

# A Technique for Generating Semantic Web Documents

Amjad Farooq

Abad Shah

Qasim Akram

Research and Development Center of computer Science  
University of Engineering and Technology, Lahore-Pakistan

**Abstract-***The huge number of available web documents makes it increasingly difficult for users to find and access required information because their semantics are not understandable by machines. Semantic Web concentrates on this issue via machine understandable metadata for web documents to make them automatic processable. XML is widely used in Web to specify structure of documents in syntactic dimension. The document structure can represent some semantic properties but still it is difficult to make them machine understandable. To generate semantic web document from XML document web ontology is recommended as core technology by W3C. Some techniques for this purpose are reported in literature but most of them are based manual processing which is very time consuming task. We propose a technique to produce semantic web document from XML documents, by relating document structure to ontology. Proposed technique transforms XML document to machine-understandable document via ontology without much manual effort.*

**Keywords:** Semantic Web, Ontology, RDF, XML

## 1. Introduction

A Semantic Web document not only offers human understandable content but also formal semantic describing the content of the document in a machine-processable way. This importance of being able to express formal semantics of web document was neglected for a long time. Now it is the vision of the Semantic Web [1] that highlights this issue. Semantic Web is an extension to the current web which concentrates on the need of machine understandable metadata [4] for the web documents. For success of the semantic web in the real world scenario as explained in [2, 3], it becomes mandatory to annotate web document with formal semantics. W3C has recommended Resource Description Framework to describe the resources on the web [6]. RDF can also be used to show the resources on the web which can not be directly retrieved from current web. The backbone technology involved in overall process is the web ontology. It is the formal specification of the concepts [10]. Actually it is the technology which provides the

descriptive knowledge in formal ways about web document's domain, its contents and relationship between those contents. Actually it formally describes about contents of web document. It organizes formal description in terms of classes, properties and their instances [12, 13]. XML is a core technology and is widely used for web documents. It specifies the structure of documents in syntactical dimension but it does not impose any formal description of the data contained in the document. Formal semantics refers to metadata understandable by machine. Different elements in the XML document are needed to be annotated with formal semantics. To accomplish this different strategies are used. One of them is the mapping based strategy, through which the XML document is annotated via web ontology. Some techniques used in [14], [15], [16], [17], [23] and [19], etc. are reported in literature, but most of them are based manual processing which is very time consuming task. The rest of the paper is organized as follows. In section 2 we give a short overview of existing approaches. We present an overview of our technique in section 3. In section 4 it is validating through case study. Paper is concluded with future work and conclusions.

## 2. Literature Review

In literature some mapping based techniques there are, for extracting the semantic markup for XML documents. Each has different considerations for annotating in order to generate the metadata description from the XML documents. In WESSA the conceptual mapping is done manually and also a tool is available for define the mapping as described in [14, 15, 16]. This mapping is central part of the WESSA technique it is very time consuming because it is done manually. In WESSA the mapping defines the RDF triples structure like subject, object and predicate with some java methods which control the ambiguity between tags. This mapping is further processed by a java method which generates RDF from the WESSA mapping.

In [24], Ontobroker has different components, one of them is 'info agent'. This component retrieves information stored in XML documents, which

