AN ONTOLOGY FOR SEMANTIC DATA DESCRIPTION IN SCIENTIFIC DOMAIN

Veska GANCHEVA, Adelina ALEKSIEVA-PETROVA, Stilyana YANEVA
Technical University of Sofia, Bulgaria

Abstract. The scientific data obtained from different experiments are heterogeneous and stored in different file formats. Discovering, integrating and distributing of such data are challenges nowadays. OWL ontology for semantic data description is proposed in this paper. The ontology structure including classes and property hierarchy and relationship of the general ontology is explained. An RDF instance of the input XML file on the basis of the ontology is obtained. This should allow efficient description, storing, discovering and interpreting of the scientific data through the semantic web. A software framework for transforming raw data obtained from various scientific experiments into RDF instances on the basis of XML data description and the proposed ontology is developed. The framework is implemented as web service and consists of several key elements: XML scientific data description; ontology describing in detail the significance of semantics and XML data; the correspondence between the XML schema and OWL ontology is determined in accordance a mapping document. A number of experiments using specific scientific dataset in the cases of physical field data as radiation and spectral measurements obtained from experimental instruments are conducted to illustrate the usage of the proposed ontology and the software framework.

Keywords: Ontology, OWL, RDF, Scientific Data, Semantic Web, XML

1. Introduction

The technologies proposed by semantic web are used in the case of common necessity of sharing data such as scientific research in order to integrate data and applications [1].

An approach for web pages mining in order to build semantic concepts describing the web contents is proposed [2]. A simplified concept presentation which includes a set of positives and negatives words is introduced. The main advantage of the proposed approach is the simplicity and the speed of building the set of concepts. An initial version of a web page mining tool is developed.

The language RDF is used in the semantic web generally for conceptual description or modelling of information in the web resources [3]. The language OWL [4] defines the types of relationships that can be expressed in RDF using an XML description to indicate the hierarchies and relationship between different resources.

The ontologies describe concepts and relationships that are important in an area called the domain, providing a vocabulary for this domain, as well as computer based specification of the meaning of terms used in the dictionary. Ontologies range from taxonomy and classifications database schemes, theories based on axioms.

In the recent years, ontologies have been adopted in many business and scientific communities as a way of sharing, reuse and exchanging knowledge about the domain. Ontologies are crucial for many applications such as scientific knowledge portals, systems for information management and integration, electronic commerce, semantic web services.

Today an important issue is how the researchers can manage, store, process and visualize scientific data obtained from different experiments. To solve this problem it is necessary to pay attention to the semantic describing the personal data within the context of researchers and their access control.

The goal of this paper is to propose an ontology for semantic description of data in scientific domain and an algorithm and software framework for ontology dependent data transformation from XML scientific data description into RDF based instances in order to using them as a complete resource on the semantic web.

The paper is structured as follows. Section II discusses the related work. The background for this research is described in section III. Section IV explains the design and implementation of the semantic scientific data ontology. Section V explains the XML to RDF data transformation algorithm and implemented software framework.

2. Related work

There are a lot of researching projects addressing the problem for semantic description of personal data as a first step in data processing.

The FOAF specification describes a language, defined as a dictionary of named properties and classes to linking people and information using the Web. FOAF integrates three kinds of network: social networks of human collaboration, friendship

and association; representational networks that describe a simplified view of a cartoon universe in factual terms, and information networks that use Web-based linking to share independently published descriptions of this inter-connected world [5]. The project describes context of Person which represents people and define Person class as a sub-class of the Agent class, since all people are considered 'agents' in FOAF.

In the research projects of the European Commission's Framework Programme, a set of ontologies have been using for modelling the information: Documentation Ontology, Event Ontology, Organization Ontology, Person Ontology and Project Ontology. These ontologies have been used in different FP6 and FP7 projects, such as NeOn (FP6), SemSorGrid4Env (FP6), SEALS (FP7), etc. The Person Ontology models knowledge of persons who work in the project and divide such concept into four different types: university staff, company staff, project officer, and student. This ontology is focused on general-purpose personal information [6].

The other project is a Personal Data Service (PDS) which controls how personal data is shared with friends and organizations that trust. A PDS is a cloud-based service and gives a central point of control for personal like interests, contact information, addresses, profiles, affiliations, friends, and so on [7].

The DBpedia data set uses a large multi-domain ontology, which has been derived from Wikipedia [8]. It also defines the Person class and contains information like an additional name, physical address, an educational organizations, a child of the person and so on.

