EXTRACTOR WEB-BASED TOOL BY PROCESSING ONTOLOGY TO ACCESS INFORMATION

Aso Mohammed Aladdin[1], Mzhda SabirAbdulkarim[2], Jaza Mahmood Abdullah[3],

University of Sulaimani, IRAQ.

[1]asoaladdin@gmail.com, [2]mzhda.sabir@gmail.com,[3]Jaza.abdullah@univsul.edu.iq

ABSTRACT

There is a rapid growth and success of public information sources on the World Wide Web (www), and it is becoming attractive to extract data from these sources and make it available for further processing by end users and application programs. It is widely acknowledged that the use of ontology is beneficial to access information. For this purpose, available data should be related in the corresponding ontology and the mechanisms of accessing data must be supplied. The common way to link data to ontology is via Resource Description Framework (RDF) representation of available data.

Keywords: Extractor, Ontology, Resource Description, Framework, (RDF), XML, SPARQL, RDF store

INTRODUCTION

There is a rapid growth and success of public information sources on the WWW, and it is becoming attractive to extract data from these sources and make it available for further processing by end users and application programs. It is widely acknowledged that the use of ontology is beneficial to access information. A formal way to describe taxonomies and classification networks is by using Ontologies. They essentially define the structure of knowledge for various domains: nouns represent classes of objects and verbs represent relations between the objects. Class hierarchies can be resembled as ontologiesin object-oriented programming but with several critical differences.Class hierarchies represent structures used in source code whereas ontologies represent information on the Internet. Similarly, ontologies are more flexible because they represent information on the Internet coming from all sorts of data sources. Class hierarchies on the other hand are fairly static and rely on less diverse and more structured sources of data (Berners-Lee, 2001).

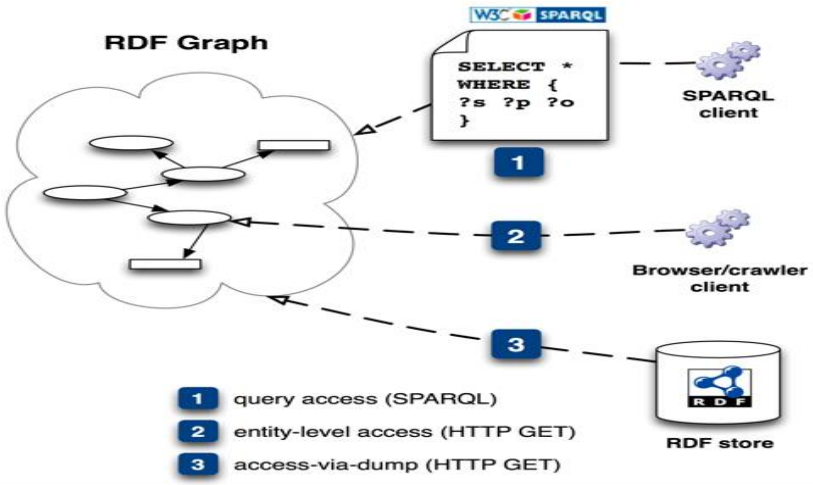

Figure 1- Extractor Web-Based Tool(EWBT)Conceptual Design(RDB2RDF, 2004)

Ontologies have a key role to support information exchange between various networks. For this purpose, available data should be related in the corresponding ontology and the mechanisms of accessing data must be supplied. The common way to link data to ontology is via Resource Description Framework (RDF) representation of available data. RDF is a syntaxdefined by the W3C to express an RDF graph as an XML document. It describes the data as items of the ontology that is represented in terms of an RDF Schema. Furthermore, RDF query languages as the familiar interface is used for interacting with ontology-based RDF data and present the SPARQL language in more detail, as shown in the above figure-1.

