Using Infomaster to Create a Housewares Virtual Catalog

by Arthur M. Keller and Michael R. Genesereth, Stanford University, USA*

Abstract

Infomaster is an information integration system that provides integrated access to multiple distributed heterogeneous information sources on the Internet, thus giving the illusion of a centralized, homogeneous information system. We say that Infomaster creates a virtual data warehouse. The core of Infomaster is a facilitator that dynamically determines an efficient way to answer the user's query using as few sources as necessary and harmonizes the heterogeneities among these sources. Infomaster handles both structural and content translation to resolve differences between multiple data sources and the multiple applications for the collected data. Infomaster connects to a variety of databases using wrappers, such as for Z39.50, SQL databases through ODBC, EDI transactions, and other World Wide Web (WWW) sources. There are several WWW user interfaces to Infomaster, including forms-based and textual. Infomaster also includes a programmatic interface and it can download results in structured form onto a client computer. Infomaster has been in production use for integrating rental housing advertisements from several newspapers (since fall 1995), and for meeting room scheduling (since winter 1996). In this paper, we illustrate how Infomaster is being used to integrate heterogeneous electronic product catalogs, specifically to create a virtual catalog of integrating several housewares catalogs.

Introduction

In recent years, there has been a dramatic growth in the number of publicly accessible databases on the Internet, and all indications suggest that this growth will continue in the years to come. Access to this data presents several complications.

The first complication is distribution. Not every query can be answered by the data in a single database. Useful relations may be broken into fragments that are distributed among distinct databases. Database researchers distinguish among two types of fragmentation. In horizontal fragmentation, the rows of a database are split across multiple databases. For example, GM will maintain its own catalog of cars separately from Ford's catalog of cars. In vertical fragmentation, the columns are split. For example, while the basic description of each car model is consistent, the price of the same model car may vary from dealer to dealer. Car model descriptions should come from the manufacturer's database, while price may come from the dealer's database. Distributed databases can exhibit mixtures of these types of fragmentation.

Arthur M. Keller (ark@cs.stanford.edu) is project manager for Stanford CIT's effort on CommerceNet, including research on searchable online catalogs, providing services over the Internet, and concurrent engineering. His research interests include interoperability of heterogeneous databases, database integration, object-oriented databases, database implementation, databases on parallel computers, federated autonomous databases, database views including updates, incomplete information and nulls, software integration and reuse, and large system integration.

Michael Genesereth (genesereth@cs.stanford.edu) is an associate professor in the Computer Science Department at Stanford University. Prof. Genesereth is most known for his work on logical systems and applications of that work in engineering automation and software interoperation. He is the current director of the Center for Information Technology at Stanford.

A second complication in database integration is heterogeneity. This heterogeneity may be notational or conceptual. Notational heterogeneity concerns access language and protocol. One source is a Sybase database using SQL while another is an Informix database using SQL and a third is an Object Store using OQL. This sort of heterogeneity can usually be handled through commercial products (such as Sybase OpenServer). However, even if we assume that all databases use a standard hardware and software platform, language and protocol, there can still be a conceptual heterogeneity, i.e., differences in relational schema and vocabulary. Distinct databases may use different words to refer to the same concept, and/or they may use the same word to refer to different concepts. Reassembling the distributed fragments of a database in the face of heterogeneity is doubly difficult.

Infomaster is an information integration tool that solves these problems. It provides integrated access to distributed, heterogeneous information sources, thus giving its users the desirable illusion of a centralized, homogeneous information system. Infomaster effectively creates a virtual data warehouse of its sources. The user does not have to be expert in accessing the diverse databases and yet the data does not have to be copied into one central location. (Data may however be cached for performance reasons.)

The next section gives some technical details about Infomaster. Section 3 is a detailed example of translation of heterogeneous sources by Infomaster. Section 4 is the conclusion.

Technical Details

The core of Infomaster is a facilitator that determines which sources contain the information necessary to answer the query efficiently, designs a strategy for answering the query, and performs translations to convert source information to a common form or the form requested by the user. Formally, Infomaster contains a model-elimination resolution theorem