Those projects defined personal data in the context of social relationships with other people, interests, hobbies, projects, competencies, etc.

This paper aims to describe the semantic data of researchers focusing on business contacts and levels of access to scientific data.

3. Background

This work is a part of a project, which offers a range of web services for management, storage, processing and visualization of data obtained from scientific experiments. A flexible, dynamic, automated approach which provides access to a range of tools and services through a service-oriented architecture is proposed [9]. The proposed architecture describes the content of scientific data, gives scientists the opportunity to process their data easier and faster, which is of critical importance to make service-oriented computing paradigm operational in a business context. The solution is composed of services, which access scientific data from different data source types with different formats, transform raw data into standard data sets that can be analyzed, processed and visualized.

An XML-based language for describing the structure, semantic and annotation of scientific data is defined [10]. The common XML Schema and the Document Type Definitions (DTDs) are provided to describe semantically and structurally scientific data from experiments and simulations. The raw data description contains three main sections (Figure 1). The general section is about the ownership, the purpose and other general dataset proprieties. The semantics section provides information about the data nature and existing relationships in the dataset. The layout section describes the physical storing scheme of the dataset.

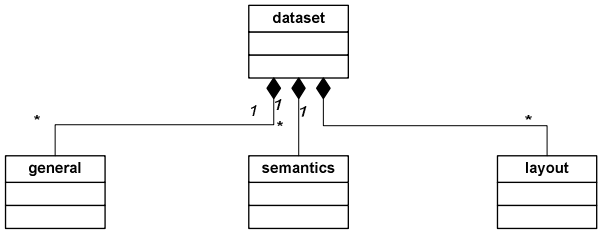

Figure 1. XML schema data description

4. Scientific data ontology structure

OWL ontology is with the greatest importance, since it describes the semantics, the significance and meaning of data: precisely these characteristics should be specified. The ontology consists of two main elements: classes and properties. The classes represent an abstract mechanism for grouping resources with similar characteristics. Each class is associated with multiple entities called class extensions. Subjects in appendices of the class are instances of the class. OWL classes are described by a class description, so that the triples of RDF output file are created. The properties are divided into two major sections: object properties and datatype properties. Object properties link an object with other object, while datatype properties define the relation between objects and data values.

The ontology defines the semantics and links between the elements of the output RDF file and also contains the meaning of each element. The existing XML file describing the general section data based on that is created the ontology is illustrated in Figure 2.

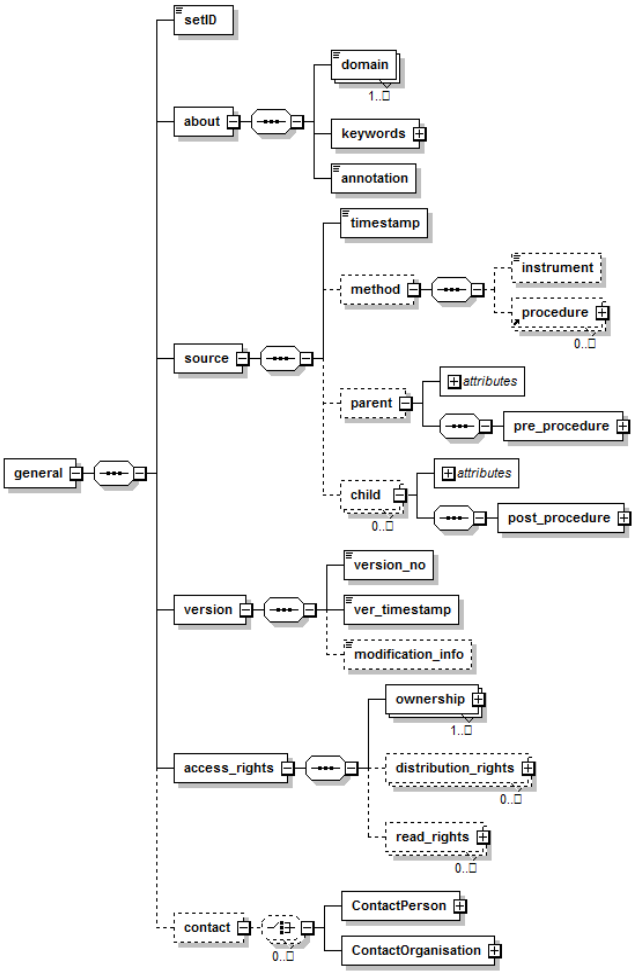

Figure 2. General element structure

After analyzing the description of the XML file structure have defined elements of the input XML file, which are considered as classes and which as properties of their respective classes in the OWL ontology. The element *general* is the main class in the ontology: all other xml tags from the input file are subclasses of the general or properties of the basic class or one of its subclasses:

```
<owl:Class rdf:about="&generalOWL;general">
        <rdfs:comment>Section for describing the
          general data about the data set as its
          annotation, owner rights, etc.
        </rdfs:comment>
</owl:Class>
```

