Journal of Computer and Communications, 2014, 2, 93-102
Published Online July 2014 in SciRes. http://www.scirp.org/journal/jcc
http://dx.doi.org/10.4236/jcc.2014.29013

Semantic Recognition of a Data Structure in Big-Data

Aïcha Ben Salem[1,2], Faouzi Boufares[1], Sebastiao Correia[2]

[1]Laboratory LIPN-UMR 7030-CNRS, University Paris 13, Sorbonne Paris Cité, Villetaneuse, France
[2]Company Talend, Suresnes, France
Email: bensalem@lipn.univ-paris13.fr, boufares@lipn.univ-paris13.fr, abensalem@talend.com, scorreia@talend.com

Received April 2014

Abstract

Data governance is a subject that is becoming increasingly important in business and government. In fact, good governance data allows improved interactions between employees of one or more organizations. Data quality represents a great challenge because the cost of non-quality can be very high. Therefore the use of data quality becomes an absolute necessity within an organization. To improve the data quality in a Big-Data source, our purpose, in this paper, is to add semantics to data and help user to recognize the Big-Data schema. The originality of this approach lies in the semantic aspect it offers. It detects issues in data and proposes a data schema by applying a semantic data profiling.

Keywords

Data Quality, Big-Data, Semantic Data Profiling, Data Dictionary, Regular Expressions, Ontology

1. Introduction

The general management and business managers must have a unified vision and usable information to make the right decisions at the right time. The data quality governance has become an important topic in companies. Its purpose is to provide accurate, comprehensive, timely and consistent data by implementing understandable indicators, easy to communicate, inexpensive and simple to calculate. In the big-data era, the quality of the information contained in a variety of data sources, is becoming a real challenge.

Data quality and semantics aspects are rarely joined in the literature [1]-[3]. Our challenge is to use semantics to improve the data quality. Indeed, misunderstanding of the data schema is an obstacle to define a good strategy to correct any anomalies in the data. Very often metadata are not enough for understanding the meaning of data.

For a given data source S, we propose a semantic data profiling to get better understanding of the data definition and improve anomalies detection and correction. No schema available to understand the meaning of data and even less to correct them. There are currently no tools [4]-[8] that bring the strings "Pékin" to "Beijing" or even "Londres" to "London". Additional semantic information is needed to know that these strings represent the same category and subcategory of information. Similarly, it is important to recognize semantically the meaning

How to cite this paper: Salem, A.B., Boufares, F. and Correia, S. (2014) Semantic Recognition of a Data Structure in Big-Data. *Journal of Computer and Communications*, **2**, 93-102. http://dx.doi.org/10.4236/jcc.2014.29013

of the string “16˚C” which is a city temperature in degree Celsius.

Let S be an unstructured data source, result of integration of multiple heterogeneous data sources. S can be seen as a set of strings, separated by semicolons (;). S can then be described by the set C of all its columns. One note S(C) the data schema. Notice that the source S has no defined structure, which can cause a problem for semantic data manipulation. S may contain inconsistences (**Figure 1**). Several questions arise such as: 1) what are the semantics of strings? 2) What are the languages used? 3) What is invalid and what is not?

Let us remark that this source has several columns. S is defined by (Col_i, i = 1;7).

In the data source S, the column Col4 should contain only cities given in English. London and Beijing are syntactically and semantically valid. While, “Pékin” and “Londres” are syntactically correct and semantically invalid. “Londre” is syntactically invalid. The Col2 column contains mostly dates. Therefore, the “13” value will be considered semantically invalid. This demonstrates the need of more semantics to understand and correct the data.

This paper is organized as follows. The second section presents the meta-information required for the semantic data structure. The semantic data profiling process is given in the third section. Our contribution and future works are given in conclusion.

2. Meta-Information

We discussed in the previous works [9] [10] various problems of data quality in particular the deduplication one. We started the development of a new kind of Big-Data ETL based on semantic aspects. It allows data profiling, data cleaning and data enrichment.

To assist the user in his quality approach, the originality of our work lies in: semantic recognition of descriptive data schema and hence fortification data themselves. We will focus, in this paper, to the data profiling step.

Data profiling presents the first step in the data quality process (DQM tool **Figure 2**). It is a quantitative analysis of the data source to identify data quality problems. It includes descriptive information such as schema, table, domain and data sources definitions. As a result, data profiling collects summaries of the data source (Number of records, attributes) [11] [12].