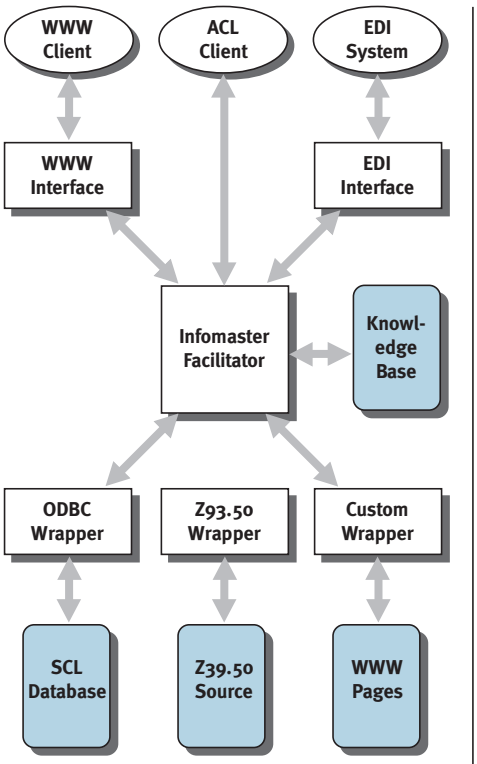


Figure 1
Infomaster Architecture showing the Infomaster Facilitator, integration engine, wrappers for ODBC, Z39.50 and custom sources, and user interfaces for WWW, EDI, and ACL.

prover as a workhorse in the planning process. Figure 1 illustrates the architecture.

There are wrappers for accessing information in a variety of sources. For SQL databases, there is a generic ODBC wrapper. There is also a wrapper for Z39.50 sources. For legacy sources and structured information available through the WWW, a custom wrapper is used. For example, we use a custom wrapper to access housing rental advertisements from several San Francisco Bay Area newspapers on the WWW. The advertisements are then accessible through Infomaster using a forms-based WWW interface that supports structured queries.

Infomaster uses rules and constraints to describe information sources and translations among these sources. These rules and constraints are stored in a knowledge base. For efficient access, the rules and constraints are loaded into Epilog, a main memory database system from Epistemics. Examples of the internal forms of these rules and constraints are given in the next section.

Infomaster includes a WWW interface for access through browsers such as Netscape's. This user interface has two levels of access: an easy-to-use, forms-based interface, and an advanced interface that supports arbitrary constraints applied to multiple information sources. However, additional user interfaces can be created without affecting the core of Infomaster.

Infomaster has a programmatic interface called Magenta, which supports ACL (Agent Communication Language) access. ACL consists of KQML (Knowledge Query and Manipulation Language), KIF (Knowledge Interchange Format), as well as vocabularies of terms.

Figure 2 shows the internal form of two of the rules that translate between the virtual data warehouse on top and the source data below. The rules given in Figure 2 are in their internal forms as interpreted by Infomaster. System maintainers are expected to use a GUI to enter their rules using a spreadsheet-like format.

Harmonizing n data sources with m uses does not require n*m sets of rules, or worse. By providing Infomaster with a reference schema, we allow database users and provides to describe their schemas without regard for the schemas of other users and providers. This strategy is shown in Figure 3. Translation rules describe how each source relates to the reference schema.

Bibliography

Duschka, Oliver M. and Genesereth, Michael R. "Query Planning in Infomaster," 1997 ACM Symposium on Applied Computing, February 1997.

Geddis, Donald F., Genesereth, Michael R., Keller, Arthur M. and Singh, Narinder P. "Infomaster: A Virtual Information System," Intelligent Information Agents Workshop, at CIKM '95, December 1995.

Genesereth, Michael R., Keller, Arthur M. and Duschka, Oliver M. "Infomaster: An Information Integration System," 1997 ACM SIGMOD, May 1997.

Item	Type	Primary Material	Interior Surface	Exterior Surface
3500117 12in Skillet	Skillet	Aluminium	Teflon	Metal

(corning-to-exterior non-stick-polished-aluminium metal)

(<= (exterior-surface ?x ?z)
(corning-material ?x ?y) (corning-to-exterior ?y ?z))

Item	Type	Material
3500117 12in Skillet	Skillet	Non Stick Polished Aluminum

Figure 2
Translation Rules showing translation of material in source database into primary material, interior surface, and exterior surface in virtual data warehouse.

