A Peer-to-Peer Service Supporting Data Quality: Design and Implementation Issues

Diego Milano, Monica Scannapieco and Tiziana Catarci

Dipartimento di Informatica e Sistemistica,
Universitá degli Studi di Roma "La Sapienza",
Via Salaria 113, Rome, Italy
{milano,monscan,catarci}@dis.uniroma1.it

Recent research has highlighted the importance of data quality issues in environments characterized by extensive data replication, such as Cooperative Information Systems (CISs). While high data quality is a strict requirement for CISs, the high degree of data replication that characterizes such systems can be exploited to improve the quality of data, as different copies of the same data may be compared in order to detect quality problems and possibly solve them.

The DaQuinCIS architecture [1] has been designed to manage data quality in cooperative contexts [2] and offers several quality-oriented services. Its core component, the *Data Quality Broker*, is in essence a data integration system that allows to access the best available quality data without having to know where such data are stored. It also allows diffusion of best quality data in the System, thus improving the overall quality of the CIS.

The Data Quality Broker is implemented as a peer-to-peer distributed service: each organization hosts a copy of the Data Quality Broker that interacts with other copies (see Figure 1, left side) playing both the roles of wrapper and mediator in the data integration architecture. Each copy of the Data Quality Broker is internally composed by four interacting modules (see Figure 1, right side).

The **Query Processor** performs query processing on the basis of a mapping specifically designed to exploit the presence of replicated data. To answer a query on the global schema, it submits appropriate data requests to multiple sources that provide overlapping sets of data. A record matching activity is performed on the global query result, and duplicated copies of the same data are compared in order to select or construct a best quality copy. Best quality copies are returned. Also, they are submitted to organizations having provided low quality copies of the same data.

The **Wrapper** translates the query from the language used by the broker to that of the specific data source. In this work the wrapper is a read-only module that accesses data and associated quality stored inside organizations without modifying them.

The **Transport Engine** is a communication facility that transfers queries and their results between the Query Processor module and data source wrappers. Another function it performs is the evaluation of the *availability* of data sources that are going to be queried for data. This feature is encapsulated into the

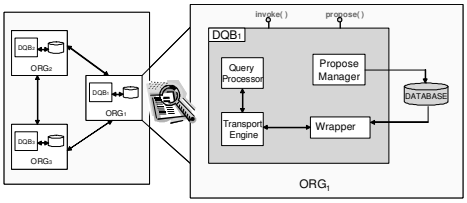

Figure 1. The Data Quality Broker as a P2P system and its internal architecture

Transport Engine as it can be easily implemented exploiting Transport Engine's communication capabilities.

The **Propose Manager** receives feedbacks sent to organizations in order to improve their data.

The choice of a P2P architecture is motivated by the need of being as less invasive as possible in introducing quality controls in a cooperative system. The P2P paradigm is able to support the cooperation without necessarily involving consistent re-engineering actions.

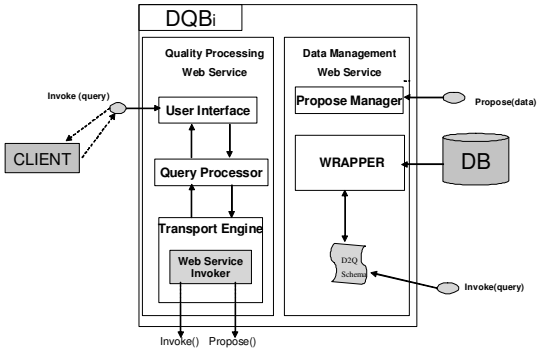

Figure 2. The Data Quality Broker system architecture

The Data Quality Broker system architecture is based on web services technologies, supported by the the J2EE 1.4 JAX-RPC API. Each copy of the Data Quality Broker is implemented by two web services, namely the Query Processing Web Service and the Data Manager Web Service as shown in Figure 2.

Some tests have been performed with two real data sets owned by Italian public administrations. The results we obtained show that the system is effective in improving the quality of data, with only a limited efficiency overhead.

References

1. The Data Quality in Cooperative Information Systems (DaQuinCIS) Project, `http://www.dis.uniroma1.it/~dq/` .
2. M. Scannapieco, A. Virgillito, M. Marchetti, M. Mecella, and R. Baldoni, *The DaQuinCIS architecture: a Platform for Exchanging and Improving Data Quality in Cooperative Information Systems*, Information Systems **29** (2004), no. 7, 551–582.