The same way SQL is the standardized query language for relational databases, SPARQL is the standardized query language for RDF. You can find some similarities between SQL and SPARQL because they shares several keywords like SELECT, WHERE, etc. In this paper, an explanation has been provided for all the tasks and issues involved to implement these project functions. Tasks include implementation of an extraction tool for retrieving data from XML documents and integrating these with data extracted from DBpedia or any URL resources. Consequently, all the extracted data, stored in a triple store implemented for this purpose and then can be queried for some specific information. Thus, this tool used as a web technology which is provided for user to enable search for specific information. In general, the Extractor Web-Based Tool (EWBT) has more aspects for semantic web depending on ontology language: Introduce and define formal semantics ontology infrastructure.

- It has a more conjectural option for some main model and richer ways to realize concepts and attributes.
- Ontology assists the developer to customize editors and inference engine.
- Flexibility is an important factor of using ontologies. Thus, it is useful for searching, extracting and maintenance information.

Background and Related Work

In the near previous years, semantic web has been developed for several domains by using various efforts and approaches; they were focusing on web-based which can run on different platforms. As seems to be the assumption is that the increasing amount of available semantic data which generates the Semantic Web. These can be utilized as a background knowledge source in ontology (Anon2009). Thus, this basis satisfies the requirements that will identified in the later sections and beneficial during limitation EWBT. Certainly, this scale basically, heterogeneous collection of semantic data supplies identified knowledge. Accordingly, it likely causes less faulty than the knowledge obtained from textual sources, however, leads to better mappings and matching(Bikakis, 2012).

A related work contains approaches to matching that depend on the background knowledge which has used. Two classes of such matchers or extractors are distinguished rely on the form of the external resource, i.e., online textual information source and ontology (Liyang, 2011).Several ontology has developed which based matching and extracting knowledge rely on a large-area generic resource, for example, Cyc. Cyc is an intelligent artificial software project and its knowledgebase with no cost to the research community. Besides, it can be used as the basis for a wide-ranging of intelligent applications that efforts to assemble a comprehensive ontology and knowledgebase such as information extraction and concept tagging or semantic database integration (Carroll, 2004).

Other intelligent semantic web example is a WordNet which is known as a lexical database using for the English language. It includes groups' English words into synonym sets which is called SynSETs, provides definitions, introduction, part of speech and examples sentence.

Leena and Luna International, Chikusei, Japan.
(株) リナアンドルナインターナショナル, 筑西市,日本
ISSN: 2186-845X ISSN: 2186-8441 Print
www.ajmse. leena-luna.co.jp

Moreover, it shows records a number of relations among these sets of synonyms. Thus, it can be seen as a dictionary combination and thesaurus which is accessible to human users via a web browser as artificial intelligence applications (Xie et al, 2004).

Significantly, the developer should obtain license the database and ontology that have been used for a software tools or use freely available schema database.

Stages of Ontology Development

An accurate development process for building ontology should prescribe the guidelines for the specification, conceptualization, formalization and implementation of ontology. The specification stage focuses on the roles and aims of the anticipated ontology as well as human person who will be using. Therefore, EWBT is also provided for user to enable looking for some specific statistics from DBpedia or any URL resources, for instance, extracting Artist information such as name of clips, photos, etc. Formerly, conceptualization phase emphasis on the design of the software tool and the necessity architecture that have used.

The formalism is used in this developing tool because software developers generally are used for object-oriented systems development. After that the implementation phase of ontology is formally represented in one of Semantic Web languages with editor ontology platforms to obtain the formal version as an extractor. Accordingly, the formal ontology must be organized onto the web programming language such as Java, C++, .NET, etc (Berners-Lee, 2001). The later sections explain how to develop and show that the types of developing plat form which have used in the EWBT.

EWBT Architecture

As argued that in the previous section, sole ontology language needs the Semantic Web's large range which holds of users and applications. Ontology Language organizes as a grown series layers of sublanguages. Thus, it assists adding a new layer because each additional layer increases a new functionality and complexity to the previous one (W3C, 2004) (Sinir, 2008).

