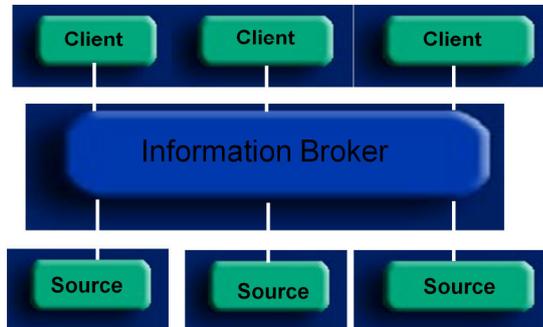


Információs bróker



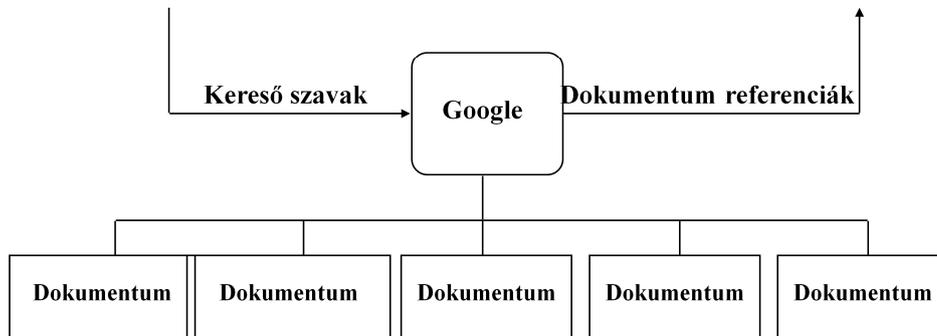
1

For these various reasons, researchers have for the last decade been investigating technology for dealing with these problems. The Holy Grail is an **information integration** system, sometimes called an information broker. It provides integrated access to fragmented, heterogeneous, distributed data sources, giving the user the illusion of a unified, homogeneous, centralized information system.

The user of an information broker interacts with the system to retrieve and update information using his own database schema while the database provider maintains data in his own schema.

The user can be a human user interacting through a web browser; it can be an application program treating infomaster as a virtual database; it can be a data warehouse using the system to update its information. The sources can be ODBC databases, XML files, LDAP systems, and so forth.

Szintaktikus kereső gépek



2

As many of you are aware, Wais is a search program to help users find textual documents on the web. It takes a set of search words as input and returns a list of documents containing those words. In order to make its determination, Wais uses a document index, which is updated periodically, usually in the wee hours of the morning. Unfortunately, as a general information system, it is not very good. Let me make my case with some examples.

Túl sok találat

Lekérdezés:

Who is older -- Jane or John?

Kereső szavak:

John

Jane

older

Dokumentum részletek:

..John is older than Jane...

Jill wants to know whether John is older than Jane...

*..John is older than Jill..
...Jim is older than Jane...*

3

Imagine, if you will, a student trying to discover who is older: John or Jane. He fires up Wais and does a syntactic search for documents containing the words John, Jane, and older and discovers the fragments pictured here. In this case, syntactic processing technology finds an appropriate document. However, it also finds numerous other documents, which do not bear on the problem. One of the problems of Wais is that it often returns too many results.

Túl kevés találat

Kérdés:

Is it the case that John is older than Jane?

Dokumentum részletek:

..John is more advanced in years than Jane...

..Jane is younger than John...

...John is the father of Jane...

4

A more serious problem is that it returns too few results (or at least does not pick them out from a vast mass of irrelevant documents). In the example, what if this document had not been available? This does not mean that John is not older than Jane, and it does not even mean that the system does not have sufficient information to answer the question. Here we see some examples of informational fragments that can be used to answer this question, fragments that do not contain this combination of keywords -- a case where a synonym is used, a case where an inverse is used, a case where knowledge of the world can be used to deduce the answer.

