Chapter 2
Data Extraction, Transformation and Integration Guided by an Ontology

Chantal Reynaud
Université Paris-Sud, CNRS (LRI) & INRIA (Saclay – Île-de-France), France

Nathalie Pernelle
Université Paris-Sud, CNRS (LRI) & INRIA (Saclay – Île-de-France), France

Marie-Christine Rousset
LIG – Laboratoire d’Informatique de Grenoble, France

Brigitte Safar
Université Paris-Sud, CNRS (LRI) & INRIA (Saclay – Île-de-France), France

Fatiha Saïs
Université Paris-Sud, CNRS (LRI) & INRIA (Saclay – Île-de-France), France

ABSTRACT
This chapter deals with integration of XML heterogeneous information sources into a data warehouse with data defined in terms of a global abstract schema or ontology. The authors present an approach supporting the acquisition of data from a set of external sources available for an application of interest including data extraction, data transformation and data integration or reconciliation. The integration middleware that the authors propose extracts data from external XML sources which are relevant according to an RDFS+ ontology, transforms returned XML data into RDF facts conformed to the ontology and reconciles RDF data in order to resolve possible redundancies.

INTRODUCTION
A key factor for the success of the Semantic Web is to provide a unified, comprehensive and high-level access to voluminous and heterogeneous data. Such an access can be provided by an ontology in integrators supporting high-level queries and information interoperation. Our work takes place in the context of a data warehouse with data defined in terms of a global abstract schema or ontology.
We advocate an information integration approach supporting the acquisition of data from a set of external sources available for an application of interest. This problem is a central issue in several contexts, data warehousing, interoperate systems, multi-database systems, web information systems. Several steps are required for the acquisition of data from a variety of sources to a data warehouse based on an ontology (1) Data extraction: only data corresponding to descriptions in the ontology are relevant. (2) Data transformation: they must be defined in terms of the ontology and in the same format. (3) Data integration and reconciliation: the goal of this task is to resolve possible redundancies.

As a vast majority of sources rely on XML, an important goal is to facilitate the integration of heterogeneous XML data sources. Furthermore, most applications based on the Semantic Web technologies rely on RDF (McBride, 2004), OWL-DL (McGuinness & Van Harmelen, 2004) and SWRL (Horrocks et al., 2004). Solutions for data extraction, transformation and integration using these recent proposals must be favoured. Our work takes place in this setting. We propose an integration middleware which extracts data from external XML sources that are relevant according to a RDFS+ ontology (RDFS+ is based on RDFS (McBride, 2004)), transforms them into RDF facts conformed to the ontology, and reconciles redundant RDF data.

Our approach has been designed in the setting of the PICSEL3 project whose aim was to build an information server integrating external sources with a mediator-based architecture and data originated from external sources in a data warehouse. Answers to users’ queries should be delivered from the data warehouse. So data have to be passed from (XML) external sources to the (RDF) data warehouse and answers to queries collected from external sources have to be stored in the data warehouse. The proposed approach has to be totally integrated to the PICSEL mediator-based approach. It has to be simple and fast in order to deal with new sources and new content of integrated sources. Finally, it has to be generic, applicable to any XML information source relative to any application domain. In Figure 1 we present the software components designed in the setting of the project to integrate sources and data. This paper focuses on the description of the content of a source, the extraction and the integration of data (grey rectangles in Figure 1). The automatic generation of mappings is out of the scope of the paper.

The extraction and transformation steps rely on correspondences or mappings between local schemas of external sources and the ontology. In a previous work, we proposed techniques to automate the generation of these mappings (Reynaud & Safar, 2009). In this chapter, we present an approach which automates the construction of
Related Content

Locally Adaptive Techniques for Pattern Classification
www.igi-global.com/chapter/locally-adaptive-techniques-pattern-classification/10970?camid=4v1a

Biological Image Analysis via Matrix Approximation
www.igi-global.com/chapter/biological-image-analysis-via-matrix/10815?camid=4v1a

Segmenting the Mature Travel Market with Data Mining Tools
www.igi-global.com/chapter/segmenting-mature-travel-market-data/11056?camid=4v1a

Data Mining for the Chemical Process Industry
www.igi-global.com/chapter/data-mining-chemical-process-industry/10860?camid=4v1a