
**INFORMATION BROKERING ACROSS
HETEROGENEOUS DIGITAL DATA**
A Metadata-based Approach

The Kluwer International Series on **ADVANCES IN DATABASE SYSTEMS**

Series Editor
Ahmed K. Elmagarmid

*Purdue University
West Lafayette, IN 47907*

Other books in the Series:

DATA DISSEMINATION IN WIRELESS COMPUTING ENVIRONMENTS, *Kian-Lee Tan and Beng Chin Ooi*; ISBN: 0-7923-7866-0

MIDDLEWARE NETWORKS: Concept, Design and Deployment of Internet Infrastructure, *Michah Lerner, George Vanecek, Nino Vidovic, Dad Vrsalovic*; ISBN: 0-7923-7840-7

ADVANCED DATABASE INDEXING, *Yannis Manolopoulos, Yannis Theodoridis, Vassilis J. Tsotras*; ISBN: 0-7923-7716-8

MULTILEVEL SECURE TRANSACTION PROCESSING, *Vijay Atluri, Sushil Jajodia, Binto George* ISBN: 0-7923-7702-8

FUZZY LOGIC IN DATA MODELING, *Guoqing Chen* ISBN: 0-7923-8253-6

INTERCONNECTING HETEROGENEOUS INFORMATION SYSTEMS, *Athman Bouguettaya, Boualem Benatallah, Ahmed Elmagarmid* ISBN: 0-7923-8216-1

FOUNDATIONS OF KNOWLEDGE SYSTEMS: With Applications to Databases and Agents, *Gerd Wagner* ISBN: 0-7923-8212-9

DATABASE RECOVERY, *Vijay Kumar, Sang H. Son* ISBN: 0-7923-8192-0

PARALLEL, OBJECT-ORIENTED, AND ACTIVE KNOWLEDGE BASE SYSTEMS, *Ioannis Vlahavas, Nick Bassiliades* ISBN: 0-7923-8117-3

DATA MANAGEMENT FOR MOBILE COMPUTING, *Evaggelia Pitoura, George Samaras* ISBN: 0-7923-8053-3

MINING VERY LARGE DATABASES WITH PARALLEL PROCESSING, *Alex A. Freitas, Simon H. Lavington* ISBN: 0-7923-8048-7

INDEXING TECHNIQUES FOR ADVANCED DATABASE SYSTEMS, *Elisa Bertino, Beng Chin Ooi, Ron Sacks-Davis, Kian-Lee Tan, Justin Zobel, Boris Shidlovsky, Barbara Catania* ISBN: 0-7923-9985-4

INDEX DATA STRUCTURES IN OBJECT-ORIENTED DATABASES, *Thomas A. Mueck, Martin L. Polaschek* ISBN: 0-7923-9971-4

DATABASE ISSUES IN GEOGRAPHIC INFORMATION SYSTEMS, *Nabil R. Adam, Aryya Gangopadhyay* ISBN: 0-7923-9924-2

VIDEO DATABASE SYSTEMS: Issues, Products, and Applications, *Ahmed K. Elmagarmid, Haitao Jiang, Abdelsalam A. Helal, Anupam Joshi, Magdy Ahmed* ISBN: 0-7923-9872-6

REPLICATION TECHNIQUES IN DISTRIBUTED SYSTEMS, *Abdelsalam A. Helal, Abdelsalam A. Heddaya, Bharat B. Bhargava* ISBN: 0-7923-9800-9

INFORMATION BROKERING ACROSS HETEROGENEOUS DIGITAL DATA

A Metadata-based Approach

by

VIPUL KASHYAP

Member of Technical Staff

Micro-electronics and Computer Technology Corporation (MCC)

3500, W. Balcones Center Drive

Austin, TX 78759

Current address:

Research Scientist

Applied Research, Telcordia Technologies

MCC-1G332R, 445 South Street

Morristown, NJ 07960

AMIT SHETH

Director, Large Scale Distributed Information Systems Laboratory

Department of Computer Science

The University of Georgia

Athens, GA 30602

KLUWER ACADEMIC PUBLISHERS

New York / Boston / Dordrecht / London / Moscow

eBook ISBN: 0-306-47028-4
Print ISBN: 0-792-37883-0

©2002 Kluwer Academic Publishers
New York, Boston, Dordrecht, London, Moscow

Print ©2000 Kluwer Academic Publishers
Boston

All rights reserved

No part of this eBook may be reproduced or transmitted in any form or by any means, electronic, mechanical, recording, or otherwise, without written consent from the Publisher