Integráció hiánya

Kérdés:

Is it the case that John is older than Jane?

Dokumentum részletek:

...John is older than Jill...

...Jill is older than Jane...

5

Here we have another example. In this case, once again, the requested information is not stored explicitly. However, unlike the previous example, getting an answer in this case is not simply a matter of transforming a single fragment. In this case, fragments from different informational resources must be combined to answer the query.

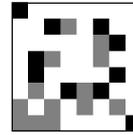
Tartalom vs Forma

Szemantikus nézet



*Those who will not reason
Perish in the act;
Those who will not act
Perish for that reason.*

Szintaktikus nézet



Thosewhowillnotr
easonPerishinthe
act;Thosehowill
notactPerishfort
hatreason.

Adatbázisok

name	manager	office	phone	
John	Jill	MJH222	38086	
Jane	Jerry	Cedar12	57493	
Jill		MJH222		
Jerry		420-032	56777	

7

Databases today come in several forms -- relational, object-oriented, and so forth. What they share is structure that enables what is often called semantic information processing (to distinguish it from the sorts of purely syntactic processing done by search tools like Alta Vista).

The personnel database shown here is a typical relational database. Each row represents a distinct individual, and each column represents a different attribute of that individual -- name, manager (for employees), office, and phone number.

Computers can search databases to answer complex questions. For example, using the information in the database shown here, it would be simple for the computer to find all people who share the same office or all people who share the same phone number but do not share the same office or all managers who share an office with one of their employees, and so forth.

Töredezettség

Horizontális töredezettség

name	manager	office	phone
John	Jill	MJH222	38086
Jane	Jerry	Cedar12	57493

name	manager	office	phone
Jill		MJH222	
Jerry		420-032	56777

Vertikális töredezettség

name	manager
John	Jill
Jane	Jerry
Jill	
Jerry	

name	office	phone
John	MJH222	38086
Jane	Cedar12	57493
Jill	MJH222	
Jerry	420-032	56777

8

Access to information, whether structured or not, is complicated by information **fragmentation**. This can occur in both databases and knowledge bases. Let us consider the case of databases; the problem is analogous for knowledge bases.

Database researchers distinguish two types of database fragmentation. In horizontal fragmentation, the rows of a database are split across multiple providers. In vertical fragmentation, the columns are split. Consider, for example, a database of information about people. Different departments may store the same type of information about different people. (This is horizontal fragmentation.) On the other hand, different corporate units may store different kinds of information about all people; for example, the personnel database may contain information about managers and employees, and the directory may contain office and phone information. (This is vertical fragmentation.)

Replikáció

Hálózati problémák

- Válaszadás késedelem

- Sávszélesség

- Megbízhatóság

Információs források kérdései

- Korlátozott elérhetőség

- Teljesítmény

- Nem tervezhető leállások

Megoldás - replikáció

- Költség? Hatékonyság?

9

Finally, there are problems of distribution. Although the WWW provides a high degree of location independence, the illusion is not perfect. Communication delays due to distance are sometimes noticeable, and there are occasional network failures. What's more, individual systems may be unavailable due to limited availability, scheduled maintenance, or unexpected failures.

A common solution to these problems is local replication of information. This can be costly; and, more significantly, it creates problems of potential inconsistency in the face of updates.

Heterogenit as

name	manager	office	phone	
John	Jill	MJH222	38086	
Jane	Jerry	Cedar12	57493	
Jill		MJH222		
Jerry		420-032	56777	

name	employee	location	telephone	
John		MJH222	7238086	
Jane		Cedar12	7257493	
Jill	John	MJH222		
Jerry	Jane	420-032	7256777	

"The biggest problem facing anyone who wants to search multiple structured databases. . . is that many organizations use different words to describe the same thing."

