I.J. Modern Education and Computer Science, **2017, 3, 56-63**
Published Online March 2017 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijmecs.2017.03.07

An Approach for Discovering and Maintaining Links in RDF Linked Data

Fatima Ardjani
EEDIS Laboratory, University of Djillali Liabès, Sidi BEL-ABBES, Algeria
Email: ardjanif@yahoo.fr

Djelloul Bouchiha
EEDIS Lab, Ctr Univ Naama, Inst. Sciences and Technologies, Dept. Mathematics and Computer Science, Algeria
Email: bouchiha.dj@gmail.com

Mimoun Malki
LabRI-SBA, Higher School of Computer Science, SBA, Algeria
Email: Malki.Mimoun@univ-sba.dz

***Abstract*—**Many datasets are published on the Web using semantic Web technologies. These datasets contain data that represent links to similar resources. If these datasets are linked together by properly constructed links, users can easily query the data through a uniform interface, as if they were querying a single dataset. In this paper we propose an approach to discover (semi) automatically links between RDF data based on the description models that appear around the resources. Our approach also includes a (semi) automatic process to maintain links when a data-change occurs.

***Index Terms*—**Linked Data, Link maintenance, Detection of links, Ontology alignment.

I. Introduction

Web of data is a collaborative movement to extend the Web by shared structured data. Its founding idea, expressed by Tim Berners-Lee in 2001 [20], is inspired by the structure of the Web pages - linked by hypertext links - to propose a new standardized representation of the data, exploitable both by the human and the machine.

The Web of data is based on the RDF framework that formats the data in the form of triplets. A triplet is composed of three elements: a subject, a predicate and an object. These triplets relate RDF resources that designate resources from the Web, the real world, or general concepts. Each RDF resource has a unique identifier, which is usually the URL of a Web page associated with the resource. The RDF schema and OWL languages are used to organize RDF resources in hierarchical classes and to define relations that can link resources [20].

These languages can also be used to make inferences from RDF data based on the description logic. The use of RDFS and OWL languages to structure data and organize them into triplets to form data graphs makes RDF bases more flexible than relational databases. The RDF bases can be interrogated by the queries of the SPARQL language [20].

The most tangible incarnation of the Web of data is the Linked Data [20], which appeared in 2008, along with SPARQL. It is a grouping of bases which concern various fields and which follow common rules for the structuring and publication of data. These bases are interconnected by equivalence relations between RDF resources designating the same elements in different bases.

The RDF resources of the bases of Linked Data are also all associated with legible Web pages by the human and all provide a means of access to their content, by downloading the base or by SPARQL queries. For nine years, Linked Data has been one of the engines of "Web of data" growth, driven by many users and contributors who make community live the hundreds of the Linked Data [20].

The Linked Data initiative aims at publishing structured and interlinked data on the Web by using Semantic Web technologies [20]. These technologies provide different languages for expressing data as graphs (RDF) and querying it (SPARQL) [11]. Linked data allow the implementation of applications that reuse data distributed on the Web. To facilitate interoperability between these applications, data issued from different providers has to be interlinked. It means that the same entity in different data sets must be identified. One of the key challenges of linked data is to deal with this heterogeneity by detecting links across datasets [10]. In such a dynamic environment, the Web of data evolves: new data are added, outdated data are removed or changed. Then, links between data have to evolve too. Since links should not be recomputed each time a change occurs, the semantic Web needs methods that consider the evolution.

Over the time, dead links can appear. Dead links are those pointing at URIs that are no longer maintained, and those that are not being set when new data is published. Too many dead links lead to a large number of unnecessary HTTP requests by client applications. A current research topic addressed by the Linked Data

community is link maintenance [6]. Fig.1. represents an overall overview of the presented issue.

