Computing and Informatics, Vol. 36, 2017, 1467–1491, doi: 10.4149/cai_2017_6_1467

HTAB2RDF: MAPPING HTML TABLES TO RDF TRIPLES

Djelloul Bouchiha, Mimoun Malki

EEDIS Laboratory, Djillali Liabes University of Sidi Bel Abbes, Algeria
e-mail: {djelloul.bouchiha, malki}@univ-sba.dz

Abdullah Alghamdi, Khalid Alnafjan

College of Computer and Information Sciences, KSU, Riyadh, Saudi Arabia
e-mail: {Ghamdi, Alnafjan}@ksu.edu.sa

Abstract. The Web has become a tremendously huge data source hidden under linked documents. A significant number of Web documents include HTML tables generated dynamically from relational databases. Often, there is no direct public access to the databases themselves. On the other hand, RDF (Resource Description Framework) gives an efficient mechanism to represent directly data on the Web based on a Web-scalable architecture for identification and interpretation of terms. This leads to the concept of Linked Data on the Web. To allow direct access to data on the Web as Linked Data, we propose in this paper an approach to transform HTML tables into RDF triples. It consists of three main phases: refining, pretreatment and mapping. The whole process is assisted by a domain ontology and the WordNet lexical database. A tool called Htab2RDF has been implemented. Experiments have been carried out to evaluate and show efficiency of the proposed approach.

Keywords: HTML tables, RDF, relational databases, Linked Data, domain ontology, WordNet

Mathematics Subject Classification 2010: 68N99

1 INTRODUCTION

The World Wide Web has enabled the creation of linked documents behind which there exists a big amount of data. As the Web becomes ever more entangled with our daily lives, there is a growing desire for direct access to data that are indirectly available through hypertext documents. Based on open standard, notably RDF, Linked Data [5] provides a paradigm to publish not only documents, but also data, on the Web.

A significant number of Web documents include HTML tables. According to Mulwad et al., table-like structures outside documents are widely used to represent and share data on the Web [29]; also, Cafarella et al. noted that there are billions of tables on the Web [6]. Often, HTML tables are generated dynamically from relational databases. Since there is no direct public access to the databases themselves, we need to transform HTML tables into RDF triples, allowing for publishing data on the Web.

In this paper we propose an approach to transform HTML tables into RDF triples. The proposed approach starts with refining HTML source code. Refining consists in deleting useless tags and keeping only useful ones. Refined HTML pages will be then presented in DOM logical format, and undergo morphological analysis. Tables in HTML pages are converted to their canonical form, and mapping rules are executed to generate an RDF graph at the last stage.

Our approach specifically focuses on Web documents that are data rich and narrow in ontological breadth. According to Embley et al., a Web document is *data rich* if it includes a set of identifiable constants, such as ID numbers, names, dates, and so on. A Web document is *narrow in ontological breadth* if it can be described with a domain ontology [11].

The rest of the paper is organized as follows: the next section covers some other similar works which aim in generally to transform Web resources into RDF graphs. Section 3 details the proposed approach. In Section 4 implemented tool, experimental results and discussion are presented. Finally, in Section 5, conclusions and perspectives are given.

2 RELATED WORK

Several similar works can be found in the literature. In this section we classify them into four categories:

1. Mapping Web resources to RDF.
2. Generating RDF graphs from relational data.
3. Generating RDF graphs from textual resources.
4. Works belonging to the ontology engineering field.

To be clear and explicit, lacks and problems of each work are formatted in *italic.*

Mapping Web resources (HTML tables in particularly) to RDF:

DBpedia is a community effort for extracting structured information from Wikipedia and to make them available on the Web. General information about the DBpedia project can be found in [22]. It includes an information extraction framework, which converts Wikipedia content to RDF. The most valuable for the DBpedia extraction are Wikipedia infoboxes. Infoboxes display the most relevant facts of an article as a table of pairs (attribute-value) on the top right-hand side of the Wikipedia page. *We note that DBpedia is not a generic approach. It is dedicated particularly to Wikipedia and does not treat any other Web application.*

Munoz et al. propose methods to recover semantics of Wikipedia's tables and extract facts from them in the form of RDF triples. Their method uses an existing Linked Data knowledge-base to find pre-existing relations between the entities in Wikipedia's tables, suggesting the same relations as holding for other entities in similar columns on different rows [30]. *Such an approach extracts RDF triples from Wikipedia's tables at a precision of only 40 %.*

Tourpedia generates an RDF catalogue from social media, notably Facebook, Foursquare, GooglePlaces and Booking [7]. *Tourpedia is limited to the tourism domain. The procedure to update datasets is still manual.*

In [31] authors propose a method to transform HTML tables to relational tables based on "Header paths" technique. In another paper, they propose to transform relational table to RDF triples [12]. *The transformation rules to RDF triples depend on the factoring process applied on HTML tables and does not consider directly HTML tables in canonical form. We note also that the proposed approach in [31] does not follow the W3C standard rules presented in [1]. Furthermore, the primary key detection depends on a preliminary step (canonical representation of table based on header paths), and does not consider directly column headers of HTML tables.*

Indirect conversion can be done to convert HTML into RDF. We can use, for example, an HTML translator to XML, like HTML Tidy[1]. Then we use XML-toRDF translator[2]. *However, these tools perform blind conversion which does not take into account important information, such as primary and foreign keys. This weakens the quality of the generated RDF document.*

Generating RDF graphs from relational data:

Konstantinou et al. propose a modular approach to generate RDF graph from metadata stored in relational database-backed digital library systems, by using a relational-to-RDF mapping engine [20]. *However, this approach does not allow intelligent queries and needs supplementary implementation on the information system.*

[1] `http://infohound.net/tidy/`

[2] `http://sourceforge.net/projects/xmltordf/`

GeoTriples [21] and Bio2RDF [10] transform databases into RDF graphs. *Both of them are limited to a specific domain. The first one focuses on the geospatial domain. The second one treats the biomedical field.*

A Direct Mapping from relational data to RDF has been presented in [1]. It takes as input a relational database (schema and data), and produces an RDF graph called the direct graph. The proposed algorithms compose a graph of relative IRIs which must be resolved against a base IRI to form an RDF graph. *Primary and foreign keys are considered in the transformation process. This causes the problem that, while primary keys are explicit in relational databases, in HTML pages, primary keys are implicit and hidden in the table columns.*

Scharffe et al. present the Datalift project, a framework and a platform for publishing datasets (CSV, XML file or relational database) on the Web of Linked Data [37]. *Datalift's users must be experts, because they need significant knowledge of the Semantic Web formalisms to perform the lifting process.*

Mulwad et al. propose to represent the content of tables as RDF, performing entity-resolution and relationship discovery by using reference knowledge-bases [28]. *This approach has been evaluated over only 15 relational tables. It reaches only 25 % for identifying relations. The algorithm of Linking Table Cells to Entities is based on a syntactic metric, notably levenshtein.* In more recent work, Mulwad et al. propose to extend their approach so that it can represent not only content but also meaning of tables as RDF [29].

Other approaches for mapping relational databases to RDF can be found in a survey presented in [36].

Generating RDF graphs from textual resources:

Rezk et al. present NLP2RDF, a tool to convert natural language sentences to RDF triples. Authors provided ontologies for Korean linguistic annotations, and they suggested an internationalization of the URI scheme of the NLP Interchange Format [35]. *This work is dedicated to only Korean language. Korean entities are linked with Wikipedia; instead, they must be linked with the linked open data cloud.*

Exner et al. introduce a framework to carry out an end-to-end extraction of DBpedia RDF triples from unstructured Wikipedia text. The proposed system is based on a pipeline of text processing modules that includes a semantic parser and a co-reference solver [13]. *Some errors stemmed from incorrect mappings and require a more detailed analysis.*

Gagnon et al. show how to use natural language processing techniques to automatically generate RDF triples from the information in the literals. Authors develop knowledge schemas to capture its information, and precise syntactic-based methods of knowledge extraction to automatically generate instances of these schemas from textual data [14]. *It was a syntactic-based method for knowledge extraction where authors look only at drug indications found in a specific Web site.*

Other similar works are more near the ontology engineering field:

> YAGO 2 [16] is a spatially and temporally enhanced knowledge base built from Wikipedia. Authors have developed an extensible approach to fact extraction from Wikipedia and other sources, and they have tapped on specific inputs that contribute to the goal of enhancing facts with spatio-temporal scope. The data format of YAGO 2 is fully RDF compliant.
>
> TANGO (Table ANalysis for Generating Ontologies) consists in generating ontologies based on table analysis [39]. It is a formalized method of processing the format and content of tables to incrementally build a relevant reusable conceptual ontology. In a later work, TANGO has been assisted to construct an ontology in the relatively narrow domain of geopolitics, with as little human intervention as possible [32].
>
> Li et al. propose rules to learn OWL ontology from a relational database [23]. Rules are defined using a combination of some formal notation and English language. *Some of the proposed rules miss some semantics of the relational schema and some rules produce specific results for inheritance and object properties that may not precisely represent concepts across domains or database modeling choices.*
>
> Astrova et al. provide rules and examples for automatic transformation of a relational schema to OWL [2]. *A number of the proposed transformations were ambiguous.*

At the last of this section, we note that there are some other efforts which aim at publishing public sector information as Linked Data. Among these works we cite: The German National Library (DNB) publishes its data as Linked Data [9]. Szekely et al. propose an approach that maps data of the Smithsonian American Art Museum to RDF Linked Open Data [38]. Jovanovik et al. provide use-case scenarios for publishing and using healthcare data in the republic of Macedonia as RDF Linked Open Data [18]. Willighagen et al. describe recent work in an ongoing project converting data from the ChEMBL database into RDF triples [40]. The works presented in [8] and [24] aim at publishing government data of USA and UK, respectively, as RDF linked open data.

3 THE PROPOSED APPROACH

In this paper, we present an end-to-end solution from HTML tables to RDF, inspired by the standard rules presented in [1]. Our starting point is a collection of HTML tables. The end point is an RDF graph. Most of the lacks of the mentioned works above are solved in our approach by a domain ontology and a lexical database assisting the mapping process.

3.1 Overview

As shown in Figure 1, the proposed approach consists of three successive phases: refining, pre-treatement and the mapping engine:

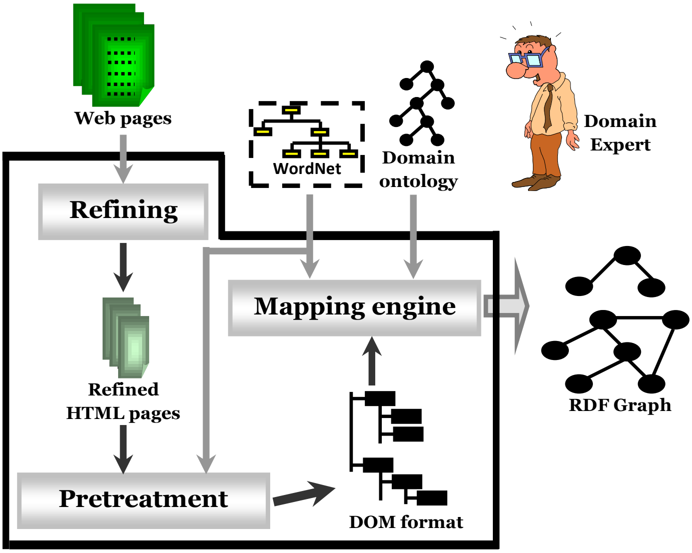


Figure 1. Mapping process

Refining: consists in browsing HTML source code, deletes useless tags such as those of layout (e.g. <b>, <i>), and keeps useful tags (e.g. <table>, <td>, <tr>, <form>, <ul>, <li>). The output of this step is a set of refined HTML pages.

Pre-treatment: first, refined HTML pages will be presented in DOM[3] logical format to facilitate their manipulation. A morphological analysis is then applied to the tables attributes. It consists in removing hyphens and keeps terms stem as they appear in WordNet[4] (e.g., morphological analysis applied to "running-away" gives "run away").

Mapping engine: at this stage, tables are converted to their canonical form (Sections 3.2 and 3.3). Then the mapping rules are executed to generate an RDF graph (Section 3.4). To identify keys at this stage, the domain ontology can

[3] DOM: Document Object Model is an API which consists in representing HTML or XML document content as a tree structure of nodes (each element of the document represents a node) [26].

[4] WordNet is a lexical database which organizes names and verbs in concepts (synsets) in is-a hierarchy of relations. Each synset is described by a short gloss [27].

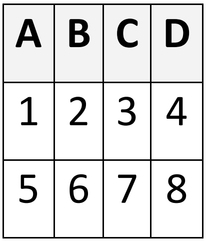
be used (Section 3.5). To solve the terms differences problem, similarity measures based on WordNet [33] can be computed between HTML elements and the attributes of the ontology concepts.

3.2 HTML Tables in a Canonical Form

To map HTML tables in Web documents to RDF triples we consider only tables in a canonical form. A canonical table is defined as follows [39]:

Definition: A schema S for a canonical table is a finite set of labels $\{L_1, \ldots, L_n\}$. Each label L_i, with $1 \leq i \leq n$, corresponds to a domain D_i. Let $D = D_1 \cup \ldots \cup D_n$. A *canonical table* T is a set of functions $T = \{t_1, \ldots, t_m\}$ from S to D with the restriction that for each function $t \in T$, $t(L_i) \in D_i$.

HTML tables are often displayed in two dimensions. In this case, the order of the labels in the schema is fixed for each function and these labels are factored to the top as column headers. Each row in the table represents the domain values for the corresponding labels in the column headers. For example, the canonical table $\{\{(A, 1), (B, 2), (C, 3), (D, 4)\}, \{(A, 5), (B, 6), (C, 7), (D, 8)\}\}$ is displayed as follows:



A	B	C	D
1	2	3	4
5	6	7	8

Figure 2. An example of a table in canonical form

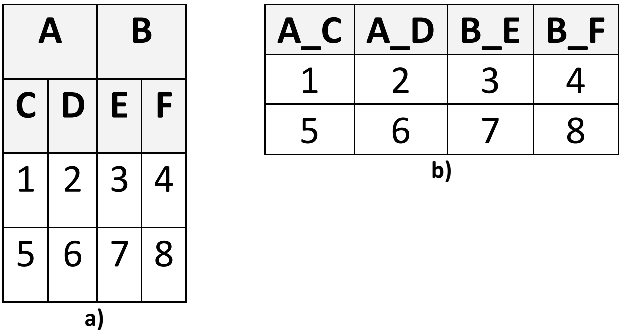
3.3 Transforming Complex Table into Canonical Form

Many tables on the Web appear essentially in canonical form [39]. However, other complex Web tables can appear with column/row headers. To transform complex tables into canonical representation, we use *Header Paths* technique [12]. This technique relates column/row headers and data cells.

A transformation example of a complex table to its canonical representation is shown in Figure 3.

The complex table is given in Figure 3 a). The *column header* is coloured in gray. The data cells are bellow the column header.

In general, the algorithm of transformation into the canonical form is as follows: data cells of the new table are kept from the complex table. Each column header



A		B	
C	D	E	F
1	2	3	4
5	6	7	8

a)

A_C	A_D	B_E	B_F
1	2	3	4
5	6	7	8

b)

Figure 3. Transforming a complex table into the canonical form

path in the new table is obtained by concatenating the labels from the high to the low header in the complex table.

Then, the canonical form of the table is given in Figure 3 b). The column header paths are A_C, A_D, B_E and B_F.

3.4 Mapping Rules

Given a set of HTML tables and a domain ontology. The corresponding RDF graph to the HTML tables is obtained as follows:

- Each row produces a triple composed as follows:

  (*) The subject is an IRI formed by concatenating the domain ontology IRI, table name, column name corresponding to the primary key in the database and its value (<Ontology_URI/Table_Name/PrimaryKey_Column_Name=PrimaryKey_Value>).

  (**) The predicate is the expression rdf:type.

  (***) The object is formed by concatenating the domain ontology IRI and the table name.

- Each row produces also a set of triples with a common subject:

  1. A common subject is formed as in (*).
  2. The predicate for each column is an IRI formed by concatenating the domain ontology IRI, the table name and the column name (Ontology_URI/Table_Name#Column_Name).
  3. The object for each column is an RDF literal corresponding to the column value.

- Each foreign key produces a triple as follows:

  1. The subject is formed as in (*).

2. The predicate is formed by concatenating the foreign key column names, the referenced table and the referenced column names.
3. The object is formed by concatenating the domain ontology IRI, name of the referenced table, column name corresponding to the primary key in the referenced table and its value (<Ontology_URI/Referenced_Table_Name/Primary Key_Referenced_Table=PrimaryKey_Value>).

