Data reliability assessment in a data warehouse opened on the Web

Sébastien Destercke and Patrice Buche and Brigitte Charnomordic

UMR IATE and MISTEA, 2 place Viala, F-34060 Montpellier Cedex 1, France
LIRMM, CNRS-UM2, F-34392 Montpellier, France
destercke@cirad.fr,{buche,bch}@supagro.inra.fr

Abstract. This paper presents an ontology-driven workflow that feeds and queries a data warehouse opened on the Web. Data are extracted from data tables in Web documents. As web documents are very heterogeneous in nature, a key issue in this workflow is the ability to assess the reliability of retrieved data. We first recall the main steps of our method to annotate and query Web data tables driven by a domain ontology. Then we propose an original method to assess Web data table reliability from a set of criteria by the means of evidence theory. Finally, we show how we extend the workflow to integrate the reliability assessment step.

1 Introduction

The huge amount of technical and scientific documents available on the Web include many data tables. In addition to local data sources, they represent big potential external data sources for the data warehouse of a company dedicated to a given domain of application. To lighten the burden laid upon domain experts when selecting data from the data warehouse for a particular application, it is necessary to give them indicative reliability evaluations. In this paper, we present a framework to estimate the reliability of data tables collected from the Web. Compared to more *ad-hoc* estimation, the presented generic method can give insights to the expert as to why a particular data table is tagged as reliable or not reliable. Due to its generic nature, this method can be reused in other data warehouses using the semantic web recommended languages.

Reliability estimation is an essential part of the Semantic Web architecture, and many research works [1] focus on issues such as source authentification, reputation, etc. For example, [2] advocates a multi-faceted approach to trust models. They propose an OWL based ontology of trust related concepts. The idea is to provide systems using the annotation power of a user community to collect information about reliability. Our approach is different, as we do not rely on users but rather on information about the Web data table origins to compute a reliability estimations. Among methods proposing solutions to evaluate trust or data quality in web applications, the method presented in [3] is close to the method presented in the paper. It uses possibility theory evidence theory, whereas we base our method on evidence theory. Another difference is that in our approach global information is obtained by a fusion of multiple uncertainty models, while in [3] global information results from the propagation of uncertainty models

through a aggregation function. Each method has its pro and cons: it is easier to integrate interactions between criteria in aggregation functions, while it is easier to retrieve explanations of the final result in our approach.

In this paper, we details our method and its integration in @Web, along with the whole workflow used in @Web. The current version of @Web (see [4, 5]), a Web-enabled data warehouse, has been implemented using the W3C recommended languages (see [6] for details about these languages): OWL to represent the domain ontology, RDF to annotate Web tables and SPARQL to query annotated Web tables.

We first recall in Section 2 the purpose and architecture of the data warehouse. Section 3 details the proposed method to assess Web data table reliability. In Section 4, we show how this reliability assessment is presented and explained to the user. Finally, in Section 5, we explain how @Web is extended to implement the reliability management.

2 @Web presentation

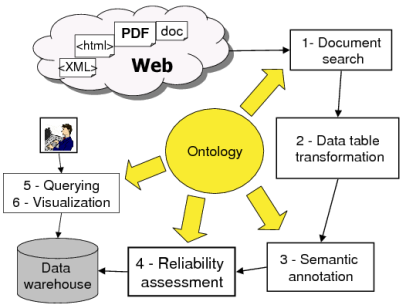

Fig. 1. Main steps of the document workflow in @Web

@Web is a data warehouse opened on the Web [4, 5] centered (in its current version) on the integration of heterogeneous data tables extracted from Web documents. The focus has been put on Web tables for two reasons: (i) experimental data are often summarized in tables, (ii) table structured data are easier to integrate than, e.g., in text or plots. The main steps of Web table integration are summarised in Fig. 1. A central role in data integration in @Web is played by the domain ontology. This ontology describes the concepts, their relations and the associated terminology of a given application domain. @Web can therefore be instantiated for any application domains (e.g., food predictive microbiology, food chemical risks, aeronautics [5]), provided a proper domain ontology is defined.

Once the ontology is built, @Web workflow includes the different steps shown in Fig. 1 to integrate new data in the warehouse. Concepts found in a data table and semantic relations linking these concepts are automatically identified. Data tables are then annotated with the identified concepts, allowing users to interrogate and query the data warehouse in an homogeneous way.

