Foreword

The problem of integrating multiple information sources into a unified data store is currently one of the most important challenges in data management. Within the field of source integration, the problem of automatically generating an integrated description of the data sources is surely one of the most relevant. The significance of the issue can be best understood if one considers the huge number of information sources that an organization has to integrate. Indeed, it is even impossible to try to do all the work by hand. Like other important issues in data management, the problem of integrating multiple data sources into a unique global system has several facets, each of which represents, "per se", an interesting research problem, and comprises, for instance, that of recognizing, at the intensional level, similarities and dissimilarities among scheme objects, that of resolving representation mismatches among schemes, and that of deciding how to obtain an integrated data store out of a set of input sources and of a semantic description of their contents. The research and application relevance of such issues has attracted wide interest in the database community in recent years. And, as a consequence, several techniques have been presented in the literature attacking one side or another of this complex and multifarious problem. However, all the results presented in the past were somehow specific to some of the aspects underlying the general problem of data source integration and no comprehensive approach had ever actually been proposed. The thesis of Domenico Ursino presents a general semi-automatic approach for the construction and the management of Cooperative Information Systems, i.e. Information Systems resulting from the integration of several information sources. From a set of input database schemes describing the information content of multiple sources, the techniques developed in the thesis yield a structured, integrated, and consistent description of the information content, represented in a suitable Data Repository. The thesis also demonstrates how to use the repository for several tasks of data management based on the integrated representation. The proposed techniques are very interesting from several points of view. They are based on a controlled use of many fundamental fields of Computer Science, such as Mining and Learning, Knowledge Representation, Databases, etc. The approach presented in the thesis has been implemented in the prototype system "Database Intensional Knowledge Extractor" (DIKE), which

has been experimented in several interesting application domains. Besides its specific technical merits, which the reader will be able to appreciate by proceeding with reading this thesis, Domenico Ursino's approach to data source integration has the characteristics of including a uniform set of techniques for data integration, thus being the first comprehensive approach to attack this problem. For this reason we believe that Domenico Ursino's PhD thesis is an excellent piece of work. It provides a complete description of the state of the art of the field, clearly describes the novel approach, and nicely illustrates the applications. In this sense, it is a unique attempt to deal with all the issues concerning the automatic derivation of semantic properties from multiple sources, and the corresponding construction of the integrated data scheme. The approach is methodologically and scientifically correct, as testified by the numerous papers already published by the author and his colleagues in several prestigious conference proceedings and journals and we think that the thesis will be very useful both for researchers investigating in the area of data integration, and for practitioners working in the field of cooperative information systems. In our role as his PhD thesis advisors, we had the privilege of being able to follow the entire development of the body of research that brought Domenico Ursino to obtain the excellent results that this thesis describes. And here, by writing this brief preface, we have the privilege to testify the quality and the continuity of Domenico's commitment to scientific research, against all odds, during these years, that allowed him to fulfill the objective we had, together with him, fixed beforehand, when all this work started.

January 2002

Prof. Luigi Palopoli,
Università "Mediterranea" di Reggio Calabria

Prof. Domenico Saccà,
Università della Calabria

Preface

This book is my PhD thesis and presents the research work I did at the Dipartimento di Elettronica, Informatica e Sistemistica, Università degli Studi della Calabria, Cosenza, from 1996 to 1999, under the supervision of Luigi Palopoli and Domenico Saccà.

My research is based on the observation that, in the last decade, the development of new technologies for data acquisition and data storing has produced an enormous growth of information available electronically. A corresponding increase in the number of models and languages used to represent and manipulate data has taken place. These two factors have induced an increasing difficulty in handling data through traditional approaches. In particular, the exploitation of pre-existing and autonomous data resources (often based on very diverse models and systems) is nowadays recognized as a key issue in the area of data management. Cooperative Information Systems (CIS) and Data Warehouses (DW) have thus been designed as the necessary solutions providing friendly and flexible access to heterogeneous information sources, yet maintaining their operational autonomy.

In order to obtain an appropriate design of both CIS and DW, the schemes of involved databases are analyzed to identify similitudes, potential replications, or inconsistencies among data. In such system re-engineering problems, the design emphasis is on the integration of pre-existing information components, where a key problem is that of deriving relations holding among objects in pre-existing schemes [6]. Then, methodologies are needed to extract properties from schemes. Most interesting, in this context, are interscheme properties, that relate objects belonging to different schemes. Indeed, an appropriate exploitation of interscheme properties is crucial for a correct synthesis of global structured dictionaries, which we will refer to as data repositories. However, in reasoning about the intensional semantics of pre-existing databases, many useful properties are not explicitly encoded in database schemes, and so they cannot be immediately exploited.