Martin Marshall, Communications Week

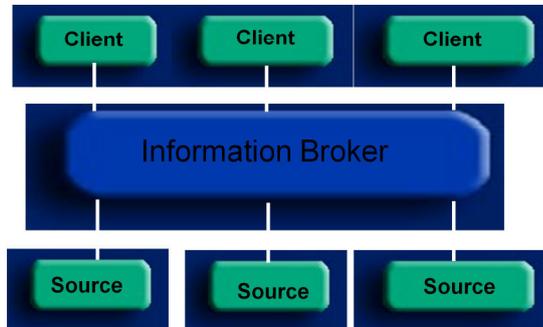
10

In order to provide these capabilities, Infomaster must overcome a variety of technical hurdles, including distribution, platform differences, and conceptual heterogeneity.

Now, there are tools in the marketplace for dealing with distribution and platform differences, but these alone are not sufficient. Even if you were to put all of your data into Oracle on a single workstation, there would still be a problem. Different people, in developing different portions of data are likely to use different vocabularies and different schemas for their data (i.e. they are likely to break their data into different tables with different rows and different columns).

Automatic Information Integration

integrated access to fragmented, heterogeneous, distributed data sources giving the illusion of a homogeneous data management system



11

For these various reasons, researchers have for the last decade been investigating technology for dealing with these problems. The Holy Grail is an **information integration** system, sometimes called an information broker. It provides integrated access to fragmented, heterogeneous, distributed data sources, giving the user the illusion of a unified, homogeneous, centralized information system.

The user of an information broker interacts with the system to retrieve and update information using his own database schema while the database provider maintains data in his own schema.

The user can be a human user interacting through a web browser; it can be an application program treating infomaster as a virtual database; it can be a data warehouse using the system to update its information. The sources can be ODBC databases, XML files, LDAP systems, and so forth.

Potential Application Areas

Corporate Logistics - Enterprise Resource Directories

Personnel, locations, organizations, equipment, orders

Electronic Commerce - Integrated Product Catalogs

Catalogs, inventories, product ratings, contracts

Health Care - Consolidated Patient Records

Doctors, nurses, lab technicians, administrators, patients

Multidisciplinary Engineering - Concurrent Engineering

Architects, engineers, construction planners

Command and Control - Situation Assessment

Commanders, intelligence, field officers, consultants

12

There are numerous applications for this technology. It is useful in any situation where multiple users must have integrated access to disparate kinds of information.

Typical application areas include electronic commerce, corporate logistics, multidisciplinary engineering, health care, and command and control.

In a while, I will tell you about two specific applications. But first a brief though slightly technical overview of the technology. As a professor, I just cannot resist the temptation to give a little technical lecture.

Question

Give me a list of 15 inch aluminum skillet with nonstick coating
rated at least 4 out of 5 by Consumer Reports
that sell for under \$30
and are currently in stock.

13

I would like you to consider these questions for a moment. These are the kinds of questions we have all wanted to answer at one time or another.

We have wanted to find products based on their **features**.

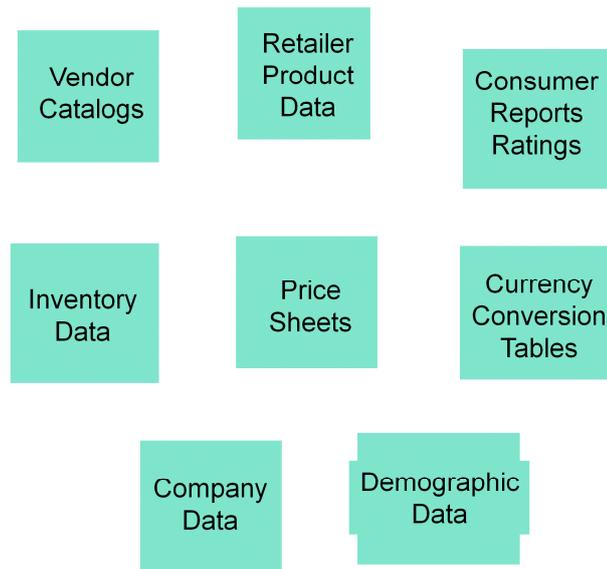
We have wanted to find products based in their **evaluations**.

