INTRODUCTION

Data quality is a complex concept defined by various dimensions such as accuracy, currency, completeness, and consistency (Wang & Strong, 1996). Recent research has highlighted the importance of data quality issues in various contexts. In particular, in some specific environments characterized by extensive data replication high quality of data is a strict requirement. Among such environments, this article focuses on Cooperative Information Systems.

Cooperative information systems (CISs) are all distributed and heterogeneous information systems that cooperate by sharing information, constraints, and goals (Mylopoulos & Papazoglou, 1997). Quality of data is a necessary requirement for a CIS. Indeed, a system in the CIS will not easily exchange data with another system without knowledge of the quality of data provided by the other system, thus resulting in a reduced cooperation. Also, when the quality of exchanged data is poor, there is a progressive deterioration of the overall data quality in the CIS. On the other hand, the high degree of data replication that characterizes a CIS can be exploited for improving data quality, as different copies of the same data may be compared in order to detect quality problems and possibly solve them.

In Scannapieco, Virgillito, Marchetti, Mecella, and Baldoni (2004) and Mecella et al. (2003), the DaQuinCIS architecture is described as an architecture managing data quality in cooperative contexts,
Measuring and Diffusing Data Quality in a Peer-to-Peer Architecture

in order to avoid the spread of low-quality data and to exploit data replication for the improvement of the overall quality of cooperative data.

In this article we will describe the design of a component of our system named as, quality factory. The quality factory has the purpose of evaluating quality of XML data sources of the cooperative system. While the need for such a component had been previously identified, this article first presents the design of the quality factory and proposes an overall methodology to evaluate the quality of XML data sources.

Quality values measured by the quality factory are used by the data quality broker. The data quality broker has two main functionalities: 1) quality brokering that allows users to select data in the CIS according to their quality; 2) quality improvement that diffuses best quality copies of data in the CIS.

As a further research contribution, this article will focus on the design and implementation features of the data quality broker as a Peer-to-Peer (P2P) system. More specifically, 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. While the functional specification of the data quality broker is not a contribution of this article, and has been presented in (Scannapieco et al., 2004; Mecella et al., 2003), its detailed design and implementation features as a P2P system are a novel contribution of this article. Moreover, we will present some results from tests made to prove the effectiveness and efficiency of our system. The data quality broker is implemented by a peer-to-peer architecture in order to be as less invasive as possible in introducing quality controls in a cooperative system. Indeed, cooperating organizations need to save their independency and autonomy requirements. Such requirements are well-guaranteed by the P2P paradigm which is able to support the cooperation without necessarily involving consistent re-engineering actions; in the section on Related Work, we will better detail this point, comparing our choice with a system that instead does not adopt a P2P architecture.

The rest of this article is organized as follows. The second section describes the main features of the quality factory and of the data quality broker. The third section presents the overall methodology and the fourth section details the architectural design of the quality factory, by focusing on the case of XML data sources. The fifth section describes the detailed design and implementation of the data quality broker as a peer-to-peer system, and each module of its component architecture. The set of performed experiments is described in the sixth section. Finally, related work and conclusions are presented in the seventh and eighth section respectively.

THE DATA QUALITY BROKER AND QUALITY FACTORY: GENERALITIES

In this section, we provide an overview of the main functionality of the data quality broker and we detail the interaction of such module with the quality factory, the design of which is provided in the third and fourth sections. The component architecture and implementation details of the data quality broker are instead described in the fifth section.

The Data Quality Broker Functionality

In the DaQuinCIS architecture, all cooperating organizations export their application data and quality data (i.e., data quality dimension values evaluated for the application data) according to a specific data model. The model for exporting data and quality data is referred to as Data and Data Quality (D²Q)