Preface

The focus of this book is effective databases for text and document management inclusive of new and enhanced techniques, methods, theories and practices. The research contained in these chapters is of particular significance to researchers and practitioners alike because of the rapid pace at which the Internet and related technologies are changing our world. Already there is a vast amount of data stored in local databases and Web pages (HTML, DHTML, XML and other markup language documents). In order to take advantage of this wealth of knowledge, we need to develop effective ways of extracting, retrieving and managing the data. In addition, advances in both database and Web technologies require innovative ways of dealing with data in terms of syntactic and semantic representation, integrity, consistency, performance and security.

One of the objectives of this book is to disseminate research that is based on existing Web and database technologies for improved information extraction and retrieval capabilities. Another important objective is the compilation of international efforts in database systems, and text and document management in order to share the innovation and research advances being done at a global level.

The book is organized into four sections, each of which contains chapters that focus on similar research in the database and Web technology areas. In the section entitled, *Information Extraction and Retrieval in Web-Based Systems*, Web and database theories, methods and technologies are shown to be efficient at extracting and retrieving information from Web-based documents. In the first chapter, "System of Information Retrieval in XML Documents," Saliha Smadhi introduces a process for retrieving relevant information from XML documents. Smadhi's approach supports keyword-based searching, and ranks the retrieval of information based on the similarity with the user's query. In "Information Extraction from Free-Text Business Documents," Witold Abramowicz and Jakub Piskorski investigate the applicability of information extraction techniques to free-text documents typically retrieved from Web-based systems. They also demonstrate the indexing potential of lightweight linguistic text processing techniques in order to process large amounts of textual data.

In the next chapter, "Interactive Indexing of Documents with a Multilingual Thesaurus," Ulrich Schiel and Ianna M.S.F. de Sousa present a method for semi-automatic indexing of electronic documents and construction of a multilingual thesaurus. This method can be used for query formulation and information retrieval. Then in the next chapter, "Managing Document Taxonomies in Relational Databases," Ido Millet addresses the challenge of applying relational technologies in managing taxonomies used to classify documents, knowledge and websites into topic hierarchies. Millet explains how denormalization of the data model facilitates data retrieval from these topic hierarchies. Millet also describes the use of database triggers to solving data maintenance difficulties once the data model has been denormalized.

Yangjun Chen and Gerald Huck, in "Building Signature-Trees on Path Signatures in Document Databases," introduce PDOM (persistent DOM) to accommodate documents as permanent object sets. They propose a new indexing technique in combination with signature-trees to accelerate the evaluation of path-oriented queries against document object sets and to expedite scanning of signatures stored in a physical file. In the chapter, "Keyword-Based Queries of Web Databases," Altigran S. da Silva, Pável Calado, Rodrigo C. Vieira, Alberto H.F. Laender and Berthier A. Ribeiro-Neto describe the use of keyword-based querying as a suitable alternative to the use of Web interfaces based on multiple forms. They show how to rank the possible large number of answers returned by a query according to relevant criteria and typically done by Web search engines. Virpi Lyytikäinen, Pasi Tiitinen and Airi Salminen, in "Unifying Access to Heterogeneous Document Databases Through Contextual Metadata," introduce a method for collecting contextual metadata and representing metadata to users via graphical models. The authors demonstrate their proposed solution by a case study whereby information is retrieved from European, distributed database systems.

In the next section entitled, *Data Management and Web Technologies*, research efforts in data management and Web technologies are discussed. In the first chapter, "Database Management Issues in the Web Environment," J.F. Aldana Montes, A.C. Gómez Lora, N. Moreno Vergara and M.M. Roldán García address relevant issues in Web technology, including semi-structured data and XML, data integrity, query optimization issues and data integration issues. In the next chapter, "Applying JAVA-Triggers for X-Link Management in the Industrial Framework," Abraham Alvarez and Y. Amghar provide a generic relationship validation mechanism by combining XLL (X-link and X-pointer) specification for integrity management and Java-triggers as an alert mechanism.

The third section is entitled, *Advances in Database and Supporting Technologies.* This section encompasses research in relational and object databases, and it also presents ongoing research in related technologies. In this section's first chapter, "Metrics for Data Warehouse Quality," Manuel Serrano, Coral Calero and Mario Piattini propose a set of metrics that has been formally and empirically validated for assessing the quality of data warehouses. The overall objective of their research is to provide a practical means of assessing alternative data warehouse designs. R. Chbeir, Y. Amghar and A. Flory identify the importance of new management methods in image retrieval in their chapter, "Novel Indexing Method of Relations Between Salient Objects." The authors propose a novel method for identifying and indexing several types of relations between salient objects. Spatial relations are used to show how the authors' method can provide high expressive power to relations when compared to traditional methods.

In the next chapter, "A Taxonomy for Object-Relational Queries," David Taniar, Johanna Wenny Rahayu and Prakash Gaurav Srivastava classify object-relational queries into REF, aggregate and inheritance queries. The authors have done this in order to provide an understanding of the full capability of object-relational query language in terms of query processing and optimization. Aphrodite Tsalgatidou and Mara Nikolaidou describe a criteria set for selecting appropriate Business Process Modeling Tools

(BPMTs) and Workflow Management Systems (WFMSs) in "Re-Engineering and Automation of Business Processes: Criteria for Selecting Supporting Tools." This criteria set provides management and engineering support for selecting a toolset that would allow them to successfully manage the business process transformation. In the last chapter of this section, "Active Rules and Active Databases: Concepts and Applications," Juan M. Ale and Mauricio Minuto Espil analyze concepts related to active rules and active databases. In particular, they focus on database triggers using the SQL-1999 standard committee's point of view. They also discuss the interaction between active rules and declarative database constraints from both static and dynamic perspectives.

The final section of the book is entitled, *Advances in Relational Database Theory, Methods and Practices.* This section includes research efforts focused on advancements in relational database theory, methods and practices. In the chapter, "On the Computation of Recursion in Relational Databases," Yangjun Chen presents an encoding method to support the efficient computation of recursion. A linear time algorithm has also been devised to identify a sequence of reachable trees covering all the edges of a directed acyclic graph. Together, the encoding method and algorithm allow for the computation of recursion. The author proposes that this is especially suitable for a relational database environment. Robert A. Schultz, in the chapter "Understanding Functional Dependency," examines whether functional dependency in a database system can be considered solely on an extensional basis in terms of patterns of data repetition. He illustrates the mix of both intentional and extensional elements of functional dependency, as found in popular textbook definitions.

In the next chapter, "Dealing with Relationship Cardinality Constraints in Relational Database Design," Dolores Cuadra Fernández, Paloma Martínez Fernández and Elena Castro Galán propose to clarify the meaning of the features of conceptual data models. They describe the disagreements between main conceptual models, the confusion in the use of their constructs and open problems associated with these models. The authors provide solutions in the clarification of the relationship construct and to extend the cardinality constraint concept in ternary relationships. In the final chapter, "Repairing and Querying Inconsistent Databases," Gianluigi Greco, Sergio Greco and Ester Zumpano discuss the integration of knowledge from multiple data sources and its importance in constructing integrated systems. The authors illustrate techniques for repairing and querying databases that are inconsistent in terms of data integrity constraints.

In summary, this book offers a breadth of knowledge in database and Web technologies, primarily as they relate to the extraction retrieval, and management of text documents. The authors have provided insight into theory, methods, technologies and practices that are sure to be of great value to both researchers and practitioners in terms of effective databases for text and document management.