Buletin Teknol. Tanaman, Bil.12, Tambahan.2 (2015): 84-89

Developing an Architecture to Publish Data Warehouses Metadata Based on the Semantic Web

Mohammad Harizi, Esmaeel Sahanalizade, Mohammad Sarkaki

Abstract

The dispersion of data warehouses in the current world of information is a challenge which causes their owners or beneficiaries to have difficulty understanding and perceiving the relevant and existing data. What defines a data warehouse is metadata. Put another way, metadata plays a vital role in introducing and identifying data warehouses. This paper was intended to develop an architecture including different standards and technologies in order to investigate how to use metadata in data warehouses in the context of the Semantic Web. The aim of the proposed architecture was to study how the metadata of data warehouses would be introduced, published and aggregated in an area for the users who intend to process them.

Key words: Data warehouse, Business Metadata, Semantic Web, RDF Syntaxes, SDMX, and Ontology.

Introduction

Consider a multinational or chain company with branches or agencies all over the world, or imagine national and international organizations and institutes which work in a common area. Each of them has a data source which is preferably called a data warehouse in this paper. A data warehouse is a sort of database with an optimal data structure which stores an extract or a part of the operational data for queries and analyses in the decision-making process. Data warehouses can be used as input in Decision Support Systems. Assuming the dispersion of data warehouses in the aforesaid scenario, how can the owners of those institutes understand the meaning of data correctly and comprehensively to make major decisions? In other words, the dispersion of data sources should not deprive the managers of a comprehensive perception of data concept in all of their data warehouses. Various standards and technologies are required to meet this challenge. This paper tried to introduce an efficient architecture for this purpose. At first, two important components are described in this regard.

Metadata

Metadata is required to understand data. In this regard, it is very crucial to use metadata existing in data warehouses to identify them. Metadata is one of the most basic elements or key factors in a data warehouse. Metadata is the central point in tracking, designing, constructing and restoring information in the data warehouse. Therefore, it can greatly influence the introduction of the data warehouse. Moreover, it is very effective in aggregating and transferring data. There is no doubt that metadata can be used to introduce the data warehouse in data processing area. It is often discussed as technical and business metadata in the data warehouse. Having a sort of internal usage, the technical metadata, which is usually structured, is used to design, develop and maintain the data warehouse. However, the business metadata, which is unstructured, is valuable for a business person. It is meant for the external use. This sort of metadata specifies the text and data semantics which a businessman can understand and perceive [1]. In this study, metadata refers to the business metadata. Despite the multiplicity of data warehouses, cooperating organizations or institutes should have a correct perception of relevant data semantics, which are not necessarily in the same format, in order to use the allowable data existing in each other's data warehouses. It is first required to identify and introduce data. This is possible with the help of business metadata existing in data warehouses. However, they should be expressed in a structured way so that publishing and aggregating them can let different parties or the decision-making organization reach a correct understanding of data in the relevant area. Thus, metadata should be published and aggregated in a way that they can be finally understood and processed by the decision-making organization.

The Semantic Web

Undoubtedly, the Semantic Web will play an important role in this architecture. The aim of the Semantic Web is to enable machines to understand data in a specific domain, besides storing and transferring them. In other words, the highest purpose of the Semantic Web is to improve the data content in the Web more, insofar as they can be published, understood and processed by all agents. In this regard, different technologies have been introduced from how metadata is dealt with to the way they are transferred, queried, and processed.Consequently, metadata is necessary to have a correct and comprehensive translation of data warehouses content, and the Semantic Web provides a platform to publish and aggregate metadata in the best way possible.

Research Literature

Some efforts have been made by researchers whether in relation to publication or aggregation of metadata. In Paper [2], a framework was introduced for data aggregation based on metadata to manage researches. In [3], a framework including the data aggregation process in the data warehouse was explained. By definition, ontology would occur. Moreover, [4] proposed a strategy to use the knowledge-based system in order to aggregate metadata in a data warehouse. In Paper [5], a conceptual framework was introduced to create a metadata for a particular type of the data warehouse. Additionally, an approach was proposed in [6] to aggregate metadata for

OLAP tools. However, the majority of the aforesaid studies used metadata aggregation only in one particular area, or the given plan lacked a complete and comprehensive framework. They also investigated different types of metadata mostly inside data warehouses, not outside their functioning boundaries.

This study was intended to introduce a complete architecture to investigate how metadata is published and aggregated outside the internal domain of activities in a data warehouses. In Part 3, an outline of the proposed architecture is first introduced. Creating a type of business metadata, then the phases and working details are explained step by step. Part 4 is dedicated to the conclusion.