All elements of the ontology belong to the namespace that defines the ontology: namespace *generalOWL*. Elements such as *about* or *source* are defined as classes: they are subclasses of the main class. Through them are described these resources, which have similar characteristics and can contain within itself other classes or properties:

```
<owl:Class rdf:about="&generalOWL;source">
        <rdfs:comment>Information about the
          data source and its history
        </rdfs:comment>
        <rdfs:subClassOf rdf:resource=
          "&generalOWL;general"/>
</owl:Class>
```

The hierarchy of the classes building up the ontology is shown in Figure 3. All classes in OWL ontology are subclasses of the main class *owl:Thing*. The general section of the XML file that is transformed into RDF is the main class in the ontology. All other classes describe the general element and its subclasses. The interesting thing here is the nested *ownership* class in the class *acces_rights*: ownership class consists of properties that characterize both *ownership* class and *acces_rights* class. These properties determine the relationship between both classes. The other classes are simple: they are constructed from data properties describing only the class to which they belong.

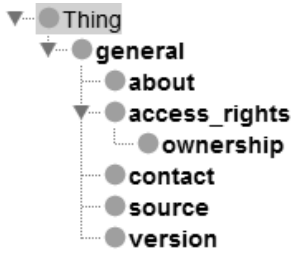

Figure 3. Classes hierarchy of the general ontology

Overall, the ontology is aimed to describe the complex elements, those that consist of one or more nested elements, such as containers (made of subelements that describe the respective main tag), i.e. they are classes of the ontology, and other simple elements that contain only data are datatype properties of the ontology. Therefore, elements such as *source*, *about*, *version*, *access_rights*, *contact*, *ownership*, etc. are subclasses of the main *general* class.

As concerns the properties of the ontology, only datatype properties are defined. Complex relations between the elements of XML file, which determine the existence of object properties are not presented in the input file. Therefore, elements such as *setID*, *keywords*, *timestamp*, *version_no*, *persAddress*, *city*, *email*, *country*, *familyName* are defined as datatype properties in the ontology. The properties hierarchy of the general ontology is ilustrated in Figure 4.

Figure 4. Properties hierarchy of the general ontology

Of course, there are relations between some of datatype properties, such as the relation between *keyword* and *keywords*, where *keywords* can appear 0 or more times in the tag *keyword*: this hierarchy in the ontology defines *keyword* as datatype properties and *keywords* (if is present in XML file) for its subpropety. The elements in tag ownership such as *Name* and *persAddress* in turn contain subpropety as *familyName*, *country*, *city*, *email:*

```
<owl:DatatypeProperty
  rdf:about="&generalOWL;familyName">
     <rdfs:subPropertyOf
       rdf:resource="&generalOWL;Name"/>
     <rdfs:domain
      rdf:resource="&generalOWL;ownership"/>
     <rdfs:domain>
       <owl:Restriction>
        <owl:onProperty
         rdf:resource="&generalOWL;familyName"/>
        <owl:minQualifiedCardinality
          rdf:datatype="&xsd;nonNegativeInteger">1
        </owl:minQualifiedCardinality>
        <owl:onDataRange
         rdf:resource="&xsd;string"/>
       </owl:Restriction>
     </rdfs:domain>
 </owl:DatatypeProperty>
```

The ontology has to be designed using the “is-a” relationship, as this is the relationship type used in most ontologies. The “is-a” relationship is used throughout and this relationship type correctly describes most of the relationships between the concepts (Figure 5).

Relations between the classes of OWL ontology and the properties of the classes are shown in Figure 6. There are several different cases:

- the class *contact* has no properties: it is described only as a subclass of class *general*;
- the classes *general* and *source* have only one data exchange property that describes them, in the case of a *general* property if *setID*, and for *source* is *timestamp* respectively;
- classes, such as *about*, *version*, *access_rights* have more than one property; some of them have subproperties, as relation *about-keyword-keywords*.