In general, it is crucial to define the relation between the principle of Infrastructure Ontology Language and RDFs. This shows that simple RDF agents should develop with infrastructure ontologies and pick up as much of their meaning as imaginable with their several limited abilities (Sinir, 2008).

According to EWBT, the fixed architecture is illustrated similarly as an RDF-based Web ontology language. Initially, we will map the original RDFS into RDFS with Fixed Architecture (FA). Thus, it is equitable to define a EWBT as Ontology and RDF web base for the following layers:

1. M Layer: The Metalanguage Layer and it has a main responsibility for define the language layer. EWBT Examples for this primitive in this layer are RDF class.
2. L Layer: The Language Layer or Ontology Language Layer and it is an instance of the M Layer. It has principal responsibility to define a language for specifying ontologies. Examples of model in this layer are rdf class and rdf property.
3. Ontology Layer: The Ontology Layer and it is an instance of Language Layer. Its primary responsibility is to define a language that describes a specific domain i.e. ontology. Examples for this layer are "Artist" and "Clip", which are instances of "hasArtist", which is an instance of property.
4. I Layer: The Instance Layer and it is an instance of Ontology Layer. It is in

charge of describing a specific domain. Examples in this layer are “Name of the Artist".

As discussed, the purpose of this tool is to create a program to retrieve information from the xml documents and DBpedia.org or other specific web pages then present it after processing it. All the tasks implemented according to the EWBT requirement. As a result, it is required to develop a program able to extract information from XML pages’ statements and query the returned information then represent it as RDF triples. The following main platforms are necessity to process this tool.

Extractor and Integrate

An extraction tool should be built to identify and excerpt information in the XML files on the site. Also, the result of the extractor tool must be an RDF triples. In spite of being a relatively new development, XML has become exactly essential for enabling data interchange between otherwise incompatible systems. Moreover, some if not most Web content might be available in the future in formats more suitable for automated processing (W3C, 2011) (Shannon, 2006).

There are two reasons for integrating information from multiple pages. First, some of the data is extracted from the XML files. Secondly, some of the information can be found querying DBpedia.org URI. In this part of the project the task is to integrate extracted data from XML files with the data retrieved from the DBpedia.org. The result is a complete set of RDF triples.

RDF Store or RDFS

Another task in this project is to develop a triple store which contains all the triples generated by the extractor. It is a database for saving and retrieving triples. Triples are data entities that composed of subject, predicate and object. An RDF store allows storage of RDF data and schema information, and provides methods to access that information.

It means that RDF is the first standardized web-based languages and includes of a data model that used for describing resources on the web, although, RDFS is a version that improved from RDF which supplies facilities for defining the basic elements of ontology. These elements contain classes and hierarchy of classes, properties, domain and range of properties []. In general, for signifying ontology, RDF uses a basic statement in the form which includes such as <S, P, O>. The representation of this statement is that S, P and O is showing subject, property and value consequently. Moreover, S and P are devoted as uniform resource identifiers (URIs) in a RDF statement, while O is either a URI or a literal value (Feigenbaum, 2007) (Shadbolt et al, 2007).

Jena API

Java ontology API is known as Jena which includes an object classes to create and manipulate RDF graphs that defined by interfaces. A RDF graph is called a model and represented with the Model interface. Consequently, it provides methods that help to save and retrieve RDF graphs to and from existed files. The Jena platform has a great specification which supports several database management systems, for instance, MySQL, Oracle, etc. It also supplied different tools including query language, I/O modules for RDF/XML output, parser for RDF/XML, etc. (Marshall et al, 2006) (Gärdenfors and Peter, 2004).

Web Interface

One of the important aspects of this project is providing a web interface for users to search for information in the triple store and to control the extractor tool and show extractor result.