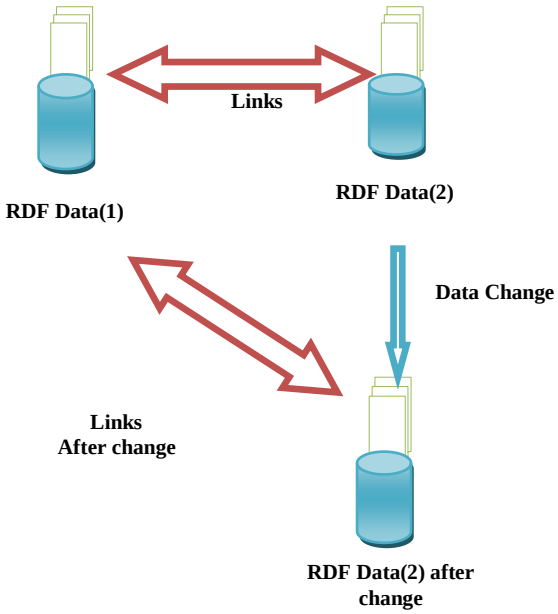

Fig.1. Discovering and maintaining links between RDF Linked Data

The basics of the Web of data contain a core of initial data created by an expert (or group of experts) in the knowledge domain of the base [20].

The updates made by contributors then increase the content of the base. In the basics of the Linked data, it is the contributors from the communities formed around the bases that make these updates. In these bases, the quality of links is particularly important because it affects the quality of the other interconnected bases of Linked Data. In order to maintain the quality of links, it is necessary to evaluate the quality of the data changes and to filter them according to the result of their evaluation. Many evaluation methods have been proposed, but none has been successful in the Linked Data.

Our paper is organized as follows: The state of art is presented in section 2. In section 3 we present the basic concepts of Linked Data. We detail our approach for discovering inter/intra-base links and maintaining links in section 4. In section 5, we conclude by reviewing our work and mentioning some perspectives.

II. State of the Art

Several solutions have been proposed to solve the problem of discovering and maintaining links: the most direct approach is to compare the attribute values of the instances to identify the links [6], but it is impossible to compare all possible pairs of attribute values [19]. Another common strategy is to compare instances according to the corresponding attributes found by instances-based ontology alignment, which allows generating attribute mappings based on instances [9]. However, it is difficult to identify similar instances across data sets because, in some cases, the values of the matching attributes are not the same. Several methods use genetic programming to construct interconnection models to compare different instances, but they suffer from long run times. Some approaches are numerical and use complex similarity measures, aggregation functions, and thresholds to construct rules. These approaches are often adapted to some set of data. In some cases, this type of rule is automatically discovered [14, 3, 2, 7]. Other approaches are based on logical rules [8, 17] that are generated automatically using the semantics of the keys or functional properties. This type of knowledge can also be used by an expert to build more complex similarity functions that take more account of the properties that are involved in the keys [12, 4]. The LiQuate tool [15] uses an evaluation approach based on Bayesian networks to identify inconsistencies, ambiguities and incomplete relations in the links between resources of different bases.

III. Basic Concepts

The Web of Data provides the framework for creating RDF bases and services based on RDF data. The Linked Data is a set of RDF bases free of access and following several common principles.

We present in this section the founding principles of Linked Data.

A. RDF-Resource Description Framework - [18]

RDF is the basis of technologies of the Web of data. It is a simple data model for representing knowledge on the Web in RDF documents. As the name suggests, RDF is used to describe RDF resources. Each RDF resource is - theoretically - unique throughout the Web of data. The RDF resource descriptions are composed of RDF triplets (subject, relation, object), which can be represented as a graph. An RDF triple expresses a relation between a subject - the described resource - and an object. The description of a resource corresponds to the set of triplets that contain the resource in subject [1]. Consider an RDF triplet (subject, relation, object):

- The subject is a resource that is explicitly identified by a URI, or by a white node. A URI - Uniform Resource Identifier - is a unique identifier on the Web.
- The relation is always an identified resource (URI);
- The object is either a resource (identified or not) or a raw data item, also called literal.

An RDF document is also called a set of RDF triplets or an RDF graph. We will assume that the base triplets are contained in the default graph, and therefore, the designations RDF data set and RDF basis are equivalent to RDF graph.

An RDF base is accessible by an HTTP access point which queries publicly (endpoint) or in the form of exporting the base contents into a downloadable file (a dump of the base).

The base should contain links to other bases of Linked Data. Most often these links are made by equivalence

relation "owl: sameAs", between resources of different bases.

