A semi-automatic solution for XML query response enrichment using a terminological domain ontology

Larbi Abdelmadjid
Computer Science Department
SidiBel Abbes D.L University
ENERGARID Lab., SimulIA Team
Bechar T.M University
Algeria
larbi.abdelmadjid@univ-bechar.dz

Malki Mimoun
ESI SidiBel Abbes
EEDIS Laboratory, ISIBD Team
Algeria
Mmalki@univ-sba.dz

Seddiki Bachir
Bechar T.M. University
Po.Box ,417 Al-IstiklalBvd
Algeria
Bachir_html@gmail.com

Larbi Ismail
TM University of Bechar
Po.Box ,417 Al-IstiklalBvd
Algeria
larbismail@gmail.com

ABSTRACT

The introduction of data semantics in various fields of science by referring to the ontological database is becoming more and more necessary. With the proliferation of domain ontologies and the large volume of data to be processed, it has become necessary to have data management systems based on ontological systems. Such a system can be exploited via the web as is the case with XML databases, which will allow us to:

- use semantic databases via the Internet.
- To enrich responses to XML queries by using domain terminology ontology.
  Also, XML present a flexible hierarchical model suitable to represent huge amounts of data with no absolute and fixed schema,

In order to highlight the usefulness of introducing the semantics in the XML database and to enrich the answers to the queries, we have proposed a semi-automatic solution that we applied it for pharmaceutical databases, and which gave very encouraging preliminary results.

CCS CONCEPTS

- **Information systems → Ontologies**
- **Information systems → Semi-structured data**
- **Theory of computation → Database query languages (principles)**

KEYWORDS

Data-centric XML, XML Database, enrichment query response, terminological ontology

ACM Reference format:

Larbi, M. Malki, B. Seddiki, I. Larbi 2017. In *Proceedings of* International Conference of Computing for Engineering and Sciences ICCES '17, July 22–24, 2017, Istanbul, Turkey.
DOI: 10.1145/3129186.3131941

ISBN 978-1-4503-5309-0/17/07…$15.00
https://doi.org/10.1145/3129186.31319411

1. INTRODUCTION

In recent years the database research field has concentrated on XML (eXtensibleMarkup Language) as a flexible hierarchical model suitable to represent huge amounts of data with no absolute and fixed schema, and a possibly irregular and incomplete structure [1].

There are two broad application areas of XML technologies. The first relates to document- centric applications, and the second to data- centric applications. Because XML can be used in so many different ways, it is important to understand the difference between these two categories: Document-Centric XML and Data- Centric XML.

1.1 The data-centric

Data-centric documents are documents that use XML [2] as a data transport. They are designed for machine consumption and the fact that XML is used at all is usually superfluous. That is, it is not important to the application or the database that the data is, for some length of time, stored in an XML document. Examples of data-centric documents are sales orders, flight schedules, scientific data, and stock quotes.

Data-centric documents are characterized by fairly regular structure, fine-grained data (that is, the smallest independent unit of data is at the level of a PCDATA-only element or an attribute), and little or no mixed content. The order in which sibling elements and PCDATA occurs is generally not significant, except when validating the document.

For example, the following sales order is data- oriented:

```
<OrdreDeVentesNumeroOrdreDeVentes="12345">
<ClientNumeroClient="543">
<ClientName>ABCIndustries</<CustomerName>
<Rue>123 Main St.</Street>
<city> Chicago </City>
<Status>IL</state>
<CodePostal> 60609</PostalCode>
</Client>
<DateOrdre>981215 </DateOrdre>
```

```
<itemNumeroItem="1">
<LotNumeroLot="123">
<Description>
<p><b>Turkey wrench:</b><br/> Stainlesssteel, one-piececonstruction, lifetime guarantee.</p>
</Description>
<Price>9.95 </price>
</Lot>
<amount>10 </Quantity>
</Item>
</SalesOrder>
```