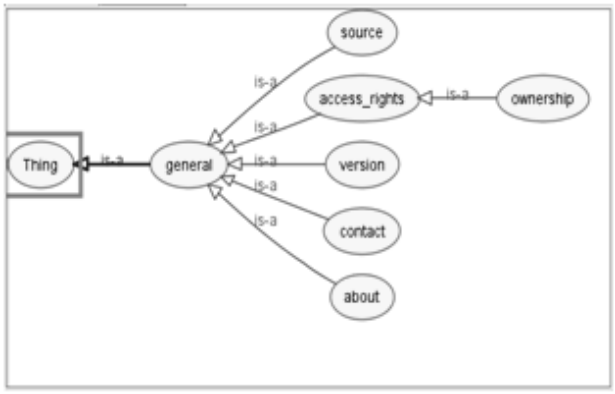

Figure 5. Classes and is-a relationship ¶

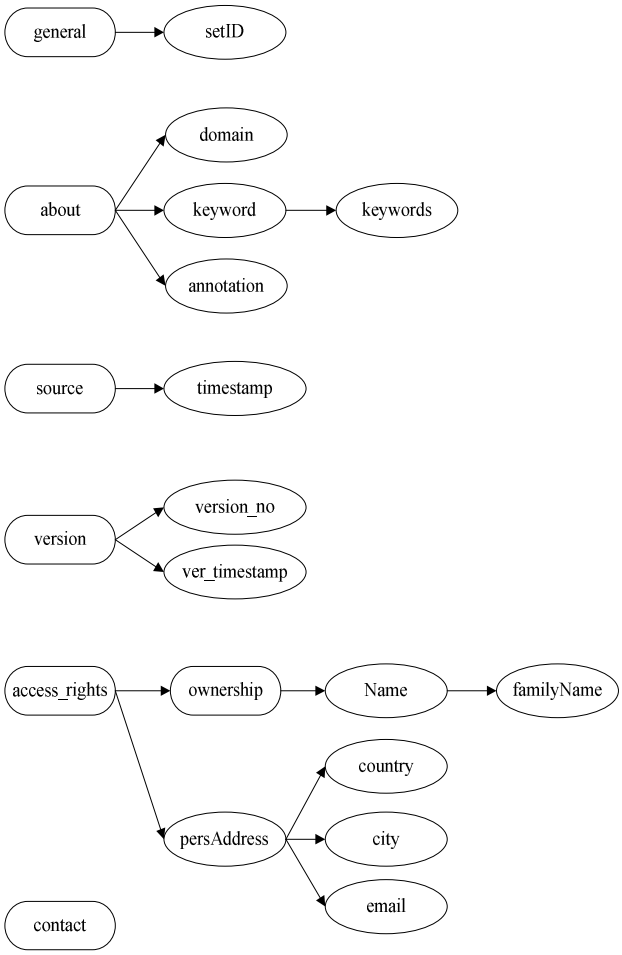

Figure 6. OWL ontology classes and properties relations

5. XML to RDF transformation

So-called ontology-dependant XML to RDF way of transforming scientific data has been proposed. A software framework, implemented as web service allowing transformation of scientific data description in XML based format into RDF format has been designed and developed. The framework consists of several key elements, each of which plays a role in the service. On the one hand is XML schema, which is accompanied by its instances: XML data. On the other hand is the OWL ontology that describes in detail the importance of semantics and XML data.

The goal is to obtain RDF instances of the OWL ontology. For this, the correspondence between the XML schema and OWL ontology is defined, which is determined in accordance mapping document.

The framework accepts as input an XML document that converts into RDF format, OWL ontology describing the data and mapping document, which performs the mapping and places the relationship between XML data and ontology. The outputs are instances of the RDF on the basis of OWL ontology.

The mapping document contains the rules by which the mapping is carried out and information about OWL vocabularies and XML files among which compliant is made. The mapping document also defines the elements of the input file transforming into RDF triples in the output file. The structure of the document it fixed so that different OWL ontologies can be mapped to it. Therefore, a mapping document can mapped some number of ontoligies that define the semantics of the source file and a global mapping document is used, i.e. once created, the mapping document can be used repeatedly (reusable mapping document).

It is important to note that the framework is independent of meta-data format used. New formats of meta data can be achieved by creating a new mapping document. The converting of XML to RDF is not limited to a specific XML schema, which forms OWL ontology. Many XML schemas can be combined to a particular ontology. It is also possible to map one or multiple XML schemas and to multiple ontologies.