However, existing data profiling tools [13]-[16] provide a statistical data profiling and do not address the semantic aspects. For that, the purpose of this paper is to introduce some semantic indicators to enrich the data profiling process and propose a semantic one.

For the semantic data profiling, we propose for each input data source S, a bug report, log for updates and a new semantic structure using some meta-information.

The bug report contains the various existing anomalies in the data source: more than one category and language used for the same column, different data formats, duplicates, null values.

Log for updates is the set of update actions to be applied to a data source such as translation in the same language, homogenization in the same format. These updates cover one column at a time. In order to make corrections

S’	(Adam; 12/11/2001; Oxford street; London; UK; 16°C; 61°F)
⊆	(Eve; 30/02/2012; Fortune Plaza; Beijing; China; 25°C; 77°F)
	(Adamsss; -; Middle Rd.; Pékin; Chines; 25°C; 77°F)
S	(Jean; 8-9-2011; -; Londres; Royaume-Uni; 16°C; 61°F)
	(Adem; 13; M; Londre; RU ; 16°C; 61°F)

Figure 1. A sample of the data source S.

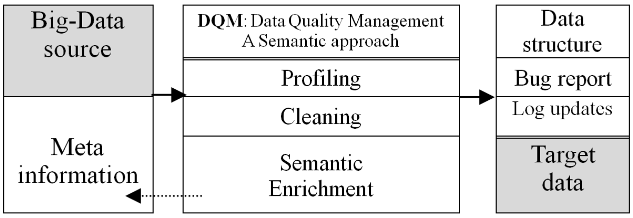


Figure 2. The DQM tool.

between columns, the concept of functional dependencies has to be applied.

This meta-information can be enriched over the time (more details will be presented in the Section 3.3).

In the following, we will be interested in the semantic data profiling process details (presented in **Figure 3**) and in particular to the meta-information.

The meta-information consists of three components: the Meta-Schema-Ontology (MSO), the Meta-Repository (MR) composed by the DD and RE and the list I of indicators.

Several tables (T_k, k = 1,7) are used to store the different artefacts corresponding to the results of the semantic data profiling process.

Let us start by defining the first component, the Meta-Schema-Ontology (MSO).

2.1. Meta-Schema-Ontology

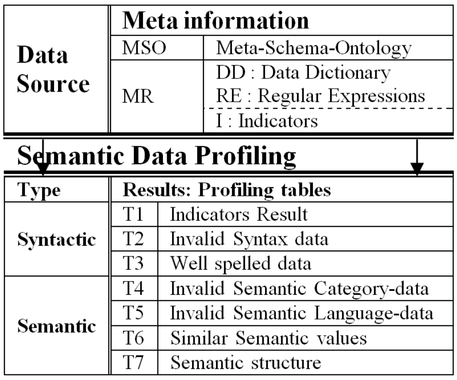
A database, as a set of information, can be described in many different ways. The difference is mainly in the name of concepts and attributes.

The idea with the MSO is to store all these equivalent descriptions in a meta-structure. The Meta is presented with the UML [17] (Unified Modeling Language) class diagram (**Figure 4**).

MSO is a set of knowledge that can be managed as ontologies [18]-[22]. Ontology is a formal language. It is a grammar that defines how terms may be used together. Ontologies allow sharing a common understanding of the information structure among people.

Many instances (knowledge) can be created from the MSO. For instance, *Person*, *Organization* and *Invoice* are three **Concepts**. Each of them may have several synonyms.

For instance, the concept *Person* can have many synonyms such as *Client*, *Student* and *Customer*. The concept *Person* is defined by some **Attributes** like *FirstName*, *Address*, *City*, *Country* and *BirthDate*. This implies that each synonym of the concept *Person* can be defined in a similar manner. The ontology is viewed with the



Data Source	Meta information	
	MSO	Meta-Schema-Ontology
	MR	DD : Data Dictionary RE : Regular Expressions
		I : Indicators

Semantic Data Profiling

Type	Results: Profiling tables	
Syntactic	T1	Indicators Result
	T2	Invalid Syntax data
	T3	Well spelled data
Semantic	T4	Invalid Semantic Category-data
	T5	Invalid Semantic Language-data
	T6	Similar Semantic values
	T7	Semantic structure

Figure 3. The semantic data profiling process.