Explaining the Proposed Architecture

The following challenges are in the way of publishing and aggregating the business metadata in three main steps:

1-In the starting point (data warehouses): how to identify and introduce the desired metadata

2-In the transition path: how to publish and aggregate metadata

3-In the destination point (decision-making organization): how to make queries and process metadata

Certain technologies and standards should be used in the proposed architecture to meet the above-mentioned challenges. RDF can be an appropriate solution to cope with the first challenge. Using RDF framework, the desired sets of metadata can be defined in data warehouses. They can also be expressed through an appropriate grammar. Doing so, the business metadata is structured into an appropriate format. The terms of Dublin Core, which is an RDF dictionary used to define metadata words, were first used in this study. Then grammars such as Tree triples and Turtle were used to express the desired RDF. Treetriples grammar is an appropriate option because it is ideal to query and access RDF content directly. Moreover, Turtle grammar can be used to query RDF documents through SPARQL. SDMX was employed to deal with the second challenge. SDMX is an ISO standard for transferring and sharing statistical data and metadata among organizations. SDMX-ML grammar, which is based on XML, was used to transfer metadata in the proposed architecture. After transferring metadata from different sources, they should be aggregated. Using ontology can be an appropriate solution to organize and aggregate metadata so that it would be possible to process and obtaining deductions. In this regard, SPARQL can be an appropriate substitute in order to make queries.

An Outline of the Architecture

Figure 1 indicates a general outline of the standards and technologies used in the proposed architecture. The final aim is to create a mechanism to extract the business metadata from different data warehouses in an area for analysis and query after transferring to a data-processing center.

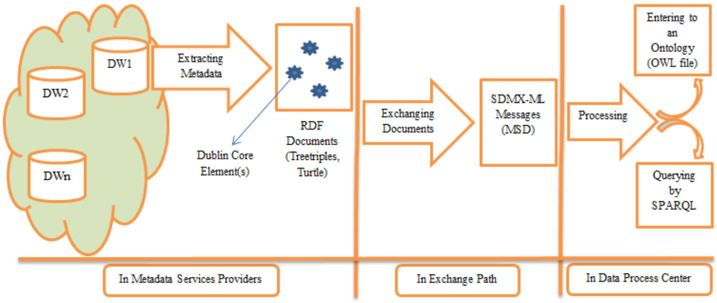

Figure 1: The general outline of the proposed architecture

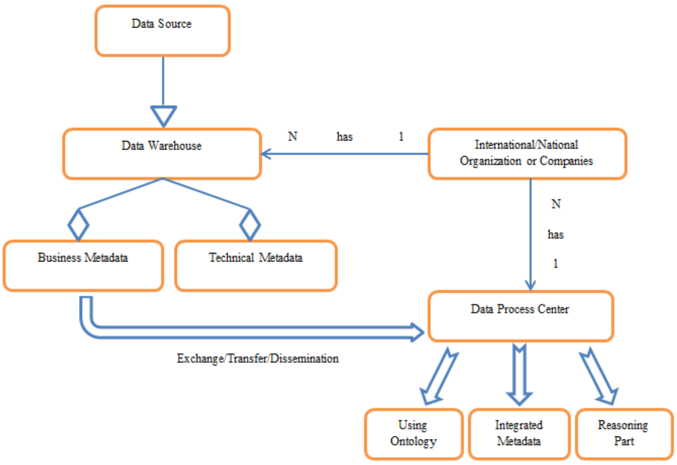

Figure 2: An outline of the concepts used in the architecture

The details pertaining to the proposed architecture are described in the following.

Constructing a Sample of the Business Metadata

As it has been pointed out, this paper used Dublin Core terms to introduce the words pertaining to metadata. DC includes 15 metadata elements to explain the source in an interdisciplinary information environment. These elements contribute to a larger set of metadata words and the technical specifications mentioned by Dublin Core Metadata Initiative [7].In the aforesaid set, each element has a Label which is used for human and a Name to be used in machine processing. The Name of each element is added to a DCMI Namespace URI so that the uniform source identifier can be organized as a unique universal identifier. Now the elements such as *source*, *identifier*, *title* and *description* are used to construct a sample business metadata. The following RDF graph indicates the business metadata pertaining to the type of *sales unit*.

