of using a threshold of 0.5 or of 0 depends on the goal of the end-users:

- a threshold of 0, i.e. high recall but lower precision, means that it is acceptable to get some relations in the annotation that are not really represented in the table, as long as all relations represented by the table are annotated;
- a threshold of 0.5, i.e. high precision but lower recall, means that nearly every relation in the annotation is correct, but that the annotation misses some of the relations actually represented by the table.

## Conclusion and perspectives

We have proposed a novel way to annotate tables so as to gather automatically data from the Web. Our annotation method gives good results for a fully-automatic way to find out what a table is about. However, there is a trade-off between precision and recall: when using the annotation system, we have to choose between missing almost nothing but getting noisy results (i.e. some of the annotated relations are false), or getting accurate results but missing some information.

Our annotation system is entirely based on the use of a controlled vocabulary, called ontology, which is used to represent the data. The richer the ontology is, the best the annotation will be, as our annotation algorithm uses word-by-word comparison between the terms used in the table and the terms already represented in the ontology. We are now considering the possibility of ontology enrichment to allow better annotation results, our method being easily adjustable to take into account the definition of synonyms.

Moreover, in its current version, the annotation process analyses only the content and the title of the table. In a very next step, we will try to take into account the information available in the sentences of the document which refer to the table. Sometimes, they contain information which is lacking in the table (for example, the name of the microorganism or the food product). We will also try to take into account the footnotes associated with the table which contain also useful information (for example, units). But it will be more difficult because the footnote management depends on the word processor used to generate the document containing the table.

Our future work will aim at allowing the querying of the annotated tables, taking into account the fact that the information is gathered automatically and thus is not completely sure. The automatically gathered data has then to be confronted with the more reliable information stored in local databases.

## Acknowlegements

Financial support from the French National Research Agency (ANR) for the project WebContent in the framework of the National Network for Software Technology (RNTL) is gratefully acknowledged.

## References

Baranyi, J., Tamplin, M., 2004. ComBase: A Common Database on Microbial Responses to Food Environments. Journal of Food Protection 67, 1834-1840.

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 Transactions on Fuzzy Systems 13, 373-383.

Buche, P., Dibie-Barthélemy, J., Haemmerlé, O., Hignette, G., 2006. Fuzzy semantic tagging and flexible querying of XML documents extracted from the Web. Journal of Intelligent Information System 26, 25-40.

Couvert, O., Augustin, J.C., Buche, P., Carlin, F., Coroller, L., Denis, C., Jamet, E., Mettler, E., Pinon, A., Stahl, V., Zuliani, V., Thuault, D, 2007. Optimising food process and formulation through Sym'Previus, managing of the food safety. Proceedings of 5th International Conference Predictive Modelling in Foods.

Le Marc, Y., Pin, C., Baranyi, J., 2005. Methods to determine the growth domain in a multidimensional environmental space. International Journal of Food Microbiology 100, 3-12.

Tamplin, M., Baranyi, J., Paoli, G., 2003. Software programs to increase the utility of predictive microbiology information. In: McKellar, R.C., Lu, X. (Eds), Modelling Microbial responses in Foods. CRC, Boca Raton.

Lin, D., 1998. An information-theoretic definition of similarity. In: ICML '98 : Proceedings of the Fifteenth International Conference on Machine Learning, 296-304. Morgan Kaufmann Publishers Inc., San Francisco.

McMeekin, 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, 181-94.

Zanibbi, R., Blostein, D., Cordy, J. R., 2004. A survey of table recognition : Models, observations, transformations, and inferences. International Journal on Document Analysis and Recognition 7, 1-16.

Figure 1: The different steps of our annotation algorithm

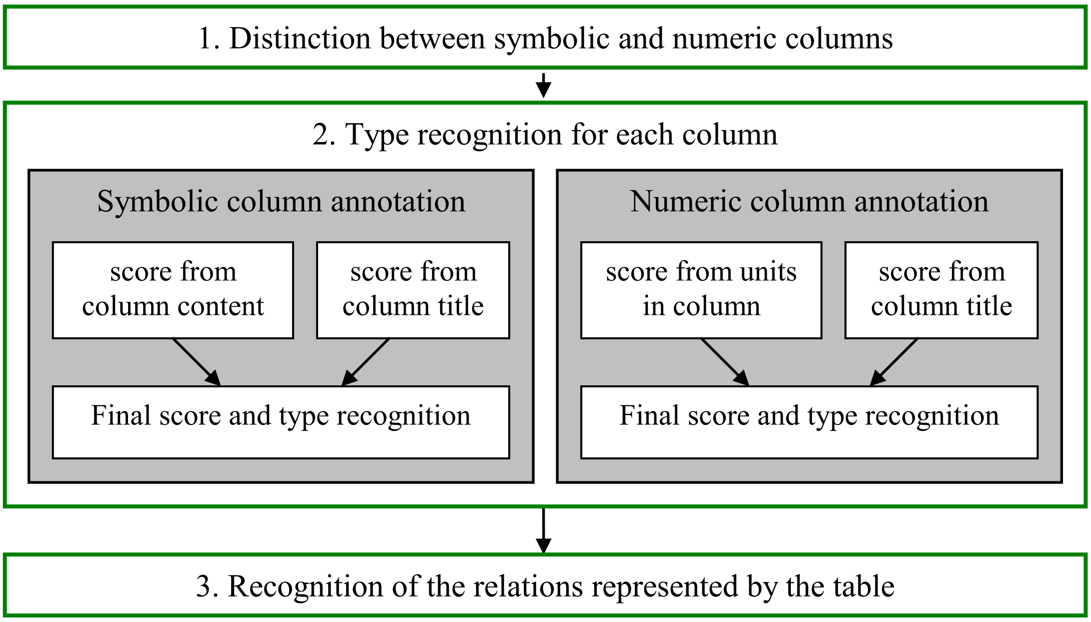
1. Distinction between symbolic and numeric columns
2. Type recognition for each column
Symbolic column annotation
score from column content
score from column title
Final score and type recognition
Numeric column annotation
score from units in column
score from column title
Final score and type recognition
3. Recognition of the relations represented by the table

