Generating Linked Data Repositories using UML Artifacts

Aqsa khan[1] and Saleem Malik[2]

[1] Virtual University of Pakistan, Lahore
[2] Federation University, Australia
ms160400333@vu.edu.pk, smalik@federation.edu.au

Abstract. The usability of diagrams and models is increasing day by day, because of this we experience problem in searching and accessing from large size repositories of diagrams and models of software systems. This research might be helpful to search and access the diagrams and models in bigger repositories. For this purpose, this research developed linked data repositories which contain UML (Unified Modeling Language) artifacts, these artifacts are being organized with using UML class model. In particular, UML is being broadly applied to data modeling in many application domains, and generating linked data repositories from the UML class model is becoming a challenging task in the context of semantic web. This paper proposes an approach, in which we will build a construction tool by joining the characteristics of RDF (Resource Description Framework) and UML. Firstly we will formally define design artifacts and linked data repositories. After that we will propose a construction tool in which we will extract UML artifacts, these UML class model further transforms into the corresponding RDFs. The generated RDF linked data then will be verified by using W3C RDF, this is a validating service used to generate and verify the RDF triples and graphs. Finally, the proposed construction tool will be implemented with few experiments and research is validated using W3C RDF validating service. The proposed approach aims to give such a design that may facilitate the users to customize linked data repositories so that diagrams and models could be examined from large size data.

Keywords: RDF (resource description framework), UML (unified modeling language), URI (uniform resource identifier).

1. Introduction

Semantic web is an addition of the present web, in which data is represented in RDF to identify in well define context, RDF description vocabulary language, RDF schema and OWL (web ontology language) are the recommendation of W3C, it is normative language to define the web resource data and their semantics (RDF-W3C, 2014). This semantic enhancement enable data to be integrated, shared and exchanged from different sources and allows applications to use information in different framework.

The need of creating linked data repositories of UML artifacts is increasing day by day, because it is not possible to search from large repositories of diagrams. Proper procedure should be developed for searching and manipulating from large number of old models and diagrams (Tong et al, 2014). In particular UML is widely applied in many application domains for data modeling, and for constructing linked data repositories of UML artifacts using UML class model is basically a procedure of constructing RDF(s) with the UML models. It is main issue, it should be solved in the framework of semantic web.

Consider the example of software house that is working from 10 to 15 years. If there are thousands of software system models saved in the memory of system, and we need to access one of specific model from thousands of models, we have to go through each and every model. It is very time-consuming process. It can be only performed with the help of linked data repositories using UML class model. This research will generate link data repositories of UML Artifacts. The core work involves constructing artifacts using UML class model (Tong et al, 2014).

1.1 Design Artifacts

Software architects, designers and developers identify certain data about the software product during the process of development. Software architecture, use cases, class description and object collaborations are examples of such data. The data can be extremely unique, the vision of the item, or exceptionally concrete, for example, the source code. In this research, design artifacts are defined as pieces of information that is all about the software product.

It is valuable to understand that there is a difference between a design artifact and its representation. The information about the software system is determined by design artifacts, and the representation decides how the information is exhibited. Sometimes UML is used to describe some design artifacts, some are described by text or by tables, and some are described in many different ways. In this approach, we will create linked data repositories of UML artifacts using UML class model.

1.2 Linked Data Repositories

Linked data repositories are collection of interrelated data sets on the web. For the purpose of creating linked data repositories, it is necessary to be available in common format of RDF. These repositories provide tools for browsing and visualizing data. There are some semantic web technologies such a RDF, SPARQL, OWL, SKOS etc., and data is conceived by the collection of these technologies (Bizer et al., 2009). Application can query about data in the environment which is provided by them.