provides a mapping between the structure and contents of XML documents and conceptual entities of an ontology. The domain model representing web ontology is used to structure the system internal knowledge base and is used to structure the XML documents before mapping. In [26] the authors introduce a mapping method from XML elements to the ontologies concepts using XSLT. The authors consider the XML document as the relational document and mapped them according to these rules. One element containing other elements in XML document is mapped to a class and object properties in ontology. A mapping is established between the XML document and ontology, this mapping is implemented using XSLT. The other aspect of this approach is that even if there is no XML document then it can be mapped to XML schema rather than document. A similar approach is presented in [27] which describes mapping between XML and RDF model. The authors assume a XML schema is available which guides the mapping process assuming that each XML document has a RDF model. In [28] authors Describe mappings from XML to RDF as well as from XML schema to OWL, but these mapping are independent of each other. That means, that OWL instances have not necessarily to suit to the OWL model, because elements in XML documents may have been mapped to different elements in OWL. This approach does not answer the question how to create the OWL model, if no XML Schema is available. In WSDM [17] the mapping is done in the task modeling phase and is done manually actually the concepts in the ontology are mapped to the object chunks of the WSDM web engineering components. In this way, it allows designers to define the meaning of the different object types and roles they introduce during conceptual modeling. This conceptual annotation is used to generate automatically the actual page annotation for the website implementation. The conceptual annotation is defined as a mapping from the different object chunk entities onto the different ontology entities. There can be three type of relationship between these entities: One-to-one mapping; an object chunk entity can be mapped in a one-to-one way onto an ontology entity; One-to-many mapping: an object chunk entity cannot be mapped onto one single ontology entity but on a combination of ontology entities; Many-to-one mapping: an object chunk entity cannot be mapped onto one single ontology entity but a combination of object chunk entities can be mapped on a single ontology entity. In [25] it was described that “because the XML schema generally does not include taxonomic information for classes and properties”. It will expect that tools that use semantic clues parsed from XML tag names will partially

remedy this, and prove useful in both DAML and ordinary XML ontology mapping. The Meaning Definition Language (MDL) [19] defines what an XML document may mean in terms of ontology, and defines how that meaning is encoded in the elements and attributes of the XML document. Once the meaning of an XML document is defined in MDL, MDLbased tools allow users and developers to interface the XML document at the level of its semantics. Tools developed for MDL enable the automatic conversion of meaning-based requests into structure-based requests, provide a Java API to the semantics to the XML content, and translate XML documents from one XML Schema into another. The MDL translation tool takes the XML Schemas from the source and target XML format, the ontology used to define the meaning, and the MDL for the source and target XML format and produces an XSLT style sheet for the transformation of XML documents from the source to the target format. Since for the RDF/XML syntax no XML Schema can be defined the translation tool cannot be used to generate RDF graphs.

In Mapping XML Fragments to Community Web Ontologies [18] an approach is presented that aims to provide a unique query interface to XML documents that are based on different document type definitions (DTD). This unique query interface is based on common domain ontology. The approach uses the DTD and XPath to establish a mapping between XML fragments and ontology concepts. When a user formulates a query, the mapping is used to reformulate the query to the DTD of the corresponding XML document. This way the heterogeneity of the XML formats are hidden from the end-users. The mapping, however, is not used to generate an RDF description from the XML documents. Both [18], [19] mapping based annotation techniques directly map the content of XML elements or attributes to concepts in the ontology. Since no further processing of the content can be specified we encounter the granularity problem.

### 3. Proposed Technique

XML allows arbitrary structure to be added in a document but says nothing about formal semantics of structure. Formal semantics are expressed by web ontology, which encodes it in sets of triples. For efficient generation of the formal semantics for new XML documents we proposed a technique which generates metadata (RDF description) and overcome some draw backs of these techniques like granularity problem, manual mapping. We propose following steps for the efficient generation of semantic web document from the XML documents.

