



**HAL**  
open science

## 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

### ► To cite this version:

Gaëlle Hignette, Patrice Buche, Olivier Couvert, Juliette Dibie-Barthelemy, David Doussot, et al..  
Semantic annotation of Web data applied to risk in food. *International Journal of Food Microbiology*,  
2008, 128 (1), pp.174-180. 10.1016/j.ijfoodmicro.2008.06.027 . hal-00557628

**HAL Id: hal-00557628**

**<https://hal.univ-brest.fr/hal-00557628>**

Submitted on 30 May 2020

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Semantic annotation of Web data applied to risk in food

Gaëlle Hignette<sup>\*1,2</sup>, Patrice Buche<sup>1</sup>, Olivier Couvert<sup>3</sup>, Juliette Dibie-Barthélemy<sup>1,2</sup>,

David Doussot<sup>1,2</sup>, Ollivier Haemmerlé<sup>4</sup>, Eric Mettler<sup>5</sup>, Lydie Soler<sup>1</sup>

<sup>1</sup>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

<sup>2</sup>AgroParisTech, UFR Informatique, 16, rue Claude Bernard, 75 231 Paris Cedex 05, France

<sup>3</sup>ADRIA Développement, Creac'h Gwen, 29196 Quimper Cedex, France

<sup>4</sup>Département de Mathématiques-Informatique, UFR Sciences, Espaces et Sociétés, Université  
Toulouse le Mirail, 5, Allée Antonio Machado, F-31058 Toulouse Cedex 1, France

<sup>5</sup>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'Previous project (<http://www.symprevious.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](http://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.

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

## 33 **Introduction**

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

43 In the framework of the Sym'Previus project (see Couvert *et al.* (2007) and  
44 <http://www.symprevius.org>), Buche, Dervin, Haemmerlé and Thomopoulos (2005) have  
45 designed a complete data integration system composed of data provided by industrial partners  
46 and data extracted from papers published in the main scientific journals of the domain. Those  
47 data have been classified by means of a predefined vocabulary, called ontology, representing  
48 the information that is relevant to food microbiology. In this sense, the database is comparable  
49 to the ComBase database described by Baranyi and Tamplin (2004). However, the data  
50 integration system that we propose has been designed in order to take into account an  
51 important characteristic of the data, their incompleteness. Data are relatively rare in the field  
52 of risk in food due to confidentiality and acquisition cost. Two solutions have been proposed  
53 to deal with that problem. The first solution relies on an extended querying system, called  
54 MIEL, which allows the user to retrieve the nearest data stored in the database corresponding  
55 to his/her selection criteria: the ontology is used in order to assess which data can be  
56 considered as “near” to the user’s selection criteria. The second solution, which is under

57 construction in the framework of the WebContent project (<http://www.webcontent.fr>), is  
58 detailed in this paper. It consists in searching data on the Web to complement the database.  
59 We have designed a semi-automatic acquisition tool, called @WEB (Annotating Tables from  
60 the WEB), which retrieves scientific documents from the Web and extracts data tables, which  
61 contain, in general, a synthesis of data published in the documents, and then annotates the  
62 tables using the ontology. The problem of finding data tables in documents has been widely  
63 addressed by the computer science research community (see for example the synthesis made  
64 by Zanibbi, Blostein, and Cordy (2004)). In this paper, we explain how the columns of the  
65 data tables are automatically annotated with the ontology. Once the tables are correctly  
66 annotated, they can be queried using the ontology in the same way as the existing database in  
67 MIEL presented by Buche, Dibia-Barthélemy, Haemmerlé and Hignette (2006).

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

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

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

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

## 102 **Materials and methods**

103 Our annotation algorithm is divided in three steps described in Figure 1. First, we distinguish  
104 between columns containing numeric data and columns containing symbolic data. Then we  
105 annotate the columns, using a different method according to whether the column is symbolic

106 or numeric. The final step of our algorithm is the recognition of the relations represented by  
107 the table.

108

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

#### 115 **Distinction between numeric and symbolic columns**

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

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

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

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

131 • else (number of words equal to number of units and numbers) the status of  $c$  is  
132 considered as unknown.