Figure 2: Evolution of precision and recall on relation recognition according to the score threshold

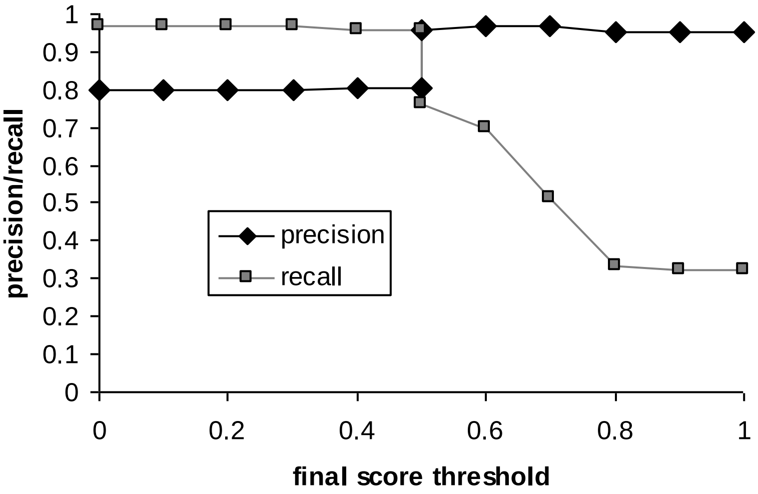

Table 1: Numeric types of the ontology.

| type name | meaning | units |
|---|---|---|
| aw Water Activity | water activity of the growth support | NONE |
| $CO_2$ | atmosphere concentration of $CO_2$ | % |
| Colony count concentration | microorganism concentration | cfu |
| D reduction | time necessary for the decimal reduction of a microorganism due to a particular treatment | mins,secs |
| EH redox potential | redox potential of the growth support | mV |
| Growth rate | growth rate in microbial growth model | h-1 |
| Lag time | lag time in microbial growth model | h |
| $N_2$ | atmosphere concentration of $N_2$ | % |
| NACL | NaCl concentration in the growth support | % |
| Number outbreaks or deaths | number of outbreaks or deaths due to a particular microorganism | NONE |
| $O_2$ | atmosphere concentration of $O_2$ | % |
| pH | pH of the growth support | NONE |
| Samples positive | prevalence: % or number of samples containing a particular microorganism | NONE,% |
| Samples tested | prevalence: number of samples tested | NONE |
| Temperature | temperature of storage | °C,°F |
| Time | time of storage | weeks,days,<br>hr,mins |
| Year | year of event (outbreak, experiment…) | NONE |
| Ymax | Ymax parameter in microbial growth model | cfu |

Table 2: Terms represented as weighted vectors.

| Term | Meaning of the vector axis | ground | meat | fresh | beef |
|---|---|---|---|---|---|
| Term from the Web | ground meat | 1 | 1 | 0 | 0 |
| Term of the ontology | fresh meat | 0 | 1 | 0.2 | 0 |
| Term of the ontology | ground beef | 0.2 | 0 | 0 | 1 |

Table 3: Redox potentials on some foods.

| *Food* | *Eh(mV)* | *pH* |
|---|---|---|
| Canned Foods “Neutral” | -130 to -550 | > 4.4 |
| Canned Foods “Acid” | -410 to -550 | < 4.4 |

Table 4: Reported prevalence of Campylobacter.

| *Product* | *Positive for Campylobacter (%)* |
|---|---|
| Chicken products | 0.07 |

Table 5: Growth of Vibrio parahaemolyticus in Trypticase-soy-broth at 21°C (7%NaCl).

| *Strain* | *Minimum pH for growth* |
|---|---|
| 284-72 | 5.5 |
| T-3765-1 | 5.2 |

Table 6: Results of the distinction between numeric and symbolic columns.

| Column manually annotated as | Total number | Classified using the ontology as | | Classified using the naïve classifier as | |
|---|---|---|---|---|---|
| | | numeric | symbolic | numeric | symbolic |
| numeric | 263 | 261 | 21 | 229 | 34 |
| symbolic | 86 | 5 | 81 | 13 | 73 |
| | | Precision : 98% | | Precision : 87% | |

Table 7: Results of the annotation of symbolic columns.

| Column manually annotated as | Total number | Classified using the ontology as | | | | Recall |
|---|---|---|---|---|---|---|
| | | *Food product* | *Micro-organism* | *Response* | *Unknown* | |
| *Food product* | 46 | 34 | - | - | 12 | 74% |
| *Microorganism* | 16 | - | 16 | - | - | 100% |
| *Response* | 1 | - | - | 1 | - | 100% |
| *Other* | 18 | 3 | 3 | - | 12 | |
| | Precision | 92% | 84% | 100% | | |