Leena and Luna International, Chikusei, Japan.
(株) リナアンドルナインターナショナル, 筑西市,日本
ISSN: 2186-845X ISSN: 2186-8441 Print
www.ajmse. leena-luna.co.jp

The interface should make it possible for the users to query about artists, albums, tracks of the albums, music venues, performances and the votes given by users. Also, the interface must provide features for starting and stopping the extractor as well as exporting the RDF triples to an RDF file.

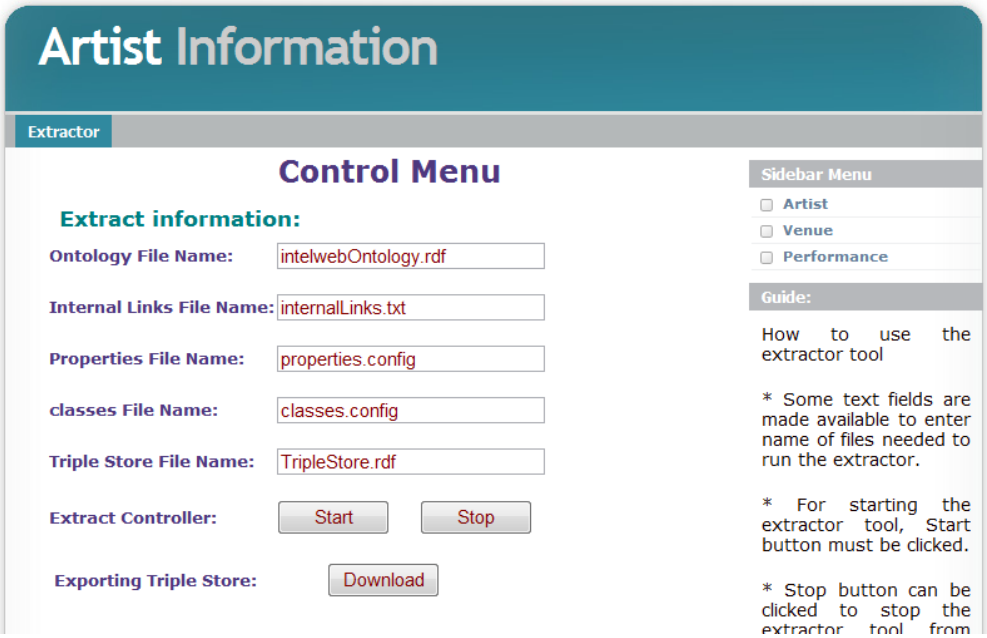

Figure 2Extractor tool control panel

Extractor System Design and Work Flow

Inspite of there are many clarifications to design this EWBT tool, Creating RDF triples design from the extracted information and organizing them in the right format was hardest task in this project. Mapping between ontology and the retrieved information and keeping track of all the matching was a significant challenge for our team. As well as comparing GeoLon and GeoLat to find distance on the map which is also the hardest task during implementing our project.

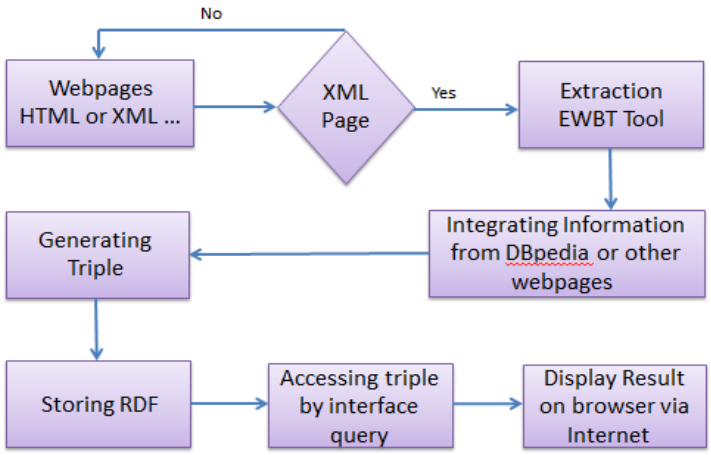

Figure 3. system work flow

This extractor tool accesses XML files on the web and extracts the information from them. It can also search the DBpedia.org for getting information which is required according to the users. The information gathered from both XML files and DBpedia.org which are then

integrated and put them in the RDF triple format. The ontology is used to represent triples generated. Look at system work flow in the above Figure 3.