- Note that no triple is generated for a NULL value.
- The union of all RDF triples obtained from rows of all tables produces the complete RDF graph corresponding to the HTML tables.

3.5 Detecting Keys

In the mapping rules described above, both primary keys and foreign keys are considered. A primary key is one or a combination of columns that uniquely identify each row in the table. A foreign key is one or a combination of columns that reference a primary key in another table. This establishes a link between two tables.

While keys are explicit in database, they are not expressed explicitly in HTML tables. To say which attribute represents a key to the HTML tables, domain ontology can be used. In OWL 2, the construct "HasKey" can be used to assign a collection of data properties as a key to a class expression. So each named instance of the class expression is identified uniquely by the set of values which these properties attain in relation to the instance [15].

As the work done in [3], also algorithms that detect keys and functional dependencies inside a given database [25, 17] can be adapted to detect keys in HTML tables.

WordNet can also help to detect keys. For example, a term equivalent to "ID" or "KEY" can be considered as key.

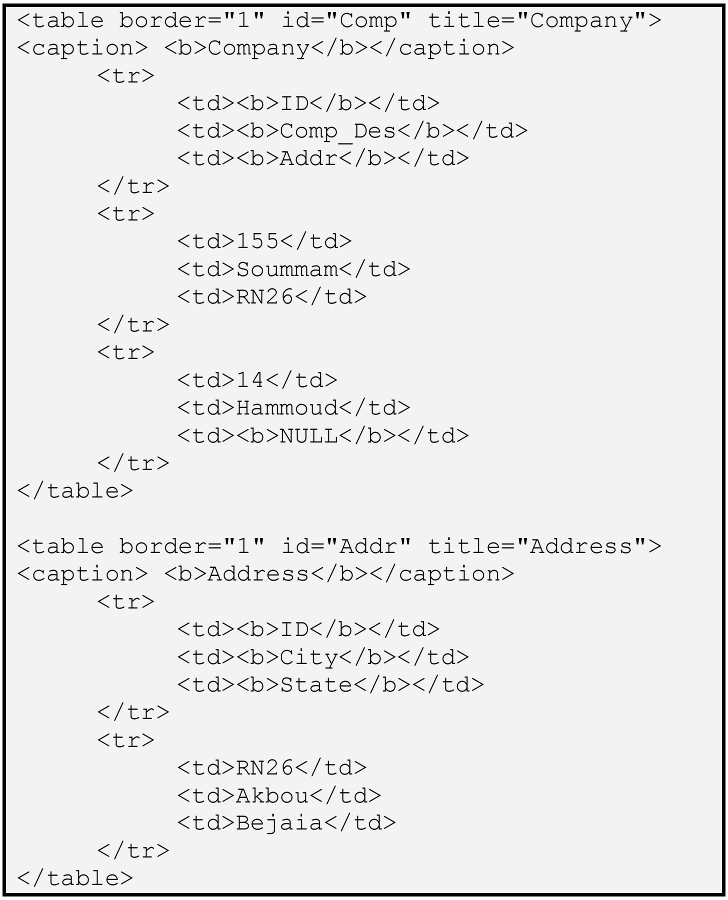
When any key is detected, we can use the heuristic that the key column is the first one. Or, a default solution consists in adding an auto-incrementally column attribute and considering it as a primary key.

Foreign keys can be detected by computing similarity measure between the detected primary keys and the columns of the other tables. When one or a set of columns are semantically equivalent to a primary key in another table, it is (or they are) considered as a foreign key.

Detecting keys is not an evident task, and the intervention of an expert remains necessary. Thus, it will be more efficient if this task will be accomplished semi-automatically.

3.6 Example

Figure 4 presents an HTML code of two HTML tables, **Company** and **Address**. Both tables have the column **ID** corresponding to Primary key. The column **Addr** in the table **Company** corresponds to the foreign key which relates the two tables.



```
<table border="1" id="Comp" title="Company">
<caption> <b>Company</b></caption>
      <tr>
           <td><b>ID</b></td>
           <td><b>Comp_Des</b></td>
           <td><b>Addr</b></td>
      </tr>
      <tr>
           <td>155</td>
           <td>Soummam</td>
           <td>RN26</td>
      </tr>
      <tr>
           <td>14</td>
           <td>Hammoud</td>
           <td><b>NULL</b></td>
      </tr>
</table>

<table border="1" id="Addr" title="Address">
<caption> <b>Address</b></caption>
      <tr>
           <td><b>ID</b></td>
           <td><b>City</b></td>
           <td><b>State</b></td>
      </tr>
      <tr>
           <td>RN26</td>
           <td>Akbou</td>
           <td>Bejaia</td>
      </tr>
</table>
```

Figure 4. HTML code of two tables, **Company** and **Address**

Given a base IRI `http://example/management.owl`. From the HTML tables of Figure 4, the RDF graph in Figure 5, expressed in RDF turtle syntax, is generated.

3.7 Other Considered HTML Forms

Other HTML forms can display the same information as tables. Even partially, a *list* can be considered as a table with one column. We can also think of the *filled-in form* as a table with one row. An *HTML form* contains labeled boxes used for the information collection: the items specified by the labels are written into the boxes, and then the form is returned to the originator.

```
@base <http://example/management.owl> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

<Company/ID=155> rdf:type <Company> .
<Company/ID=155> <Company#ID> 155 .
<Company/ID=155> <Company#Comp_Des> "Soummam" .
<Company/ID=155> <Company#Addr> "rn26" .

<Company/ID=155> <Company#ref-addr> <Address/ID=rn26> .

<Company/ID=14> rdf:type <Company> .
<Company/ID=14> <Company#ID> 14 .
<Company/ID=14> <Company#Comp_Des> "Hammoud" .

<Address/ID=rn26> rdf:type <Address> .
<Address/ID=rn26> <Address#ID> "rn26" .
<Address/ID=rn26> <Address#City> "Akbou" .
<Address/ID=rn26> <Address#State> "Bejaia" .
```

Figure 5. RDF graph corresponding to the HTML tables example

4 EXPERIMENTAL RESULTS AND DISCUSSION

To evaluate our mapping rules, a tool called Htab2RDF has been developed.

4.1 Htab2RDF Tool

Htab2RDF[5] converts HTML tables into RDF graphs. Its UI allows to upload HTML pages and eventually a domain ontology. It refines pages as described in Section 3.1. It detects tables through DOM trees. It stores then all information about tables: names, attributes and data. Htab2RDF detects then primary and foreign keys. Finally, the tool produces an RDF graph expressed in RDF turtle syntax [4].

Extracting tables' names with Htab2RDF tool: The table name is obtained from one of the following items sorted by a priority order:

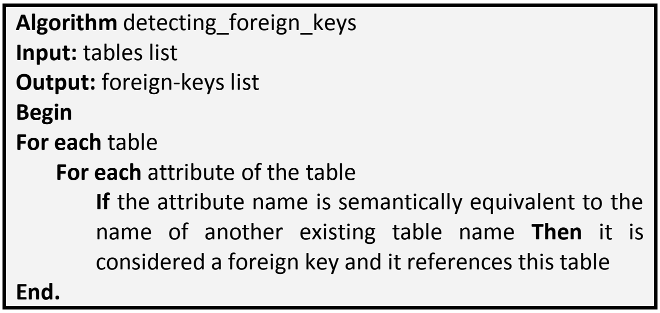
1. The "id" attribute value of the table tag.
2. The "title" attribute value of the table tag.
3. The footer section in the table.
4. The "caption" tag in the table element.
5. The page name plus the table number.

Detecting primary keys with Htab2RDF tool: The primary key can be detected from several sources sorted by a priority order as follows:

[5] `http://www.cuniv-naama.dz/infoteam/tools/htab2rdf/`

1. From the domain ontology, when the construct "HasKey" is used.
2. An equivalent attribute to "ID" is considered a primary key.
3. The heuristic that a primary key is often the first attribute of the table.
4. Else, a domain expert intervention (manually) is necessary to detect the primary key.

Detecting foreign keys with Htab2RDF tool: To detect foreign keys, the algorithm in Figure 6 is used.



```
Algorithm detecting_foreign_keys
Input: tables list
Output: foreign-keys list
Begin
For each table
    For each attribute of the table
        If the attribute name is semantically equivalent to the
        name of another existing table name Then it is
        considered a foreign key and it references this table
End.
```

Figure 6. Algorithm of detecting foreign keys

Two concepts are semantically equivalent if the similarity between the two terms identifying these concepts exceeds a certain threshold suggested by the system user. Similarity measure aims to quantify how much two terms are alike. In particularly, Htab2RDF uses WordNet based similarity measures [33]. A threshold is a value between "zero" and "one". "One" indicates that there is a total semantic equivalence.