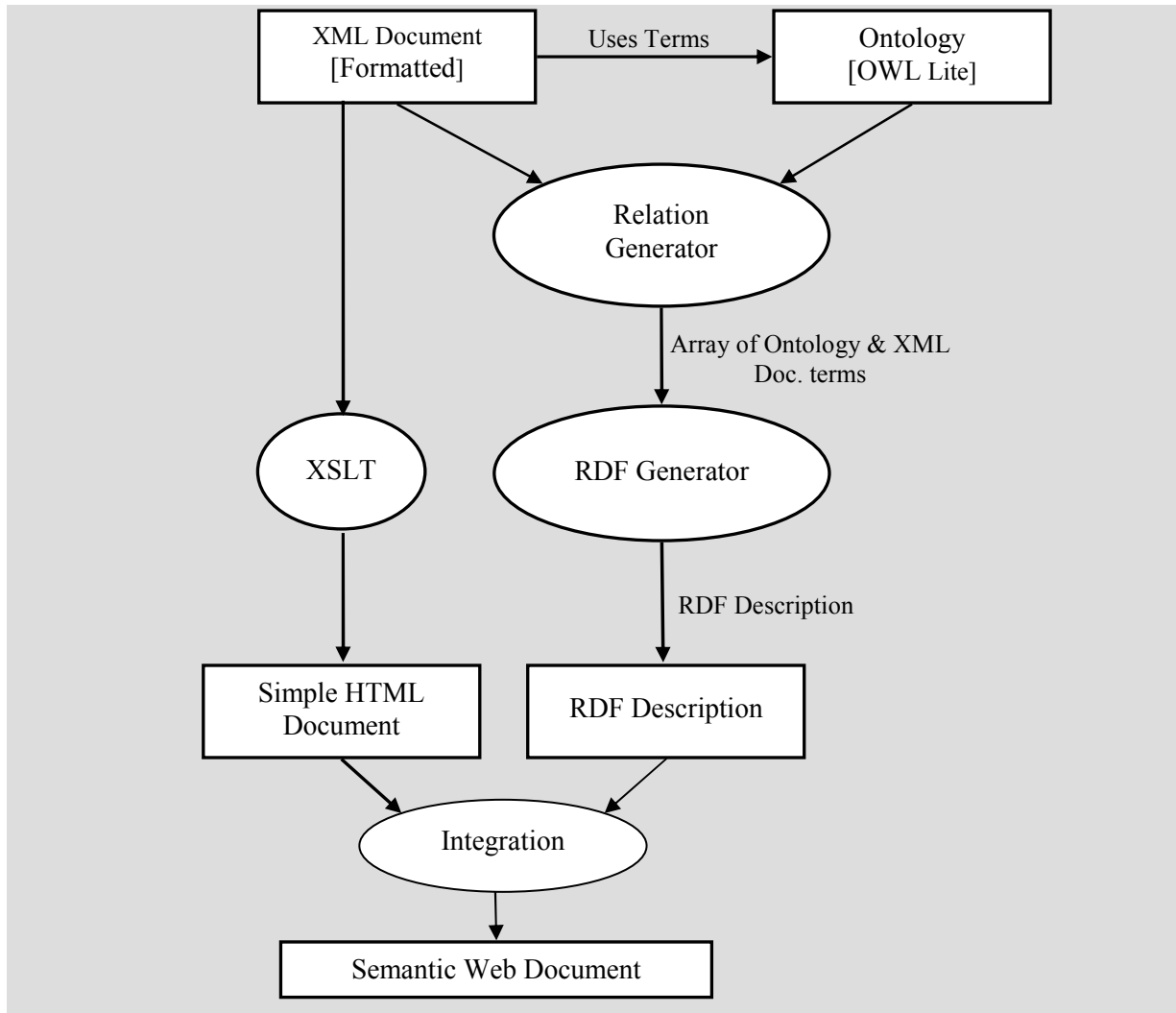


Figure 1: Sketch of proposed technique Architecture

### 3.1 Formatted XML Document

The first requirement for our technique is a specific formatted XML document which is used as input for our technique. It is not so problematic restriction on the structure of the XML document but we just have to add a <def> and <cont> tag for every corresponding term used in ontology let suppose if we have a entry of name table is our ontology of some XML document then we must specify the <def> and <content> tags for it in our XML document. e.g. < def > table </def> <cont> stored data in combination of rows and columns </cont>. The above tags are inserted in the XML document and the definition of the table is inserted in content tag. Here we apply these tags to the table element because the table tag is also existed in the ontology of the document. By following the above rule we have to insert the definition tag and content for each term of ontology in our XML document.

### 3.2 Ontology in the OWL Lite format

The second step in our technique is to take ontology of the required document in the OWL Lite format. Here we are using the Lite format for our research work. We are starting from ontology because we want to use ontology concepts in our XML document. This will ensure that the RDF description generated will be according to the concepts of the Ontology and each time the ontology changes, the corresponding RDF Description will also change. This solves the granularity problem which exists in existing techniques in which the concepts of the ontology does not match with the RDF Description.