We have wanted to find out who sells products at an attractive **price**.

We have wanted to find out whether products are in **stock** or on display.

Note that the issue here is not natural language. The same kinds of questions can be asked with form-based interfaces in HTML. What I am illustrating is our need for the **kinds of information** expressed in these questions.

Data Sources



14

The **good news** is that the information needed to answer these questions is available today in the form of online databases and knowledge bases. And more databases are becoming available everyday, with help of products like Infoserver from Epsistemics, Step Search from Saqqara, and Krakatoa from Cadis, and others.

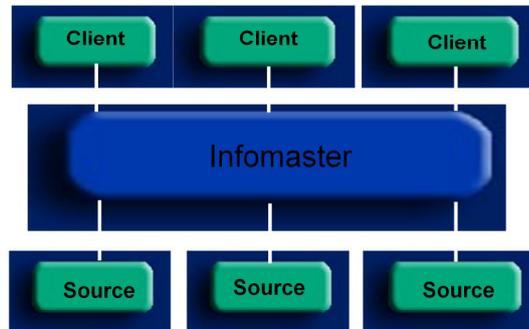
The **bad news** is that the job of finding the right databases, manually extracting the relevant information, and integrating it with information from others sources is time-consuming and error-prone.

What is needed is a tool that does this for us automatically. Unfortunately, such tools have just not been available...

Until now.

Infomaster

Data Integration System - integrated access to heterogeneous data sources giving the illusion of a homogeneous data management system



"Infomaster creates an environment that makes it easier for information consumers to get the information they need to answer their questions, while making it easier for owners to publish and share their databases."

Dennis Rayer, Manager, Data Warehouse,
Stanford University

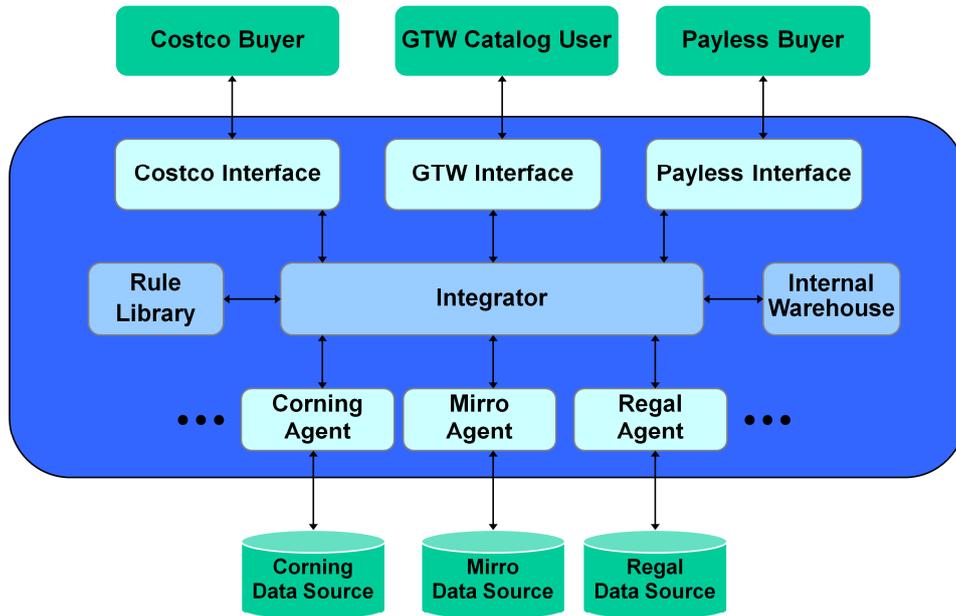
15

Infomaster is our solution to this problem. It is a database integration system. As such, It provides integrated access to conceptually heterogeneous data sources, giving the user the illusion that he is interacting with a conceptually homogeneous database system.