1.2 The document-centric documents

Document-centric XML document creation is a process [3] of marking up textual content rather than typing text in a predefined structure. It turns out that, although the final document has to be valid with respect to the DTD/Schema used for the encoding, the "in-progress" document is lmost never valid. At the same time, it is important to ensure that at each moment of time, the editor is working with an XML document that can be enriched with further markup to become valid.
The following document, for example, describes a product and it is oriented document:

```
<¨Produit>
    <Intro>The<ProductName>Turkey Wrench</ProductName>from<Developer>Full Fabrication Labs,Inc.</Developer> is<Summary>likea monkey wrench, butnotasbig.</Summary>
    </Intro>
    <Description>
    <Para>The turkey wrench,whichcomes in <i>both right-andlefthanded versions(skyhookoptional) </i>, ismade of the<b>finest stainlesssteel</b>.TheReadi-grip rubberizedhandle quicklyadapts toyour hands,evenin the greasiestsituations. Adjustmentis possiblethrough avarietyof customdials.</Para>
    <Para>Youcan:</Para>
    <Liste>
    <Item><LinkURL="Order.html">Order your own turkey wrench</Link></Item></Liste>
    <Para>The turkey wrench costs<b>just $19.99</b>and, if you ordernow, comeswith a<b>hand-craftedshrimp hammer</b>asa bonusgift.</Para>
    </Description>
</Produit>
```

Query languages used in this case are of two types:
i. The query languages based on models: A Native XML database (NXD in English) is a database that is based on the data model provided by XML. It typically uses XML query languages such as XPath or XQuery.
ii. The query language based on SQL query languages based on SQL SELECT statements use modified the results are transformed into XML.
In this work, we are more interested in data- centric XML or database that are more or less structured with a regular structure to differentiate the XML documents that are characterized by a less regular or even irregular. The need to clarify the semantics of the data in various science fields (biology, medicine, geography, engineering...) is introduced by definition data referring to ontologies, also called ontological data base. With the proliferation of domain ontologies, and the large volume of data to be handled, the need emerged for systems that manage data based on ontological large, such management systems are called systems management database based ontological (OBDB). We can operate such system via the web if we think represent them as XML.
Which allows us to:

- To exploit OBDB via the Internet
- To enrich the answers to XML queries (or other OBDB BD in XML format) using domain the ontologies terminology.

1.3 Defining a native XML database

We require a native XML database to have the following two properties [4,5] as well:

- The XML data model (either in the XML Infoset or the XQuery/XPath Data Model) is the fundamental logical data model both used internally by the database and exposed to database users when XML is the data type.
- The XML data model is the fundamental unit of physical storage of all XML data, without mapping to a different data model.

This narrowed definition means that XML is more than an externalized data type - it is how the data is handled both logically and physically. The data is represented as XML right down to its physical storage schema on the disk. This model is the best for efficient searching of the XML data.
This vision conforms to the logic model can consider XML easily using defined standards around XML (XQuery, XPath, XSLT, XUpdate) to access and process the data in the database.

1.4 Characteristics of a native XML DB

We can briefly some characteristics of databases Native XML. This should help the reader by giving an idea of features available today and those expected in the future.

2. THE ONTOLOGY CONCEPT

Gruber [6] introduced the notion of ontology as "an explicit specification of a conceptualization" Borst. A. has slightly modify the definition by combination of the two definitions can be summarized as follows: "an explicit and formal specification of a shared conceptualization"
This definition explains: “explicit” means that the "type of concepts and constraints on their use are explicitly defined ". “Formal” refers to the fact that the specification should be readable by a machine, refers to the shared notion that an ontology "captures consensual knowledge, which is not specific to an individual but by a group validated" conceptualization refers to "an abstract model of some

phenomenon in the world based on identification of relevant concepts of this phenomenon".
An ontology provides a solid foundation for communication between machines but also between humans and machines in defining the meaning of First of all objects through symbols (words or phrases) that and characterized means and then through a structured representation or formal role in the field.
Ontologies are used in many fields. Areas identified in 1998 by Guarino are engineering knowledge, qualitative modeling, language engineering, databases design, information retrieval, information extraction, management and the organization of knowledge.
Since then, thanks to the growth of the Web, they are used in the domain of ecommerce and are at the heart of the Semantic Web, the future version of the current Web.
One of the biggest projects based on the use of ontologies is add to a real web layer enabling knowledge research information at the semantic level and at the simplest lexical and / or syntactical level. Ultimately, it is expected that applications deployed on the Internet may lead reasoning using the knowledge stored on the web.