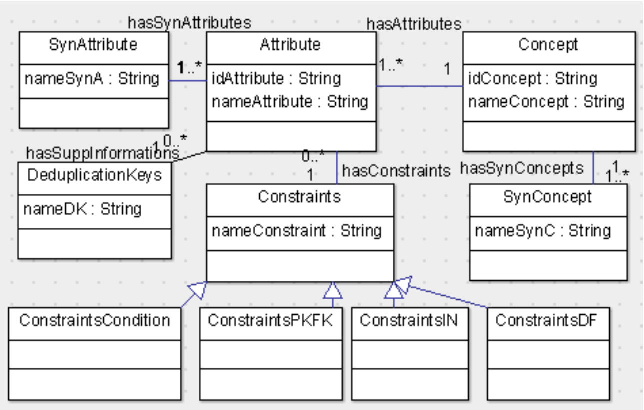


Figure 4. The Meta-Schema-Ontology UML class diagram.

open source Protégé tool [23] (**Figure 5**).

This knowledge can evolve over the time according to different descriptions of the databases and it can be represented as a meta-repository.

2.2. Meta-Repository

The meta-repository is a set of knowledge describing the data dictionary (set of categories in different languages), regular expressions and a list of indicators (statistical, syntactic and semantic).

2.2.1. Data Dictionary

Valid strings (syntactically and semantically) can be grouped into categories. Categories describe concepts. These descriptions (strings) can be in several languages. They may also contain sub-categories. The set of categories Catext can be seen as a data dictionary. For example, the monument category will contain all valid strings describing the airports, universities, hospitals, museums and castles names. The names of cities, countries and continents where are these monuments, are also part of data dictionary (DD).

Let Cat_{ext} be the set of categories defined by extension: Cat_{ext} = {Cat_i, i = 1;n} with Cat_i belongs to {FirstName, Country, City, Civility, Gender, Email, Web Site, Phone Number}. For each Cat_i, a set of sub-categories SubCat = {Cat_{ij}, j = 1;m} can be defined. In this study, language is used as a sub-category. The set of languages used is *Lang* = {English, French, German, Italian, Portuguese, Spanish}.

We define the DD as a set of triplets of (Category, Information, Language). A category Cat_i is then defined by extension where Information is a valid string, Category $\in$ Cat_{ext} and Language $\in$ Lang.

Note that, as mentioned in the **Figure 6**, the information "France" can refer to two categories in the same time: Country and FirstName. Other exceptions may exist.

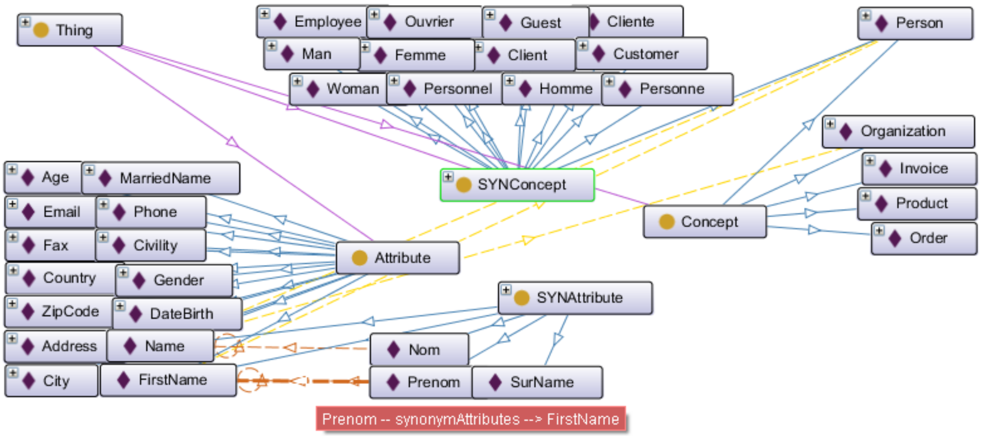
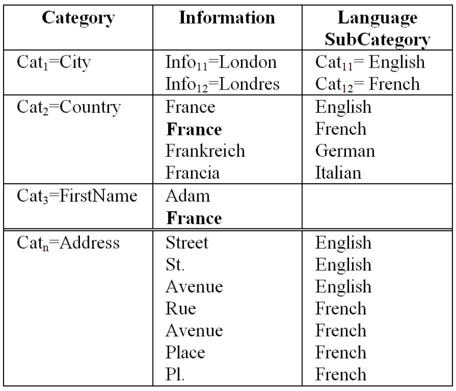


Figure 5. An instance of the Meta-Schema-Ontology under Protégé.