Some papers (e.g., [32, 65, 126]) put into evidence the need for the adoption of formal languages to describe and manipulate intensional knowledge about data. In particular, [32] proposes a logic formalism largely based on Description Logics to express interscheme properties in CIS and DW.

When the number or the size of database schemes involved in the integration process is large and/or when the set of information resources changes quite frequently over time, manual design of CIS and DW can be very expensive and difficult. Therefore, the construction of semi-automatic integration tools appears to be necessary.

In this thesis, we illustrate a general approach to semi-automatically constructing and managing Cooperative Information Systems and Data Warehouses. The input to our method is the set of source database schemes constituting the base of the Cooperative Information Systems and the Data Warehouses. The output is a structured, integrated, and consistent description of information available in the Cooperative Information Systems or in the Data Warehouses and their properties in the form of a data repository. The data repository is used as the core structure of either the Mediator module of a Cooperative Information System or the reconciled level of a three-level Data Warehouse architecture.

The proposed approach is mainly based on the automatic derivation of properties holding among objects belonging to different input schemes. It consists of a number of steps: *(i)* the enrichment of scheme descriptions, obtained by the semi-automatic extraction of interscheme properties, i.e., terminological and structural properties between objects belonging to different schemes; *(ii)* the exploitation of derived interscheme properties for obtaining, in a data repository, an integrated and abstracted view of available data; *(iii)* the design of both a mediator-based Cooperative Information System and a three-level Data Warehouse having, as their core, the derived data repository. The techniques we have developed have been implemented in a prototype system called D.I.K.E. (Database Intensional Knowledge Extractor).

It is a pleasure to thank the people who have helped me most during this work. First of all, my gratitude goes to my advisors Luigi Palopoli and Domenico Saccà. I would like to thank Luigi not only for his support in my research activity, but also, more importantly, for being a true friend and a precious source of advice and suggestions. He provided the right environment for me to freely carry out my research and to fulfill my objectives. In particular, I owe Luigi special thanks for his support during all the phases of this thesis' submission to the LNCS series. I would like to thank Domenico for his constant push towards in-depth analysis of problems, that taught me to uncover their inner structure in order to find the way to attack them, which is the very nature of scientific research.

I owe very special thanks to my great friend Giorgio Terracina to whom I must express the whole of my gratitude since he helped me during the research activities and was a co-author of many papers, spending many hours together with me working towards the achievement of the results presented here. He has also been a true friend and provided precious support during some difficult moments of my life.

I would like to express my gratitude to my great friend Giampiero Dattilo, who has provided me with his constant support during many years. He has been a true friend to me, constantly helping me to face small and great difficulties of everyday life, even when this has meant sacrifices for him.

I wish to thank all the other people who have collaborated with me during my years of research, particularly Angela Bonifati, Larid Guga, Elisa Iezzi, Massimo La Camera, Fabio Lamberti, Francesco Locane, Alessandro Longo, Alfredo Pellicanò, Tiziana Pugliese, Salvatore Rotundo, Gregorio Sorrentino, Biagio Tramontana, Pasquale Viola.

I would like to thank all the people of the "Dipartimento di Elettronica, Informatica e Sistemistica", in particular the database group who have always been ready to talk about issues related to the thesis, particularly, Stefano Basta, Mario Cannataro, Mario Ettorre, Domenico Famularo, Sergio Flesca, Gianluigi Folino, Sergio Greco, Giovambattista Ianni, Nicola Leone, Elio Masciari, Clara Pizzuti, Luigi Pontieri, Pasquale Rullo, Francesco Scarcello, Giandomenico Spezzano, Domenico Talia, Ester Zumpano.

For about a year now, I have been working with the "Dipartimento di Informatica, Matematica, Elettronica e Trasporti" - Università Mediterranea di Reggio Calabria, and I would like to thank my new colleagues Francesco Buccafurri, Gianluca Lax, Domenico Rosaci, and Giuseppe Maria Luigi Sarnè for helping me in reviewing the last version of this manuscript.

I would like to thank Salvatore Capria, Giovanni Costabile, and Francesco De Marte for their help on several occasions.

I wish to thank Marco Cadoli for being the first to suggest that I submit my PhD thesis to LNCS.

Finally, I would like to gratefully acknowledge the support of the Italian Information System Authority for Public Administration (AIPA). They kindly provided the schemes of the Italian Central Governmental Office databases and the technical support to go with them.

January 2002 Domenico Ursino

List of Figures

List of Tables