1
Introduction

1.1 Motivation

Statistical data are considered one of the major sources of information and are essential in many fields, such as in governmental, scientific and business domains. A statistical data set is comprised of a collection of observations made at some points across some logical space, such as a data set that describes life expectancy by region, age and time (Tennison 2012).

Statistical data are mostly stored in relational databases. The final data can be disseminated on the Web using different types of formats, such as HTML pages and PDF documents. These formats are human readable, which means that machines are not able to easily process the content of the data published.

In the process of analyzing statistical data, some requirements are essential to ensure that data are consumed in a simple but efficient way, such as: (i) the data should be published in a simple machine-readable format, to be reused and processed by automated tools; (ii) the data should be contextualized with other existing data to enrich the quality of the statistics.

The analysis of statistical data is never an easy task due to the volume of the data stored. Applications dealing with these types of data usually include Online Analytical Processing (OLAP), a set of tools and algorithms for querying large statistical databases. In OLAP, data are perceived as multidimensional structures known as *data cubes* (Etcheverry and Vaisman 2012a), which can be seen as a star schema view of the relational database. The main advantage of using such structures is based on the possibility of obtaining different perspectives of the data cubes according to the needs. That is the main reason why we represent statistical data as *data cubes* in this work.

In this context, the Linked Data Principles (Berners-Lee 2007) can be profitably applied to statistical data, in the sense that the principles offer a strategy to provide the missing semantics of the data. Intuitively, if followed, the Linked Data Principles include the data in a context, i.e., connect statistical data with

related data sources, creating a globally interconnected data space that enables a rich analysis of the data (Richard Cyganiak et al. 2011). One example is linking the data collected from a demographic region with the information about the region already existing in other databases, such as DBpedia (Auer et al. 2007).

To represent the data, the Linked Data Principles recommend using RDF (Resource Description Framework), a simple and flexible model which can be serialized using machine-readable formats in such a way that it can be mixed, exposed, and shared across different applications, thereby facilitating data interoperability (Manola and Miller 2013).

The Linked Data design principles focus on the use of URI-identified resources and their interlinkage, that is, they are concerned with publishing and retrieving data. In this context, it is intuitive to combine both Linked Data and REST Principles to provide a uniform interface for data cube access and manipulation that simplifies the overall system architecture (Stadtmüller and Harth 2012).

Due to its simplicity, the REST approach has inspired the design of several services on the Web and serves as an alternative to expose the data generated from relational databases through customized mapping files. The REST Principles, which take advantage of the architectural basis of the Web, such as the HTTP protocol, when properly applied, help to offer a uniform interface to consume both cube metadata and data cube observations.

1.2 Goal and Contributions

This work introduces the LDC Mediator, that offers an interface to consume linked data cubes and exports their metadata and their observations as RDF triples, generated on demand from the underlying data sources. The data cubes are accessed by HTTP methods using REST Principles. Therefore, this work takes advantage of both Linked Data and REST Principles to describe and consume linked data cubes in a simple but efficient way.

1.3 Dissertation Structure

This dissertation is structured as follows. Chapter 2 presents the basic concepts related to this work: the architecture of the Wide World Web, the Linked Data approach, the data cube representation and the REST approach. Chapter 3 shows related work, divided into two approaches: OLAP data approaches and the RESTful Linked Data approaches. Chapter 4 presents the architecture of the *OLAP2DataCube Catalogue On Demand*, a broader architecture to help describing and consuming statistical data, exposed as RDF triples, but stored in relational databases. The *LDC Mediator* (Linked Data Cubes Mediator), the component of this architecture that offers an interface to access the linked data cube descriptions, is the main focus of this dissertation and is presented and discussed in Chapter 5. Finally, Chapter 6 presents the conclusions, the limitations of this work and possible future work.