- We call resources nodes, the resources in subjects and objects in triplets. We define two sets for all RDF documents, the set of resources and the set of node resources.
- Let a set of RDF triplets S. The set R_S containing subjects, relations and objects of the triplets of S is called the set of resources of S.
- The set N_S containing subjects and objects of the triplets of S is called the set of the node resources of S.
- The set K_S containing subjects and objects of the triplets of S which are white nodes is called the set of white nodes resources of S.
- We call the related RDF document, an RDF document in which there is a path connecting each node resource of the document to all the others. In other words, if the graph that represents the RDF document is a related graph.
- We call the degree of a node resource, the number of triplets that contain it in an RDF database. In other words, its degree in the graph that represents the RDF base.
- The content of an RDF document can be seen as the sum of the descriptions of the resources it contains. In RDF terminology, a description of a resource designates in particular the set of triplets around a given node resource, that is, its neighborhood. This neighborhood is not limited to the triplets containing the resource; it can be extended to include all neighboring triplets up to a rank n.
- Let a set of RDF triplets L, $r \in N_L$ and $n \in N*$.
- The $neighborhoodfunction_L^n(r)$: $N_L \to L$ such that:

$$neighborhood_L^1(r) = \{t \mid t \in L,\ r \in N\{t\}\} \quad (1)$$

- For $n > 1$,

$$neighborhood_L^n(r) = neighborhood_L^{n-1}(r) \cup \bigcup_{k \in N^{neighborhood_{L(r)}^{n-1}}} neighborhood_L^1(k) \quad (2)$$

Descriptive graph of a resource: We define the descriptive graph of a resource, grouping a resource and its neighborhood. Let an RDF base L, a resource $r \in R_L$ and $n \in N*$, we call a descriptive graph at the rank *n* of the resource *r* in *L* the pair $(r,\ neighborhood_L^n(r))$, and r is the center of the descriptive graph.

The descriptive graph of a resource is composed of the resource and its description in the RDF base, that is, the neighborhood of the resource up to a given rank. In row 1, the descriptive graph forms a star graph centered on r. We want to obtain the common points of several descriptive graphs, in the form of a description model of descriptive graphs.

B. Description models

A description model is composed of triplets that contain elements common to the descriptive graphs of several resources. The node resources of a description model are either descriptive graph resources at its origin or white nodes. Let a non-empty RDF base L, a set of resources $R \subseteq R_L$ called set of descriptive centers, with $R = \{r_1,\ r_2,\ ...,\ r_n\}$ and G the set of descriptive graphs of the resources of R of the same rank in L with $G = \{(r_1, S_{r1}), (r_2, S_{r2}),\ ..., (r_n,\ S_{rn})\}$. The set of non-empty and related triplets M, with $r \in K_M$, is a description model if, for each element (r_i, S_i) of G we can obtain *M* from a subset of S_i by applying the following rules:

- The resource r_i is transformed into r;
- A node resource is transformed into a white node.

The resources of *R* are called descriptive centers of M and the node *r* is called the root node of *M*.

A triplet of a description model is composed of elements that appear in the same way in a triplet of each descriptive graph (being subject, relation or object) or white nodes. The descriptive centers of a model are the elements at the origin of the model.

Thus, a description model always contains a white node called the root node of the model in the triplets, which contain the common resources of the triplets, which contain the descriptive centers in their respective descriptive graphs.

From a graphical point of view, the neighborhood of the root of the description model is structurally similar to the (at a part of) neighborhood of each descriptive center.

Note that in a description model, not all resources common to the descriptive graphs are necessarily present.

Occurrence, size and description value of a model: Let *R* be a set of resources, *M* a description model of *R* for a rank $n \in N$ and $r \in K_M$ the root of *M*.

The function *occ(M)* such that $occ(M) = |R|$ is called the occurrence of *M*.

The size function (M) such that $size(M) = |M|$ is called the size of *M*.

The occurrence of a description model is the number of descriptive centers at the origin of the model.

The size of a description model is the number of triplets it contains.

C. Ontology

These many bases of Linked Data use an exclusive ontology. There are also several ontologies common to a large number of bases. It should be noted that the RDFS language is currently the most used. The common ontologies in the Linked Data increase the possibilities of using and merging data from different bases. Among

these common ontologies, we find mainly the following four:

- Dublin Core: Dublin core is an ontology defined in RDFS to allow the description of digital documents. It is maintained by the Dublin Core Metadata Initiative.
- FOAF: Friend Of A Friend is an ontology defined in OWL and designed to summarily describe and connect people, organizations and documents.
- SKOS: Simple Knowledge Organization System is an ontology defined in OWL that allows describing knowledge - taxonomies - in a complementary way to OWL. It allows the definition of generalization/specialization relations between individuals.
- The W3C defines the VoID ontology [5] to inform the metadata about the RDF bases in the RDF bases. This ontology was created to facilitate the discovery and listing of Web bases. It is designed to allow the description of the data inside the base and its links with other bases. It uses some Dublin Core ontology relation to describe the basics as digital documents and FOAF to describe their sites. It also makes it possible to describe the information around the base, such as the license of the base, all the links to other bases or the technical details for its access. The VoID description of an RDF base must be placed in a file following the Turtle syntax, named void.ttl, accessible at the root of the base domain.