Implementation platform: Htab2RDF is implemented in JAVA. It interacts with DOMSAX API to parse HTML documents. It interacts also with Protege-OWL API[6] for Ontological parsing and Java WordNet Similarity Library [34] for computing similarity measures. It provides a set of features for personalizing the calculations performed during the mapping process.

4.2 Experiments

To evaluate our approach, we have got a corpus of 200 tables[7] imported from 10 large statistical data sites, most with a geopolitical orientation, in the US and overseas. As domain ontology, we used a geopolitical ontology [19]. It is available in OWL version[8]. We installed also WordNet 2.0[9] for our experiments.

[6] http://protege.stanford.edu/plugins/owl/api/guide.html

[7] http://tango.byu.edu/data/

[8] http://aims.fao.org/aos/geopolitical.owl

[9] http://wordnetcode.princeton.edu/2.0/WordNet-2.0.exe

From the 200 tables we select 145 files. All of them are in canonical form. The available tables were in MS-Excel format. So we convert them into HTML format. This is done by a simple "save as" command from the "File" menu of MS-Excel User Interface.

Before generating the final RDF graph, we apply our algorithms on HTML tables one by one and we note the anomalies in:

1. Table name: this anomaly is reported when a wrong name is given by our system to the considered table.
2. Detecting key: this anomaly appears when not the right attribute is mentioned as primary key.
3. Attributes names: this anomaly is noted when two or more attributes have the same name.

Table 1 summarizes anomalies for each HTML table as follows: The "Table Code" column corresponds to the HTML file name in the used corpus. In columns "(1)", "(2)" and "(3)", anomalies are noted as explained above. Anomaly is noted with "X". The cases where no anomaly was detected are colored in *green*.

After analysing the results, we have seen that the first anomaly is detected because of the used technique to extract the table name from a phrase. This technique consists in extracting a simple term (or a composed term) which exists in WordNet starting from the beginning of the analysed phrase. For example the table "C10193" is described by the phrase "North American Trade". While the most important term in this statement is "Trade", the table was identified incorrectly by "North American". This problem can be solved by NLP techniques to extract the most important term(s) as a table name.

For the second anomaly, we have seen that any information about keys exists in the chosen domain ontology. This can be improved by using a richer domain ontology.

As the first anomaly, the third one is caused by the used technique to extract attributes names.

Table 2 presents success rate for each considered criteria.

Now, to generate the RDF graph we kept only 47 tables; those with no anomalies. So we obtain an RDF graph of 4644 lines. Next is an excerpt of the generated RDF graph.

We note that 18 foreign keys have been detected, among which 2 are correct, 4 reflect a poor relation between the two corresponding tables and 12 are incorrect. This low result can be explained by the fact that the detecting foreign keys process depends on detecting tables names and attributes; so it inherits all their lacks. Therefore, to improve this result, the algorithm of detecting foreign keys must be independent of any other algorithm.

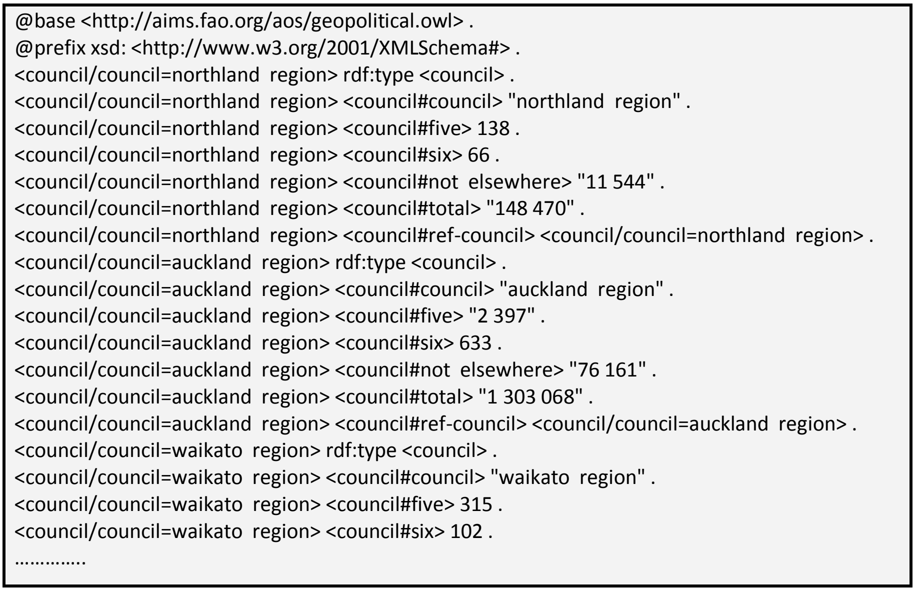
All results above have been obtained with a threshold = 0.7 and the Lin measure. With more accuracy, i.e. threshold > 0.7, we will have more satisfying results.

Table Code	(1)	(2)	(3)	Table Code	(1)	(2)	(3)	Table Code	(1)	(2)	(3)
C10002			X	C10003			X	C10004			
C10007			X	C10008				C10009			
C10010			X	C10012				C10013	X		
C10014	X			C10015	X			C10016			
C10017	X		X	C10018			X	C10020	X		X
C10021	X			C10023	X		X	C10027			X
C10028	X			C10029			X	C10030			
C10031	X			C10032			X	C10033			
C10038			X	C10040	X			C10041			X
C10042	X		X	C10043			X	C10044		X	
C10045	X			C10046			X	C10048			X
C10050				C10054				C10056	X		X
C10059				C10060	X		X	C10061	X		X
C10062	X			C10063			X	C10064			X
C10065			X	C10066				C10067			X
C10068				C10069				C10070			X
C10071			X	C10072	X		X	C10073	X		X
C10074	X			C10075			X	C10076			
C10077				C10080				C10081			
C10084	X		X	C10085			X	C10086			X
C10087	X			C10088			X	C10090			X
C10093				C10095				C10096			
C10097	X		X	C10098				C10100			
C10101	X			C10103	X		X	C10104	X	X	
C10105				C10106			X	C10107			X
C10108	X			C10109	X	X		C10111	X		
C10112				C10115				C10116			
C10117	X			C10118			X	C10119			X
C10122				C10123				C10124			
C10125			X	C10126				C10127			X
C10128			X	C10129			X	C10130			
C10131			X	C10132			X	C10134	X		
C10135			X	C10136				C10137			X
C10138				C10139		X		C10141			
C10142	X			C10143			X	C10144			X
C10145			X	C10146	X			C10147			X
C10148			X	C10149				C10150			X
C10151				C10152			X	C10153			
C10154				C10155				C10156	X		
C10159	X			C10160				C10161	X		X
C10162				C10164			X	C10165	X		
C10166				C10168				C10170	X		
C10172	X		X	C10173	X		X	C10174			X
C10175			X	C10177				C10178			
C10179		X		C10182			X	C10184	X		
C10186	X			C10188			X	C10189	X		
C10190				C10191				C10193	X	X	
C10195			X	C10196			X	C10197	X		
C10198			X								

Table 1. Anomalies in detecting table name, primary key and attributes name

Criteria	**Success Rate**
Extracting table name	69.65 %
Detecting primary key	95.86 %
Extracting attributes names	54.48 %
All criteria at the same time	32.41 %

Table 2. Success rate for each considered criteria



```
@base <http://aims.fao.org/aos/geopolitical.owl> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
<council/council=northland  region> rdf:type <council> .
<council/council=northland  region> <council#council> "northland  region" .
<council/council=northland  region> <council#five> 138 .
<council/council=northland  region> <council#six> 66 .
<council/council=northland  region> <council#not  elsewhere> "11 544" .
<council/council=northland  region> <council#total> "148 470" .
<council/council=northland  region> <council#ref-council> <council/council=northland  region> .
<council/council=auckland  region> rdf:type <council> .
<council/council=auckland  region> <council#council> "auckland  region" .
<council/council=auckland  region> <council#five> "2 397" .
<council/council=auckland  region> <council#six> 633 .
<council/council=auckland  region> <council#not  elsewhere> "76 161" .
<council/council=auckland  region> <council#total> "1 303 068" .
<council/council=auckland  region> <council#ref-council> <council/council=auckland  region> .
<council/council=waikato  region> rdf:type <council> .
<council/council=waikato  region> <council#council> "waikato  region" .
<council/council=waikato  region> <council#five> 315 .
<council/council=waikato  region> <council#six> 102 .
.............
```

Figure 7. Excerpt of the resulting RDF graph

To determine which measure and threshold give best results in our approach, several tests have been done with different measures (Lin, Jiang, Pirro-Seco and Resnik) and progressive threshold values $(0, 0.1, 0.2, 0.3, \ldots, 0.9, 1)$.