### 3.3 Relation Generator

This is the most important part of our framework it takes the previous two documents as input and creates

relations between the concepts of ontology and the XML document terms. These relations are necessary for generating RDF description for XML document. The terms in the ontology are traversed against terms in the XML document and check if those terms exist in the XML document or not. This is done using the X-Path language for XML documents. The steps of the Relation generator algorithm is as follow: a) Read ontology file and get the information about the domain of the web document by traversing that ontology file. b) Search XML document for that specific kind of the information for the terms used in the ontology file. c) Extract contents from XML document against that domain information. d) Put that information into RDF objects and maintain a vector array of those RDF objects for RDF generation.

### 3.4 RDF Generator

The relations generated by the previous steps are further processed by the RDF generator algorithm which generates RDF Description from the vector array maintained in the previous step by follow the below mentioned steps: a) Get objects from vector and fill RDF tags against it. b) Write into RDF file. The output of this phase i.e. the RDF document can be tested using W3C validation service available at [22].

### 3.5 Integration

In the last step of our framework we associate the RDF description into the head tag of the html document and finally we have the (XML+RDF) page, which is the part of semantic web document.

## 4. Case Study

Here we are validating our proposed technique by taking a case study of the book catalogue and step by step go through each step of our case study.

### Step No.1 Formatted XML Document Book Catalogue

```
<?xml version="1.0"?>
<Catalogues xmlns
="http://example.com/catalogue#">
<def>book1</def>
<cont>
<book>
<author>aslam</author>
<Title>Software Engineering</Title>
<Editor>Noman</Editor>
</book>
</cont>
```

```
<def>book2</def>
<cont>
<book>
<author>Qammar</author>
<Title>Infomormation Reterival</Title>
<Editor>Noman</Editor>
</book>
</cont>
</Catalogues>
```

### Step No. 2 Ontology OWL Lite Book Catalogue

```
<?xml version="1.0"?>
<rdf:RDF
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-
syntax-ns#"
xmlns="http://example.com/ontology#"
xmlns:rdfs="http://www.w3.org/2000/01/rdf-
schema#"
XMLNs:owl="http://www.w3.org/2002/07/owl#"
xml:base="http://example.com/ontology">
<owl:Ontology rdf:about=""/>
<owl:Class rdf:ID="Catalogue"/>
<owl:Class rdf:ID="Book"/>
<owl:ObjectProperty rdf:ID="Author">
<rdfs:range rdf:resource="#Book" />
<rdfs:domain rdf:resource="#Catalogue"/>
</owl:ObjectProperty>
<owl:ObjectProperty rdf:ID="Title">
<rdfs:range rdf:resource="#Book" />
<rdfs:domain rdf:resource="#Catalogue"/>
</owl:ObjectProperty>
<owl:ObjectProperty rdf:ID="Editor">
<rdfs:range rdf:resource="#Book" />
<rdfs:domain rdf:resource="#Catalogue"/>
</owl:ObjectProperty>
</rdf:RDF>
```

The above ontology is checked by the Altova Semantic Works [29] to qualify as the OWL Lite Ontology.

### Step No.3 Relation Generator

Now we have the formatted XML document and related owl lite ontology file for our case study in this step our algorithm searches the corresponding entries of ontology and search them in XML document. And get the relevant data to the ontology terms from the XML document content tags using XPath expression. Now for example we have a property name author in our ontology we have to search for the relevant def tag which contains the author and fetches its relevant data and fill the RDF triples using XPath. Here Book is subject Author is predicate and Aslam is object. The related subject predicate and object are shown in table 1, generated by w3c validation service [22].