Category	Information	Language SubCategory
Cat_1=City	Info_{11}=London Info_{12}=Londres	Cat_{11}= English Cat_{12}= French
Cat_2=Country	France **France** Frankreich Francia	English French German Italian
Cat_3=FirstName	Adam **France**	
Cat_n=Address	Street St. Avenue Rue Avenue Place Pl.	English English English French French French French

Figure 6. A sample of the data dictionary.

2.2.2. Regular Expressions

A category Kat$_i$ can also be defined by intention using regular expressions (RE). These are used to validate the syntactic and semantic of strings. Let Kat_{int} be the set of these categories.

RE can be defined as a set of pairs Catregex (Category, Regular-Expressions).

RE = {Catregex$_i$/Catregex$_i$ (Kat$_i$, Regex$_{ij}$); i = 1...p, j = 1...q}. Some instances of categories are presented in **Figure 7**.

2.2.3. Indicators

The semantic data profiling is based on a set I of p indicators applied to the data source. Most of the existing tools are interested only in quantitative summaries of the source data. Few tools focus on semantic analysis. For that, we propose semantic indicators. I is composed of three types of indicators (**Figure 8**): statistic indicators {I_{stati}, i = 1;p}, two syntactic indicators ($I_{SYN1,2}$) and two semantic ones ($I_{SEM1,2}$).

After presenting in this paragraph, the input data for semantic data profiling, we will outline below, the process itself.

3. Semantic Data Profiling Process

Let us give some notations and definitions used in the algorithm of the semantic data profiling process.

Each column C$_i$, belonging to the data source S, has a set of values v$_i$ (i = 1...n). Each v$_i$ has a data type such as {String, Number, Date, Boolean, list or range of values}.

Definition 1: *Syntactic validity of a value v*

A value v is syntactically valid if and only if (iff) v ∈ RE or v ≈ w ∈ DD. (≈means similar using similarity distances [5] [6]).

Definition 2: *Syntactic invalidity of a value v*

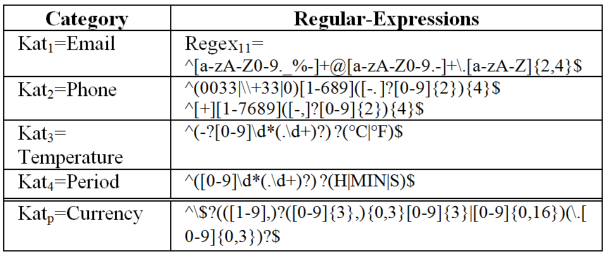
A value v is syntactically invalid iff v ∉ RE and v ∉ DD.

Definition 3: *Dominant Category*

Let Cat$_i$(v) be the number of syntactically correct values for a given attribute.

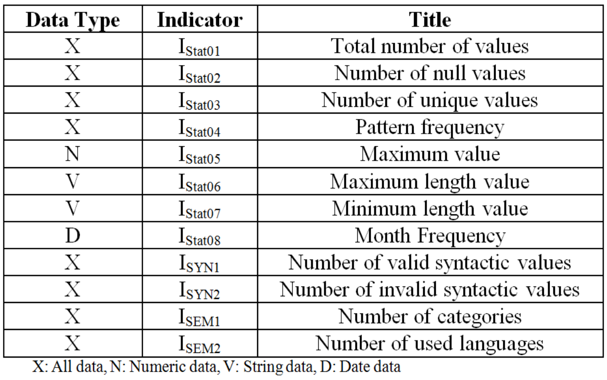
A Cat$_i$ is a dominant category iff Cat$_i$(v) > Cat$_j$(v) with i ≠ j.

The "Number of categories" indicator defines the number of categories detected.



<table>
<tr><th>Category</th><th>Regular-Expressions</th></tr>
<tr><td>Kat$_1$=Email</td><td>Regex$_{11}$=
^[a-zA-Z0-9._%-]+@[a-zA-Z0-9.-]+\.[a-zA-Z]{2,4}$</td></tr>
<tr><td>Kat$_2$=Phone</td><td>^(0033|\\+33|0)[1-689]([-.]?[0-9]{2}){4}$
^[+][1-7689]([-,]?[0-9]{2}){4}$</td></tr>
<tr><td>Kat$_3$=
Temperature</td><td>^(-?[0-9]\d*(.\d+)?) ?(°C|°F)$</td></tr>
<tr><td>Kat$_4$=Period</td><td>^([0-9]\d*(.\d+)?) ?(H|MIN|S)$</td></tr>
<tr><td>Kat$_p$=Currency</td><td>^\$?(([1-9],)?([0-9]{3},){0,3}[0-9]{3}|[0-9]{0,16})(\.[0-9]{0,3})?$</td></tr>
</table>

Figure 7. A set of regular expressions.