Linked data lies at the core of semantic web: Reasoning, integration at large scale and web of data (Hassanzadeh, 2011). A framework and collection of technologies that enable the web of data is defined bellow:

- RDF

- Many formats of data interchange(For example RDF/XML, turtle, N3, N-Triples)
- Notations are also included such as RDF schema (RDFS) and Web ontology language (OWL) (Hassanzadeh, 2011).
- All proposed to provide a concepts, formal description, relationships and terms within a given domain of knowledge

1.3 RDF

The Resource Description Framework (RDF) is a framework for defining resources which is recommended by W3c. It is a standard model which is used for interchanging data on the Web(RDF-W3C, 2014).

The language of RDF is a component of the Semantic Web Activity of W3c's. W3C's "Semantic Vision of web" has some future goals involve: Information of web will be arranged by exact meaning, information of web can be understandable by computers and computers can integrate and collect data from the web. RDF provide a framework that is common for defining information, so it can exchanged between different applications.

1.7 OntoGen (Ontology Editor)

Nowadays, Modern content management system is faced the challenge of increasing web pages, document, textual content of document. It is very difficult to manage lot of web pages and textual documents. Ontologies are play very important rule for them .Ontologies provides help to minimize the information which is overloaded for a specific domain. The main purpose of ontology for user is to provide easily access the information. However, there are many ontology editors i.e Onto Studio and Protégé. There are manual and provide help to the users for constructing ontology. This ontology editor has many drawbacks. To overcome the drawback of manual ontology editor, OntoGen is a ontology editor which is introduce to help the user to construct ontology and provide user interface.

A new method of building ontology is introduced with the help of knowledge discovery and text mining. This method is used to provide help to build ontology. An example of using this method ontology generation is OntoGen. OntoGen is a "data driven and semi-automatic "system which is used for generating topic Ontologies. The old version of OntoGen is only used to edit the different types of ontology which was connected with different kinds of relations.

It is difficult for user because user spend lot of time to edit ontology. But the new version of OntoGen is providing attractive interface for user which reduces time and complexity for user. The system is attractive and provides helps for user during the ontology construction process. It give suggests about concepts and relations between concepts and automatically assigns instances to concepts. For the construction process of ontology in OntoGen, data is provided by user. OntoGen is providing support for

user to give automatic concepts and also describe the relationship between these concepts according to data.

1.4 Motivation

This paper proposes an approach, in which we will build a construction tool by joining the characteristics of RDF (Resource Description Framework) and UML. Firstly, we will formally define design artifacts and linked data repositories. After that we will propose a construction tool in which we will extract UML artifacts, these UML class model further transforms into the corresponding RDFs. The generated RDF linked data then will be verified by using W3C RDF, this is a validating service used to generate and verify the RDF triples and graphs. Finally, the proposed construction tool will be implemented with few experiments and research is validated using W3C RDF validating service. The proposed approach aims to give such a design that may facilitate the users to customize linked data repositories so that diagrams and models could be examined from large size data.

This paper also contributes to explore new ways for developing linked data repositories of large sized graphical models by which we can easily search and manipulate from large size of diagrams and data models (RDF-W3C, 2014). To achieve this goal following set of research objectives are target such as theory development for the generation of RDF based Linked data from graphical representations of software artifacts and implementation of the theory for the transformation of graphical models into linked data repository (Bizer et al, 2009). The experiments to test the performance of the presented approach with the help of examples and real world problems is also discussed in this paper.

2. Review of Literature

Tong and Cheng (2017) represented web information using standard languages RDF and RDF schema. The main issue of research is extracting RDF(S) from the existing data sources. They deal with uncertain and imprecise information and fuzzy data models in real world applications. At the end experiments evaluated that tool and approach are viable. However, they do not perform any experiments on verifying the RDF(s) data models.

Chakkarwar and Joshi (2016) described that redundancy in information is rapidly increasing due to the advancement of information on the web. They described an idea to combine semantic web and web mining. The extraction of information has been done in such a way that top ranked pages are shown to user. This research used three areas such as RDF data, semantic web and ontology. They performed experiments both on RDF datasets and other standards, it showed efficiency, scalability, and portability of solution across RDF engines. However, this research is limited to keyword search. It does not provide sentence search, image search and video search.