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

### 139 **Symbolic column annotation**

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

155 The similarity between a term from the Web and a term from the ontology is computed as the

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

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

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

$$161 \quad sim(ground\ meat, fresh\ meat) = \frac{1 \times 0 + 1 \times 1 + 0 \times 0.2 + 0 \times 0}{\sqrt{(1^2 + 1^2) \times (1^2 + 0.2^2)}} \approx 0.57$$

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

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

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

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



174 lower than the specified threshold, then the cell  $c$  is considered as of unknown type.

175 **Example:** We consider the first column of Table 3. The first cell of the column contains the  
176 term «Canned foods “Neutral”» which has common words with several terms from the *Food*  
177 *product* taxonomy: «Baby foods» (similarity of 0.258), «Deep frozen foods» (similarity of  
178 0.236), «Hospital food» (similarity of 0.516), «Food products» (similarity of 0.516) and «Rice  
179 baby food» (similarity of 0.192). The sum of similarities of the type *Food product* for the cell  
180 is then 1.718, while the other symbolic types (i.e. *Microorganism* and *Response*) have sums of  
181 similarities of 0. The cell is thus considered as having the type *Food product*. The second cell  
182 in the column contains the term «Canned foods “Acid”». This term has the same similarity  
183 measures with the terms from the *Food product* taxonomy as the term in the first cell of the  
184 column, but it also has common words with some terms from the *Microorganism* taxonomy:  
185 «Lactic acid bacteria» (similarity of 0.333), «Lactic acid microorganisms» (similarity of 0.333)  
186 and «Acidophilic lactic acid microorganisms» (similarity of 0.289). The sum of similarities of  
187 the type *Food Product* for the cell is 1.718 and the sum of similarities of the type  
188 *Microorganism* for the cell is 0.955. The proportional advantage of *Food Product* for the cell  
189 is then  $(1.718-0.955)/1.718 = 44.4\%$ . If this is higher than the specified threshold, then the cell  
190 is considered as having the type *Food product*.

191

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

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

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

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

199 The type of the column is then the type that has the best final score for this column, provided  
 200 that this score is sufficiently higher than the second best score according to the proportional  
 201 advantage measure: the proportional advantage is computed in the same way as described in  
 202 equation (2), replacing the sum of similarities with the final score of the type for the column.

203 The type with the best final score is then considered as the type of the column if its  
 204 proportional advantage is higher than a specified threshold. If the proportional advantage is  
 205 lower than the specified threshold, then the column is considered as of unknown type.

206 **Example:** We consider the first column of Table 3. Assuming that the threshold of  
 207 proportional advantage to adopt a type for a cell is lower than 44,4% (see preceding example),  
 208 the score of the type *Food product* for the column according to the column contents is 1 (two  
 209 cells over two are classified as *Food product*). The scores of the types *Microorganism* and  
 210 *Response* for the column according to the column contents are both 0. The title of the column  
 211 is the term «Food»: the score of the type *Food product* according to the column title is thus  
 212 0.577, while the scores of *Microorganism* and *Response* according to the column title are both  
 213 0. The final score of the type *Food product* for the column is computed as  $1 - (1 - 1) * (1 - 0.577) =$   
 214 1, while the final score of the two other symbolic types is 0. The column is then considered as  
 215 having the type *Food product*.

### 216 **Numeric column annotation**

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

222 score of the numeric type *type* for the column *col* according to the units presented in the  
223 column is:

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

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

228  
229 We also compute the score of a numeric type *type* for the column *col* according to the column  
230 title, noted  $score_{title}(type, col)$ , as the cosine similarity measure between the column title and  
231 the type name.