Data Type	Indicator	Title
X	I_{Stat01}	Total number of values
X	I_{Stat02}	Number of null values
X	I_{Stat03}	Number of unique values
X	I_{Stat04}	Pattern frequency
N	I_{Stat05}	Maximum value
V	I_{Stat06}	Maximum length value
V	I_{Stat07}	Minimum length value
D	I_{Stat08}	Month Frequency
X	I_{SYN1}	Number of valid syntactic values
X	I_{SYN2}	Number of invalid syntactic values
X	I_{SEM1}	Number of categories
X	I_{SEM2}	Number of used languages

X: All data, N: Numeric data, V: String data, D: Date data

Figure 8. A set of indicators.

Definition 4: *Semantic validity of a value v*

A value v is semantically valid iff v ∈ Cat_i, and Cat_i is the dominant category.

Definition 5: *Semantic invalidity of a value v*

A value v is semantically invalid iff v ∉ Cat_i, and Cat_i is the dominant category.

3.1. Profiling Algorithm

The principle of semantic data profiling algorithm (**Figure 9**) is to check if a value v belongs to the meta-repository. The aim is to verify the syntactic and semantic validity of v.

Given the data source S and the meta-information as inputs, the algorithm returns several tables (T_k, k = 1,7). These contain indicators results, invalid syntactic data, valid syntactic data, invalid semantic category-data, invalid semantic language-data and the new semantic structure.

The **statistic Indicators** function consists on applying different statistical indicators for a general summary (total number of values, number of duplicate values, pattern frequency) or according to the data type such as year Frequency, Maximum Length, Minimum Length.

The role of the **semantic Recognition Structure** function is trying to find a category and language for each data (v) using RE or DD. The three steps below will describe the principle of this function. Note that if v is a string, several possibilities are considered. Two types of research are used according to the presence or absence of keywords.

The first step is to check if v satisfies the definition 1. v is then considered syntactically valid. Then, we check the semantic validity (definition 4) using the dominant category concept (definition 3). This step allows obtaining the category and language for each column.

The second step deals, in one hand, with semantically invalid values (definition 5), remind that they are syntactically correct. In the other hand, this step processes with syntactically invalid ones (definition 2).

In the third step, the syntactically correct and semantically incorrect values are handled in several ways. According to their membership to the dominant category and the selected language, updates are automatically proposed such as homogenization, translation and standardization.

Whenever, the syntactically invalid values are well spelled (satisfy some regular expressions), they can be used to enrich the DD.

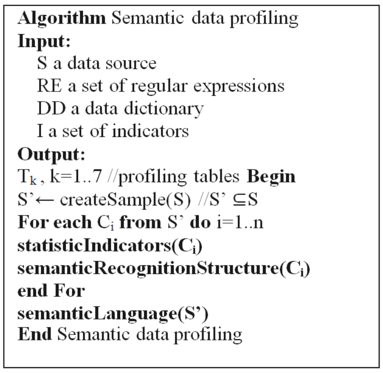
As there may be several languages for each column, not only one has to choose the dominant language column but also the dominant language of the source studied. The principle is presented in the semantic Language function.

The details of these functions (statistic Indicators, semantic Recognition Structure, semantic Language) are presented in Appendix (**Figure A1**).

The following paragraph will present the intermediate results.

3.2. Profiling Results

Several tables are used to store the different artefacts corresponding to the results of the semantic data profiling process.



```
Algorithm Semantic data profiling
Input:
   S a data source
   RE a set of regular expressions
   DD a data dictionary
   I a set of indicators
Output:
T_k , k=1..7 //profiling tables Begin
S'← createSample(S) //S' ⊆S
For each C_i from S' do i=1..n
statisticIndicators(C_i)
semanticRecognitionStructure(C_i)
end For
semanticLanguage(S')
End Semantic data profiling
```

Figure 9. Semantic data profiling algorithm.