Our case study uses the @Web instance implemented in the Sym'Previus [7] decision support system whose aim is to simulate the growth of a pathogenic microorganism in a food product. Semantic relations in this system include for instance the *GrowthRate* that links a given microorganism within a given food product to a specific growth rate and its associated parameters. Data retrieved from tables can then be used to define the parameters of numerical growth oriented simulated models.

2.1 @Web generic ontology

The current OWL ontology representation used in the @Web system is composed of two main parts: a generic part, called *core ontology*, which contains the structuring concepts of the Web table integration task, and a specific part, commonly called *domain ontology*, which contains the concepts specific to the considered domain. The *core ontology* is composed of symbolic concepts, numeric concepts and relations between these concepts. It is separated from the definition of the concepts and relations specific to a given domain, i.e., the *domain ontology*. All the ontology concepts are materialized by OWL classes. For example, in the microbiological ontology, the respectively symbolic and numeric concepts *Microorganism* and *pH* are represented by OWL classes, respectively subclass of the generic classes *SymbolicConcept* and *NumericConcept*. An excerpt of an OWL class organization for symbolic concepts is given in Figure 2.

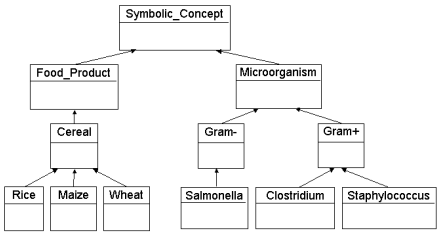

Fig. 2. Excerpt of OWL class hierarchy for symbolic concepts in the microbial domain

2.2 @Web workflow

The first three steps of @Web workflow (see Fig. 1) are as follows. The first task consists in retrieving relevant Web documents for the application domain, using key-words extracted from the domain ontology. It does so by defining queries executed by different crawlers. In the second task, data tables are extracted from the retrieved documents and are semi-automatically translated into a generic XML format. The Web tables are then represented in a classical and generic way – i.e., a set of lines, each line being a set of cells. In the third task, the Web tables are semantically annotated according to the domain ontology. The semantic annotation process of a Web table consists in identifying which semantic relations of the domain ontology can be recognized in each row of the Web table (see [5] for details). This process generates RDF descriptions.

Organism	a_w minimum	a_w optimum	a_w maximum
Clostridium	0.943	0.95-0.96	0.97
Staphylococcus	0.88	0.98	0.99
Salmonella	0.94	0.99	0.991

Fig. 3. Example of a Web table

Example 1. Fig. 3 presents an example of a Web table in which the semantic relation *GrowthParameterAwMin* has been identified. The domain of this relation is a kind of Microorganism and its range is food product water activity (a_w). The first row indicates that *Clostridium* requires a minimal food product a_w of 0.943 to be able to grow.

Example 2. Figure 4 presents the main part of the RDF descriptions corresponding to the recognition of the relation *GrowthParameterAwMin* in the first row (denoted *uriRow1*) of the Web table given by Fig. 3. Starting from the left part of the figure, we see that the row is annotated by the relation *GrowthParameterAwMin*, abbreviated as *GPaw1*. The domain of the relation *GrowthParameterAwMin* is an instance of the symbolic concept *Clostridium*. The range of the relation is an instance of the numerical concept *Aw* and has for value 0.943.

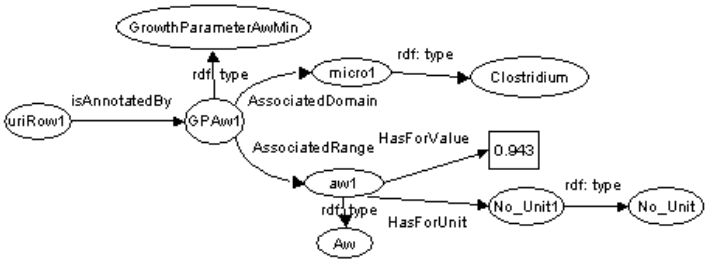

Fig. 4. Example of RDF annotations generated from the Web table of Figure 3

2.3 SPARQL querying of RDF graphs

In the XML/RDF data warehouse, the querying is done through MIEL++ queries. We briefly recall how MIEL++ queries are executed in the current version of @Web (For details, see [4]). A MIEL++ query is asked in a view that corresponds to a relation of the ontology (e.g., the relation *GrowthParameterAwMin*). A MIEL++ query is an instanciation of a view by the end-user, who specify among the set of querryable attributes of the view what are the selection attributes and their searched values, and what are the projection attributes (with the meaning of the relational model). An important specificity of a MIEL++ query is that searched values may be expressed as fuzzy sets (see [8–10]), which use allows end-users to represent their preferences in a gradual way.