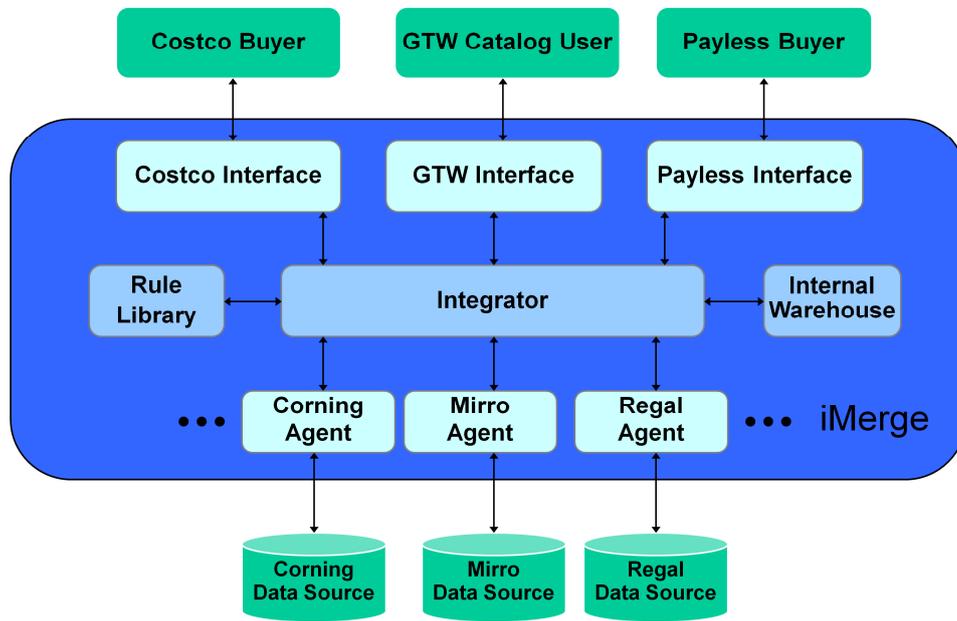
The user interacts with Infomaster to retrieve and update information using his own database schema while the database provider maintains data in his own schema.

The Infomaster client can be a human user interacting through a web browser; It can be an application program treating infomaster as a virtual database; it can be a data warehouse using infomaster to update its information. The sources can be ODBC databases, XML files, LDAP systems, and so forth.

