Chapter 7

Creation and Integration of Reference Ontologies for Efficient LOD Management

Mariana Damova
Ontotext AD, Bulgaria

Atanas Kiryakov
Ontotext AD, Bulgaria

Maurice Grinberg
Ontotext AD, Bulgaria & New Bulgarian University, Bulgaria

Michael K. Bergman
Structured Dynamics, USA

Frédéric Giasson
Structured Dynamics, USA

Kiril Simov
Ontotext AD, Bulgaria & Bulgarian Academy of Sciences, Bulgaria

ABSTRACT

The chapter introduces the process of design of two upper-level ontologies—PROTON and UMBEL—into reference ontologies and their integration in the so-called Reference Knowledge Stack (RKS). It is argued that RKS is an important step in the efforts of the Linked Open Data (LOD) project to transform the Web into a global data space with diverse real data, available for review and analysis. RKS is intended to make the interoperability between published datasets much more efficient than it is now. The approach discussed in the chapter consists of developing reference layers of upper-level ontologies by mapping them to certain LOD schemata and assigning instance data to them so they cover a reasonable portion of the LOD datasets. The chapter presents the methods (manual and semi-automatic) used in the creation of the RKS and gives examples that illustrate its advantages for managing highly heterogeneous data and its usefulness in real life knowledge intense applications.

DOI: 10.4018/978-1-4666-0188-8.ch007
INTRODUCTION

Linking Open Data (LOD) (Linking Open Data, 2011) facilitates the emergence of a Web of linked data by publishing and interlinking open data on the web in RDF (Brickley & Guha, 2004). The current 203 datasets in LOD cover a wide spectrum of subject domains – biomedical, science, geographic, generic knowledge, entertainment, government, etc. (State of the LOD Cloud, 2011). As they constantly grow, we face the problem of conveniently accessing, manipulating, and further developing them. It is believed that this large set of interconnected data will enable new classes of applications, making use of more sophisticated querying, knowledge discovery, and reasoning. This calls for approaches for their efficient use and better integration.

At the same time, LOD are characterized by heterogeneity and inconsistency, which makes their use in automated ways via algorithms difficult. A lot of research effort nowadays has been focused on looking for methods to cope with and preserve the diversity of LOD, which can scale and manage their increasing growth rates. These methods bring experimental results, which show that the state of the art is still far from the performance necessary for real life applications.

Another perspective to LOD management, which we adopt in this chapter, relates to Master Data Management (MDM) as understood in the business enterprise and DBMS worlds (Wolter, & Haselden, 2006; Withbrock, 2007; Wikipedia, 2011a; Wikipedia, 2011b). In enterprise settings, the homogeneity of the data is a fundamental requirement, e.g. the entities in the data model and the tables in a physical database have to be identical with respect to their properties, behavior, and management needs (Wolter, & Haselden, 2006). Master data are data that are shared and used by many applications within the organization. MDM aims at ensuring consistency and control of the ongoing maintenance and use of this non-transactional information, critical for the business operation of the organization. Moreover, MDM develops a shared view across the organization by creating and maintaining consistent and accurate lists of master data. Master data usually include reference data, e.g. any kind of data that is used solely to categorize other data found in a database or to relate the data in a database to information outside the enterprise.

Highly heterogeneous contexts such as LOD and the Web need similar mechanisms to ensure consistency based on a set of data agreed upon or commonly acceptable, shared by various datasets, and make them interconnected.

Our main claim in this chapter is that a reference layer, consisting of ontologies with different degrees of generality built on top of LOD and interlinked with their schemata and instances, is a viable and optimal solution for coping with LOD heterogeneity at the present time. In our opinion, such an approach will lead to more efficient LOD management and dataset integration while preserving the diversity of the data. In the Semantic Web, the idea of having integrated global ontology extracting information from the local ontologies and providing a unified view through which users can query the local ontologies is unrealistic, as it is practically impossible to maintain this global ontology in a highly dynamic environment.

The reference layer we propose here, called the Reference Knowledge Stack (RKS), will provide reference points that will serve as bridges between the various views about things, described in the LOD cloud and on the Web.

The idea of building reference structures at the schema level has been advocated previously (e.g. see Jain et al., 2010). Jain et al. (2010) state that it would be valuable to have a schema describing the subject domain of the datasets in LOD. Moreover, the three big players in the Web space—Bing, Google, and Yahoo—recently embraced the same initiative and joint forces to build the so-called Web of Objects (Bing Google Yahoo, 2011).