2.1 Ontologies Classification

Several classifications have been defined and are based on different criteria. We retain the classifications proposed by G.VanHeijst 97, N.Guarino 97 and O. Lassila and D.McGuinness. Van Heijst proposed two types of ontologies classification [7] according different criteria.
The first classification is based on the types and wealth of structures used in the ontology. According to these criteria, there are three categories of ontologies:
2.1.1Terminological ontologies. That are used to specify the terms of the vocabulary of a knowledge field.
*2.1.2Information ontologies.*That specify the structure / schema of a database to enable the storage of information.
*2.1.3Ontologies that model knowledge.*That offer internal structures that are richer and more defined according to their uses such information sharing. It also proposes a classification of ontologies that relies on decision account of "objectives" of modeling. It holds four categories ontologies according to this criteria:
*2.1.4Ontologies applications.*That specify the necessary information to one or more particular applications.
2.1.5Domain ontologies. That express the conceptualization of a particular domain knowledge.
2.1.6 Generic ontologies. That model of transverse knowledge to different areas. Typically generic ontologies define concepts such as the concepts of state, event, action, etc.
*2.1.7Ontologies representation.*That seek to explain the conceptualizations underlying the knowledge representation formalisms. They represent the real-world entities without a priori, to "neutral" way.

This classification is illustrated in Figure1

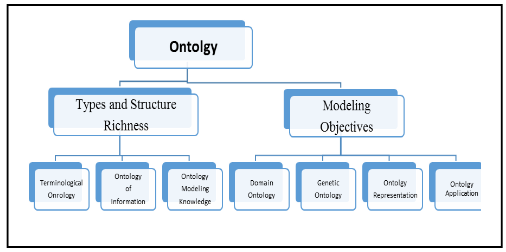

Figure 1:Ontology classification

2.2 Ontology languages

The Ontologies play a very important role in theSemantic Web proposed by Tim Berners-Lee to specify are formally vocabularies for describing Web content. Thus, the information Web can be accessed and understood by both humans and by machines. Several languages [8] have been proposed to represent ontologies with a higher or lower expressiveness and complexity of reasoning more or less. We are interested in three languages representation of ontologies or recommended subject to recommendation by W3C (World Wide Web Consortium) RDF / RDFS, OWL and SWRL.

2.2.1 *RDF and RDFS languages.* RDF (Resource Description Framework). It is a language for describing the semantics of data in an understandable way by machines. The RDF data model is composed of three types of components:

2.2.1.1 *Topic. I*t is necessarily identified by a URI.

2.2.1.2 *Predicate or property. I*t is a property used to characterize and describe a resource. A predicate is necessarily identified by a URI.

2.2.1.3*Re.* Is data or another resource (identified by a URI). Using the RDF, information resources are described by a set of RDF statements in the form of triples: (05) (Subject, Predicate, Object).
This notation is called triplet N-TRIPLE. XML syntax has been defined to express RDF statements, it is called RDF / XML.

2.2.2 *OWL.* It is, like RDF, an XML language taking advantage of the syntactic universality of XML.Based on the RDF / XML syntax, OWL offers a way to write web ontologies. OWL Differs from the pair RDF / RDFS in that, unlike RDF, it is precisely an ontologies language.

We can find: OWL Lite, OWL DL, and OWL Full

3. STATE OF ART

There are many works including the ontology for treating the problem of including the semantic to the query [9]

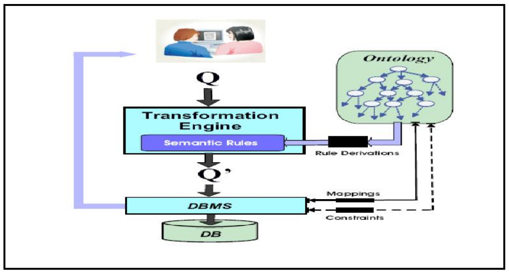

Figure 2: Semantic query transformation using ontologies [9]

or for using semantic query over heterogeneous databases [10],

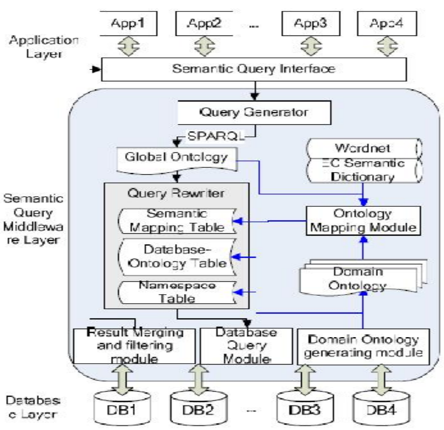

Figure 3: An ontology-based system for semantic query over heterogeneous databases [10]