From the 145 tables used in the experiment above, we kept 51 tables; those which the process of detecting primary key depends strongly on the similarity measure. Thus, the experiment presented in the section above has been reproduced 44 times again, and the obtained results have been reported in Table 3.

	Success Rate (%)															
	LIN				**JIANG**				**RESNIK**				**PIRRO-SECO**			
Th	**(1)**	**(2)**	**(3)**	**(4)**	**(1)**	**(2)**	**(3)**	**(4)**	**(1)**	**(2)**	**(3)**	**(4)**	**(1)**	**(2)**	**(3)**	**(4)**
0.0	64.7	41.17	66.6	21.56	64.7	43.13	66.6	23.52	64.7	43.13	66.6	21.56	64.7	41.17	66.6	21.56
0.1	64.7	43.13	66.6	21.56	64.7	43.13	66.6	23.52	64.7	43.13	66.6	21.56	64.7	41.17	66.6	21.56
0.2	64.7	45.09	66.6	21.56	64.7	45.09	66.6	23.52	64.7	45.09	66.6	21.56	64.7	43.13	66.6	21.56
0.3	64.7	72.54	66.6	33.33	64.7	49.01	66.6	25.49	64.7	76.47	66.6	33.33	64.7	54.90	66.6	27.45
0.4	64.7	78.43	66.6	33.33	64.7	74.50	66.6	33.33	64.7	78.43	66.6	33.33	64.7	78.43	66.6	33.33
0.5	64.7	78.43	66.6	33.33	64.7	78.43	66.6	33.33	64.7	78.43	66.6	33.33	64.7	78.43	66.6	33.33
0.6	64.7	80.39	66.6	33.33	64.7	80.39	66.6	33.33	64.7	76.47	66.6	33.33	64.7	80.39	66.6	33.33
0.7	64.7	88.23	66.6	35.29	64.7	88.23	66.6	35.29	64.7	82.35	66.6	35.29	64.7	88.23	66.6	35.29
0.8	64.7	88.23	66.6	41.17	64.7	88.23	66.6	41.17	64.7	88.23	66.6	41.17	64.7	88.23	66.6	41.17
0.9	64.7	88.23	66.6	41.17	64.7	88.23	66.6	41.17	64.7	88.23	66.6	41.17	64.7	88.23	66.6	41.17
1	64.7	88.23	66.6	41.17	64.7	88.23	66.6	41.17	64.7	88.23	66.6	41.17	64.7	88.23	66.6	41.17

Th: Threshold
Criteria: (1) Extracting table name, (2) Detecting primary key,
(3) Extracting attributes names, (4) All criteria at the same time (without any anomaly).

Table 3. Success rates with different threshold values, different similarity measures and different quality criteria

The graphs in Figure 8 represent the evolution of the success rate according to the chosen threshold and similarity measure for each criterion.

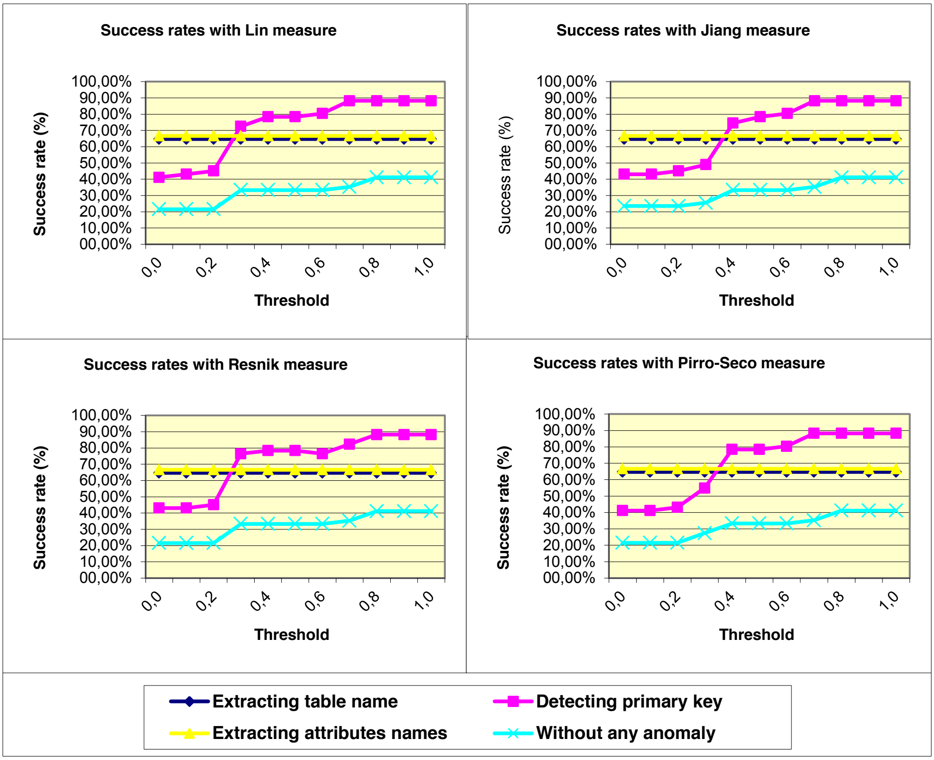


Figure 8. Curves of the success rates evolution for each criterion according to the chosen threshold and similarity measure

For both criteria, Extracting table name and Extracting attributes names, the success rates are stable, 64.70 % and 66.66 %, respectively. This can be justified by the fact that the two processes are independent of the similarity measure.

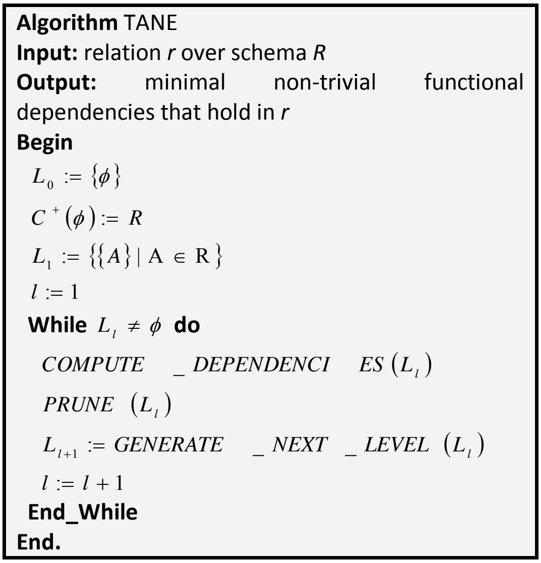
The process of detecting primary keys, reaches its best results (88.23 %) with a threshold of 0.7, except for the Resnik metric, where the process reaches its best with a threshold of 0.8.

Whatever the similarity measure chosen, the maximum success rate, to have results without any anomaly, is achieved with a threshold of 0.8.

4.3 Comparison Study

In this section we focus on the primary keys detection step, and we compare our approach with the TANE algorithm proposed in [17].

TANE is an efficient algorithm to detect primary keys. The process is based on discovering functional dependencies from large databases.



```
Algorithm TANE
Input: relation r over schema R
Output: minimal non-trivial functional dependencies that hold in r
Begin
  $L_0 := \{\phi\}$
  $C^+(\phi) := R$
  $L_1 := \{\{A\} \mid A \in R\}$
  $l := 1$
  While $L_l \neq \phi$ do
    $COMPUTE\_DEPENDENCIES(L_l)$
    $PRUNE(L_l)$
    $L_{l+1} := GENERATE\_NEXT\_LEVEL(L_l)$
    $l := l + 1$
  End_While
End.
```

Figure 9. TANE algorithm [17]

Since there is no available tool, we developed a prototype tool supporting the TANE algorithm.

As dataset we kept 145 files from the corpus described above. All of them contain tables in canonical form.