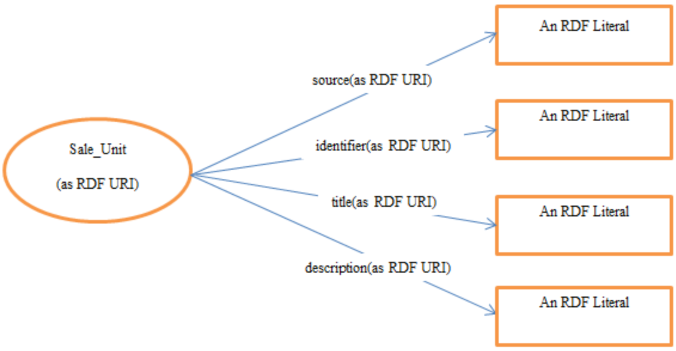

Figure 3: The sample graph of the business metadata

This graph can be expressed in the form of DC-Text as follows

```
@prefix          dcelements : <http://purl.org/dc/elements/1.1/>.
DescriptionSet(
Description(
ResourceURI( <http://example.org/Sale_Unit> )
Statement(
                 PropertyURI( dcelements : source )
                 LiteralValueString( "DW1"
                 Language( "en" )
                 )
      )
Statement(
                 PropertyURI( dcelements : identifier )
                 LiteralValueString( "IRR"
                 Language( "en" )
                 )
      )
Statement(
                 PropertyURI( dcelements : title )
                 LiteralValueString( "Rial"
                 Language( "en" )
                 )
      )
Statement(
                 PropertyURI( dcelements : description )
                 LiteralValueString( "It is the currency of Iran"
                 Language( "en" )
                 )
      )
)
)
```

In the next part of constructing the business metadata, it is necessary to express it based on RDF grammars. First, the equivalent of RDF/XML metadata grammar can be seen as follows:

```
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
         xmlns:dcelements="http://purl.org/dc/elements/1.1/">
       <rdf:Description   rdf:about="http://example.org/Sale_Unit">
                 <dcelements:source                     xml:lang="en">DW1</dcelements:source>
<dcelements:identifier  xml:lang="en">IRR</dcelements:identifier>
<dcelements:title                 xml:lang="en">Rial</dcelements:title>
<dcelements:description           xml:lang="en">It is the currency of Iran
</dcelements:description>
       </rdf:Description>
</rdf:RDF>
```

Moreover, the equivalents of Treetriples and Turtle grammars can be considered as follows:

```
<!DOCTYPE           rdf [
<!ENTITY            rdf       "http://www.w3.org/1999/02/22-rdf-syntax-ns#">
<!ENTITY            dcelements          "http://purl.org/dc/elements/1.1/">
]>
<rdf  xmlns="http://djpowell.net/schemas/treetriples/1/">
      <s        id="http://example.org/Sale_Unit">
                <p        id="&dcelements;source">
                          <o        xml:lang="en">DW1</o>
                </p>
                <p        id="&dcelements;identifier">
                          <o        xml:lang="en">IRR</o>
                </p>
                <p        id="&dcelements;title">
                          <o        xml:lang="en">Rial</o>
                </p>
                <p        id="&dcelements;description">
                          <o        xml:lang="en">It is the currency of Iran</o>
                </p>
      </s>
</rdf>

@prefix         rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#>.
@prefix         dcelements:http://purl.org/dc/elements/1.1/>.
<http://example.org/Sale_Unit>

                dcelements:source             "DW1" @en;
                dcelements:identifier         "IRR"   @en;
                dcelements:title              "Rial"   @en;
                dcelements:description        "It is the currency of Iran"  @en;
```

The Use of SDMX for Transferring

SDMX is a set of standards used to exchange and publish data/metadata [8]. SDMX provides a method for modeling statistical data, structured metadata and data-transferring processes. In other words, SDMX has introduced some standard formats for data and metadata along with content-oriented guidelines and an IT architecture for transferring data and metadata.It also defined a model for additionally explanatory metadata, called the Reference Metadata considered in this paper, in the text format. In SDMX, the reference metadata is a large set of concepts explaining and describing the statistical group of data and processes completely. The reference metadata usually depends on the entire set of data or even the institute presenting data, not on a particular aspect or series of data. This set can include the conceptual metadata – which would describe the concepts which have been used and taken into account – and methodological and qualitative metadata.For the reference metadata, SDMX introduced a format based on a Metadata Structure Definition. An MSD is designed and developed in the following way [9]:

1- Analyzing the set of metadata to identify “Concepts”

2- Determining the structure of “Metadata Report” with respect to the concepts which were used, their hierarchy, and the way they were displayed (Code List/Text)

3- Specifying “Object Type” to which metadata are connected.