The first one contains indicators results. For each column, we have some statistical summaries (e.g. percentage of null values), the number of invalid syntax values, the number of valid syntax values, the number of detected categories and number of detected languages.

The misspelled values are automatically added to the invalid syntax table (second table).

The third table contains the values, syntactically correct, which do not belong to Meta-Repository. They will be designated unknown categories.

For each column of the data source, we can have more than one category. So, to validate the dominant category, we choose the one with the greater percentage. The percentage is calculated based on the number of values that belong to this category. If we have two categories with the same percentage, we choose another sample from the data source and apply the semantic data profiling.

The values that do not belong to the dominant category are stored in the table T4 as semantic invalid category-value. In the same way, values that do not belong to the dominant language are stored in the table T5 as semantic invalid language-value.

Note that each column C_i of the source S is seen initially as a string. The goal is to recognize its semantic meaning (**Figure 10**). The dominant category and language are used to define the semantic structure for a data source.

Data source may contain similar columns, noted $Col_i \le Col_j$. For instance, Temperature_1 and Temperature_2 columns are similar categories ($Col_6 \le Col_7$). When two columns Col_i and Col_j belong to the same semantic category and have the same content ($Col_i = Col_j$), one of the two columns should be deleted.

3.3. Semantic Enrichment

As mentioned before, the meta-information must be enriched with new information. Both the data dictionary and the Meta-Schema-Ontology can be enriched.

The content of the DD may evolve using the values in T3, which must exist in some lexical databases suchas WordNet [24] and WOLF [25]. Similarly, when new categories are discovered after the semantic data profiling, the Meta-Schema-Ontology is expanded using new **Attributes** and their synonyms **synAttributes**.

Users can also enrich the meta-information with new regular expressions.

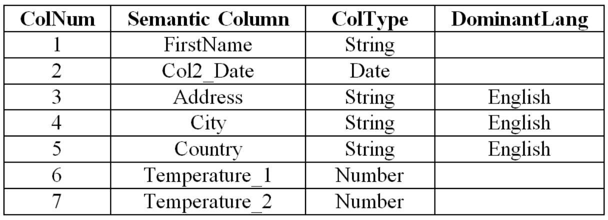
4. Conclusions and Contribution

Big data often have even less metadata than usual databases and that's a problem when the data scientist wants to perform analyses on these data. The use of our DQM tool would help the data scientist in recognizing data types (integer, dates, strings) and data semantics (Email, FirstName, Phone). The semantics would then be useful to automatically suggest views on data with a semantic meaning or to find matches between heterogeneous structures in big data.

DQM tool that we are currently developing is a contribution to new generation of Big-Data ETL based on semantics. Our goal is to guide the user in his quality approach.

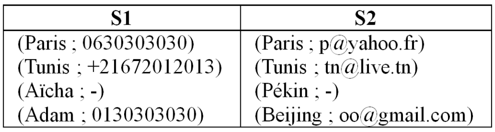
In the case of the absence of the data structure, we help the user:

1) To understand more the definition of manipulated data. Indeed, during the integration process for the union or the join operations, it is essential to differentiate synonyms and homonyms to succeed semantic data integration. Existing tools [14]-[16] [26] do not take into account semantic aspects. Only the syntactic ones are considered. For instance, in the case of the data integration process, user can choose to join two columns syntactically equivalent but semantically not S1.Col1 and S2.Col1 can be synonyms or homonyms (**Figure 11**). The union of S1 and S2 is semantically meaningless, while existing tools allow this operation. DQM tool alerts users to



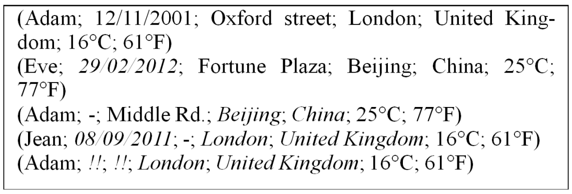
ColNum	Semantic Column	ColType	DominantLang
1	FirstName	String	
2	Col2_Date	Date	
3	Address	String	English
4	City	String	English
5	Country	String	English
6	Temperature_1	Number	
7	Temperature_2	Number	

Figure 10. Semantic structure for the data source S.



S1	S2
(Paris ; 0630303030)	(Paris ; p@yahoo.fr)
(Tunis ; +21672012013)	(Tunis ; tn@live.tn)
(Aïcha ; -)	(Pékin ; -)
(Adam ; 0130303030)	(Beijing ; oo@gmail.com)

Figure 11. Integration of the data sources S1 and S2.



