Chapter 11
Querying Conflicting Web Data Sources

Gilles Nachouki, Mohamed Quafafou, Omar Boucelma, and François-Marie Colonna

Abstract. Over the last twenty years, information integration has received considerable efforts from both industry and academia. Approaches to information integration developed so far can be categorized as follows: (1) first-generation approaches, that require the definition of a global schema and a semantic integration which should be performed upfront (before query execution); (2) second-generation approaches, well illustrated by the dataspace management concept, which promote a pay-as-you-go data integration. The first category has led to well known mediation approaches such as GAV (Global as View), LAV (Local as View), GLAV (Generalized Local As View), BAV (Both As View), and BGLAV (BYU Global-Local-as-View). Approaches pertaining to the second category are geared towards the development of dataspace management systems and are currently gaining a lot of attention. In this chapter we are interested in exploiting both types of approaches in querying conflicting data spread over multiple web sources. To this aim, first we show how an XML-based BGLAV approach can handle these conflicting data sources, then we describe how the same problem can be addressed by using the Multi Fusion Approach (MFA), an approach pertaining to second-generation techniques. Both BGLAV and MFA are illustrated in using genomic data sources accessible through the Web.

Gilles Nachouki
LINA-UMR CNRS 6241, Nantes University, France
e-mail: Gilles.Nachouki@univ-nantes.fr

Mohamed Quafafou · Omar Boucelma
LSIS-UMR CNRS 6168, Aix-Marseille University, France
e-mail: {mohamed.quafafou,omar.boucelma}@lsis.org

François-Marie Colonna
Institut Supérieur de l’Electronique et du Numérique, France
e-mail: francois-marie.colonna@isen.fr
11.1 Introduction

Over the past two decades, the database communities (both industry and academia), have intensively addressed data integration problems. A first-generation category of approaches and systems has been developed, and significant contributions has been made in various subtopics such as: source descriptions, schema mappings, query reformulation, and incomplete information modeling. Approaches pertaining to this first category require upfront semantic integration, that is a global/mediation schema needs to be supplied beforehand. As an example, in mediation approaches such as GAV (Global as View) [12], LAV (Local as View) [23], GLAV (Generalized Local As View) [11] or BAV (Both As View) [5], the mediator provides the user with a global schema and allows him/her to access heterogeneous data sources providing the illusion to access a single local database. The wrapper, another component of mediation systems, plays the role of an interface between the mediator and the data sources: it receives queries from a mediator and uses its own knowledge (source descriptions and mapping rules) in order to access the data sources.

As an example, in LAV the content of each data source is expressed in terms of a view over the global schema. Mapping rules associate a query over the global schema to each element of the local sources. On the other hand, in GAV each element of the global schema is expressed in terms of a view over the data sources. Each mapping rule associates a query over a local source to each element in the global schema.

The main component of first-generation data integration systems is the query rewriting module; it explores a set of mappings in order to rewrite queries, expressed upon the global schema, in terms of local sources’ schemas. The complexity of the query rewriting phase depends on how the global schema is defined. For example, GAV query rewriting is very simple since the elements in the global schema are defined in terms of the source schemas [22]. In this case, query rewriting simply consists in unfolding the definitions of the elements in the global schema. However, in this case, adding a new source to the data integration system is not trivial. The new source may indeed have an impact on the definition of various elements of the global schema, whose associated views need to be redefined. On the other hand, LAV query rewriting is not straightforward (exponential time complexity) and many rewriting algorithms have been developed; among them, we recall the Bucket, Inverse Rules and MiniCon algorithms [14], to cite a few. At the same time, however, the LAV approach favors the extensibility of the system: adding a new source simply means enriching the mapping with a new assertion, without other changes.

In general, query rewriting works well assuming that the schema of the local source is known a-priori and static. Unfortunately, such assumptions are not satisfied by data managed in several new processing environments, where data sources to be integrated are selected and combined on-demand. New second-generation data integration approaches have therefore been proposed. Dataspaces and Dataspaces Management Systems (DSMSs), described in Chapter 12, represent a significant