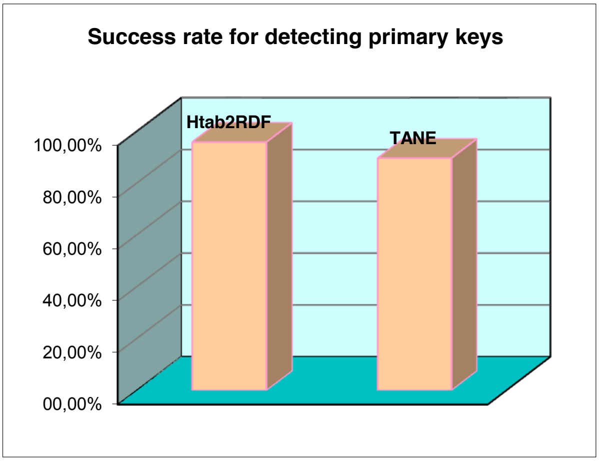
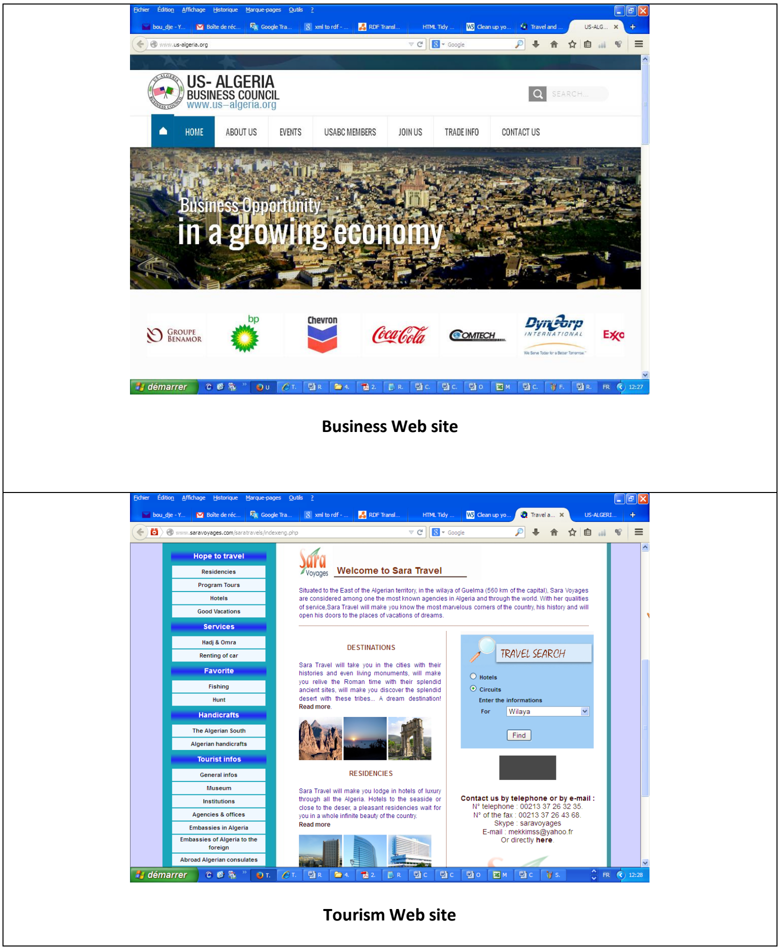


Figure 10. Htab2RDF's detecting keys algorithm vs. TANE algorithm

With a threshold = 0.7 and the Lin measure, Htab2RDF reaches a success rate of 95.86 % in detecting primary keys. For the same dataset (145 HTML files), 15 errors have been committed by the TANE algorithm in detecting primary keys, which gives a success rate of 89.65 %.

4.4 Case Study



Business Web site

Welcome to Sara Travel

Situated to the East of the Algerian territory, in the wilaya of Guelma (560 km of the capital), Sara Voyages are considered among one the most known agencies in Algeria and through the world. With her qualities of service,Sara Travel will make you know the most marvelous corners of the country, his history and will open his doors to the places of vacations of dreams.

DESTINATIONS

Sara Travel will take you in the cities with their histories and even living monuments, will make you relive the Roman time with their splendid ancient sites, will make you discover the splendid desert with these tribes... A dream destination! Read more.

RESIDENCIES

Sara Travel will make you lodge in hotels of luxury through all the Algeria. Hotels to the seaside or close to the deser, a pleasant residencies wait for you in a whole infinite beauty of the country. Read more

TRAVEL SEARCH

○ Hotels
● Circuits

Enter the informations

For: Wilaya

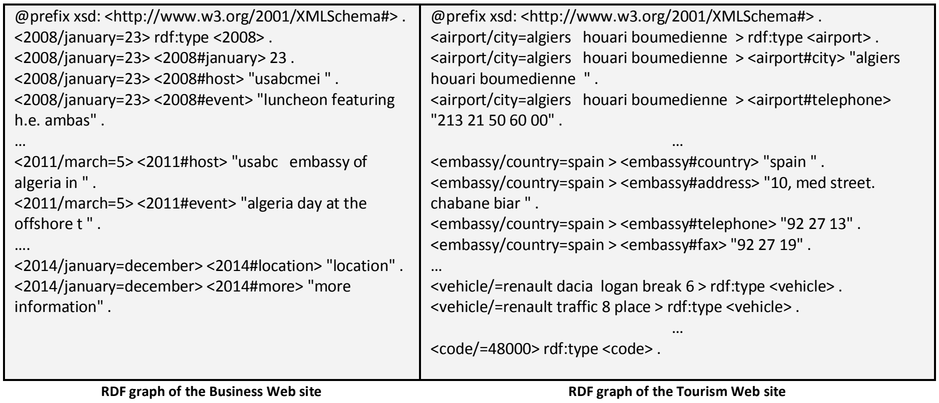
Find

Contact us by telephone or by e-mail :
N° telephone : 00213 37 26 32 35.
N° of the fax : 00213 37 26 43 68.
Skype : saravoyages
E-mail : mekkimss@yahoo.fr
Or directly here.

Tourism Web site

Figure 11. Snapshots of Business and Tourism Web sites

After testing our approach on a large corpus, we move up to the next step; it is to check our algorithms on real existing Web sites. For this, we choose two domains,



```
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
<2008/january=23> rdf:type <2008> .
<2008/january=23> <2008#january> 23 .
<2008/january=23> <2008#host> "usabcmei " .
<2008/january=23> <2008#event> "luncheon featuring
h.e. ambas" .
...
<2011/march=5> <2011#host> "usabc   embassy of
algeria in " .
<2011/march=5> <2011#event> "algeria day at the
offshore t " .
....
<2014/january=december> <2014#location> "location" .
<2014/january=december> <2014#more> "more
information" .
```

RDF graph of the Business Web site

```
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
<airport/city=algiers   houari boumedienne  > rdf:type <airport> .
<airport/city=algiers   houari boumedienne  > <airport#city> "algiers
houari boumedienne  " .
<airport/city=algiers   houari boumedienne  > <airport#telephone>
"213 21 50 60 00" .
                                        ...
<embassy/country=spain > <embassy#country> "spain " .
<embassy/country=spain > <embassy#address> "10, med street.
chabane biar " .
<embassy/country=spain > <embassy#telephone> "92 27 13" .
<embassy/country=spain > <embassy#fax> "92 27 19" .
...
<vehicle/=renault dacia  logan break 6 > rdf:type <vehicle> .
<vehicle/=renault traffic 8 place > rdf:type <vehicle> .
                                        ...
<code/=48000> rdf:type <code> .
```

RDF graph of the Tourism Web site

Figure 12. Excerpts of RDF code generated from the two Web sites, with the Lin measure and a threshold of 0.8

notably business[10] and tourism[11]. Thus, two Web sites have undergone our tests. The first one is a Web site describing tourism in Algeria[12]. The second is a Web site describing business activities between Algeria and USA[13].

Snapshots of the two Web sites are in Figure 11.

From the first Web site, an RDF file of more than 800 triples has been generated from seven tables. From the second Web site, an RDF file of more than 550 triples has been generated from six tables. Figure 12 presents excerpts of the two RDF files.

5 CONCLUSION AND PERSPECTIVES

The Semantic Web allows putting data and links on the Web, so that a person or machine can explore the Web of data; we speak about Linked Data [5] expressed as RDF graphs. RDF allows representing data on the Web based on a Web-scalable architecture for identification and interpretation of terms [1].

Data on the Web are often displayed as tables. To allow a direct access to this data, we proposed an approach to transform HTML tables into RDF graph. The proposed approach is supported by a set of mapping rules. It consists of three main phases: refining, pre-treatment and mapping. A tool has been implemented and a set of experiments have been carried out to show the effectiveness of our approach.