(Adam; 12/11/2001; Oxford street; London; United Kingdom; 16°C; 61°F)
(Eve; *29/02/2012*; Fortune Plaza; Beijing; China; 25°C; 77°F)
(Adam; -; Middle Rd.; *Beijing*; *China*; 25°C; 77°F)
(Jean; *08/09/2011*; -; *London*; *United Kingdom*; 16°C; 61°F)
(Adam; *!!*; *!!*; *London*; *United Kingdom*; 16°C; 61°F)

Figure 12. Target data with cleaning actions.

incompatible semantic integration operations.

2) Throughout the laborious cleaning step. Transformation and homogenization that we propose will allow better elimination of duplicate or similar tuples. In fact, recalling that no method of calculating similarity distance permits the approximation between *Pékin* and *Beijing*, for example, because information on the language used is not taken into account. Our approach allows this reconciliation.

The originality of our approach is to infer the semantics of the data source structure using on one hand, the data itself and on the other hand, instances of the Meta-Schema-Ontology. Furthermore, our approach allows us to automatically propose cleaning actions on unstructured data. This constitutes part of our current and future work using MapReduce concepts [13] [27].

The results of the data profiling process are: 1) a data structure for better understanding of the semantic content of Big Data, 2) a set of updates for the correction of invalid data.

The semantic structure of the Big-Data source is:
S (Col1_FirstName: String, Col2_Date: Date,
Col3_Address: String, Col4_City: String,
Col5_Country: String, Col6_Temperature_1: Number,
Col7_Temperature_2: Number).

The target data after the cleaning actions should be for instance (**Figure 12**).

References

[1] Becker, J., Matzner, M., Müller, O. and Winkelmann, A. (2008) Towards a Semantic Data Quality Management—Using Ontologies to Assess Master Data Quality in Retailing. *Proceedings of the Fourteenth Americas Conference on Information Systems* (*AMCIS*'08), Toronto.

[2] Madnick, S. and Zhu, H. (2005) Improving Data Quality through Effective Use of Data Semantics. Working Paper CISL#2005-08, 1-19.

[3] Wang, X., Hamilton, J-H. and Bither, Y. (2005) An Ontology-Based Approach to Data Cleaning. Technical Report CS-2005-05, 1-10.

[4] Köpcke, H. and Rahm, E. (2009) Frameworks for Entity Matching: A Comparison. *Data Knowledge Engineering* (*DKE*'09), Leipzig, 197-210.

[5] Bilenko, M. and Mooney, R.J. (2003) Adaptive Duplicate Detection Using Learnable String Similarity Measures. *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery, and Data Mining*, Washington DC, 39-48. http://dx.doi.org/10.1145/956750.956759

[6] Koudas, N., Sarawagi, S. and Srivastava, D. (2006) Record Linkage: Similarity Measures and Algorithms. In: *ACM SIGMOD*'06, *International Conference on Management of Data*, Chicago, 802-803.

[7] Cohen, W.W. and Richman, J. (2004) Iterative Record Linkage for Cleaning and Integration. *Proceedings of the* 9th *ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery* (*DMKD*'04), Paris, 11-18.

[8] Monge, A.E. and Elkan, C.P. (1997) An Efficient Domain-Independent Algorithm for Detecting Approximately Duplicate Database Records. *Proceedings of the Second ACM SIGMOD Workshop Research Issues in Data Mining and Knowledge Discovery* (*DMKD*'97), 23-29.

[9] Boufarès, F., Ben Salem, A., Rehab, M. and Correia, S. (2013) Similar Elimination Data: MFB Algorithm. *IEEE* 2013 *International Conference on Control*, *Decision and Information Technologies* (*CODIT*'13), Hammamet, 6-8 May 2013, 289-293.

[10] Boufarés, F., Ben-Salem, A. and Correia, S. (2012) Qualité de données dans les entrepôts de données: Elimination des similaires. 8*èmes Journées francophones sur les Entrepôts de Données et l'Analyse en ligne* (*EDA*'12), Bordeaux, 32-41.

[11] Berti-Équille, L. (2007) Quality Awereness for Managing and Mining Data. HDR, Rennes.

[12] Tamraparni, D., Theodore, J., Muthukrishnan, S. and Vladislav, S. (2002) Mining Database Structure; or, How to Build a Data Quality Browser. *Proceedings of the ACM SIGMOD International Conference on Management of Data*, (*SIGMOD*'02), Madison, 2002, 240-251.