Definition 1 A fuzzy set μ defined on a space $\mathscr{A}$ is a function $\mu : \mathscr{A} \rightarrow [0,1]$ with $\mu(x)$ the membership degree of x. The support $S(\mu)$ and the kernel $K(\mu)$ of a fuzzy sets are the sets $S(\mu) = \{x \in \mathscr{A} | \mu(x) > 0\}$ and $K(\mu) = \{x \in \mathscr{A} | \mu(x) = 1\}$.

Example 3. Let us define a MIEL++ query Q expressed in the view *GrowthParameter-AwMin* as follows:

$$Q = \{\text{Microorganism, aw} \mid (\text{GrowthParameterAwMin(Microorganism, aw)} \wedge (\text{Microorganism} \approx \text{MicroPreferences}) \wedge (\text{aw} \approx \text{awPreferences})\}.$$

The discrete fuzzy set MicroPreferences, which is equal to {(Gram+,1.0), (Gram-,0.5)}, means that the end-user is firstly interested in microorganisms which are Gram+ and secondly Gram-. The trapezoidal fuzzy set awPreferences that has the characteristic points [0.9, 0.94, 0.97, 0.99], means that the end-user is first interested in aw values in the interval [0.94, 0.97] (the kernel of the fuzzy set), but that he/she accepts to enlarge the querying till the interval [0.9, 0.99] (the support of the fuzzy set).

Since fuzzy sets are not supported in a standard SPARQL query, a complete solution to translate a MIEL++ query into a standard SPARQL query is presented in detail in [4]. In this paper, we only recall how is measured the satisfaction of a MIEL++ query. The satisfaction of a selection criterion $att \approx attPref$ is measured by the membership degree $\mu_{attPref}(x)$ of the corresponding value x expressed in the RDF graph (x is supposed to be a crisp value in this paper). As selection criteria are considered to be conjunctive, a global adequation degree, denoted *ad*, is computed using the t-norm *min*.

Example 4. The answers to the SPARQL query associated with the MIEL++ query of Example 3 compared with the Web table presented in Figure 3 is given below:

ad	$\mu_{MicroPref}(x)$	$\mu_{AwPref}(x)$	*Microorg*	*aw*
1.0	1.0	1.0	*Clostridium*	0.943
0.5	0.5	1.0	*Salmonella*	0.94
0.0	1.0	0.0	*Staphilococcus*	0.88

3 Reliability evaluation

This section describes the method we propose to evaluate the reliability of Web tables.

3.1 A model for reliability evaluation

We assume that reliability takes its value on a finite ordered space $\Theta = \theta_1, \ldots, \theta_N$ such that $\theta_i < \theta_j$ iff $i < j$. θ_1 corresponds to total unreliability, while θ_N corresponds to total reliability. We denote by $I_{a,b} = \{\theta_a, \ldots, \theta_b\}$ a set such that $a \leq b$ and $\forall\ c$ s.t. $a \leq c \leq b$, $\theta_c \in I_{a,b}$. Such sets include all values between their minimum value θ_a and maximum value θ_b, and using a slight stretch of language we call them *intervals*.

The evaluation will be based on the values taken by S groups $A_1, \ldots, A_S$ of criteria. Note that a group may be composed of multiple criteria, e.g., number of citation $\times$ publication date.The group constitution ensures that the impact of each group A_i on the

reliability evaluation can be judged (almost) independent of the impact of any other group A_j. Each group A_i can assume C_i distinct values on spaces $\mathscr{A}_i = \{a_{i1}, \ldots, a_{iC_i}\}$.

For each possible value of each criteria group $A_1, \ldots, A_S$, a domain expert is asked to give its opinion about the corresponding data reliability. To facilitate expert elicitation, a linguistic scale with a reasonable number of terms is used, for instance the five terms *very unreliable*, *slightly unreliable*, *neutral*, *slightly reliable* and *very reliable*. These opinions are modeled as fuzzy sets that describes some ill-known value of reliability.

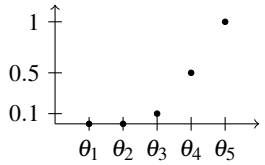

Fig. 5. Fuzzy set corresponding to the term *very reliable* defined on Θ with $N = 5$.