Demonstration Architecture

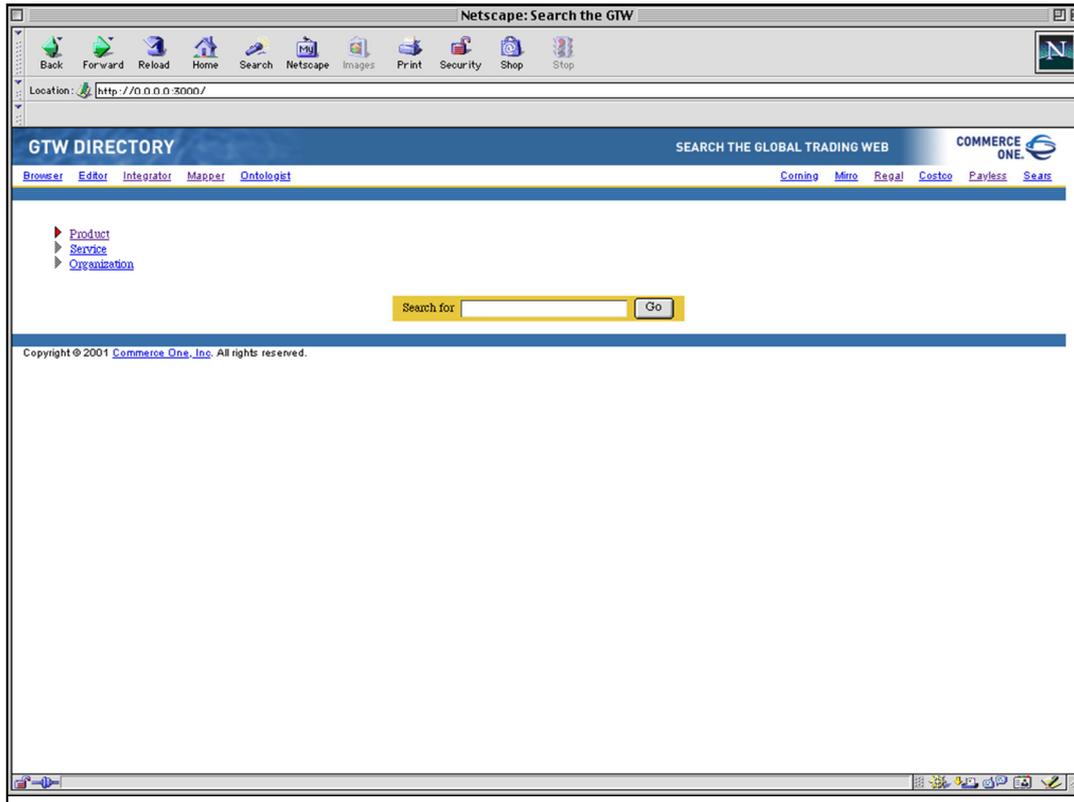


Demonstration Architecture



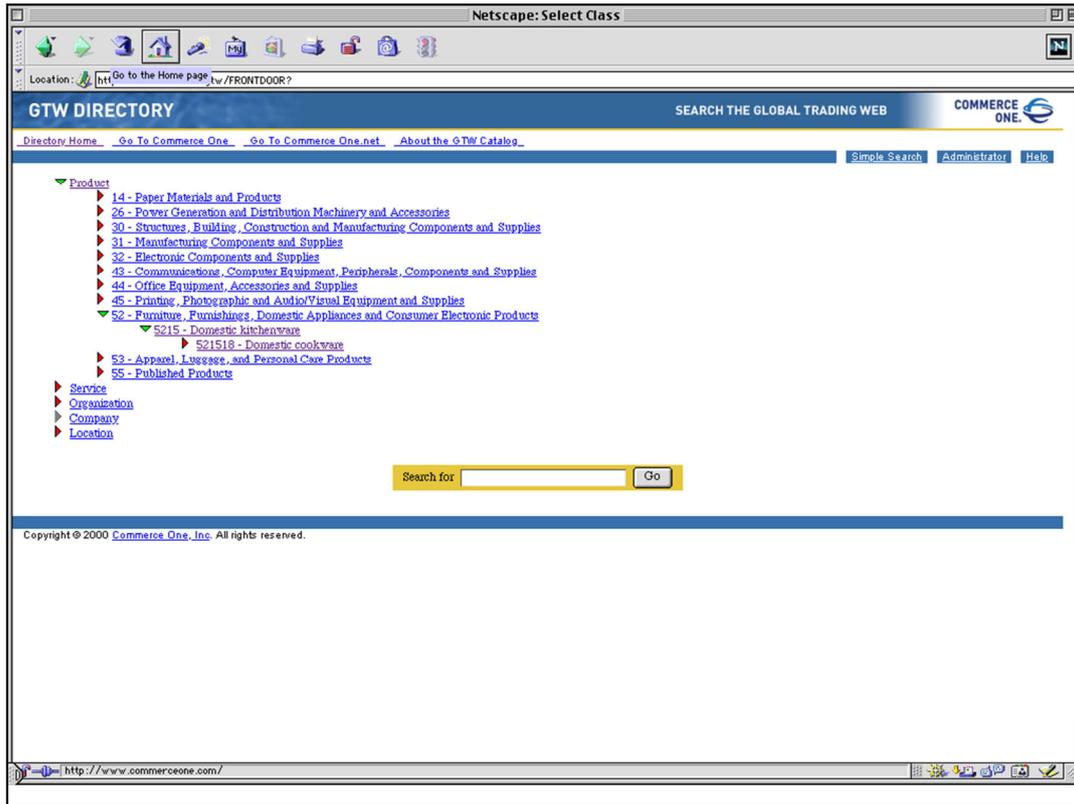
17

This is a demonstration of Infomaster in action. This is a system built in 1996 at the behest of the National Housewares Manufacturers Association. The goal was to a virtual catalog for cookware, drawing information from the catalogs of manufacturers (in this case, Corning, Mirro, and