Created in the United States of America

Visit Kluwer Online at: <http://kluweronline.com>
and Kluwer's eBookstore at: <http://ebooks.kluweronline.com>

*To my parents, Yogendra and
Vibha, whose blessings and
encouragement have played
an active life-long role in my
work, and my wife Nitu for
being patient with me.*

-Vipul

*To the members of the
InfoHarness and InfoQuilt
team at the LSDIS Lab, who
helped us in thinking about
metadata and semantics.*

-Amit and Vipul

Contents

List of Figures	xi
List of Tables	xiii
Preface	xv
Acknowledgments	xvii
Foreword by Michael Huhns	xix
1. INTRODUCTION	1
1. Information Overload	2
1.1 Heterogeneity	3
1.2 Globalization	4
2. Information Brokering: Handling the Information Overload	6
2.1 Information Brokering: Stakeholder and Beneficiaries	7
2.2 Levels of Information Brokering	8
2.3 The Role of Facilitators on the GII	13
2.4 Information Brokering: Relation to other Approaches	14
3. Book Organization	15
2. METADATA AND ONTOLOGIES	17
1. Perspectives on Metadata Management	18
1.1 The Application Scenarios Perspective	18
1.2 The Information Content Perspective	19
2. Language and Vocabulary Issues for Metadata Construction	22
2.1 Language for Metadata Representation	22
2.2 Vocabulary for Metadata Expressions	23
2.3 Constructing Intensional Metadata Descriptions	24
2.4 Design and Use of Ontologies	25
3. Summary	28
3. METADATA-BASED ARCHITECTURES	29
1. An (abstract) Metadata-based Architecture	29
1.1 The Vocabulary Brokering Component	31
1.2 The Metadata Brokering Component	33
2. Properties of Information Brokering Architectures	38

2.1	Issues of Scalability	39
2.2	Issues of Extensibility	41
2.3	Issues of Adaptability	42
3.	Architectural Evolution and Properties	43
3.1	Federated Multidatabase Systems	43
3.2	Mediator-based Systems	46
3.3	Agent-based Brokering Systems	48
4.	Summary	50
4.	METADATA-BASED BROKERING FOR DIGITAL DATA	51
1.	A Multimedia Information Request	51
2.	The InfoHarness System	54
2.1	Metadata-based Encapsulation and Brokering	54
2.2	The InfoHarness System Architecture	57
3.	Issues of Metadata and Architecture in the MIDAS System	63
3.1	Role of Metadata in MIDAS	64
3.2	The MIDAS System Architecture	66
3.3	Properties of the MIDAS System Architecture	72
4.	The InfoSleuth Text Agent	75
4.1	Metadata-based View of the Information Space	75
4.2	Mapping Domain Specific Metadata to Textual Data	76
4.3	Translating Queries into Topic Expressions	78
5.	Metadata-based Correlation in the MIDAS system	80
5.1	Scalability v/s Extensibility: Space/Time Trade-Offs	80
5.2	Scalability: The Stepping Stone to Adaptability	83
6.	Summary: Metadata as Schema for Digital Data	86
5.	CAPTURING INFORMATION CONTENT IN STRUCTURED DATA	89
1.	Schematic Heterogeneities across Multiple Databases	90
1.1	Domain Definition Incompatibility	91
1.2	Entity Definition Incompatibility	94
1.3	Data Value Incompatibility	97
1.4	Abstraction Level Incompatibility	98
1.5	Schematic Discrepancies	100
2.	Capturing the Information Content of Database Objects	102
2.1	Semantic Proximity: Capturing Information Content	103
2.2	C-Contexts: A Partial Representation	108
2.3	Association of Mappings with Contexts: An Algebra	116
2.4	Advantages of Context Representation	120
3.	Summary	123
6.	THE INFOSLEUTH SYSTEM	129
1.	The InfoSleuth Agent-based Architecture	130
2.	Metadata Brokering in InfoSleuth	132
3.	InfoSleuth: A Summary	135
7.	VOCABULARY BROKERING IN THE OBSERVER SYSTEM	137
1.	Architecture of OBSERVER	138