Denote by $\mathscr{F}(\Theta)$ the set of all fuzzy sets defined over a domain Θ. For each group A_i, we define a mapping $\Gamma_{A_i} : \mathscr{A}_i \to \mathscr{F}(\Theta)$ according to the expert opinions, such that $\Gamma_{A_i}(a)$ with $a \in \mathscr{A}_i$ is the interpretation on Θ of the information provided by $A_i = a$ about the reliability. We denote by μ_a the fuzzy set $\Gamma_{A_i}(a)$.

The expert may select from a limited number of linguistic terms as well as combination of them, using "or" disjunctions[1]. An additional term allows to express total ignorance. A fuzzy set on Θ is then associated to each term. Fig. 5 provides an illustration of a fuzzy set corresponding to the term *very reliable*.

Example 5. Consider the two groups $A_1 =$ *source type* and $A_2 =$ *experience repetition* such that

$$A_1 = \{a_{11} = \text{journal paper}, a_{12} = \text{governmental report}, a_{13} = \text{project report}, a_{14} = \text{other}\}$$
$$A_2 = \{a_{21} = \text{repetitions}, a_{22} = \text{no repetition}\}$$

The expert then provides his opinion about the reliability value for the different values of these two criteria. These opinions are summarised below

$$\mu_{a_{11}} = \text{very reliable}, \mu_{a_{12}} = \text{slightly reliable}, \mu_{a_{13}} = \text{neutral}, \mu_{a_{14}} = \text{slightly unreliable};$$
$$\mu_{a_{21}} = \text{very reliable}, \mu_{a_{22}} = \text{slightly unreliable}.$$

3.2 Global reliability information through merging

For a given data table each group A_i takes a particular value, hence S different fuzzy sets are provided as pieces of information. We propose to use evidence theory [11] to merge these information in a global reliability assessment. Indeed, this theory comes with a rich choice of merging rules [12], together with a good compromise between

[1] In this case, fuzzy sets are combined by the classical t-conorm max

expressiveness and tractability. It encompasses fuzzy sets and probability distributions as special cases. We recall here the basics of the theory and its links with fuzzy sets.

A basic belief assignment (*bba*) m on a space Θ is a mapping from the power set $2^{|\Theta|}$ of Θ onto the unit interval $[0,1]$, such that $\sum_{E\subseteq\Theta} m(E) = 1$ and $m(\emptyset) = 0$. Sets E such that $m(E) > 0$ are called **focal elements**. We denote by $\mathscr{F}_m$ the set of focal elements of m. The mass $m(E)$ can be interpreted as the probability that the most precise description of what is known about a particular situation is of the form "$x \in E$". From this mass assignment, Shafer [11] defines two set functions, called *belief and plausibility functions*, for any event $A \subseteq \Theta$:

$$Bel(A) = \sum_{E,E\subseteq A} m(E); \qquad Pl(A) = 1 - Bel(A^c) = \sum_{E,E\cap A\neq\emptyset} m(E),$$

where the belief function measures the certainty of A (i.e., sums all masses that cannot be distributed outside A) and the plausibility function measures the plausibility of A (i.e., sums all masses that it is possible to distribute inside A).

A fuzzy set μ with M distinct membership degrees $1 = \alpha_1 > \ldots > \alpha_M > \alpha_{M+1} = 0$ defines a *bba* m having, for $i = 1,\ldots,M$, the focal elements E_i with masses $m(E_i)$ [13]:

$$\begin{cases} E_i = \{\theta \in \Theta | \mu(\theta) \geq \alpha_i\} = A_{\alpha_i}, \\ m(E_i) = \alpha_i - \alpha_{i+1}. \end{cases} \tag{1}$$

Therefore, each fuzzy set provided by experts during information collection can be mapped into an equivalent *bba*.

Example 6. Consider the fuzzy set depicted in Fig. 5. Its equivalent *bba* m is such that

$$m(E_1 = \{\theta_5\}) = 0.5, \quad m(E_2 = \{\theta_4, \theta_5\}) = 0.4, m(E_3 = \{\theta_3, \theta_4, \theta_5\}) = 0.1.$$

When S groups of criteria (called sources in the sequel) provide pieces of information modelled as *bbas* $m_1,\ldots,m_S$ over a same space Θ, it is necessary to merge them into a global model. Two main issues related to merging rules are the handling of (i) dependence [14] and of (ii) conflict [12] between sources.