D. Integrity constraints

Integrity constraints are a concept derived from databases. They are not natively defined in the Web of data because they rely on the hypothesis of a closed world for data, while RDF is based on an open world. As described in [13] for the Web of data, integrity constraints correspond to the rules generated from certain definitions of the ontology whose respect is required to maintain the consistency of the RDF database. Some ontological definitions can be considered as integrity constraints for the description of resources. They are used when reasoning in classification description logics (organizing the hierarchy of types in the ontology) and when instantiating (assigning the most specific type to an individual or a literal one). In practice, these constraints define rules to be respected when creating data for maintaining the consistency. According to [16], integrity constraints derived from an ontology are to be classified similarly to those used in relational databases: Typing constraints, which require that the resources linked by a given relation with a specific type, are generated by the domains and co-domains of the properties as well as by the disjunctions. Uniqueness constraints, which require that a resource cannot be present in the same way in more than one triplet containing the same relation, are generated by the functional properties. Definition constraints, which require that one resource linked to another by a triplet containing a relation or specific node resources, are generated by the class definitions with restriction in OWL.

IV. Proposed Approach

In the Web of data, ontologies are used to frame the data structure of an RDF base. The ontology defines the set of classes and relations that can be used in the base. The definitions of classes and relations are made in such a way as to frame their uses by constraints in order to preserve the consistency of the data in the base. Over the time, dead links can appear. Dead links are those pointing at URIs that are no longer maintained, and those that are not being set when a new data is published. Too many dead links lead to a large number of unnecessary HTTP requests by the users' applications. A current research topic addressed by the Linked Data community is link maintenance. With each change, it is necessary to check if they do not lead to the appearance of new inconsistencies. It is also necessary to check whether the equivalences between the resources of different bases are maintained at each evolution. It is also necessary to keep track of the changes. The goal of our approach is to detect the correct links and erroneous links in the same base (inter-base links) as well as in a basic set (intra-base links), and give a maintenance method to avoid any heterogeneity.

A. Process of Discovering "Inter-base" links

The method deals with data that is inconsistent with the ontology. To identify links that respect ontology or not, we use integrity constraints. Integrity constraints allow to extract the resource base that respects or not the ontology. To determine whether some of these resources are similar between them, the description models give a summary of the common points of a set of resources in the form of a set of triplets. Positive models are those that respect a constraint, and negative models are those that do not respect it. Positive models are those that respect a constraint, i.e. a set of triplets which respect the integrity constraint of an ontology; so one can extract the number of correct links in this set. Negative models are those that do not respect a constraint, i.e. a set of triplets that do not respect the integrity constraint of an ontology; so one can extract the number of erroneous links in this set. The method is shown in Fig.2.