Weconsidertheproblemofcomputinganswerstoqueries byUsingmaterializedviews. Asidefromitspotentialinoptimizingqueryevaluation,theProblemalsoarisesinapplicationssuchasGlobalInformationSystems,MobileComputingandmaintaining physicaldataindependence. We consider the problem of finding a rewriting of a query that uses the materialized views, the problem of finding minimal rewritings, and finding complete rewritings[11]

4. DESIGN OF THE PROPOSED SOLUTION

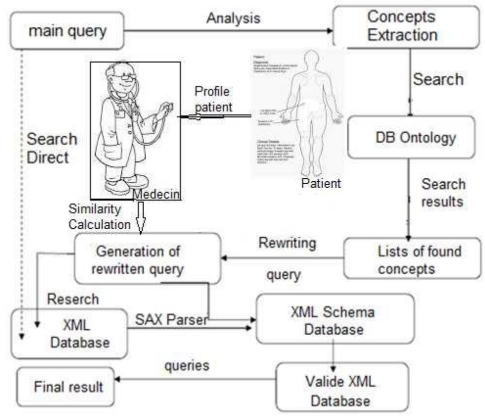

Figure 4: Query rewrite model based on ontology

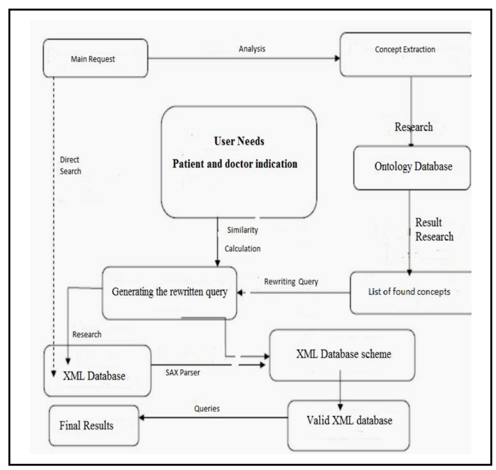

Figure 5: Graph rewriting query in an XML database by use of a pharmaceutical terminology ontology

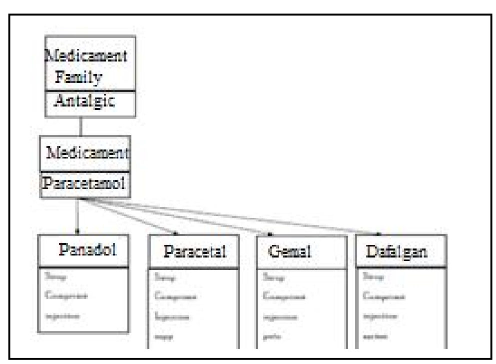

Figure 6: Creating classes of our ontology

With the help of the editor Protégé 2000 we can create our ontology following the information given by the expert.

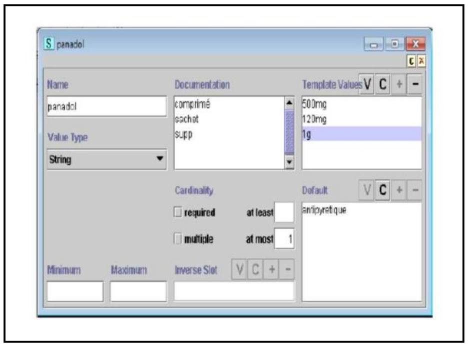

Figure 7: Creating instances of our ontology

The sequence diagrams are used to represent interactions between objects in a temporal point of view. The focus is on the chronology of sending messages.

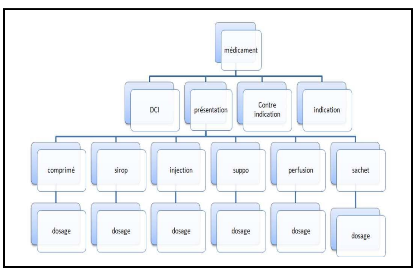

Figure 8:Classes Diagram and their hierarchies of PANALGAN drug.

The application supports multiple paths to get results in this example there are two voices:

4.1 *Direct Request.* In this case the application directly accesses the database.

4.2 *Semantics Query.* Accesses the ontology query: Terminology and regenerate other queries against the XML database in this case the resulting query will closer to the original query in a field and in the user selection.

