Chapter 14
Data Warehouses and the Semantic Web

The availability of enormous amounts of data from many different domains is producing a shift in the way data warehousing practices are being carried out. Massive-scale data sources are becoming common, posing new challenges to data warehouse practitioners and researchers. The semantic web, where large amounts of data are being stored daily, is a promising scenario for data analysis in a near future. As large repositories of semantically annotated data become available, new opportunities for enhancing current decision-support systems will appear. In this scenario, two approaches are clearly identified. One focuses on automating multidimensional design, using semantic web artifacts, for example, existing ontologies. In this approach, data warehouses are (semi)automatically designed using available metadata and then populated with semantic web data. The other approach aims at analyzing large amounts of semantic web data using OLAP tools. In this chapter, we tackle the latter approach, which requires the definition of a precise vocabulary allowing to represent OLAP data on the semantic web. Over this vocabulary, multidimensional models and OLAP operations for the semantic web can be defined. Currently, there are two proposals in this direction. On the one hand, the data cube vocabulary (also denoted QB) follows statistical data models. On the other hand, the QB4OLAP vocabulary follows closely the classic multidimensional models for OLAP studied in this book.

In this chapter, we first introduce in Sect. 14.1 the basic semantic web concepts, including the RDF and RDFS data models, together with a study of RDF representation of relational data and a review of R2RML, the standard language to define mappings from relational to RDF data. In Sect. 14.2, we give an introduction to SPARQL, the standard query language for RDF data. In Sect. 14.3, we discuss the representation and querying of multidimensional data in RDF, including an in-depth discussion of the QB and QB4OLAP vocabularies. We continue in Sect. 14.4 showing how the Northwind data cube can be represented using both vocabularies. We conclude in Sect. 14.5 by showing how to query the QB4OLAP representation of the Northwind data warehouse in SPARQL.
14.1 Semantic Web

The semantic web is a proposal oriented to represent web content in a machine-processable way. The basic layer for data representation on the semantic web recommended by the World Wide Web Consortium (W3C) is the resource description framework (RDF). In a semantic web scenario, domain ontologies are used to define a common terminology for the concepts involved in a particular domain. These ontologies are expressed in RDF or in languages defined on top of RDF like the Web Ontology Language (OWL)\(^1\) and are especially useful for describing unstructured, semistructured, and text data. Many applications attach metadata and semantic annotations to the information they produce (e.g., in medical applications, medical images, and laboratory tests). We expect that, in the near future, large repositories of semantically annotated data will be available, opening new opportunities for enhancing current decision-support systems.

14.1.1 Introduction to RDF and RDFS

The resource description framework (RDF)\(^2\) is a formal language for describing structured information. One of the main goals of RDF is to enable the composition of distributed data to allow data exchange over the web. To uniquely identify resources, RDF uses internationalized resource identifiers (IRIs). IRIs generalize the concept of universal resource locators (URLs) since they do not necessarily refer to resources located on the web. Further, IRIs generalize the concept of the uniform resource identifiers (URIs): while URIs are limited to a subset of the ASCII character set, IRIs may contain Unicode characters.

RDF can be used to express assertions over resources. These assertions are expressed in the form of subject-predicate-object triples, where subject are resources or blank nodes, predicate are resources, and object are resources or literals (i.e., data values). Blank nodes are used to represent resources without an IRI, typically with a structural function, for example, to group a set of statements. A set of RDF triples or RDF data set can be seen as a directed graph where subjects and objects are nodes and predicates are arcs.

RDF provides a way to express statements about resources using named properties and values. However, sometimes it is also needed to define kinds or classes of resources and the specific properties describing those resources. A set of reserved words, called RDF Schema (RDFS),\(^3\) is used to define

\(^1\)http://www.w3.org/2004/OWL/
\(^2\)http://www.w3.org/TR/rdf11-concepts/
\(^3\)http://www.w3.org/TR/rdf-schema/