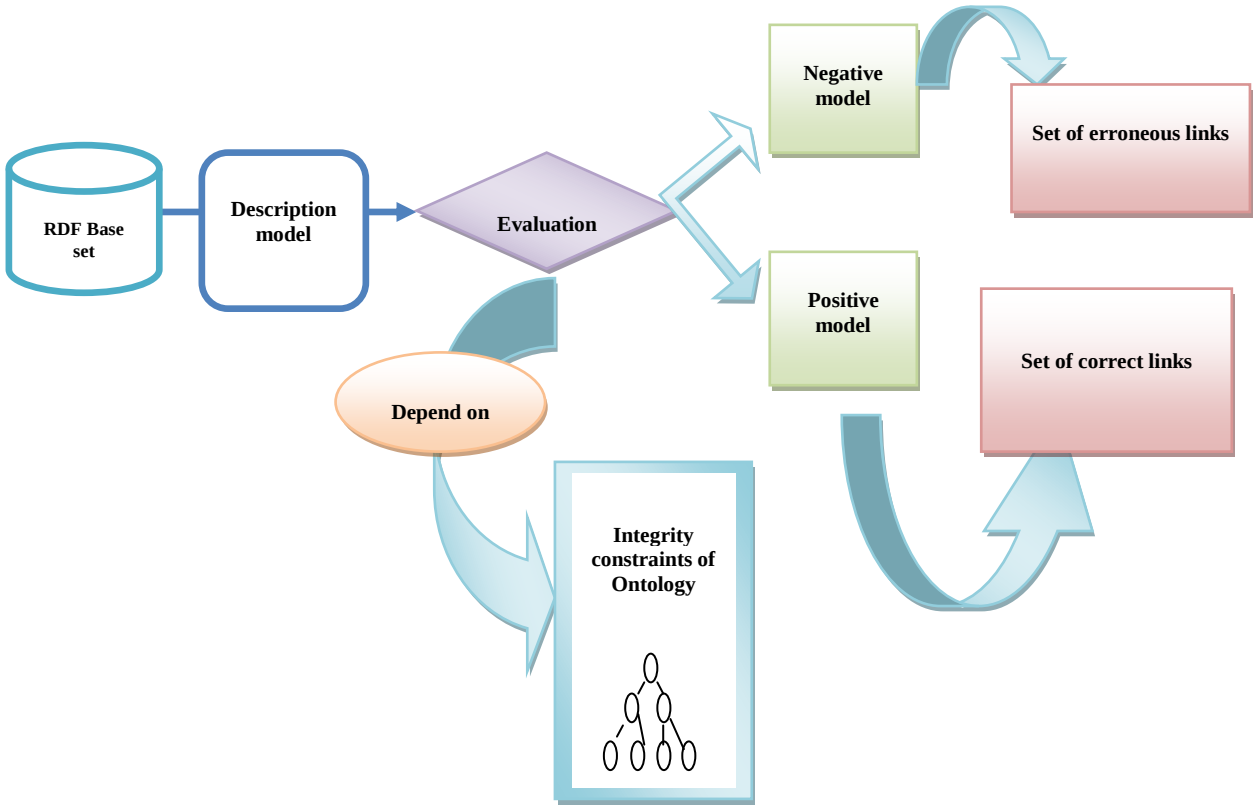

Fig.2. Proposed approach for discovering inter-base links

B. Process of Discovering "Intra-base"links

This method is based on the data of two bases in accordance with two different ontologies. To identify the links, we use the description models and the alignment of two instances-based ontologies.

* Positive models of base1 are those that respect a constraint of ontology1, i.e. a set of triplets that respect the integrity constraint of the ontology1, so one can extract the number of the correct links in this set. * Negative models of base1 are those that do not respect it, i.e. a set of triplets that do not respect the integrity constraint of the ontology1, so we can extract the number of erroneous links in this set, which do not respect the integrity constraint.

* Positive models of base2 are those which respect a constraint of the ontology2, i.e. a set of triplets which respect the integrity constraint of the ontology1, so we can extract the number of the correct links in this set.

* Negative models of base2 are those that do not respect it, i.e. a set of triplets which do not respect the integrity constraint of the ontology2, so one can extract the number of the erroneous links in this set or that do not respect the integrity constraint.

* Positive models of base1-base2 are those that respect the correspondence of the alignment of two ontologies, i.e. a set of triplets that respect the correspondence of the alignment of two ontologies, so we can extract the correct number of links in this set.

* Negative models of base1-base2 are those who do not respect the alignment of matches, i.e. a set of triplets that do not respect the correspondence of the alignment of two ontologies, so we can extract the number of the erroneous links in this set So :

- ☞ The number of correct links = the number of correct links of base1 + the number of correct links of base2 + the number of correct links of base1-base2.
- ☞ The number of erroneous links = the number of erroneous links of base1 + the number of erroneous links of base2 + the number of erroneous links of base1-base2.