Extractor Test and Result

The EWBT tool has been test on real sample, that were a collection of webpages using crawler tool whichwas created specifically for this purpose, each page was contained XML file that holds information about Artists, Venue and their performances. As it is explained in the previous section, the user only sees the final phase of system work flow which is the results on the browser.

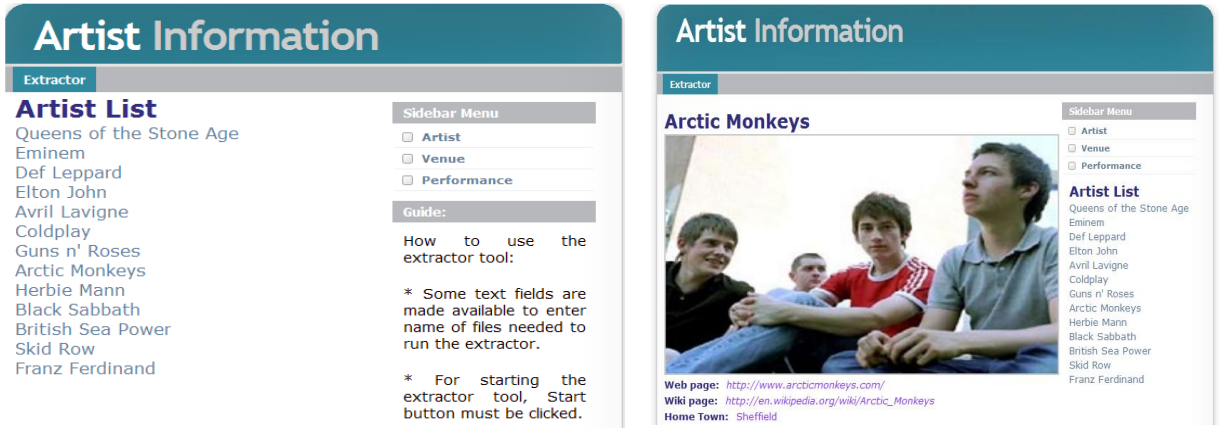

Figure 4 user can choose artist to look at all allbums and tracks

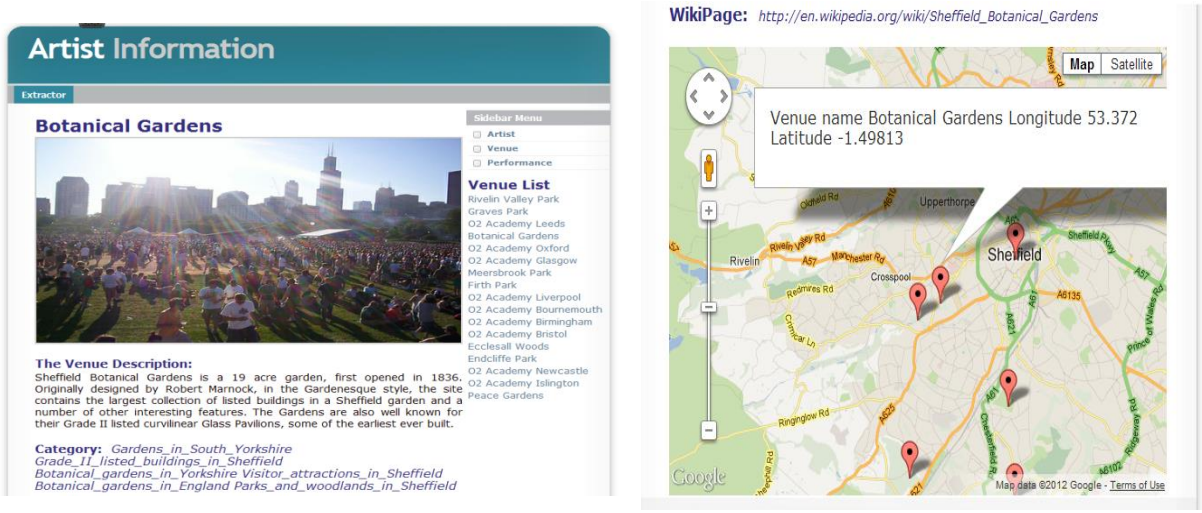

Figure 5 user may looks at Venue in the list, also located on the google map

CONCLUSİON AND FEATURE WORKS

To conclude, the report introduces a web data extraction model based on XML. By implementing thisproject, a semantic web data extraction model was provided. The extracted data from both xml documents and DBpedia are processed and stored in a triple store after changing its format to RDF format. In addition, the retrieved data is integrated based on ontology to map between different data schemas. There is also an interface to enable user to interact with the system and query about all information extracted. Moreover, SPARQL

Leena and Luna International, Chikusei, Japan.
(株) リナアンドルナインターナショナル, 筑西市,日本
ISSN: 2186-845X ISSN: 2186-8441 Print
www.ajmse. leena-luna.co.jp

queries are used to access information and search in the RDF triples store. A facility has been provided to user to control the extractor tool by starting and stopping it.

The model we introduce is designed for XML document. But data on the web are not limited in XML document; there are other forms, such as databases, logs, and files. How to extract data from these sources is a great challenge.

REFERENCES

[1] Anon. (2009) Advantage of semantic data. Retrieved from http://www.semagix.com/advantage-of-semantic-data.htm.