[10] http://www.getopt.org/ecimf/contrib/onto/REA/index.html

[11] http://protege.cim3.net/file/pub/ontologies/travel/travel.owl

[12] http://www.saravoyages.com/saratravels/indexeng.php

[13] http://www.us-algeria.org/

The obtained results were satisfactory and encouraging, and show that the approach provides a suitable starting point for semantic Web development.

The work presented in this paper can serve institutions, organizations, offices and agencies that are active on the Web and want to provide public access to their data. In many areas, such as health, economy, finance, education, tourism and employment, all interested persons may interrogate Linked Data, download them for a future reuse, create links to these data or create innovative services.

The implemented tool is still under development. Some stages, such as converting tables into canonical form, are still done manually and must be accomplished automatically. Also, it should have a support for semantic micro tags such as RDFa which are becoming more and more common in Web sites.

A full comparative study will be done, covering all stages of the proposed approach.

As a future work, we shall generalise our approach so that a full reengineering process from legacy Web application to Linked Data can be performed.

REFERENCES

[1] Arenas, M.—Bertails, A.—Prud'hommeaux, E.—Sequeda, J.: A Direct Mapping of Relational Data to RDF. W3C Recommendation 27 September 2012. Available at: `http://www.w3.org/TR/2012/REC-rdb-direct-mapping-20120927/`.

[2] Astrova, I.—Korda, N.—Kalja, A.: Rule-Based Transformation of SQL Relational Databases to OWL Ontologies. Proceedings of the 2nd International Conference on Metadata and Semantics Research, October 2007.

[3] Atencia, M.—David, J.—Scharffe, F.: Keys and Pseudo-Keys Detection for Web Datasets Cleansing and Interlinking. Proceedings of the 18th International Conference on Knowledge Engineering and Knowledge Management (EKAW 2012). Lecture Notes in Computer Science, Vol. 7603, 2012, pp. 144–153, doi: 10.1007/978-3-642-33876-2_14.

[4] Beckett, D.—Berners-Lee, T.: Turtle – Terse RDF Triple Language. W3C Team Submission 28 March 2011. Available at: `http://www.w3.org/TeamSubmission/turtle/`.

[5] Berners-Lee, T.: Linked Data – Design Issues. 2006. Available at: `http://www.w3.org/DesignIssues/LinkedData.html`.

[6] Cafarella, M. J.—Halevy, A. Y.—Wang, D. Z.—Wu, E.—Zhang, Y.: WebTables: Exploring the Power of Tables on the Web. Proceedings of the VLDB Endowment, Vol. 1, 2008, No. 1, pp. 538–549, doi: 10.14778/1453856.1453916.

[7] Cresci, S.—D'Errico, A.—Gazze, D.—Lo Duca, A.—Marchetti, A.—Tesconi, M.: Towards a DBpedia of Tourism: The Case of Tourpedia. Proceedings of the ISWC 2014 Posters and Demonstrations Track, 13th International Semantic Web Conference (ISWC 2014), Riva del Garda, Italy, October 21, 2014, Vol. 1272, pp. 129–132.

[8] Ding, L.—Lebo, T.—Erickson, J. S.—DiFranzo, D.—Williams, G. T.—Li, X.—Michaelis, J.—Graves, A.—Zheng, J. G.—Shangguan, Z.—Flores, J.—McGuinness, D. L.—Hendler, J. A.: TWC LOGD: A Portal for Linked Open Government Data Ecosystems. Journal of Web Semantics: Science, Services and Agents on the World Wide Web, Vol. 9, 2011, No. 3, pp. 325–333.

[9] DNB: The Linked Data Service of the German National Library: Modelling of Bibliographic Data. German National Library (Leipzig, Frankfurt am Main), last update 15-09-2014. Available at: `http://www.dnb.de/SharedDocs/Downloads/EN/DNB/service/linkedDataModellierungTiteldaten.pdf`.

[10] Dumontier, M.—Callahan, A.—Cruz-Toledo, J.—Ansell, P.—Emonet, V.—Belleau, F.—Droit, A.: Bio2RDF Release 3: A Larger Connected Network of Linked Data for the Life Sciences. Proceedings of the ISWC 2014 Posters and Demonstrations Track, 13th International Semantic Web Conference (ISWC 2014), Riva del Garda, Italy, October 21, 2014, Vol. 1272, pp. 401–404.

[11] Embley, D. W.—Campbell, D. M.—Jiang, Y. S.—Liddle, S. W.—Lonsdale, D. W.—Ng, Y. K.—Smith, R. D.: Conceptual-Model-Based Data Extraction from Multiple-Record Web Pages. Data and Knowledge Engineering, Vol. 31, 1999, No. 3, pp. 227–251, doi: 10.1016/S0169-023X(99)00027-0.

[12] Embley, D. W.—Krishnamoorthy, M.—Nagy, G.—Seth, S.: Factoring Web Tables. Modern Approaches in Applied Intelligence (IEA/AIE 2011). Springer, Lecture Notes in Computer Science, Vol. 6703, 2011, pp. 253–263, doi: 10.1007/978-3-642-21822-4_26.

[13] Exner, P.—Nugues, P.: Entity Extraction: From Unstructured Text to DBpedia RDF Triples. Proceedings of the Web of Linked Entities Workshop (WoLE 2012), Boston, USA, November 11, 2012.

[14] Gagnon, M.—Barrière, C.—Charton, E.: Full Syntactic Parsing for Enrichment of RDF Dataset. Proceedings of the 12th International Semantic Web Conference, Sydney, Australia, October 21–25, 2013.

[15] Hitzler, P.—Krotzsch, M.—Parsia, B.—Patel-Schneider, P. F.—Rudolph, S.: OWL 2 Web Ontology Language Primer (Second Edition). W3C Recommendation 11 December 2012. Available at: `http://www.w3.org/TR/owl2-primer/`.

[16] Hoffart, J.—Suchanek, F. M.—Berberich, K.—Weikum, G.: YAGO2: A Spatially and Temporally Enhanced Knowledge Base from Wikipedia. Artificial Intelligence Journal, Vol. 194, 2013, pp. 28–61, doi: 10.1016/j.artint.2012.06.001.

[17] Huhtala, Y.—Kärkkäinen, J.—Porkka, P.—Toivonen, H.: Tane: An Efficient Algorithm for Discovering Functional and Approximate Dependencies. Computer Journal, Vol. 42, 1999, No. 2, pp. 100–111, doi: 10.1093/comjnl/42.2.100.

[18] Jovanovik, M.—Najdenov, B.—Trajanov, D.: Linked Open Drug Data from the Health Insurance Fund of Macedonia. Proceedings of the 10th Conference for Informatics and Information Technology (CIIT 2013), Bitola, Macedonia, April 18–21, 2013, pp. 56–61.

[19] Kim, S.—Iglesias-Sucasas, M.—Viollier, V.: The FAO Geopolitical Ontology: A Reference for Country-Based Information. Journal of Agricultural and Food Information. Vol. 14, 2013, No. 1, pp. 50–65.

[20] Konstantinou, N.—Spanos, D. E.—Houssos, N.—Mitrou, N.: Exposing Scholarly Information as Linked Open Data: RDFizing DSpace Contents. International Journal for the Application of Technology in Information Environments (Electronic Library), Vol. 32, 2014, No. 6, pp. 834–851, doi: 10.1108/EL-12-2012-0156.

[21] Kyzirakos, K.—Vlachopoulos, I.—Savva, D.—Manegold, S.—Koubarakis, M.: GeoTriples: A Tool for Publishing Geospatial Data as RDF Graphs Using R2RML Mappings. Proceedings of the ISWC 2014 Posters and Demonstrations Track, 13th International Semantic Web Conference (ISWC 2014), Riva del Garda, Italy, October 21, 2014, Vol. 1272, pp. 393–396.

[22] Lehmann, J.—Isele, R.—Jakob, M.—Jentzsch, A.—Kontokostas, D.—Mendes, P. N.—Hellmann, S.—Morsey, M.—van Kleef, P.—Auer, S.—Bizer, C.: DBpedia – A Large-Scale, Multilingual Knowledge Base Extracted from Wikipedia. Semantic Web, Vol. 6, 2015, No. 2, pp. 167–195, doi: 10.3233/SW-140134.