Pham et al. (2015) proposed an idea to stimulate and define methods that enable to discover an “emergent” RDF data relational schema. Author defined solutions of

semantic challenges that include short naming, humans will find conceptual to these columns and tables that are emergent, it also involve link and relationships between these tables. However, it does not explore other ways semantically from an RDF dataset, it should structurally optimized the relational schema to form it accurate.

Sherif and Ngomo (2015) portrayed the idea of datasets of semantic Quran. It includes almost all chapters of Quran written in 43 languages. This dataset is basically a RDF collection which is represented in several languages. The limitation of paper is not improving the ease of access of the sets of data. It should be developed in such a way, so it can easily acquire sensible information form datasets.

Faye, Cure and Blin (2012) discussed the different features of techniques for storing RDF data. They described ideas of Semantic Web in which we understand and easily discover the Web by using computers with the help of different Web principles. With the increase of data on the Internet RDF has been a pervasive data format for the Semantic Web. As the scale and number of Semantic Web increased in real world with the usage of application, so, there is a real need to retrieve and efficiently store RDF data. Scalability becomes more important as datasets grow larger and more datasets are linked together. Query processing and efficient data storage also discussed in this research, and possibly schema-less data has become an important topic of research.

Korthaus, Schwind and Seedorf (2005) proposed idea to use technologies of semantic web, for the purpose of semantic integration it is basically a business component specification. The limitation of paper is that Author defined a example, it is proposed component specification in UML to derive an RDF graph. In that case, there is a need to explain a profile of UML and a conversion from the serialization of XMI into RDF.

Decker et al. (2005) described that TRIPLE was defined for data manipulation applications as a practical rule language. They first initiate the TRIPLE's design principles, it presents some language that has been applied for some application. It should define the context notion and it is necessary for different applications.

Cranefield (2001) examined about innovation to help the utilization for ontologies of UML and area information in the Semantic Web. The two mappings have been characterized and actualized utilizing XSLT to deliver the Java classes and a RDF outline from a metaphysics spoke as a class of UML graph and it is encoded utilizing the XMI. An instrument should likewise present for showing object diagram has absent or inadequate information.

3. Used Approach

This paper proposes an approach, in which we will build a construction tool by joining the characteristics of RDF (Resource Description Framework) and UML. Firstly, we will formally define design artifacts and linked data repositories. After that we will propose a construction tool in which we will extract UML artifacts, these UML class model further transforms into the corresponding RDFs. The generated RDF linked data then will be verified by using W3C RDF, this is a validating service used to generate and verify the RDF triples and graphs. Finally, the proposed construction tool will be

implemented with few experiments and research is validated using W3C RDF validating service. The proposed approach aims to give such a design that may facilitate the users to customize linked data repositories so that diagrams and models could be examined from large size data. Main modules of the construction process are shown in Fig. 1.

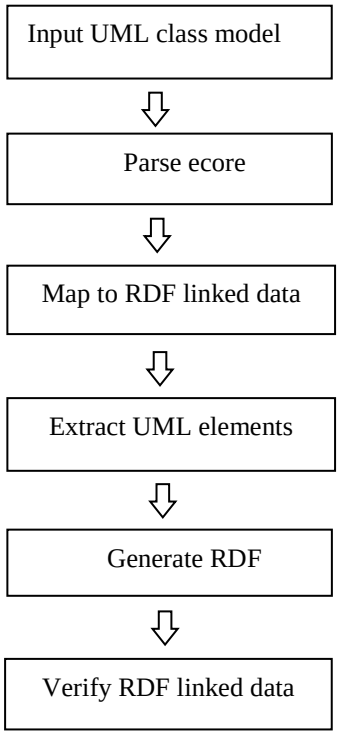

Fig. 1. Modules of constructing linked data repositories

There will be following modules of tasks involved for constructing linked data repositories of UML artifacts as defined in Figure 1. It is displayed in figure that involves few main modules, i.e. the .ecore module, parse module, UML element module, mapping module and constructing RDF module:

3.1. Input UML class model

First of all, a UML class model will be generated in Eclipse EMF (Eclipse Modeling Framework) and then it will be given as input (Budinsky et al., 2004). Figure 2 depict a UML class diagram, which includes classes, attributes and relationships of UML. In this diagram Principle is a class and student and teacher are sub classes of principle. There is an association Head used between principle and student. There is an

association employee between Principle and sub class teacher. Student and teacher classes have teaching association. Since, a UML class model is a diagram and it is difficult to parse a UML diagram. However, XML representation of a UML diagram can be easy to parse and machine process. So, a XML representation of a UML class diagram file is generated in the .ecore format exported by Eclipse EMF. This .ecore format will be easy to parse and can be processed in Java to extract the metadata of the underlying UML class model.

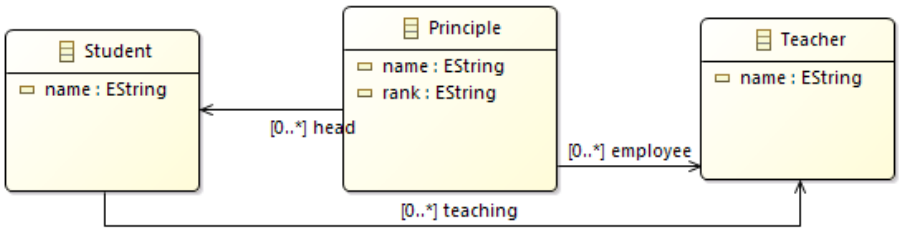

Fig. 2. UML class diagram

3.2. Parse .ecore

EMF is basically a framework for modeling your data model and creating Java code from it. The EMF tools allows you to create UML diagrams. EMF diagrams can be exported In .ecore format that is a XML based format for interchange of diagrams metadata. To parse .ecore format, a set of EMF libraries will be used in Java to extract metadata of a UML class diagram drawn in EMF model. Figure 3 shows some EMF libraries which is used in parsing process.

We need metadata of UML class model. For this purpose, we have to parse the XML coded file. In this step we will parse the input file which is UML class diagram file which is XML-coded, then it will store the parsed information.

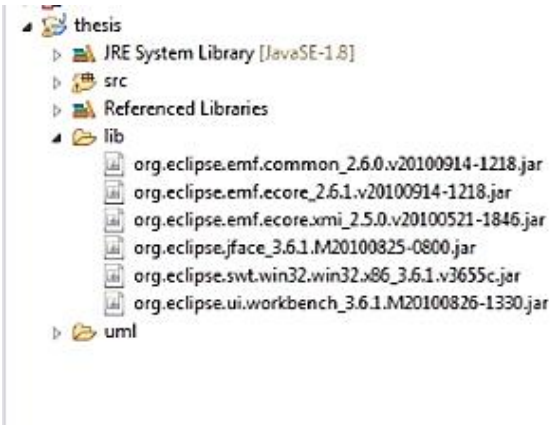

Fig. 3. Used EMF Libraries

3.3. Extract UML Elements

This step is about extracting some features of the class diagram of UML, and it is in the XML coded file. It includes classes, attributes, methods, association and relationships between them, it is represented by the formalization of the UML class diagram. After parsing the ecore file, we extract the UML elements as shown in Figure3.

Fig. 4. Extracted UML elements

3.4. Map to RDF Linked Data

Many complex UML diagrams which include many attributes that can be change into RDF(S) by mutually using this methodology. In mapping modules, it briefly summarizes the mapping in UML and RDF(S) (Cranefield, 2001). It defines mapping between main elements of UML and RDF, for example class, attribute, association etc. to resources, properties, triples etc.

3.5. Generate RDF

This step transforms the parsed output of UML class model into the corresponding concepts RDFs. It will finally produce the resulting RDFs that is stored as text file, and it is displayed on the tool screen as shown in figure 5. And it also displayed the input UML class diagram file that are XML coded and the parsing results on the tool screen. Jena library will be used to generate RDF metadata and triples. Apache Jena (shortly Jena) is a framework of java that is free and open source, and it is used for

creating semantic web and application of linked data (Jena, 2011). This framework is basically created of different APIs that are interacting together to process few RDF data.