1.1	Architecture of the Metadata System	139
1.2	The Inter-Ontologies Relationships Manager (IRM)	143
1.3	Properties of the OBSERVER Architecture	146
2.	Vocabulary Brokering by the Query Processor	148
2.1	Semantics Preserving Vocabulary Brokering	149
2.2	Vocabulary Brokering with Loss of Information	157
3.	OBSERVER: A Summary	167
8.	AN ILLUSTRATIVE EXAMPLE	181
1.	Ontologies and Construction of Metadata	181
2.	Vocabulary Brokering	183
3.	Metadata Brokering	184
4.	Summary	187
9.	RELATED WORK	189
1.	The SIMS Project	189
2.	The TSIMMIS Project	191
3.	The Information Manifold Project	193
4.	The KMed Project	194
5.	The Conceptual Indexing/Retrieval Project	196
6.	HERMES: A Heterogeneous Reasoning and Mediator System	197
7.	InfoScopes: Multimedia Information Systems	199
8.	The Context Interchange Network Project	201
9.	A Comparison of Brokering Systems	203
9.1	Level of Information Brokering	203
9.2	Metadata: Types, Languages and Computation	204
9.3	Architectural Properties	206
10.	Summary	208
10.	CONCLUSION	209
	References	213
	Index	221

List of Figures

1.1	A Taxonomy of Heterogeneity and Interoperability	2
1.2	Dimensions of Information Brokering	6
1.3	Information Brokering: A Stakeholder Perspective	8
2.1	Examples of Generalization and Aggregation Hierarchies for Ontology Construction	27
2.2	The Bibliographic Data Ontology (DARPA Knowledge Sharing Effort)	27
2.3	The WordNet 1.5 Thesaurus	28
3.1	A High Level View of the Architecture	30
3.2	The Vocabulary Brokering Level of the Architecture	31
3.3	The Metadata System	34
3.4	Different Possibilities for a Metadata Repository	37
3.5	Federated Architecture for Multidatabase Interoperability	44
3.6	A Standard Mediator Architecture	47
3.7	Agent-based Information Brokering	49
4.1	The Domain Ontology for the Multimedia Query	52
4.2	Metadata Brokering for Multimedia Information Requests	53
4.3	Logical structuring of the Information Space	56
4.4	Attribute Based Access in InfoHarness	57
4.5	The InfoHarness System Architecture	58
4.6	Metadata Extraction for a C Program	59
4.7	Architecture of the MIDAS System	67
4.8	Extraction of Land Cover Information in MIDAS	70
4.9	Parameterized Routines as Procedural Fields in the Metadata Repository	71
4.10	Vocabulary Subset Supported by the Text Agent	76
4.11	Mapping Domain Specific Metadata to Media Specific Metadata	77

4.12	Translating Information Requests into Information Retrieval Operations	78
4.13	Hierarchies Describing a Domain Vocabulary	84
5.1	Schematic Heterogeneities Across Data in Multiple Databases	90
5.2	Heterogeneities Arising Out of Domain Incompatibility	91
5.3	Heterogeneities Arising Out of Incompatible Entity Descriptions	94
5.4	Heterogeneities Arising Out of Inconsistencies in Data Values	97
5.5	Heterogeneities Arising Out of Differing Levels of Abstraction	99
5.6	Heterogeneities Arising Out of Schematic Discrepancies	100
5.7	Capturing Information Content Using Semantic Proximity	103
5.8	Attribute and Object Domains in a Database	106
5.9	A Classification of Predicates and Concepts	113
5.10	Association Between Global and Database Objects	119
5.11	Information Focusing Based on Inferences on C-contexts	123
5.12	Incorporating Constraints from the Query	123
6.1	InfoSleuth's Agent Based Architecture	130
6.2	The Competitive Intelligence Ontology	132
7.1	Architecture of the OBSERVER System	138
7.2	Architecture of the OBSERVER Metadata System	140
7.3	Integration of Two Component Ontologies	158
7.4	Integration of the WN and Stanford-I Ontologies	161
7.5	A Composite Measure for Loss of Information	162
7.A.1	WN: A Subset of the WordNet 1.5 Ontology	169
7.A.2	Stanford-I: A Subset of the Bibliographic Data Ontology	170
7.A.3	Stanford-II: A Subset of the Bibliographic Data Ontology	172
7.A.4	The LSDIS Ontology	173
8.1	User Ontology: A Subset of WordNet 1.5	182
9.1	The Architecture of the SIMS System	190
9.2	The Architecture of the TSIMMIS System	192
9.3	Architecture of the Information Manifold System	193
9.4	Architecture of the KMed System	195
9.5	The HERMES System Architecture	198
9.6	A Four Layered Data Model	199
9.7	Infoscopes Architecture	200
9.8	The Context Interchange Network Architecture	202