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

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

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

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

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

### 250 **Finding the semantic relations represented by the table**

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

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

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

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

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

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

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

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

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

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

282 **Example:** We consider the example presented in Table 5. The first column is of unknown  
 283 type, while the second has been recognised as of type *pH*. The only relations of our ontology  
 284 having *pH* as result type are *Growth parameter – pH* (access type: *Microorganism*) and  
 285 *Product property – pH* (access type: *Food product*). For both these relations, only one over  
 286 the two types of the signature is recognised: the scores of these relations for the table  
 287 according to the column types are both 0.5. The table title contains the word “growth” which

288 is in common with the name of the relation *Growth parameter – pH* (score of the relation  
289 according to the table title: 0.218), while the table title has no common word with the relation  
290 *Product property – pH* (score of the relation according to the table title: 0) . The final score of  
291 the relation *Growth parameter – pH* for the table is computed as:  $1-(1-0.5)*(1-0.218) = 0.609$   
292 while the final score of the relation *Product property – pH* for the table is computed as:  $1-(1-$   
293  $0.5)*(1-0) = 0.5$ . The table is then annotated with the relation *Growth parameter – pH*.

## 294 **Experimental approach**

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

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

307 The quality of our method to distinguish between symbolic and numeric columns was  
308 assessed against a “naive” classifier: in that classifier, the units defined in the ontology, as  
309 well as the list of “no result indicators” are not used. In the naive classifier, a cell is  
310 considered as numeric if and only if it contains a number, and a column is numeric if at least  
311 half of its cells are numeric (else the column is symbolic). The quality of the rest of the  
312 annotations is assessed using two common measures: precision and recall. Precision is the

313 ratio of correct computed annotations over the total number of computed annotations (correct  
314 and wrong). Recall is the ratio of correct computed annotations over the number of manual  
315 annotations.

## 316 **Results and discussion**

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

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

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

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

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

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

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

## 354 **Conclusion and perspectives**

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

361 Our annotation system is entirely based on the use of a controlled vocabulary, called ontology,



362 which is used to represent the data. The richer the ontology is, the best the annotation will be,  
363 as our annotation algorithm uses word-by-word comparison between the terms used in the  
364 table and the terms already represented in the ontology. We are now considering the  
365 possibility of ontology enrichment to allow better annotation results, our method being easily  
366 adjustable to take into account the definition of synonyms.

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

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

## 379 **Acknowledgements**

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

## 383 **References**

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

386 Buche, P., Dervin, C., Haemmerlé, O., Thomopoulos, R., 2005. Fuzzy querying of  
387 incomplete, imprecise, and heterogeneously structured data in the relational model using  
388 ontologies and rules. *IEEE Transactions on Fuzzy Systems* 13, 373-383.

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

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

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

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

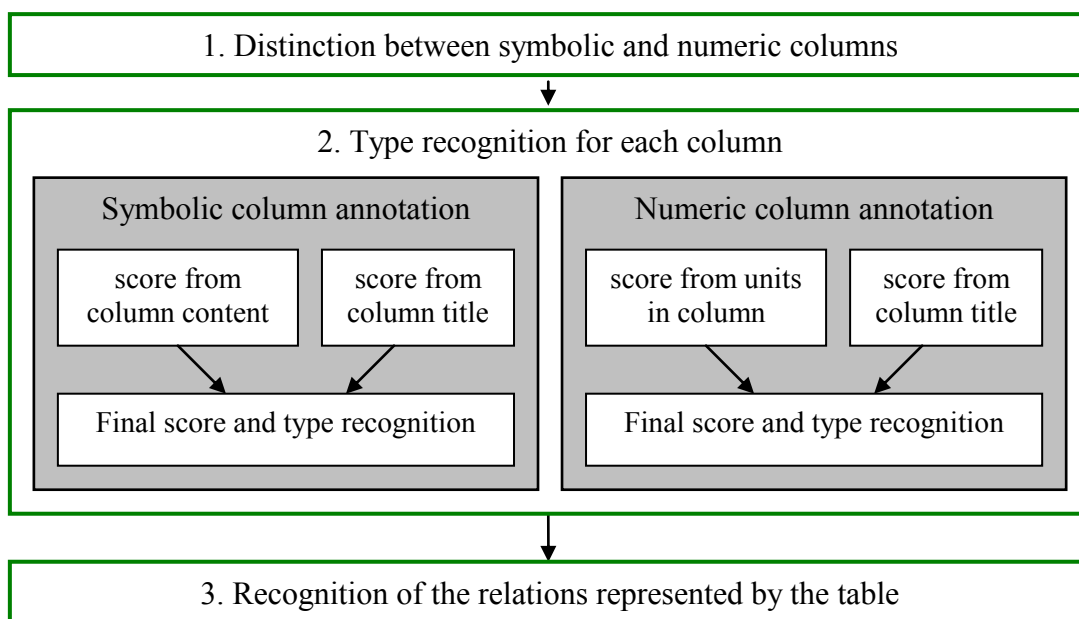
401 Lin, D., 1998. An information-theoretic definition of similarity. In: *ICML '98 : Proceedings*  
402 *of the Fifteenth International Conference on Machine Learning*, 296-304. Morgan Kaufmann  
403 *Publishers Inc., San Francisco*.