A sample MSD structure accompanied by its message format, which is in the form of an XML file, is indicated here. According to the following table, DC properties were used in this structure.

Table 1: The MSD Structure of the business metadata

Attribute	Concept	Sub Attribute	Type	Concept Id *	Format
Sale_Unit	Unit	Source	Property	source	Text
		identifier	Property	identifier	text
		title	Property	title	text
		description	Property	description	text

```
* is equal to the Name property in DC.
<?xml version="1.0" encoding="UTF-8"?>
<Structure xmlns="http://www.SDMX.org/resources/SDMXML/schemas/v2_0/message"
xmlns:structure="http://www.SDMX.org/resources/SDMXML/schemas/v2_0/structure"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.SDMX.org/resources/SDMXML/schemas/v2_0/message SDMXMessage.xsd">
 <Header>
  <ID>REGISTRY_RESPONSE</ID>
  <Test>true</Test>
  ....
</Header>
 <MetadataStructureDefinitions>
  <structure:MetadataStructureDefinition id="MY_MSD_ID" agencyID="MY_ORGANIZATION_1" version="1.0"
isFinal="true" validFrom="2015-09-15" validTo="2015-09-30">
   <structure:Name xml:lang="en">My MSD</structure:Name>
   <structure:TargetIdentifiers>
    <structure:FullTargetIdentifier id="MY_TARGRT_ID">
     <structure:Name xml:lang="en">My Target</structure:Name>
    </structure:FullTargetIdentifier>
   </structure:TargetIdentifiers>
   <structure:ReportStructure id="MY_REPORT_STRUCTURE_ID" target="MY_TARGRT_ID">
    <structure:Name xml:lang="en">My Report Structure</structure:Name>
    <structure:MetadataAttribute conceptRef="SALE_UNIT_ID" conceptSchemeRef="MY_CONCEPT_SCHEME_ID"
conceptSchemeAgency="MY_ORGANIZATION_1" conceptSchemeVersion="1.0" usageStatus="Mandatory">
     <structure:TextFormat textType="String" />
     <structure:MetadataAttribute conceptRef="SOURCE_ID" conceptSchemeRef="MY_CONCEPT_SCHEME_ID"
conceptSchemeAgency="MY_ORGANIZATION_1" conceptSchemeVersion="1.0" usageStatus="Mandatory">
      <structure:TextFormat textType="String" />
      <structure:Annotations>
       <common:Annotation
xmlns:common="http://www.SDMX.org/resources/SDMXML/schemas/v2_0/common">
        <common:AnnotationTitle />
        <common:AnnotationType>ATTRIBUTE_NAME</common:AnnotationType>
        <common:AnnotationURL />
        <common:AnnotationText xml:lang="en">Source of sale unit</common:AnnotationText>
       </common:Annotation>
      </structure:Annotations>
     </structure:MetadataAttribute>
     .....
    </structure:MetadataAttribute>
   </structure:ReportStructure>
  </structure:MetadataStructureDefinition>
 </MetadataStructureDefinitions>
</Structure>
```

The above MSD message can be spread and analyzed in the next phases.

Metadata Query and Analysis

Forming and organizing the published metadata, they can be queried and analyzed. Inserting metadata into an ontology through mapping the file, including metadata in the form of XML in the previous part, onto an OWL file, one solution is to aggregate them for deduction and analysis [10].Using SPARQL, another approach is to query the transferred metadata. However, the metadata expressed with respect to Turtle grammar is required here. There has been a position considered for it as specified in the architecture.SPARQL has a query language attribute for RDF which uses a grammar similar to Turtle to express its patterns. However, the query result of SPARQL can still be converted into XML on Turtle [11].Finally, the following figure indicates the interactions required among different parts of the proposed architecture:

