Chapter 86
Quantifying the Connectivity of a Semantic Warehouse and Understanding Its Evolution Over Time

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. In this paper the authors describe the corresponding requirements and challenges, and they focus on the aspects of quality and value of the warehouse. For this reason they introduce various metrics (or measures) for quantifying its connectivity, and consequently its ability to answer complex queries. The authors demonstrate the behaviour of these metrics in the context of a real and operational semantic warehouse, as well as on synthetically produced warehouses. 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.

DOI: 10.4018/978-1-5225-5191-1.ch086

Consequently, these metrics can be used for advancing data/endpoint profiling and for this reason the authors use an extension of VoID (for making them publishable). Such descriptions can be exploited for dataset/endpoint selection in the context of federated search. In addition, the authors show how the metrics can be used for monitoring a semantic warehouse after each reconstruction reducing thereby the cost of quality checking, as well as for understanding its evolution over time.

1. 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 *warehouse*, offering thereby more complete and efficient browsing and query services (in comparison to those offered by the underlying sources). We use the term *Semantic Warehouse* (for short warehouse) to 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, we can distinguish *domain independent* warehouses, like the Sindice RDF search engine (Oren, et al., 2008) or the Semantic Web Search Engine (SWSE) (Hogan, et al., 2011), but also *domain specific*, like TaxonConcept[1] and the MarineTLO-based warehouse (Tzitzikas, et al., 2013, November). Domain specific semantic warehouses aim to serve particular needs, for particular communities of users, consequently their “quality” requirements are more strict. 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 the scope of such a warehouse, 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, how to keep the warehouse fresh, i.e., how to automate its construction or refreshing. In this paper we focus on the following questions:

- How to measure the value and quality (since this is important for e-science) of the warehouse?
- 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?

We have encountered these questions in the context of a real semantic warehouse for the *marine* domain which harmonizes and connects information from different sources of marine information[2]. Most past approaches have focused on the notion of conflicts (Michelfeit & Knap, 2012), and have not paid attention to *connectivity*. We use the term *connectivity* to express the degree up to which the contents of the semantic warehouse form a connected graph that can serve, ideally in a correct and complete way, the query requirements of the semantic warehouse, while making evident how each source contributes to that degree. Moreover connectivity is a notion which can be exploited in the task of dataset or endpoint selection.

To this end, in this paper we introduce and evaluate upon real and synthetic datasets several metrics for quantifying the connectivity of a warehouse. What we call *metrics* could be also called *measures*, i.e.