Here, sources are selected to remain as independent as possible, therefore tackling the first issue. We are thus left with the problem of properly handling conflicting information. Given the fact that sources are independent, the merging of *bbas* $m_1,\ldots,m_S$ can be written,

$$\forall E \subseteq \Theta \quad m(E) = \sum_{E_i\in\mathscr{F}_i}^{\oplus_{i=1}^S(E_i)=E} \prod_{i=1}^{S} m_i(E_i), \tag{2}$$

with $\mathscr{F}_i$ the focal elements of m_i, and $\oplus_{i=1}^S(E_i) = E$ an aggregation operator on sets. Note that TBM conjunctive rule and the disjunctive rule [12] are retrieved when $\oplus = \cap$ and $\oplus = \cup$, respectively. However, the former is not adapted to the case of conflicting information, while the latter often results in a very imprecise model.

To deal with the problem of conflicting information, we propose a merging strategy based on maximal coherent subsets (MCS). Given a set of conflicting sources, MCS consists in applying a conjunctive operator within each non-conflicting subset of

sources, and then using a disjunctive operator between the partial results [15]. Consider $N = \{I_{a_1,b_1}, \ldots, I_{a_k,b_k}\}$ a set of k intervals. Using the MCS method on such intervals consists in taking the intersection over subsets $\overline{K_j} \subset N$ s.t. $\cap_{i \in K_j} I_{a_i,b_i} \neq \emptyset$ that are maximal with this property, and then in considering the union of these intersections as the final result (i.e. $\cup_j \cap_{i \in K_j} I_{a_i,b_i}$). We denote by $\oplus_{MCS}$ the MCS aggregation operator. In general, detecting MCS is NP-hard, however in the case of intervals over an ordered space (our case here), the algorithm proposed in [15] reduce this complexity drastically.

An application of MCS on four (real valued) intervals I_1, I_2, I_3, I_4 is shown in Fig. 6. The two MCS are (I_1, I_2) and (I_2, I_3, I_4) and the final result is $(I_1 \cap I_2) \cup (I_2 \cap I_3 \cap I_4)$. Note that, if all intervals are consistent, conjunctive merging is retrieved, while disjunction is retrieved when every pair of intervals conflicts. As we shall see, the groups of intervals forming maximal coherent subsets may be used as elements explaining the result. Applying MCS in our case comes down to apply Eq. (2) with $\oplus = \oplus_{MCS}$ to the fuzzy sets $\mu_{a_{ij}}$, $a_{ij} \in A_i$ once they have been transformed into bbas (thanks to Eq. (1)).

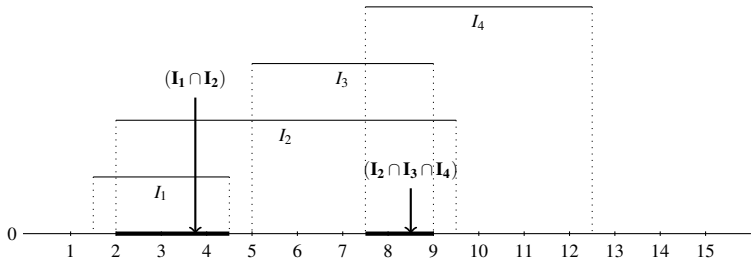

Fig. 6. Illustration of maximal coherent subsets merging.

Example 7. Consider the two groups of Example 5. Now, assume that the retrieved data come from a journal paper ($A_1 = a_{11}$) but that the experiment has not been repeated ($A_2 = a_{22}$). Value a_{11} corresponds to "very reliable", while a_{22} corresponds to "slightly unreliable". Each group A_i thus provides an individual *bba* corresponding to the criterion value. These *bbas* are given in the following table:

	$E_{11} = \{\theta_3, \theta_4, \theta_5\}$	$E_{12} = \{\theta_4, \theta_5\}$	$E_{13} = \{\theta_5\}$
$m_{a_{11}}$	0.1	0.4	0.5

	$E_{21} = \{\theta_1, \theta_2, \theta_3\}$	$E_{22} = \{\theta_2, \theta_3\}$	$E_{23} = \{\theta_2\}$
$m_{a_{22}}$	0.1	0.4	0.5

We denote by j_i the index of the criterion value for a given data item, and m_g the *bba* obtained by merging $m_{a_{1j_1}}, \ldots, m_{a_{Sj_S}}$ through Equation (2) with $\oplus = \oplus_{MCS}$. The merging result of the *bbas* given in Example 7 is summarised in Table 1.