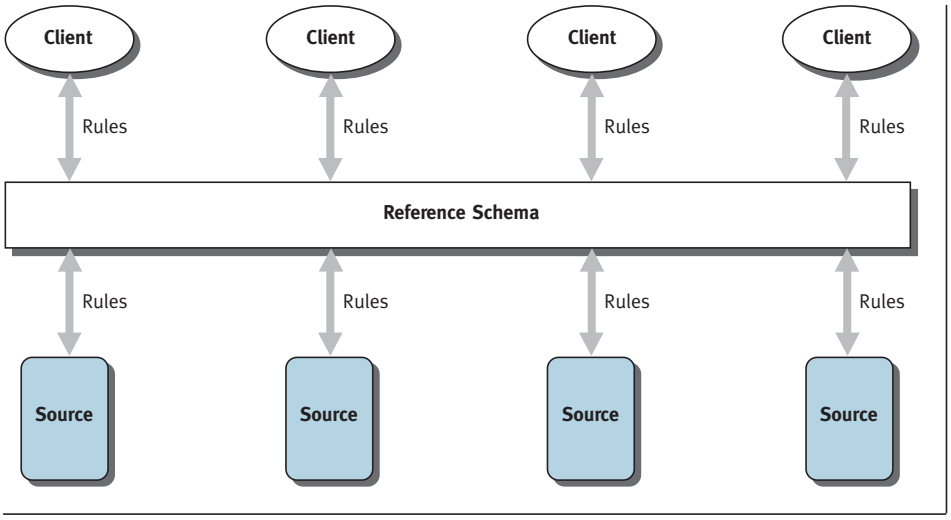


Figure 3 Use of Reference Schema to support efficient and maintainable translation rules harmonizing heterogeneous information sources and providing information as needed by each user.

These translation rules are bidirectional whenever possible, so information stored in one source's format may be accessed through another source's format. The same type of rules are used to describe how clients want to access data. These rules are combined and interpreted by Infomaster during query optimization and query processing. Because these translations reference each other essentially through the reference schema, entry and maintenance of translation rules is enhanced.

All rules are entered separately and incrementally with each data source. Infomaster assembles these into an efficient knowledge base for interpretation when a query is handled.

The example in the next section shows how Infomaster was used to integrate heterogeneous electronic catalogs stored as databases. However, Infomaster actually uses a peer-to-peer connection architecture. The rules can be used to direct updates from the clients to the relevant information sources, when permitted. Clients can register their interests and be notified when an information source changes in a way that intersects the registered interest.

Example of Harmonizing Source Heterogeneity

In this section, we illustrate the use of Infomaster with an example of three heterogeneous product catalogs of housewares, a database of products stocked by Sears, and a database of companies, and a virtual catalog integrating these five databases, designed for use by the assortment planners at Sears. The three product catalogs are stored in three separate heterogeneous database representing the three manufacturers, Corning, Mirro, and Regal. Figure 4 illustrates the structure of the virtual housewares catalog.

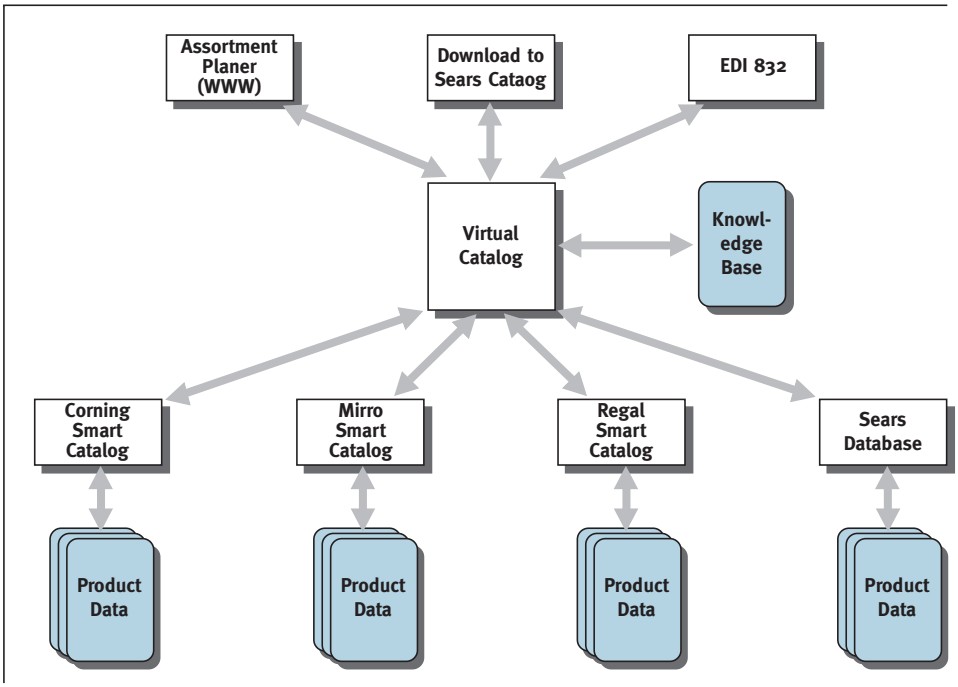
Table 1 shows the result of a query to the virtual housewares catalog for aluminum skillets with Teflon coating stocked by Sears. Sears specifically asked that all non-stick coatinges be referred to as Teflon to illustrate multiple levels of translation. The columns are the individual skillets, and the rows are the characteristics type, primary material, interior surface, exterior surface, color, diameter, SKU, UPC, Stock Item, and Manufacturer.