[13] Dean, J. and Ghemawat, S. (2004) MapReduce: Simplified Data Processing on Large Clusters. 6*th Symposium on Operating System Design and Implementation* (*OSDI*'04), San Francisco, 6-8 December 2004, 137-150.

[14] Data Cleaner, Reference Documentation, 2008-2013, datacleaner.org.

[15] (2011) Oracle Warehouse Builder Data Modeling, ETL, and Data Quality Guide, Performing Data Profiling. http://docs.oracle.com/cd/E11882_01/owb.112/e10935/data_profiling.htm#WBETL18000

[16] Datiris Profiler. http://www.datiris.com/

[17] UML. http://www.uml.org/

[18] Noy, N.F. and McGuinness, D.L. (2001) Ontology Development 101: A Guide to Creating Your First Ontology. Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880, 1-25.

[19] Bechhofer, S. (2012) Ontologies and Vocabularies. *Presentation at the* 9*th Summer School on Ontology Engineering and the Semantic Web* (*SSSW*'12), Cercedilla.

[20] Hauswirth, M. (2012) Linking the Real World. *Presentation at the* 9*th Summer School on Ontology Engineering and the Semantic Web* (*SSSW*'12), Cercedilla.

[21] Herman, I. (2012) Semantic Web Activities@W3C. *Presentation at the* 9*th Summer School on Ontology Engineering and the Semantic Web* (*SSSW*'12), Cercedilla.

[22] Kamel, M. and Aussenac-Gilles, N. (2009) Construction automatique d'ontologies à partir de spécification de bases de données. *Actes des 20èmes Journées Francophones d'Ingénierie des Connaissances* (*IC*), Hammamet, 85-96.

[23] Protégé Tool. http://protege.stanford.edu/

[24] Wordnet Database. http://wordnet.princeton.edu/

[25] WOLF Database. http://alpage.inria.fr/~sagot/wolf-en.html

[26] Talend Data Profiling. http://fr.talend.com/resource/data-profiling.html

[27] MapReduce (2013) The Apache Software Foundation. MapReduce Tutorial.

Appendix

```
Function statisticIndicators (Column C)
//return statistical indicators results
Begin
For each $I_d$ from I do //d=1..18
Add($I_d$ (C), $T1_c$)
        //statistic indicators: total number of values, number of null
            values…
end for
EndstatisticIndicators
```

```
Function semanticLanguage (Data Source S’)
//return the dominant language
Begin
For each$Language_i$ from T7 (i=1..n) //T7 is the semantic structure
$n_i$:= Count the number of occurrences ($Language_i$)
End for
DominantLanguage := Language where Max($n_i$)
End semanticCategories
```

```
Function semanticCategories (Column C)
//return syntactic and semantic indicators results and semantic structure

Begin
For each $v_j$from C do //j=1..m (m number of tuples)
If$v_j \in$   RE
then$add(v_j, Cat_j, Lang_j)$ // $v_j \in Cat_j$and$v_j \in Lang_j$
elseif$v_j$checkSpelling=true
//verifies some regular expressions for strings
then if $v_j \approx$  $w \in$   DD //w a value from DD
then$add(v_j, Cat_{j'}, Lang_{j'})$//$v_j \in Cat_{j'}$
and$v_j \in Lang_{j'}$; $j' \neq j$
else $add(v_j, Cat_{UNKNOWN})$
//$v_j \in$Unknown Category
$add(v_j, T3_c)$ //$v_j$ is a candidate to enrich DD
end if
                      else $add(v_j, T2_c)$
end if
end If
End for
$add(I_{sem1}(C), T1_c)$ //number of used categories
$add(I_{sem2}(C), T1_c)$ //number of used languages
$add(I_{syn1}(C), T1_c)$ //number of valid syntax value
$add(I_{syn2}(C), T1_c)$ //number of invalid semantic value
$add((Cat_{dom}, Lang_{dom}), T7_c)$ where $\%Cat_{dom} = Max(\%Cat_p)$ //p=1..x
and $\%Lang_{dom} = Max(\%Lang_q)$ //q=1..y
$add(Cat_{p'}, T4_c)$where $p' \neq p$
$add(Lang_{q'}, T5_c)$where $q' \neq q$
EndsemanticCategories
```

Figure A1. Functions of the semantic data profiling algorithm.