4 Reliability presentation and explanation

A look at Table 1 tells us that the merging result is hard to read, and that it is necessary to provide tools that summarize this information in a digestible representation. Given a set

Criteria in MCS	MCS focal sets	Focal set	Mass of focal set
$\{A_1, A_2\}$	$E_{11} \cap E_{21} + E_{11} \cap E_{22}$	$\{\theta_3\}$	0.05
$\{A_1\}$ and $\{A_2\}$	$E_{11} \cup E_{23} + E_{12} \cup E_{22}$	$\{\theta_2, \ldots, \theta_5\}$	0.21
$\{A_1\}$ and $\{A_2\}$	$E_{12} \cup E_{21}$	$\{\theta_1, \ldots, \theta_5\}$	0.04
$\{A_1\}$ and $\{A_2\}$	$E_{12} \cup E_{23}$	$\{\theta_2, \theta_4, \theta_5\}$	0.20
$\{A_1\}$ and $\{A_2\}$	$E_{13} \cup E_{21}$	$\{\theta_1, \theta_2, \theta_3, \theta_5\}$	0.05
$\{A_1\}$ and $\{A_2\}$	$E_{13} \cup E_{23}$	$\{\theta_2, \theta_5\}$	0.25
$\{A_1\}$ and $\{A_2\}$	$E_{13} \cup E_{22}$	$\{\theta_2, \theta_3, \theta_5\}$	0.2

Table 1. Example of merging independent information using MCS

$D = \{e_1, \ldots, e_d\}$ of d data, we propose three complementary means to summarise their reliability evaluations: by ordering them, by providing a summarising (quantitative) interval and by explaining the main reasons for the reliability evaluation.

In this section, we use the notion of lower and upper expectations of a function $f : \Theta \to \mathbb{R}$ induced by a *bba* m_g. These lower and upper expectations are defined as

$$\underline{\mathbb{E}}_g(f) = \sum_{A \subseteq \Theta} m(A) \min_{\theta \in A} f(\theta) \quad \text{and} \quad \overline{\mathbb{E}}_g(f) = \sum_{A \subseteq \Theta} m(A) \max_{\theta \in A} f(\theta). \tag{3}$$

They correspond to the infimum and supremum values of all expectations of f w.r.t. probability measures dominating the belief function induced by m_g.

4.1 Comparing, evaluating and ordering data

Let $m_{g_1}, \ldots, m_{g_d}$ be the global *bbas* representing our knowledge about the reliability of $e_1, \ldots, e_d$. We propose to induce an order between them by using numerical comparison of interval-valued estimations, using a particular function in Eq. (3). We propose to consider $f_\Theta : \Theta \to \mathbb{R}$ such that $f_\Theta(\theta_i) = i$ (each θ_i receives its rank as value), and to summarize the reliability of data item e_i by the interval $[\underline{\mathbb{E}}_{g_i}(f_\Theta), \overline{\mathbb{E}}_{g_i}(f_\Theta)]$ (obtained by using Eq. (3)).

Example 8. Consider the three *bbas* $m_{g_1}, m_{g_2}, m_{g_3}$ respectively representing the reliability of e_1, e_2, e_3 (e.g. resulting from the merging process illustrated in Example 7), defined over $\Theta = \{\theta_1, \ldots, \theta_5\}$ such that

$$m_{g_1}(\{\theta_1, \theta_2, \theta_3\}) = 0.3, m_{g_1}(\{\theta_2, \theta_3\}) = 0.7; \quad m_{g_2}(\{\theta_3, \theta_4\}) = 0.5, m_{g_2}(\{\theta_4, \theta_5\}) = 0.5;$$
$$m_{g_3}(\{\theta_1\}) = 0.4, m_{g_3}(\{\theta_5\}) = 0.4, m_{g_3}(\{\Theta\}) = 0.2.$$

Corresponding reliability intervals are:

$$[\underline{\mathbb{E}}_{g_1}(f_\Theta), \overline{\mathbb{E}}_{g_1}(f_\Theta)] = [1.7, 3]; [\underline{\mathbb{E}}_{g_2}(f_\Theta), \overline{\mathbb{E}}_{g_2}(f_\Theta)] = [3.5, 4.5];$$
$$[\underline{\mathbb{E}}_{g_3}(f_\Theta), \overline{\mathbb{E}}_{g_3}(f_\Theta)] = [2.6, 3.4].$$

Partial order: We propose to order the *bbas* according to the (partial) order $\leq_{\mathbb{E}}$ s.t. $m_g \leq_{\mathbb{E}} m_{g'}$ iff $\underline{\mathbb{E}}_g(f_\Theta) \leq \underline{\mathbb{E}}_{g'}(f_\Theta)$ and $\overline{\mathbb{E}}_g(f_\Theta) \leq \overline{\mathbb{E}}_{g'}(f_\Theta)$. In Example 8, we have $e_1 <_{\mathbb{E}} e_3 <_{\mathbb{E}} e_2$ (further on, we make no difference between a datum e_i and its bba m_{g_i}), obtaining in this case a complete order among the objects. However, as $\leq_{\mathbb{E}}$ is in general a partial order, we propose an algorithm allowing to build from it a complete (pre-)order, so that users are provided with an ordered list, easier to understand and interpret.