List of Tables

2.1	Metadata for Digital Media	21
4.1	Metadata Types in the InfoHarness System	57
4.2	Metadata Storage Using the Attribute-Value Approach	61
4.3	Geo-Spatial Metadata Stored Using the Entity-Attribute Approach	63
4.4	Newsgroup Metadata Stored Using the Entity-Attribute Approach	63
4.5	Additional Metadata Types in the MIDAS System	64
4.6	Storage of Pre-computed Metadata in the MIDAS System	69
4.7	Parameterized Routines Stored in the Metadata Repository	71
4.8	Querying Metadata at Different Levels of Abstraction	85
5.1	Mapping Between Marks and Grades	93
7.1	Details of Ontologies and Underlying Repositories	144
9.1	A Comparison of Various Brokering Systems	205
9.2	A Comparison of Brokering System Properties	206

Preface

Information intermediation is the foundation stone of some of the most successful Internet companies, and is perhaps second only to the Internet Infrastructure companies. On the heels of information integration and interoperability, this book on information brokering discusses the next step in information interoperability and integration.

The emerging internet economy based on burgeoning B2B and B2C trading, will soon demand semantics-based information intermediation for its feasibility and success. Even as we speak, new B2B ventures are involved in the “rationalization” of new vertical markets and construction of domain specific product catalogs. In this book we provide approaches for re-use of existing vocabularies and domain ontologies as a basis for this rationalization and provide a framework based on inter-ontology interoperation. Infrastructural trade-offs that identify optimizations in performance and scalability of web sites will soon give way to information based trade-offs as alternate rationalization schemes come into play, and the necessity of interoperating across these schemes is realized.

The complex issues of intermediation as conceived by information brokering involve the following:

- Construction of information content descriptions that resolve information “impedance” or mismatch between what is required by the information consumer and presented by the information provider.
- Using “information brokering” to provide intermediation to resolve impedance at multiple levels. Architectures for brokering/intermediation should provide a framework for resolution of impedance at different levels, and may be multi-level in nature. New paradigms such as agent based architectures need to be evaluated for their applicability to information intermediation.

- Exploration of the critical role of Metadata. Information content descriptions may be at different levels, such as content-independent, representation dependent, representation independent or domain specific. The descriptions enable resolution of impedances at various levels. Domain specific metadata descriptions are crucial for intermediation based on the semantics (meaning and use) of the information content.
- Using ontology or vocabulary to construct metadata descriptions for aligning the “world views” of consumers and providers. Techniques for re-using existing vocabularies/ontologies and interoperation across them need to be developed to perform this type of intermediation.
- Characterizing loss of information that inevitably arises in semantics-based intermediation due to differences in world views between the consumers and providers. Information-based trade-offs need to be characterized and computed to provide effective semantics-based intermediation.

This book’s intended readers are researchers, software architects and CTOs, and advanced product developers dealing with information intermediation issues in the context of e-commerce (B2B and B2C), information technology professionals in various vertical markets (e.g., geo-spatial information, medicine, auto), and practically all librarians at higher-education, technical institutions and universities.

VIPUL KASHYAP AND AMIT SHETH

Acknowledgments

Amit Sheth and I have thoroughly enjoyed intense discussion sessions with each other which we often had in the course of working on this book. Many new research directions and possibilities emerged from those discussions and have played a critical role in enhancing and adding value to this book. This was supplemented by our interactions with members of the InfoHarness and InfoQuilt projects, that helped us in our thinking about metadata and semantics.

Special thanks goes to Eduardo Mena, the heated discussions with whom resulted in the OBSERVER system. Mention needs to be made of the members of the InfoSleuth project team that I worked with. Dr. Marek Rusinkiewicz officially my manager at Micro-electronics and Computer Technology Corporation (MCC), has been a great source of inspiration and help in my research endeavors. I am also indebted to my managers at Telcordia, Sid Dalal and Gardner Patton for their encouragement and advice. Gardner's comments on the draft manuscript were very useful and helped in improving the readability and presentation of content in this book.