Table 2 shows the source entries for the selected skillets from Mirro. This table is very similar in format to the desired virtual catalog. However, if one searched the Mirro catalog for teflon skillets using the

Keller, Arthur M. "Smart Catalogs and Virtual Catalogs," in Readings in Electronic Commerce, Ravi Kalakota and Andrew Whinston (eds.), Addison-Wesley, 1997, pp.259-271.

Keller, Arthur M. and Genesereth, Michael R. "Multivendor Catalogs: Smart Catalogs and Virtual Catalogs," in EDI Forum, The Journal of Electronic Commerce (9:3), September 1996, pp. 87-93.

Figure 4 Architecture of the Virtual Housewares Catalog.



	3500117 12in Skillet	27069 Saute	27199 Grillpan	92732 Saute	J5817 12 In Super Skillet
Type	Skillet	Skillet	Skillet	Skillet	Skillet
Primary Material	Aluminium	Aluminium	Aluminium	Aluminium	Aluminium
Interior Surface	Teflon	Teflon	Teflon	Teflon	Teflon
Exterior Surface	Metal	Porcelain	Porcelain	Porcelain	Metal
Color	Silver	Black	Blue	Black	Grey
Diameter	12	12	12	12	12
SKU	008 7020 000	008 7031 000	008 7032 000	008 7033 000	008 7019 000
UPC	050035 001176	071108 270695	071108 271999	071108 927322	078008 018563
Stock Item	Yes	Yes	Yes	Yes	Yes
Manufacturer	Corning	Mirro	Mirro	Mirro	Regal

Table 1 Virtual Housewares Catalog display of aluminum teflon skillets stocked by Sears.

	27069 Saute	27199 Grillpan	92732 Saute
Mirro Type	Saute	Grillpan	Saute
Mirro Line	Concentric Air Channelon	Concentric Air	Concentric Air Channelon
Mirro UPC	071108 270695	071108 271999	071108 927322
Mirro Primary Material	Aluminium	Aluminium	Aluminium
Mirro Interior	Maxalon X2000	Maxalon X2000	Maxalon X2000
Mirro Exterior	Porcelain w/Silkscreen Bottom	Porcelain w/Silkscreen Bottom	Porcelain w/Silkscreen Bottom
Mirro Handle Material	Phenolic	Phenolic	Phenolic
Mirro Diameter	12	12	12
Mirro Color	Black	Blue	Black

Table 2 Source data of selected skillets from Mirro.

	J5817 12 In Super Skillet
Regal Type	Super Skillet
Regal Diameter	12
Regal UPC	078008 018563
Regal Interior	Arc Sprayed
Regal Exterior	Grey Nonstick
Regal Gauge	8 Ga
Regal Pan Construction	Aluminium
Regal Cover Construction	Tempered Glass

Table 3 Source data of J5817 12 in Super Skillet from Regal.

	3500117 12in Skillet
Corning Type	Skillet
Corning Diameter	12in
Corning Material	Non Stick Polished Aluminium
Corning Shape	Round

Table 4 Source data of 3500117 12in Skillet from Corning.

exact characters, none would be found. But by using the translations in Infomaster, these 3 skillets are found.

Table 3 shows the source entry for the selected skillet from Regal. Here the exterior is mapped to color Grey and exterior Teflon.

Table 4 shows the source entry for the selected skillet from Corning.

Conclusion

Infomaster is an information integration system originally developed at the Center for Information Technology of Stanford University.

Infomaster has been in use since fall 1995 for searching housing rentals in the San Francisco Bay Area, and since winter 1996 for room scheduling at Stanford.

Infomaster is the basis for the current Stanford Information Network (SIN) project that is integrating numerous structured information sources on the Stanford campus.

Infomaster is also the basis for the Housewares Virtual Catalog, a proof of concept with participants from Corning, Mirro, Regal, Sears, GE Information Services, National Housewares Manufacturers Association, National Retail Federation, Stanford University, and Epistemics.

Stanford's Infomaster service can be found at http://infomaster.stanford.edu

Infomaster is now being commercialized by Epistemics. Epistemics can be reached at info@epistemics.com or http:// www.epistemics.com.

Acknowledgements

The authors would like to thank the members of the Future Electronic Catalog Project team and the sponsoring and participating companies for their assistance. The authors would also like to thank the members of the Center for Information Technology for their assistance and feedback in developing Infomaster.