Building groups: The next step is to order data by groups of decreasing reliability according to the order $\leq_{\mathbb{E}}$, i.e., to build an ordered partition $\{D_1,\ldots,D_O\}$ of D, where D_1 corresponds to the most reliable data. Given a subset $F \subseteq \{e_1,\ldots,e_d\}$, denote by $opt(\mathbb{E},F)$ the set of optimal data in the sense of reliability, i.e. not dominated w.r.t. $\leq_{\mathbb{E}}$

$$opt(\mathbb{E},F) = \{e_i \in F |\ \nexists e_j \in F,\ such\ as\ e_i \leq_{\mathbb{E}} e_j\}.$$

The partition $\{D_1,\ldots,D_O\}$ can now be defined recursively as follows:

$$D_i = opt(\mathbb{E}, (\{e_1,\ldots,e_d\} \setminus \bigcup_{j=0}^{i-1} D_j)) \text{ with } D_0 = \emptyset. \quad (4)$$

4.2 Explaining the results

Another interest of MCS is that they give insights about the reasons that have led to a particular reliability assessment, providing the user with some possibly useful explanations. Indeed, according to our method, the more often a subgroup F of MCS appears in $\oplus_{i=1}^{S}(E_i) = E$ (see Eq. (2)), the more important its impact is on the global reliability score m_g. Therefore, we propose to measure the importance $w(F)$ of a MCS F in m_g by summing all the masses of m_g for which it has been a maximal coherent subset, that is

$$w(F) = \{\sum \prod_{i=1}^{S} m_i(E_i) | F \text{ is an MCS of } \oplus_{i=1}^{S}(E_i)\}$$

Example 9. In Table 1, the impact w of the different encountered MCS is evaluated as follows: $w(\{A_1\}) = 0.95$, $w(\{A_2\}) = 0.95$, $w(\{A_1,A_2\}) = 0.05$, from which it can be inferred that the two criteria $\{A_1\}$ and $\{A_2\}$ appear often alone and do not agree with each other. This means that the imprecision in the final reliability representation can be explained by the conflict between criteria A_1 and A_2.

In this example, the analysis is straigthforward. However, when dealing with thousands of data and half a dozen of criteria groups, such tools may help users to perform a quick analysis and retain the data that best serve their purposes.

5 Extending @Web for data reliability management

This section describes the change made to @Web to add a reliability estimation to each Web table and to use them in the display of a user query result. As data from a same table often come from a same experiment, table level has been retained to model reliability.

5.1 Extending the ontology to include reliability criteria

Some criteria retained for reliability estimation are part of the domain knowledge. For example, measurements methods to count micro-organisms all roughly have the same precision, while the accuracy of methods to appreciate wheat grain size greatly varies.

Therefore, it is natural to include criteria in the domain ontology. This solution allows designers to adapt the choice of the criteria associated with each domain of application, preserving the @Web generic approach at the same time.

In the extended version of the ontology integrating reliability criteria, the *core ontology* is enriched with corresponding symbolic and numeric criteria. The *domain ontology* is completed by the definition of the criteria selected to evaluate the reliability, together with their possible values. For example, the respectively symbolic and numeric criteria *SourceType* and *CitationNumber* are represented by OWL classes and belong to the *domain ontology*. They are subclasses of the generic classes *SymbolicCriterion* and *NumericCriterion*, respectively, which belong to the *core ontology*. As for symbolic concepts, the possible values associated with a symbolic criterion are represented by OWL classes that are subclasses of the OWL class representing the criterion.

5.2 Storing data reliability criteria in RDF graphs

In extended @Web, an additional fourth task concerns the reliability management (see Figure 1). Users manually enter the values associated with the reliability criteria for each Web table. This information is stored in a RDF graph associated with the table.

Example 10. Fig. 7 presents the RDF descriptions representing the reliability criteria and values associated with the Web table of Fig. 3. They express that the table (having the *uriTable1* identifier within the XML document) has for associated criteria the same values than in Example 7: journal paper and no repetitions of experiments.