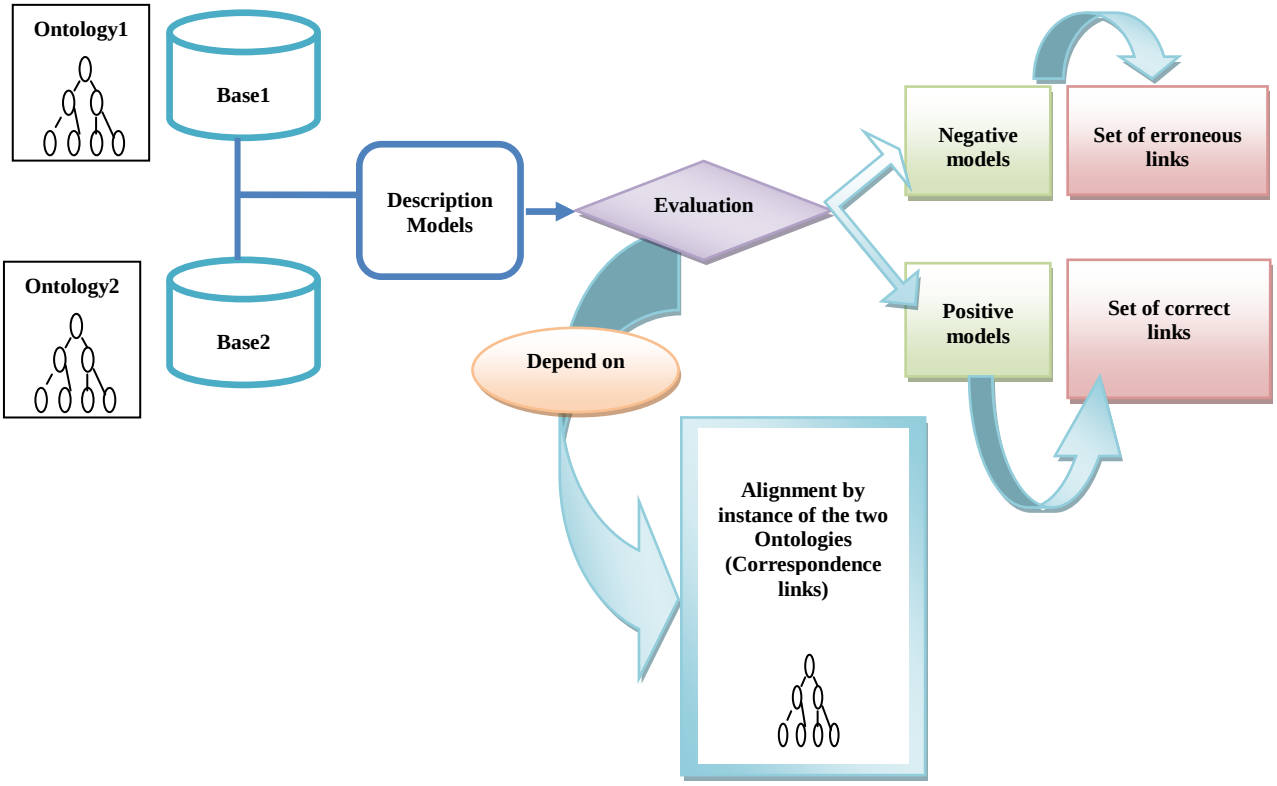

Fig.3. Proposed approach for discovering intra-base links

Our method of detecting "intra-base" links is shown schematically in Fig.3.

C. Maintenance Process

To maintain the quality of the linked data, it is necessary to evaluate the quality of the changes in data and to filter them according to the result of their evaluation. When an update does not respect the positive models, it implies that there is an inconsistent with the ontology, so the change is refused.

