Chapter 1
Connectivity, Value, and Evolution of a Semantic Warehouse

Michalis Mountantonakis
FORTH-ICS, Greece & University of Crete, Greece

Nikos Minadakis
FORTH-ICS, Greece

Yannis Marketakis
FORTH-ICS, Greece

Pavlos Fafalios
FORTH-ICS, Greece & University of Crete, Greece

Yannis Tzitzikas
FORTH-ICS, Greece & University of Crete, Greece

ABSTRACT

In many applications, one has to fetch and assemble pieces of information coming from more than one source for building a semantic warehouse offering more advanced query capabilities. This chapter describes the corresponding requirements and challenges, and focuses on the aspects of quality, value and evolution of the warehouse. It details various metrics (or measures) for quantifying the connectivity of a warehouse and consequently the warehouse's ability to answer complex queries. The proposed metrics allow someone to get an overview of the contribution (to the warehouse) of each source and to quantify the value of the entire warehouse. Moreover, the paper shows how the metrics can be used for monitoring a warehouse after a reconstruction, thereby reducing the cost of quality checking and understanding its evolution over time. The behaviour of these metrics is demonstrated in the context of a real and operational semantic warehouse for the marine domain. Finally, the chapter discusses novel ways to exploit such metrics in global scale and for visualization purposes.

INTRODUCTION

An increasing number of datasets are already available as Linked Data. For exploiting this wealth of data, and building domain specific applications, in many cases there is the need for fetching and assembling pieces of information coming from more than one sources. These pieces are then used for constructing a *Semantic Warehouse*, offering thereby more complete and efficient browsing and query services (in comparison to those offered by the underlying sources). The term *Semantic Warehouse* (for short warehouse) refer to a read-only set of RDF triples fetched (and transformed) from different sources that aims at serving a particular set of query requirements. In general, there exists *domain independent* warehouses, like the Sindice (Oren, et al., 2008) and SWSE (Hogan, et al., 2011), but also *domain specific*, like TaxonConcept (n.d.) and the MarineTLO-based warehouse (Tzitzikas, et al., 2013, November). Domain specific warehouses aim to serve particular needs, for particular communities of users, consequently their "quality" requirements are stricter. It is therefore worth elaborating on the process that can be used for building such warehouses, and on the related difficulties and challenges.

In brief, for building such a warehouse one has to tackle various challenges and questions, e.g., how to define the objectives and its scope, how to *connect* the fetched pieces of information (common URIs or literals are not always there), how to tackle the various issues of provenance that arise, and how to keep the warehouse fresh (i.e., how to automate its reconstruction or refreshing). This chapter has focused on the following questions:

- How to measure the value and quality of the warehouse (since this is important for e-science)?
- How to monitor its quality after each reconstruction or refreshing (as the underlying sources change)?
- How to understand the evolution of the warehouse?
- How to measure the contribution of each source to the warehouse, and hence deciding which sources to keep or exclude?

These questions have been encountered in the context of a real semantic warehouse for the *marine* domain which harmonizes and connects information from different sources of marine information[1]. Most past approaches have focused on the notion of conflicts (Michelfeit & Knap, 2012), and have not paid attention to *connectivity*. The term *connectivity* express the degree up to which the contents of the warehouse form a connected graph that can serve, ideally in a correct and complete way, the query requirements of the warehouse, while making evident how each source contributes to this degree. Besides, connectivity is a notion which can be exploited in the task of dataset or endpoint selection.

To this end, this chapter summarizes the methods and metrics introduced in Tzitzikas et al. (2014, March) and Mountantonakis et al. (2016) for quantifying the connectivity of a warehouse, reports their implementation on real datasets, and discusses interesting and novel works that exploit them. What the authors call *metrics* could be also called *measures*, i.e. they should not be confused with distance functions. These metrics allow someone to get an overview of the contribution (to the warehouse) of each source (enabling the discrimination of the important from the non-important sources) and to quantify the value (benefit) of such a warehouse. In a nutshell, this chapter presents:

Related Content

Enterprise Tomography: Maintenance and Root-Cause-Analysis of Federated ERP in Enterprise Clouds
Jan Aalmink, Timo von der Dovenmühleand Jorge Marx Gómez (2013). *Advancing Information Management through Semantic Web Concepts and Ontologies (pp. 133-153).*
www.irma-international.org/chapter/enterprise-tomography-maintenance-root-cause/71853

OOPS!: A Pitfall-Based System for Ontology Diagnosis
María Poveda-Villalón, Asunción Gómez-Pérezand Mari Carmen Suárez-Figueroa (2018). *Innovations, Developments, and Applications of Semantic Web and Information Systems (pp. 120-148).*
www.irma-international.org/chapter/oops/196437

Using a Natural Language Understanding System to Generate Semantic Web Content
Akshay Java, Sergei Nirneburg, Marjorie McShane, Timothy Finin, Jesse Englishand Anupam Joshi (2007). *International Journal on Semantic Web and Information Systems (pp. 50-74).*
www.irma-international.org/article/using-natural-language-understanding-system/2842

Visual Ontology Modeling and the Semantic Web
Gilbert Paquette (2010). *Visual Knowledge Modeling for Semantic Web Technologies: Models and Ontologies (pp. 198-223).*
www.irma-international.org/chapter/visual-ontology-modeling-semantic-web/44932

The Berlin SPARQL Benchmark
Christian Bizerand Andreas Schultz (2011). *Semantic Services, Interoperability and Web Applications: Emerging Concepts (pp. 81-103).*
www.irma-international.org/chapter/berlin-sparql-benchmark/55042