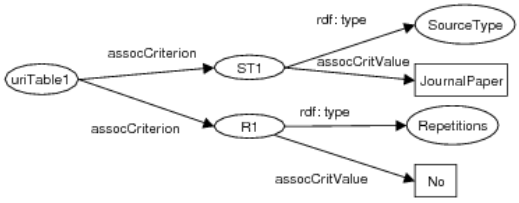

Fig. 7. Example of RDF annotations associated with the Web table of Figure 3

The fourth task output of the extended @Web system is an XML/RDF data warehouse composed of a set of XML documents which represent Web tables, together with the RDF annotations corresponding to the recognized semantic relations and the reliability criteria values.

5.3 SPARQL queries and data reliability

To evaluate the reliability of the answers associated with a MIEL++ query, the following post-processing is executed. Reliability criteria associated to a Web table are retrieved thanks to SPARQL queries (generated by using the ontology). Each answer associated with a given row of a Web table is then associated to its reliability interval thanks to its URI which links it to its original table. Answers are then compared and ordered according to methods of Sec. 4.

6 Conclusion and perspectives

In this paper, we have proposed a method that evaluates reliability of Web data tables by using sets of criteria concerning the data origins. This method, based on evidence theory, is generic and can be applied to any domain once proper criteria have been defined. Special attention has been given to tractability and ease of use. A first possible perspective of this work should be to take account of possible uncertainty in the criteria values. In the present paper, we have considered that criteria were known. It would be desirable to consider the case where some criteria are ill-known (using *bbas* to describe this uncertainty). A second possible perspective would be to extend our approach to cope with multiple experts providing (possibly) different opinions about the same criteria.

References

1. Gil, Y., Artz, D.: Towards content trust of web resources. In: WWW'06 : Proceedings of the 15th international conference on World Wide Web, New York, NY, USA (2006) 565–574
2. Quinn, K., Lewis, D., O'Sullivan, D., Wade, V.: An analysis of accuracy experiments carried out over a multi-faceted model of trust. Int. J. of Information Security **8** (2009) 103–119
3. Denguir-Rekik, A., Montmain, J., Mauris, G.: A possibilistic-valued multi-criteria decision-making support for marketing activities in e-commerce: Feedback based diagnosis system. European Journal of Operational Research **195**(3) (2009) 876–888
4. Buche, P., Dibie-Barthélemy, J., Chebil, H.: Flexible sparql querying of web data tables driven by an ontology. In: Proceedings of FQAS. Volume 5822 of Lecture Notes in Computer Science. (2009) 345–357
5. Hignette, G., Buche, P., Dibie-Barthélemy, J., Haemmerlé, O.: Fuzzy annotation of web data tables driven by a domain ontology. In: Proceedings of ESWC. Volume 5554 of Lecture Notes in Computer Science. (2009) 638–653
6. Antoniou, G., van Harmelen, F.: A semantic Web primer. The MIT Press, Cambridge, Massachusetts (2008)
7. Buche, P., Couvert, O., Dibie-Barthélemy, J., Hignette, G., Mettler, E., Soler, L.: Flexible querying of web data to simulate bacterial growth in food. Food Microbiology **28**(4) (2011) 685–693
8. Zadeh, L.: Fuzzy sets. Information and control **8** (1965) 338–353
9. Buche, P., Haemmerlé, O.: Towards a unified querying system of both structured and semi-structured imprecise data using fuzzy views. In: Proceedings of the 8th International Conference on Conceptual Structures. Volume 1867 of Lecture Notes in Artificial Intelligence. (2000) 207–220
10. Thomopoulos, R., Buche, P., Haemmerlé, O.: Different kinds of comparison between fuzzy conceptual graphs. In: Proceedings of the 11th International Conference on Conceptual Structures. Volume 2746 of Lecture Notes in Artificial Intelligence. (2003) 54–68
11. Shafer, G.: A mathematical Theory of Evidence. Princeton University Press (1976)
12. Smets, P.: Analyzing the combination of conflicting belief functions. Information Fusion **8** (2006) 387–412
13. Dubois, D., Prade, H.: On several representations of an uncertain body of evidence. In Gupta, M., Sanchez, E., eds.: Fuzzy Information and Decision Processes. North-Holland (1982) 167–181
14. Denoeux, T.: Conjunctive and disjunctive combination of belief functions induced by non-distinct bodies of evidence. Artificial Intelligence **172** (2008) 234–264
15. Dubois, D., Fargier, H., Prade, H.: Multi-source information fusion: a way to cope with incoherences. In Cepadues, ed.: Proc. of French Days on Fuzzy Logic (LFA), Cepadues (2000) 123–130