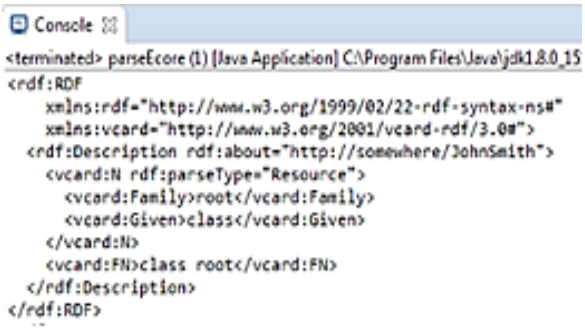

Fig.5. Generated RDF on tool screen

3.6. Verify RDF Linked Data

In our last step, we will verify the created RDF for each linked data. W3C RDF validating service will be used to generate RDF triples and graph. This is validating service to generate the RDF triples and graphs. If this service will be successful in the process, then the generated RDF will be selected as correct one. Figure 4 shows the verified RDF by validating service of W3C.

Validation Results

Your RDF document validated successfully.

Triples of the Data Model

Number	Subject	Predicate	Object
1	http://somewhere/JohnSmith	http://www.w3.org/2001/vcard-rdf/3.0#N	genid:A1503
2	genid:A1503	http://www.w3.org/2001/vcard-rdf/3.0#Family	"class"
3	genid:A1503	http://www.w3.org/2001/vcard-rdf/3.0#Given	"Student"
4	http://somewhere/JohnSmith	http://www.w3.org/2001/vcard-rdf/3.0#FN	"Student class"

Fig. 6. Validated RDF triples

4. Results

We carried out experiments of construction using our implementation tool, with a PC (CPU core i5/2.53 ghz, RAM 4.0 GB and windows 7 system). We choose many UML

class diagrams, that includes important features of UML mentioned in section 3.1. Many more complex diagrams of UML which consist of different features can be converted into RDF by using our tool and approach. There are many types of UML class diagram used in our test, e.g., class of school domain, university system and webpages diagrams. We created many diagrams manually in the EMF(Eclipse Modeling Framework) of different sizes and scales, scales of UML class diagram denotes different number of classes, attributes, roles, association and relations of UML diagrams. The results of generating RDF and verifying RDF that we explained in methodology section show that our approach actually work, and the time complexity of our construction tool is linear with the UML diagrams scales, and it is also consistent with the theoretical analysis. Hence the time complexity of our approach depends upon structure of UML diagrams. Suppose that scale of any UML diagram is $N= N_c + N_a + N_s + N_r + N_{agg} + N_{dep}$, where N_c, N_a, $N_{s,}$, N_r, N_{agg} and N_{dep} denotes the cardinality of the collection of classes, attributes, association roles, aggregation and dependency relations. Figure 5 shows the ratio of process time of UML class diagrams of different sizes during different construction phases. In methodology section we provided construction example of our approach, and the experiment proves that the approach is feasible. Our research bridge the gap between applications of UML and the semantic web.