The algorithm for data transforming includes several key points. Briefly the algorithm can be described as follows:

1. For each element of the input XML file that defines a class in OWL ontology, a separate method is defined;
2. All elements that are defined as properties in the ontology are defined in the methods of classes that determine;
3. Subclasses are also defined in the methods defining the classes of the ontology;
4. Regarded as specifying the input parameters consistently are called methods, which define the classes of ontology that in turn is described by subclasses and their properties;
5. The elements having comments in *rdfs:comment* element of OWL ontology are transformed into elements *general:comment* (defined element of the source file corresponding to the content in *rdfs:comment* element, which content in turn corresponds to that in *xs:documentation* element of XML Schema describing input data);
6. The value of the respective element in a class, subclass or property of ontology, which is in a tag *match* of mapping document is used to obtain a rdf element describing the general namespace: *<general: match_value> value </ general: match_value>*;
7. The values contained in XML elements are transformed into values describing RDF elements of the output file: *<general: match_value> value of input XML file </ general: match_value>;*
8. The OWL ontology location is imported as namespace in the header of the RDF file in order during uploading the output file as a resource in the Semantic Web, to upload also the ontology describing it, and desire the users to view the relations between elements in the RDF file and its ontology.

The algorithm for transformation of XML description into RDF file as pseudo code is as follows.

For all classes of the OWL Ontology do
Search the corresponding method
Then get XML element data AND
get mapping element match data AND
get owl element rdfs:comment AND
Then Transform them into an RDF Triple

6. Conclusion

An ontology for semantic data description in scientific domain is proposed in this paper. The ontology is aimed to be used for transformation process of XML based scientific data description into RDF instances in order to using them as a complete resource on the semantic web. The ontology structure including classes and property hierarchy and relationship of the general ontology is explained.

The transforming process includes: creating an ontology that describes the data semantic and links between the elements of the output RDF file; creating a mapping document that defines the relation between the input XML schema file and the OWL created ontology.

The proposed ontology has been implemented in a software framework aimed to transformation of XML based scientific data description into RDF instances. The framework has been applied successfully as web service implementation and verified by utilizing existing specific scientific experimental datasets obtained from experimental instruments in order to illustrate the usage.

Acknowledgement: This work is supported by the National Scientific Fund of Bulgarian Ministry of Education and Science, grant DO 02-175/2008.

References

1. Georgieva, J., Gancheva, V. (2008) *Technologies Used to Integrate Applications on the Semantic Web*. Proceedings of the Fourth International Conference on Computer Science, p. 749-755, ISBN: 978-954-580-256-0, Kavala, Greece
2. Momtchev, I. (2008) *A Web Page Mining Approach*. Proceedings of the Fourth International Conference on Computer Science, p. 956-959, ISBN: 978-954-580-256-0, Greece
3. Klyne, G., Carroll, J., editors (2004) *Resource Description Framework (RDF): Concepts and Abstract Syntax, W3C Recommendation, World Wide Web Consortium*. Available from: http://www.w3.org/TR/rdf-concepts/, Accessed: 28/11/2011
4. McGuinness, D., van Harmelen, F., editors (2004*) OWL Web Ontology Language: Overview, W3C Recommendation, World Wide Web Consortium.* Available from: http://www.w3.org/TR/owl-features/, Accessed: 28/11/2011
5. Brickley, D., Miller, L., *FOAF Vocabulary Specification*. Available from: http://xmlns.com/foaf/spec, Accessed: 28/11/2011
6. *FP Research project ontologies*. Available from: http://mayor2.dia.fi.upm.es/oeg-upm/index.php/en/ontologies/81-research-proj-ontologies, Accessed: 28/11/2011
7. *Personal Data Service.* Available from: http://www.eclipse.org/higgins, Accessed: 28/11/2011
8. *DBpedia*. Available from: http://dbpedia.org, Accessed: 28/11/2011
9. Shishedjiev, B., Goranova, M., Georgieva, J., Gancheva, V. (2009) *Processing and Managing Scientific Data in SOA Environment*. Proceedings of the 9th WSEAS International Conference on Applied Informatics and Communications (AIC '09), p. 25-30, ISSN: 1790-5109, ISBN: 978-960-474-107-6, Moscow, Russia
10. Shishedjiev, B., Goranova, M., Georgieva, J. (2010) *XML-Based Language for Specific Scientific Data Description.* Proceedings of the Fifth International Conference on Internet and Web Applications and Services (ICIW), p. 345-350, Barcelona, Spain

Received in February 2012