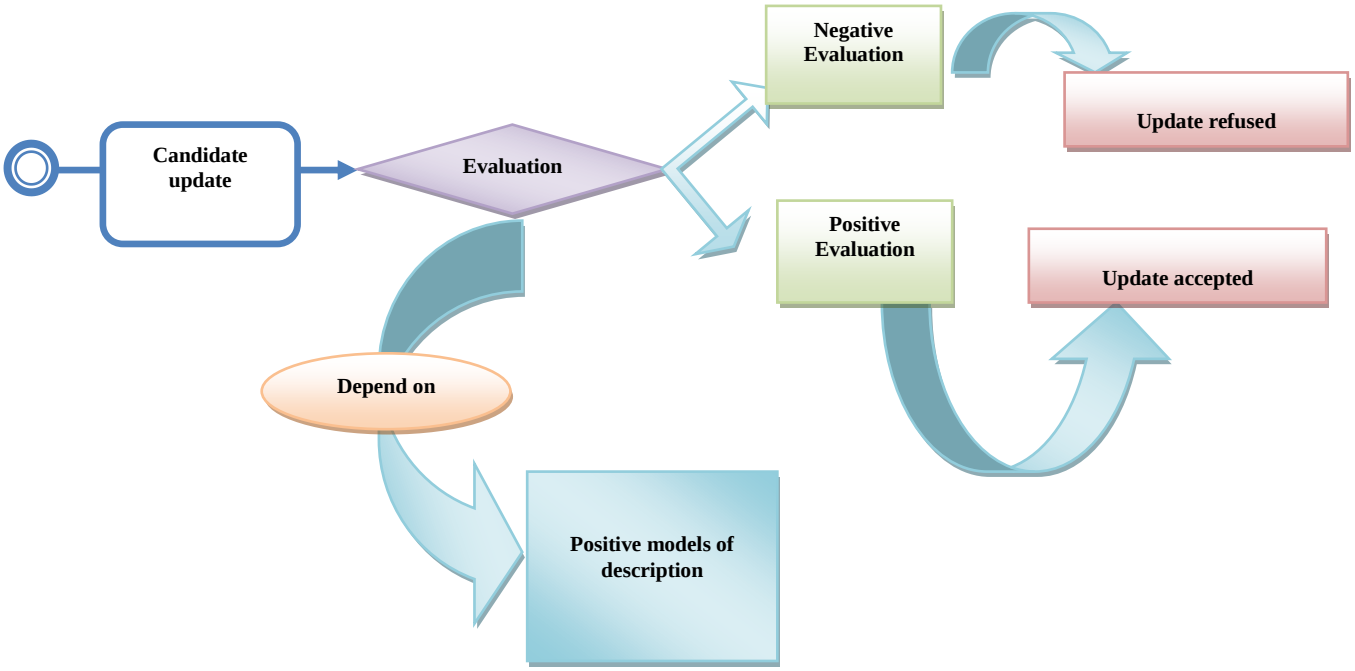

Fig.4. Proposed approach for maintenance

V. Conclusion and perspectives

Linked Data is the most visible and known part of the Web of data. It is a set of RDF bases in various domains, which respect several publication rules for sharing data. The quality of the data available on the Web of data is not well assured. The number of RDF bases and their sizes are constantly growing. This accumulation of data is done without a standardized or recognized mechanism to evaluate the quality of linked data. So the quality assessment is a crucial problematic in the Linked Data. It depends mainly on the quality of links between data resources.

Contributions of this article give original answers to several recurring problems which appear throughout the life of data in the Web of data. They respond mainly to three problems: discovering inter-base links, discovering intra-base links and link maintenance.

"The discovery of inter-base links" deals with data that is inconsistent with the ontology. To identify the links that respect ontology or not, we define the description models, which give a summary of the common points of a set of similar resources, based on integrity constraints.

"The discovery of intra-base links" is based on data from two bases in accordance with two different ontologies. To identify the links, we use the description models and the alignment of two ontologies based on instances.

"The link maintenance" is based on the quality of the linked data. It is necessary to evaluate the quality of changes in data and to filter them according to the result of their evaluation. When an update does not respect the positive models, it implies that there is an inconsistent with the ontology, so the change is refused.

This first work will be experimented on real data as "DBPedia", which is one of the most important bases of Linked Data in terms of content and community. Our results will be compared to other related work.

References

[1] S. Albagli, R. Ben-Eliyahu-Zohary, and S.E. Shimony, “Markov network based ontology matching”, J. Comput. Syst. Sci. 78(1), 105–118, 2012.

[2] A. Nikolov, M. d’Aquin, and E. Motta, “Unsupervised learning of link discovery configuration” In 9th Extended Semantic Web Conference (ESWC), pages 119–133, Berlin, Heidelberg, 2012.

[3] A. Cyrille, N. Ngomo, and K. Lyko, “Eagle: Efficient active learning of link specifications using genetic programming”, In 9th Extended Semantic Web Conference (ESWC), pages 149–163, 2012.

[4] A. Cyrille, N. Ngomo, and S. Auer, “Limes a time-efficient approach for large-scale link discovery on the Web of data”, In IJCAI, pages 2312–2317, 2011.

[5] R. Cyganiak, J. Zhao, M. Hausenblas, and K. Alexander, “Describing linked datasets with the VoID vocabulary”, W3C note, W3C, 2011.