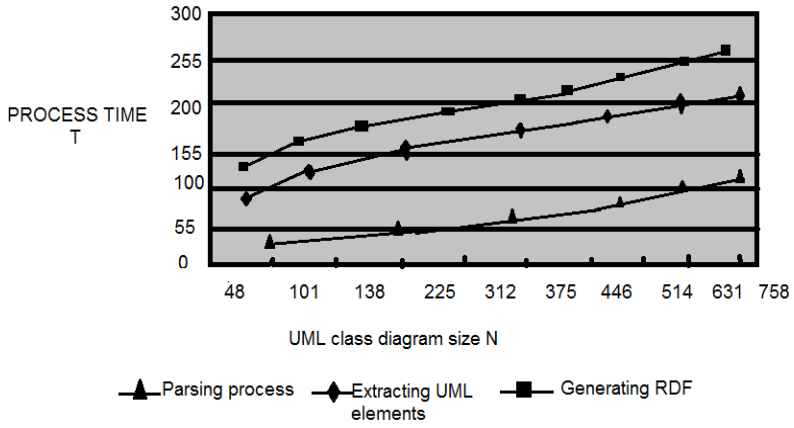

Fig. 7. Results ratio in different phases

This paper also contributes to explore new ways for developing linked data repositories of large sized graphical models by which we can easily search and manipulate from large size of diagrams and data models (RDF-W3C, 2014). To achieve this goal following set of research objectives are target such as theory development for the generation of RDF based Linked data from graphical representations of software artifacts and implementation of the theory for the transformation of graphical models

into linked data repository (Bizer et al, 2009). The experiments to test the performance of the presented approach with the help of examples and real world problems is also discussed in this paper.

5. Conclusion

This research contributes to explore new ways for developing linked data repositories of large sized graphical models by which we can easily search and manipulate from large size of diagrams and data models. This approach works with a set of main modules of construction process. By using this methodology, we are able to generate linked data repositories using UML class model. Because of this we are able to tackle the problem of searching and accessing from large size repositories of diagrams and models of software systems. This research facilitates the use of diagrams and models in bigger repositories.

Acknowledgements

This research would be impossible without patient help, support and encouragement of many people. I thanks all individuals who contributed to this effort.

I acknowledge the loving support of my parents who encouraged and helped me during this research. My warmest thanks go to my friends, teacher and supervisor.

References

1. World Wide Web Consortium. (2014). RDF 1.1 concepts and abstract syntax.
2. Tong, Q., Zhang, F., & Cheng, J. (2014). Construction of RDF (S) from UML class diagrams. Journal of computing and information technology, 22(4), 237-250.
3. Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked data-the story so far. International journal on semantic web and information systems, 5(3), 1-22.
4. Hassanzadeh, O. (2011). Introduction to Semantic Web Technologies & Linked Data. *University of Toronto*.
5. Tong, Q., & Cheng, H. (2017). RDF (S) extraction from fuzzy entity-relationship models. Journal of Intelligent & Fuzzy Systems, 33(5), 2783-2792.
6. Chakkarwar, V. A., & Joshi, A. A. (2016). Semantic Web Mining using RDF Data. International Journal of Computer Applications (0975–8887) Volume.
7. Pham, M. D., Passing, L., Erling, O., & Boncz, P. (2015, May). Deriving an emergent relational schema from RDF data. In Proceedings of the 24th International Conference on World Wide Web (pp. 864-874). International World Wide Web Conferences Steering Committee.
8. Sherif, M. A., & Ngonga Ngomo, A. C. (2015). Semantic Quran. Semantic Web, 6(4), 339-345.
9. Faye, D. C., Cure, O., & Blin, G. (2012). A survey of RDF storage approaches. Revue Africaine de la Recherche en Informatique et Mathématiques Appliquées, 15, 11-35.
10. Korthaus, A., Schwind, M., & Seedorf, S. (2005). Semantic integration of business component specifications with RDF schema. In International Workshop on Semantic Web Enabled Software Engineering (SWESE).

11. Decker, S., Sintek, M., Billig, A., Henze, N., Dolog, P., Nejdl, W., ... & Miklós, Z. (2005). Triple-and rdf rule language with context and use cases.
12. Cranefield, S. (2001). UML and the Semantic Web.
13. Budinsky, F., Steinberg, D., Ellersick, R., Grose, T. J., & Merks, E. (2004). *Eclipse modeling framework: a developer's guide*. Addison-Wesley Professional.
14. Cranefield, S. (2001). Networked knowledge representation and exchange using UML and RDF. Journal of Digital information.
15. Jena, A. P. I. (2011). Jena–A Semantic Web Framework for Java. Talis Systems.