Figure 9: presentation of our XML-based ontology database

It can be seen from the example that there is a correct result ie represents the execution of the main query directly on the basis of XML data, and other results represent the query execution generated by the use of ontology**.**

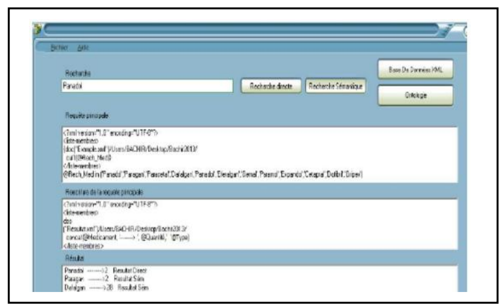

Figure 10: Display Semantic results on our application

5 CONCLUSION

Our query expansion mechanism exploits the richness of relationships semantics provided by ontologies. Its adaptation to the user by the through prototypicality (represented by the terminology, by weights on the terms) allows you to customize both the extension query input the amount of results provided.

Our solution allows the time to customize the user query by patient requirements and directed by the doctor on one hand and according to the concepts provided by the ontology Terminology in the other (validated by pharmaceutical expert) in order to have a comprehensive response. The first results are encouraging and it is estimated continuity of this project in the future.

As perspectives of this work, we project:

- Replace the doctor's intervention with a database containing the profile of the patient defined and updated by the doctor in order to make the solution automatic.
- Deepen the study of the drugs similarity based on the

intervention of a pharmacology expert.

REFERENCES

[1] Mazuran M., Quintarelli E. and Tanca L., 2012. Data mining for xml query-answeringsupport, IEEE Transactions on Knowledge and Data Engineering(Volume: 24, Issue: 8, Aug. 2012), Page(s): 1393 – 1407. DOI: 10.1109/TKDE.2011.80, , Print ISSN: 1041-4347

[2] Bourret R., 1999-2003, XML and databases, , Felton, CA 95018 USA, Key: citeulike:8973847

[3] Ionut E. I., 2006. On potential validity of document-centric xml documents, data engineering workshops, Proceedings. 22nd International Conference on Data Engineering Workshops, DOI: 10.1109 University of Kentucky

[4] Kolar P. and Loupal P., 2006.Comparison of Native XML Databases and Experimenting with INEX, Proceedings of the "Dateso" Workshop Databases, Texts Specifications, and Objects. Dept. Of computer science and engineering fee, Czech technical university.

[5] MeierW., 2002. Exist: An open source native XML database. Web, Web-Services, and Database Systems NODe 2002 Web- and Database-Related Workshops Erfurt, Germany, LNCS, volume 2593

[6] GruberT.R., 1993.A translation approach to portable ontology specifications. Knowledge acquisition, 5, (2):199-220,

[7] GuarinoN. andGiaretti P., 1995. Ontologies and knowledge bases: Towards a terminological clarification. Towards very large knowledge bases. N. J. In mars, ed., IOS Press:25-32, 1995

[8] Hachemi F., 2007. Description sémantique des objets d'apprentissage à base de modèles de contenu, Mémoire de magistère, Université Abou BekrBelkaid. Tlemcen- Algérie.

[9] Freytag N. C.B. and Johann-Christoph, 2005. Semantic Query Transformation Using ontologies, 9th International Database Engineering and Application Symposium IDEAS, DOI: 10.1109/IDEAS.2005.51, Print ISSN: 1098-8068.

[10] Donglin C., 2009. An ontology-based system for semantic query over heterogeneous databases ,Proceedings of the 2009 WRI World Congress onSoftware Engineering - Volume 02 , Pages 50-53

[11] Levy Y., Mendelzon A.O. and SagivY., 1995.Answering Queries Using Views, Proceedings of the fourteenth ACMISBN:0-89791-730-8doi:10.1145/212433.220198

[12] Barbier F.; Henderson-Sellers B. and Le Parc-Lacayrelle A. ; J.-M. Bruel. Formalization of the whole-part relationship in the unified modelling language, IEEE transactions on software Engineering, 29(5), 459-470. DOI: 10.1109/TSE.2003.1199074 , 2003

[13] Valente G. and CarpioG., 2009. A query expansion methodology in a cooperation of information systems based on ontologies. WEBIST.