[23] Li, M.—Du, X.—Wang, S.: Learning Ontology from Relational Database. Proceedings of the Fourth International Conference on Machine Learning and Cybernetics, Vol. 6, 2005, pp. 3410–3415.

[24] Maali, F.—Cyganiak, R.—Peristeras, V.: A Publishing Pipeline for Linked Government Data. Proceedings of the 9th Extended Semantic Web Conference (ESWC 2012), Heraklion, Crete, Greece, May 27–31, 2012. Lecture Notes in Computer Science, Vol. 7295, 2012, pp. 778–792.

[25] Mannila, H.—Raiha, K. J.: Algorithms for Inferring Functional Dependencies from Relations. Data and Knowledge Engineering, Vol. 12, 1994, No. 1, pp. 83–99, doi: 10.1016/0169-023X(94)90023-X.

[26] MDN: Mozilla Developer Network and individual contributors. DOM Developer Guide. 2013. Available at: `https://developer.mozilla.org/en-US/docs/Web/Guide/API/DOM`.

[27] Miller, G. A.: WordNet: An On-Line Lexical Database. International Journal of Lexicography, 1990, pp. 235–312, doi: 10.1093/ijl/3.4.235.

[28] Mulwad, V.—Finin, T.—Syed, Z.—Joshi, A.: T2LD: Interpreting and Representing Tables as Linked Data. Proceedings of the ISWC 2010 Posters and Demonstrations Track: Collected Abstracts, Shanghai, China, November 9, 2010.

[29] Mulwad, V.—Finin, T.—Joshi, A.: Semantic Message Passing for Generating Linked Data from Tables. Proceedings of the 12th International Semantic Web Conference, Sydney, Australia, October 21–25, 2013, doi: 10.1007/978-3-642-41335-3_23.

[30] Munoz, E.—Hogan, A.—Mileo, A.: Using Linked Data to Mine RDF from Wikipedia's Tables. Proceedings of the 7th ACM International Conference on Web Search and Data Mining (WSDM '14), New York City, USA, February 24–28, 2014, pp. 533–542.

[31] Nagy, G.—Embley, D. W.—Machado, S.—Seth, S.—Jin, D.—Krishnamoorthy, M.: Data Extraction from Web Tables: The Devil is in the Details. Proceedings of the International Conference on Document Recognition (ICDAR '11), Beijing, September 2011, doi: 10.1109/ICDAR.2011.57.

[32] Padmanabhan, R. K.—Jandhyala, R. C.—Krishnamoorthy, M.—Nagy, G.—Seth, S.—Silversmith, W.: Interactive Conversion of Web Tables. In: Ogier, J.-M., Liu, W., Lladós, J. (Eds.): Graphics Recognition. Achievements, Challenges, and Evolution (GREC 2009). Lecture Notes in Computer Science, Vol. 6020, 2010, pp. 25–36.

[33] Pedersen, T.—Patwardhan, S.—Michelizzi, J.: WordNet::Similarity – Measuring the Relatedness of Concepts. Proceedings of the Nineteenth National Conference on Artificial Intelligence (AAAI-04), 2004, pp. 1024–1025, doi: 10.3115/1614025.1614037.

[34] Pirro, G.—Euzenat, J.: A Feature and Information Theoretic Framework for Semantic Similarity and Relatedness. Proceedings of the 9th International Semantic Web Conference (ISWC 2010). Springer, Lecture Notes in Computer Science, Vol. 6496, 2010, pp. 615–630.

[35] Rezk, M.—Park, J.—Yongun, Y.—Lim, K.—Larsen, J.—Hahm, Y.—Choi, K.-S.: Korean Linked Data on the Web: Text to RDF. Proceedings of the Joint International Semantic Technology Conference (JIST 2012), Nara, Japan, 2012. Lecture Notes in Computer Science, Vol. 7774, 2012, pp. 368–374.

[36] Sahoo, S. S.—Halb, W.—Hellmann, S.—Idehen, K.—Thibodeau Jr., T.—Auer, S.—Sequeda, J.—Ezzat, A.: A Survey of Current Approaches for Mapping of Relational Databases to RDF. Technical Report, W3C, 2009. Available at: `http://www.w3.org/2005/Incubator/rdb2rdf/RDB2RDF_SurveyReport.pdf`.

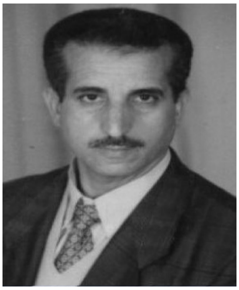
[37] Scharffe, F.—Atemezing, G.—Troncy, R.—Gandon, F.—Villata, S.—Bucher, B.—Hamdi, F.—Bihanic, L.—Képéklian, G.—Cotton, F.—Euzenat, J.—Fan, Z.—Vandenbussche, P.—Vatant, B.: Enabling Linked Data Publication with the Datalift Platform. Proceedings of AAAI Workshop on Semantic Cities, 2012.

[38] Szekely, P.—Knoblock, C. A.—Yang, F.—Zhu, X.—Fink, E. E.—Allen, R.—Goodlander, G.: Connecting the Smithsonian American Art Museum to the Linked Data Cloud. Proceedings of the 10th International Conference (ESWC 2013), Montpellier, France, May 26–30, 2013. The Semantic Web: Semantics and Big Data. Lecture Notes in Computer Science, Vol. 7882, 2013, pp. 593–607.

[39] Tijerino, Y. A.—Embley, D. W.—Lonsdale, D. W.—Ding, Y.—Nagy, G.: Towards Ontology Generation from Tables. World Wide Web: Internet and Web Information Systems, Vol. 8, 2005, No. 3, pp. 261–285.

[40] Willighagen, E. L.—Waagmeester, A.—Spjuth, O.—Ansell, P.—Williams, A. J.—Tkachenko, V.—Hastings, J.—Chen, B.—Wild, D. J.: The ChEMBL Database as Linked Open Data. Journal of Chemiformatics, Vol. 5, May 2013, doi: 10.1186/1758-2946-5-23.

Djelloul Bouchiha received his Engineer and M.Sc. degrees in computer science from Sidi Bel Abbes University, Algeria, in 2002 and 2005, respectively, and Ph.D. in 2011. In the scope of years 2005 and 2010, he joined the Department of Computer Science, Saida University, Algeria, as Lecturer. He became Assistant Professor since January 2011. Currently, he is Assistant Professor at the University Center of Naama. His research interests include semantic web services, web reverse-engineering, ontology engineering, knowledge management and information systems.



Mimoun Malki graduated with Engineer degree in computer science from National Institute of Computer Science, Algiers, in 1983. He received his M.Sc. and Ph.D. in computer science from the University of Sidi Bel-Abbes, Algeria, in 1992 and 2002, respectively. He was Associate Professor in the Department of Computer Science at the University of Sidi Bel-Abbes from 2003 to 2010. Currently, he is Full Professor at Djillali Liabes University of Sidi Bel-Abbes, Algeria. He has published more than 50 papers in the fields of web technologies, ontology and reverse engineering. He is the Head of the Evolutionary Engineering and Distributed Information Systems Laboratory. Currently, he serves as an editorial board member for the International Journal of Web Science. His research interests include databases, information systems interoperability, ontology engineering, web-based information systems, semantic web services, web reengineering, enterprise mash up and cloud computing.

Abdullah Alghamdi is Full Time Professor, SWE Department, College of Computer and Information Sciences, KSU, Riyadh, KSA. He holds his Ph.D. in software engineering from the Department of Computer Science, Sheffield University, UK, 1997. He got a Post-Doc certificate from University of Ottawa, Canada, where he conducted a joint research at the MCRLab during academic year 2004/2005. He worked as Full and Part Time Consultant with governmental and private organizations in the field of IS strategic planning and defense systems and headed a number of committees inside and outside KSU. He recently published a number of papers in C4I and Enterprise Architecture Frameworks fields. Currently he is Head of Software Engineering Department, KSU and Director of the national C4I Center for Advanced Systems (C4ICAS).

Khalid Alnafjan received his Bachelor's degree in information systems from King Saud University, Saudi Arabia, in 1991, his M.Sc. in computer science from Sheffield University, United Kingdom, in 1994, and Ph.D. in computer science in 1998 from the same university. From 1999 to 2009 he joined the Department of Computer Technology, Riyadh College, GOTEVOT as Assistant Professor, and from 2010 to date he is Associate Professor in the Department of Software Engineering, King Saud University. His research interests include software engineering education, software quality assurance, C4i quality assurance and semantic web services.