404 McMeekin, T.A., Baranyi, J., Bowman, J., Dalgaard, P., Kirk, M., Ross, T., Schmid, S.,  
405 Zwietering, M.H., 2006. Information systems in food safety management. *International*  
406 *Journal of Food Microbiology* 112, 181-94.

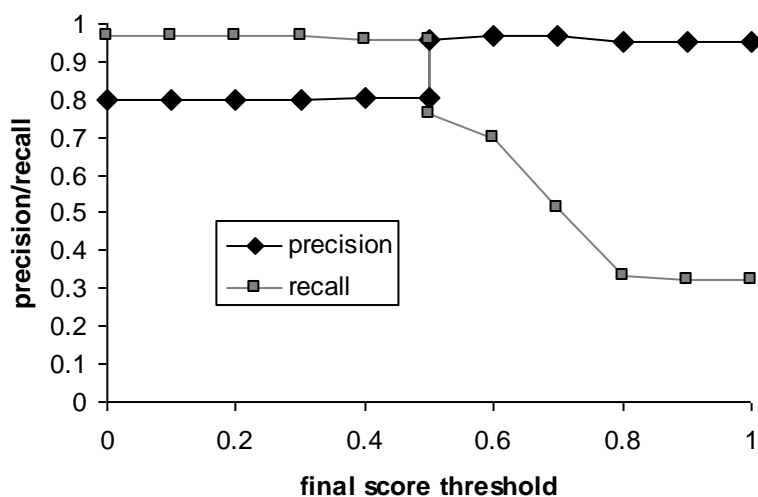
407 Zanibbi, R., Blostein, D., Cordy, J. R., 2004. A survey of table recognition : Models,  
408 observations, transformations, and inferences. *International Journal on Document Analysis*  
409 *and Recognition* 7, 1-16.

410

411 Figure 1: The different steps of our annotation algorithm



412  
413 Figure 2: Evolution of precision and recall on relation recognition according to the score  
414 threshold



415

416

417

418 Table 1: Numeric types of the ontology.

<b>type name</b>	<b>meaning</b>	<b>units</b>
aw Water Activity	water activity of the growth support	NONE
CO <sub>2</sub>	atmosphere concentration of CO <sub>2</sub>	%
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 <sub>2</sub>	atmosphere concentration of N <sub>2</sub>	%
NACL	NaCl concentration in the growth support	%
Number outbreaks or deaths	number of outbreaks or deaths due to a particular microorganism	NONE
O <sub>2</sub>	atmosphere concentration of O <sub>2</sub>	%
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, hr,mins
Year	year of event (outbreak, experiment...)	NONE
Ymax	Ymax parameter in microbial growth model	cfu

419

420 Table 2: Terms represented as weighted vectors.

	Meaning of the vector axis	ground	meat	fresh	beef
Term					
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

421

422 Table 3: Redox potentials on some foods.

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

423

424 Table 4: Reported prevalence of Campylobacter.

<i>Product</i>	<i>Positive for Campylobacter (%)</i>
Chicken products	0.07

425

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

<i>Strain</i>	<i>Minimum pH for growth</i>
284-72	5.5
T-3765-1	5.2

427

428 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%	

429

430

431 Table 7: Results of the annotation of symbolic columns.

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

432

433

434