[2] Berners-Lee, T., Hendler, J., & Lassila, O. (2001). *The semantic web*. USA: Scientific American Magazine.

[3] Bikakis, N. (n.d.). XML and semantic web W3C standards timeline. Retrieved from http://www.dblab.ntua.gr/~bikakis/XMLSemanticWebW3CTimeline.pdf.

[4] Carroll, J., & Stickler, P. (2004). RDF triples in XML. Retrieved from http://www.hpl.hp.com/techreports/2003/HPL-2003-268.pdf.

[5] Feigenbaum, L. (2007). The semantic web in action. USA: Scientific American.

[6] Gärdenfors, P. (2004). *How to make the Semantic Web more semantic: Formal ontology in ınformation systems* (Proceedings of the third international conference). Sweden: Lund University.

[7] Liyang, Y. (2011). *A developer's guide to the Semantic Web*. USA: Springer.

[8] Marshall, C. C., & Shipman, F. M. (2003). Which semantic web? (PDF). Retrieved from *cis.k.hosei.ac.jp/~jianhua/course/net/Reference/**semantic**7.pdf*.

[9] RDB2RDF. (2004). Mapping relational data to RDF. Retrieved from http://www.w3.org/2001/sw/rdb2rdf/status/2010/#(4).

[10] World Wide Web Consortium (W3C). (2004). *Semantic web standards*. USA: World Wide Web Consortium (W3C).

[11] Shadbolt, N., Hall, W., & Berners-Lee, T. (2006). T*he semantic web revisited*. USA: IEEE Intelligent Systems.

[12] Shannon, V. (2006). *A 'more revolutionary' web*. USA: International Herald Tribune.

[13] Sinir, S. (2008). Ontology Mapping Survey. Retrieved from http://www.authorstream.com/Presentation/aSGuest757-94524-ontology-mapping-survey-science-technology-ppt-powerpoint/.

[14] World Wide Web Consortium. (2011). *W3C "W3C semantic web activity*. USA: World Wide Web Consortium.

[15] Xie, W., Zeng, C., Lin, Y., & Chen, Y. (2004). A web data extraction technique based on XML. Retrieved from http://metronu.ulb.ac.be/imacs/papers/T3-I-89-0323.pdf.