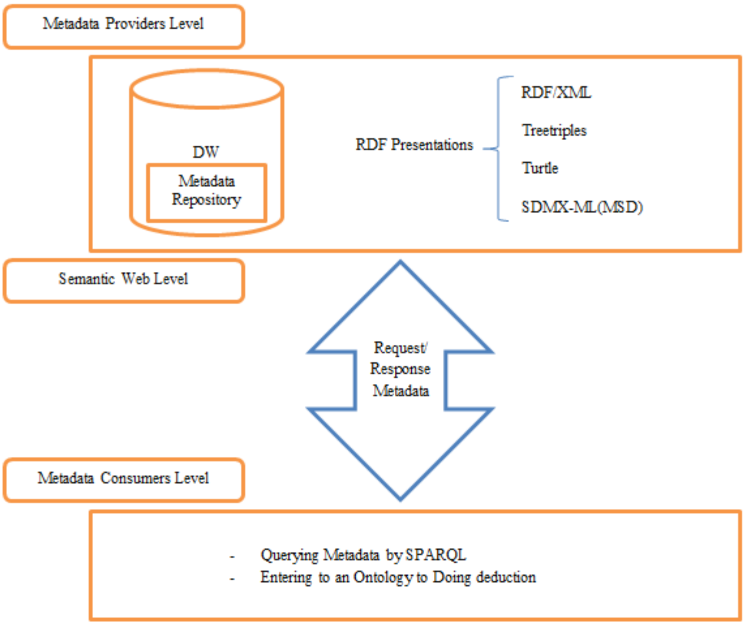

Figure 4: Interactions between different parts of the architecture

As observed, the above figure is comprised of three different levels including the metadata providing level, the Semantic Web level and the metadata consuming level. The metadata providing level pertains to how metadata is prepared to be published. At this level, different types of displays can be taken into account to express metadata. The Semantic Web level deals with how to publish and transfer the type of metadata. Finally, at the metadata consuming level, the query and processing operations would be done on metadata.

Conclusion

Based on the Semantic Web standards and technologies, this paper was intended to propose a complete and comprehensive architecture in order to introduce and publish the business metadata from data warehouses. In the proposed architecture, RDF played the major role as the fundamental framework of the Semantic Web. Likewise, SDMX standard was taken into account in order to publish metadata. Finally, the use of ontology was proposed as a solution to analyze and deduct the published metadata. Another solution was to use SPARQL in order to query and process the documents containing metadata.

Acknowledgement

This article is extracted from a research project entitled "Comparison of Metadata Design Algorithms on Data Warehouse", and the was funded by Ramhormoz Branch, Islamic Azad University, Khuzestan, Iran.

Resources

1. Inmon, W., O'Neil, B., Fryman, L. Business Metadata. Burlington, MA: Morgan Kaufinann. 2008.
2. Zhilong Chen, Dengsheng Wu, Jingxiu Lu, Yuanping Chen. Metadata-based information resource integration for research management. Procedia Computer Science 17. 54 – 61. 2013.
3. ALBERTO SALGUERO, FRANCISCO ARAQUE, CECILIA DELGADO. Ontology based framework for data integration. WSEAS TRANSACTIONS on INFORMATION SCIENCE & APPLICATIONS. Issue 6, Volume 5, June 2008. 953-962.
4. Dan Wu, Anne Hakansson. Applying a Knowledge Based System for
5. Metadata Integration for Data Warehouses. KES 2010, Part IV, LNAI 6279, 60–69, 2010.
6. M. Laxmaiah, A. Govardhan. CONCEPTUAL METADATA FRAMEWORK FOR SPATIAL DATA WAREHOUSE. International Journal of Data Mining & Knowledge Management Process (IJDKP), Vol.3, No.3, May 2013. 63-73.
7. Jesus Pardillo, Jose-Norberto Mazon, Juan Trujillo. Model-Driven Metadata for OLAP Cubes from the Conceptual Modelling of Data Warehouses. DaWaK 2008, LNCS 5182, 13–22, 2008.
8. J. Kunze, T. Baker. Dublin Core Metadata. http://tools.ietf.org/pdf/rfc5013. August 2007.
9. SDMX 2.1 User Guide: Version 0.1. SDMX 2.1 documentation. 2012.
10. Chris Nelson. Reference Metadata Support in SDMX Version 2.0. Metadata Working Group, 7-8 June 2007.
11. Yahia Nora, Mokhtar Sahar A., Ahmed AbdelWahab. Automatic Generation of OWL Ontology from XML Data Source. International Journal of Computer Science Issues (IJCSI), Mar2012, Vol. 9 Issue 2. 2012.
12. Dave Beckett, Jeen Broekstra. SPARQL Query Results XML Format. http://www.w3.org/TR/rdf-sparql-XMLres/. W3C Recommendation 21 March 2013.

Mohammad Harizi, Department of Computer, Ramhormoz Branch, Islamic Azad University, Ramhormoz, Iran
Corresponding Author Email: Mohammad.harizi@iauramhormoz.ac.ir
Esmaeel Sahanalizade, Mohammad Sarkaki, Department of Computer, Ramhormoz Branch, Islamic Azad University, Ramhormoz, Iran