Preface

Over the past 20 years, the role of the database, and especially of database techniques, has changed dramatically. We have moved from a world in which an enterprise or organization had one central, relatively closed database for all record-keeping to a Web-dominated world in which many different databases and other sources of structured information must interact and interoperate, ideally in a way that gives users a fully integrated view of the world.

This book focuses on that latter world. It shows how database ideas have been broadened and deepened to accommodate external sources of information, to handle the distributed aspects of the Web and the issues that arise from mutual information sharing, and especially to deal with heterogeneity and uncertainty. We see such topics as a natural extension of the topics covered in a typical university-level database course. Hence, the book is primarily intended as a text for an advanced undergraduate or graduate course that follows the undergraduate database class present in many curricula. Additionally, the book is suitable as a reference and tutorial for researchers and practitioners in the database and data integration fields.

The book is divided into three main parts. Part I builds upon the topics covered in a database course and focuses on the foundational techniques used in data integration: techniques for manipulating query expressions, for describing data sources, for finding matches across heterogeneous data and schemas, for manipulating schemas, for answering queries, for extracting data from the Web, and for warehousing and storing integrated data. Part II focuses on extended data representations that capture properties not present in the standard relational data model: hierarchy (XML), ontological constructs from knowledge representation, uncertainty, and data provenance. Part III looks at novel architectures for addressing specific integration problems, including diverse Web sources, keyword queries over structured data that have not been fully integrated, peer-to-peer methods of data integration, and collaboration. We conclude with a brief overview of promising future directions for the field.

A range of supplementary, Web-based material is available for the book, including problem sets, selected solutions, and lecture slides.

Acknowledgments

Many people gave us helpful feedback on earlier versions of the book. We are extremely grateful to Jan Chomicki and Helena Galhardas, who were among the first to use the book in their courses and to give us feedback. Others who gave us fantastic feedback included Mike Cafarella, Debby Wallach, Neil Conway, Phil Zeyliger, Joe Hellerstein, Marie-Christine Rousset, Natasha Noy, Jayant Madhavan, Karan Mangla, Phil Bernstein, Anish Das Sarma, Luna Dong, Rajeev Alur, Chris Olston, Val Tannen, Grigoris Karvounarakis, William Cohen, Prasenjit Mitra, Lise Getoor, and Fei Wu. The students in the CS 784 course on advanced data management

at Wisconsin and the CIS 550 course on database and information systems at Penn have read various chapters of the book over several semesters and provided great comments. We especially thank Kent Chen, Fei Du, Adel Ardalan, and KwangHyun Park for their help.

Portions of the content in this book were derived from the authors' work in collaboration with many colleagues and students. We would like to acknowledge and thank them for their contributions.

Finally, we are extremely grateful to our families, who sacrificed a good deal of quality time so that we could complete this manuscript.

AnHai Doan
Alon Halevy
Zachary Ives