Finally, I am grateful to my wife, Nitu, who was gracious enough to donate her time (which we would have otherwise spent together) towards the editing and proof reading of this book.

Foreword

Computing is fast becoming ubiquitous and pervasive. It is *ubiquitous* because computing power and access to the Internet is being made available everywhere; it is *pervasive* because computing is being embedded in the very fabric of our environment. Xerox Corp. has recently coined the phrase “smart matter” to capture the idea of computations occurring within formerly passive objects and substances. For example, our houses, our furniture, and our clothes will contain computers that will enable our surroundings to adapt to our preferences and needs. New visions of interactivity portend that scientific, commercial, educational, and industrial enterprises will be linked, and human spheres previously untouched by computing and information technology, such as our personal, recreational, and community life, will be affected.

However, when there is information everywhere and all manner of things are interconnected, there arise the problems of information overload and misunderstandings. Dogbert, from Scott Adams’ *Dilbert*, describes the situation as “Information is gushing toward your brain like a fire hose aimed at a teacup”. This book analyzes the problems of information overload and misunderstandings and provides a solution: a way for all of the different devices, components, and computers to understand each other, so that they will be able to work together effectively and efficiently. This is a powerful and important advance.

Dr. Vipul Kashyap and Professor Amit Sheth have long been leaders in the area of information system semantics and are widely known as two of the area’s most dedicated, productive, and insightful researchers. Together in this book they have crafted a coherent vision of the single most important element of a distributed heterogeneous information system: the information broker. Previously, Kashyap has been the architect of a broker-based multiagent system for cooperative information access. Sheth has been an innovator of metadata-based approaches to the integration of heterogeneous semantics for database systems and workflow systems. Together, their expertise is complementary and forms the unique perspective of this book.

The essential agent-based architecture that they describe and analyze is becoming canonical. Agents are used to represent users, resources, middleware, security, execution engines, ontologies, and brokering. As the technology advances, we can expect such specialized agents to be used as standardized building blocks for information systems. Two trends lend credence to such a prediction.

First, software systems in general are being constructed with larger components, such as ActiveX and JavaBeans, which are becoming closer to being agents themselves. They have more functionality than simple objects, respond to events autonomously, and, most importantly, respond to system builders at development-time, as well as to events at run-time.

Second, there is a move toward more cooperative information systems, in which the architecture itself plays an important role in the effectiveness of the system, as opposed to traditional software systems where effectiveness depends on the quality of the individual components. These are the architectures of standardized agents that Kashyap and Sheth elucidate. Architectures based on standardized agent types should be easier to develop, understand, and use. Perhaps most important of all, these architectures will make it easier for separately developed information systems to interoperate.

Among the reasons why agents are attractive, there are two main ones of interest here. One, agents enable the construction of modular systems from heterogeneous pieces, potentially created by any number of vendors. Two, the agents themselves embody diverse knowledge, reasoning approaches, and perspectives. This diversity is sometimes essential, because the agents represent people or business interests that have different goals and motivations. Diversity can sometimes be added in by design: it can make an agent system more robust by enabling a variety of viewpoints to be represented and exploited.

However, agents are typically complex pieces of software, so the question arises whether a set of different agents would unnecessarily add to a system's complexity. The more kinds of agents there are, the harder it might be to build and maintain them. Fortunately, this turns out to be a false concern. The agents have to be diverse in content, e.g., knowledge, reasoning techniques, and interaction protocols, but not in the form in which that content is realized, e.g., the language or toolkit with which they are constructed. Problems arise through unnecessary heterogeneity in construction; the cost of necessary heterogeneity in content is more than recovered through the flexibility it offers.

In summary, the results in this book are applicable not only to the huge amount of information available globally over the World-Wide Web, but also to the diverse information soon to be available locally over household, automobile, and environment networks. I am excited by the possibilities for new applications and uses for information that are engendered by this book, as well

as by the challenges that remain. This book provides a solid foundation for advances in distributed heterogeneous information systems.

Professor Michael N. Huhns,
Director, Center for Information Technology,
University of South Carolina, Columbus, SC