Table 1: RDF Triple generated by relation generator

No	Subject	Predicate	Object
1	<a href="http://www.w3.org/RDF/Validator/run/1144397122751#book1">http://www.w3.org/RDF/Validator/run/1144397122751#book1</a>	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://example.com/ontology#Book
2	<a href="http://www.w3.org/RDF/Validator/run/1144397122751#book">http://www.w3.org/RDF/Validator/run/1144397122751#book</a>	http://example.com/ontology#Author	"aslam"
3	<a href="http://www.w3.org/RDF/Validator/run/1144397122751#book1">http://www.w3.org/RDF/Validator/run/1144397122751#book1</a>	http://example.com/ontology#Title	"Software Engineering"
4	<a href="http://www.w3.org/RDF/Validator/run/1144397122751#book1">http://www.w3.org/RDF/Validator/run/1144397122751#book1</a>	http://example.com/ontology#Editor	"Noman"
5	<a href="http://www.w3.org/RDF/Validator/run/1144397122751#book2">http://www.w3.org/RDF/Validator/run/1144397122751#book2</a>	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://example.com/ontology#Book
6	<a href="http://www.w3.org/RDF/Validator/run/1144397122751#book2">http://www.w3.org/RDF/Validator/run/1144397122751#book2</a>	http://example.com/ontology#Author	"qammar"
7	<a href="http://www.w3.org/RDF/Validator/run/1144397122751#book2">http://www.w3.org/RDF/Validator/run/1144397122751#book2</a>	http://example.com/ontology#Title	"Infomration Reterival"
8	<a href="http://www.w3.org/RDF/Validator/run/1144397122751#book2">http://www.w3.org/RDF/Validator/run/1144397122751#book2</a>	http://example.com/ontology#Editor	"Noman"

#### Step No. 4 RDF Generator

In this phase we transform the array of objects maintained in the previous into a formal way i.e. in the form of RDF document, using a software module.

```
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF
xmlns:rdf=
"http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns="http://example.com/ontology#">
<Book rdf:ID="book1">
<Author>aslam</Author>
<Title>Software Engineering</Title>
<Editor>Noman</Editor>
</Book>
<Book rdf:ID="book2">
<Author>qammar</Author>
<Title>Infomration Reterival</Title>
<Editor>Noman</Editor>
</Book>
</rdf:RDF>
```

#### Step No. 5 Integration

In this last step the RDF file is just embedded to the head tag of the HTML page as follows: <link rel "meta" type="application/XML+rdf" href="catalogue..rdf">

## 5. Conclusion and future work

XML and the web ontology are the core technologies involved in the semantic web. Some techniques have been proposed to develop semantic web documents based on these technologies, but mostly they are manual and more complicated. Our proposed

technique is more efficient and less time consuming. Proposed technique eliminates the most of the manual work like mapping in previous approaches which makes them very costly and time consuming. This technique can be incorporated in web development methodologies to upgrade them for Semantic Web.

## Acknowledgement

We like to thank Higher Education Commission (HEC), Government of Pakistan for supporting this research work.

## References

- [1] Tim Berners Lee, J. Hendler, O. Lassila, "The Semantic web: A new form of web content that is meaningful to computers will unleash a revolution of new possibilities", Scientific American 2001: 5(1).
- [2] Kal Ahmed, Danny Ayers, Mark Birbeck, Jay Cousins, David Dodds, Josh Lubell, Miloslav Nic, Daniel Rivers-Moore, Andrew Watt, Robert Worden, and Ann Wrightson. Professional XML Meta Data. Wrox Press, 2001.
- [3] Tim Berners-Lee. A roadmap to the Semantic Web. W3C homepage, September 1998. <http://www.w3.org/DesignIssues/Semantic.html>
- [4] Tim Berners-Lee, James Hendler, and Ora Lassila. The SemanticWeb. Scientific America, 284(5): 2001.
- [5] Dan Brickley and Ramanathan V. Guha eds. RDF Vocabulary Description Language 1.0:

- RDF Schema. W3C Recommendation, 10 February 2004. <http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/>.
- [6] Graham Klyne and Jeremy J. Carroll eds. Resource Description framework (RDF): Concepts and Abstract Syntax. W3C Recommendation, 10 February 2004. <http://www.w3.org/TR/2004/REC-rdf-schema-20040210/>.
- [7] RDF Primer ( W3C Recommendations ) <http://www.w3.org/TR/2004/REC-rdf-Primer-20040210>
- [8] RDF Concepts and Abstract Syntax ( W3C Recommendations ) <http://www.w3.org/TR/2004/REC-rdf-Concepts-20040210>
- [9] RDF Syntax Specifications (Revised) (W3CRecommendations) <http://www.w3.org/TR/2004/REC-rdf-syntax-grammar-20040210>
- [10] OWL Web Ontology Language Overview (W3CRecommendations) <http://www.w3.org/TR/2004/REC-owl-features-20040210>
- [11] James Clark and Steve DeRose. XML Path Language (XPath) Version 1.0. W3C Recommendation, 16 November 1999. <http://www.w3.org/TR/xpath>.
- [12] Mike Dean and Guus Schreiber eds. OWL Web Ontology Language Reference. W3C Recommendation, 10 February 2004 <http://www.w3.org/TR/owl-ref/>.
- [13] Deborah L. McGuinness and Frank van Harmelen eds. OWL Web Ontology Language Overview. W3C Recommendation, 10 February 2004. <http://www.w3.org/TR/webont-features/>.
- [14] Gerald Reif, Harald Gall, and Mehdi Jazayeri. Towards semantic web engineering: WEESA - Mapping XML Schema to ontologies. In Workshop on Application Design, Development and Implementation Issues in the Semantic Web at the 13th International World Wide Web Conference, New York, USA, May 2004. CEUR Workshop Proceedings. <http://CEUR-WS.org/Vol-105/>.
- [15] Gerald Reif, Harald Gall, and Mehdi Jazayeri. Using WEESA - to Semantically Annotate Cocoon Web Applications. Technical Report TUV-1841-2005-31, Distributed Systems Group, Vienna University of Technology, 2005. <http://www.infosys.tuwien.ac.at/weesa/TUV->
- [16] Gerald Reif, Harald Gall, and Mehdi Jazayeri. WEESA - Web Engineering for Semantic Web Applications. In Proceedings of the 14th International World Wide Web Conference, pages 722-729, Chiba, Japan, May 2005.
- [17] Peter Plessers, Olga De Troyer Web Design for the Semantic Web In workshop on application design, development and designing issues in semantic web on 13th international world wide web Conference New York, USA, March 2004.
- [18] Bernd Amann, Irimi Fundulaki, Michel Scholl, Catriel Beeri, and Anne-Marie Vercoustre. Mapping XML fragments to community web ontologies. In Proceedings 4th International Workshop on the Web and Databases, 2001.
- [19] Robert Worden. Meaning Definition Language (MDL), Version 2.06, July 2002. <http://www.charteris.com/XMLToolkit/Downloads/MDL206.pdf>.
- [20] Schema web homepage, Last visited February 2006. <http://www.schemaweb.info>
- [21] <http://www.w3c.org> Last visited February 2006
- [22] <http://www.w3.org/RDF/Validator/ARPServlet>
- [23] Michael Erdmann, Rudi Studer. Ontologies as Conceptual Models for XML Documents. Institut für Angewandte Informatik und Formale Beschreibungsverfahren (AIFB) University of Karlsruhe (TH) D-76128 Karlsruhe (Germany)
- [24] Ontobroker: Or How to Enable Intelligent Access to the WWW. <http://ksi.cpsc.ucalgary.ca/KAW/KAW98/fensell/>
- [25] Stephen L. Reed, Douglas B. Lenat. Mapping Ontologies into Cyc (2002) [http://www.cyc.com/doc/white\\_papers/mapping-ontologies-into-cyc\\_v31.pdf](http://www.cyc.com/doc/white_papers/mapping-ontologies-into-cyc_v31.pdf)
- [26] Hannes Bohring, Soren Auer, Mapping XML to OWL Ontologies. [www.informatik.uni-leipzig.de/~auer/publication/XML2owl.pdf](http://www.informatik.uni-leipzig.de/~auer/publication/XML2owl.pdf)
- [27] Steve Battle. Round-tripping between XML and RDF. In International Semantic Web Conference (ISWC), Hiroshima, Japan, November 2004. Springer, 2004.
- [28] Matthias Ferdinand, Christian Zirpins, and D. Trastour. Lifting XML Schema to OWL. In 4th International Conference, ICWE 2004, Munich, Germany.
- [29] SemanticWorks™ 2006 [http://www.altova.com/products\\_semanticworks.html](http://www.altova.com/products_semanticworks.html)