[6] J. Euzenat, and P. Shvaiko, “Ontology matching”, Springer Verlag, Heildelberg (DE), 2013.

[7] F.M. Suchanek, S. Abiteboul, and P. Senellart, “Paris: Probabilistic alignment of relations, instances, and schema”, The Proceedings of the VLDB Endowment (PVLDB), 5(3):157–168, 2011.

[8] F. Saïs, N. Pernelle, and M.C. Rousset, “Combining a logical and a numerical method for data reconciliation”, Journal on Data Semantics, 12:66–94, 2009.

[9] F. Ardjani, D. Bouchiha, and M. Malki, "Ontology-Alignment Techniques: Survey and Analysis", IJMECS, vol.7, no.11, pp.67-78, 2015.DOI: 10.5815/ijmecs.2015.11.08.

[10] A. Ferrara, A. Nikolov, and F. Scharffe, “Data Linking for the Semantic Web”. Int. J. Semantic Web Inf. Syst., 7(3), 46,76, 2011.

[11] P. Hitzler, M. Krötzsch, and S. Rudolph, “Foundations of semantic Web technologies”, Chapman & Hall/CRC, 2009.

[12] J. Volz, C. Bizer, M. Gaedke, and G. Kobilarov, “Discovering and maintaining links on the Web of data”, In Proceedings of the 8th International Semantic Web Conference(ISWC), ISWC ’09, pages 650–665, Berlin, Heidelberg, 2009. Springer-Verlag.

[13] B. Motik, I. Horrocks, and U. Sattler, “Bridging the gap between owl and relational databases”, Web Semantics : Science, Services and Agents on the World Wide Web, 7(2) :74–89, 2009.

[14] R. Isele, and C. Bizer, “Learning expressive linkage rules using genetic programming”, PVLDB, 5(11):1638–1649, 2012.

[15] E. Ruckhaus, M.-E. Vidal, S. Castillo, O. Burguillos, and O. Baldizan, “Analyzing linked data quality with liquate”. In The Semantic Web: ESWC 2014 Satellite Events, 2014.pages 488–493. Springer.

[16] E. Sirin, and J. Tao, “Towards integrity constraints in owl”. In OWLED, volume 529, 2009.

[17] W. Hu, J. Chen, and Y. Qu,” A self-training approach for resolving object coreference on the semantic Web”, In WWW, pages 87–96, 2011.

[18] D. Wood, M. Lanthaler, and R. Cyganiak, “RDF 1.1 Concepts and Abstract Syntax”, W3C Recommendation, W3C, 2014.

[19] Y. Atig, A. Zahaf, and D. Bouchiha, "Conservativity Principle Violations for Ontology Alignment: Survey and Trends", International Journal of Information Technology and Computer Science(IJITCS), Vol.8, No.7, pp.61-71, 2016. DOI: 10.5815/ijitcs.2016.07.09

[20] C. Bizer, T. Heath, and T. Berners-Lee, “Linked Data-The Story So Far”. Int. J. Semantic Web Inf. Syst, 5(3), 1-22, 2009.

Authors’ Profiles

Ardjani Fatima received his engineering degree in computer science from the University of Saida, Algeria, in 2007, and M. Sc. In Computer Science from the University of Oran, Algeria, in 2010, and in 2012, she joined the Department of Computer Science, University Center of Naama, Algeria. Currently, she is Assistant Professor at the University Center of El Bayadh. Her research interests include semantic Web, ontology alignment, and optimization methods.

Djelloul Bouchiha received his Engineer degree in computer science from Sidi Bel Abbes University, Algeria, in 2002, and M. Sc. in computer science from Sidi Bel Abbes University, Algeria, in 2005, and Ph. D. in 2011. Between 2005 and 2010, he joined the Department of Computer Science, Saida University, Algeria, as a Lecturer. He became an Assistant Professor since January 2011. Currently, he is an Associate 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 an Associate Professor in the Department of Computer Science at the University of Sidi Bel-Abbes from 2003 to 2010. Currently, he is a 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.

How to cite this paper: Fatima Ardjani, Djelloul Bouchiha, Mimoun Malki,"An Approach for Discovering and Maintaining Links in RDF Linked Data", International Journal of Modern Education and Computer Science(IJMECS), Vol.9, No.3, pp.56-63, 2017.DOI: 10.5815/ijmecs.2017.03.07