Regal) and making the information available in integrated form to retailers (here Payless and Sears).

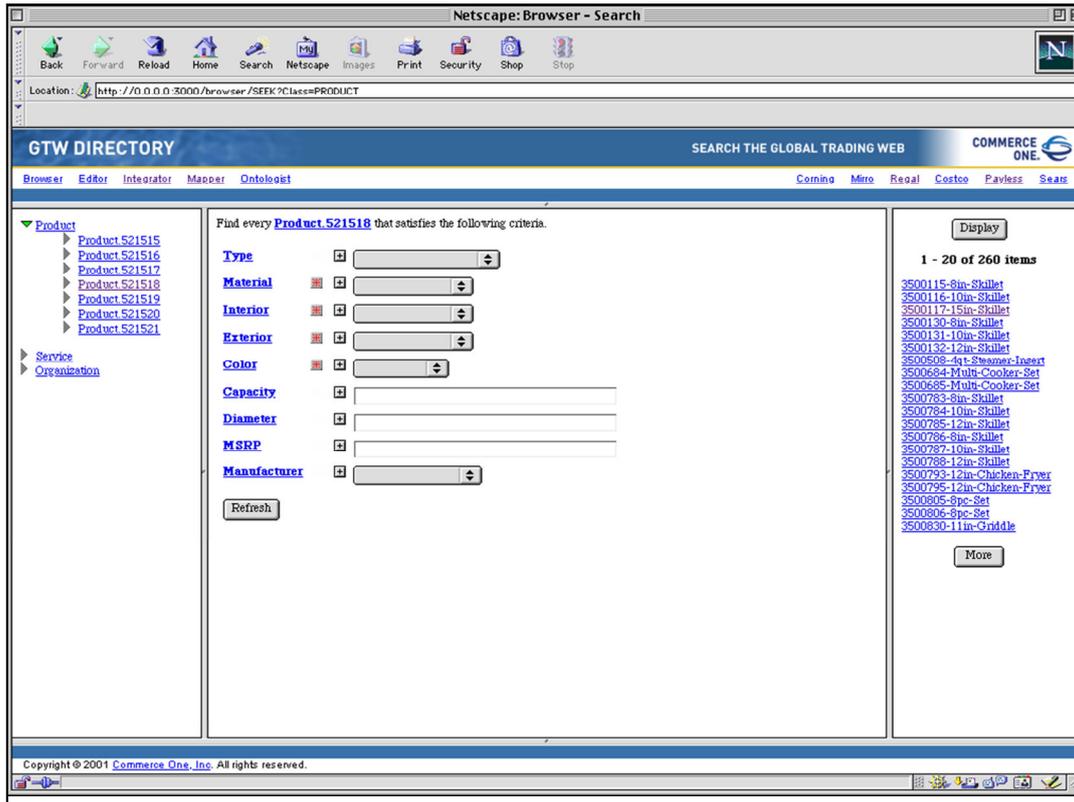


The Global Trading Web Directory provides associated buyers with integrated access to the catalogs of associated suppliers. In this case, there are just three suppliers, Corning, Mirro, and Regal; and there are just three buyers, viz. Costco, Payless, and Sears.

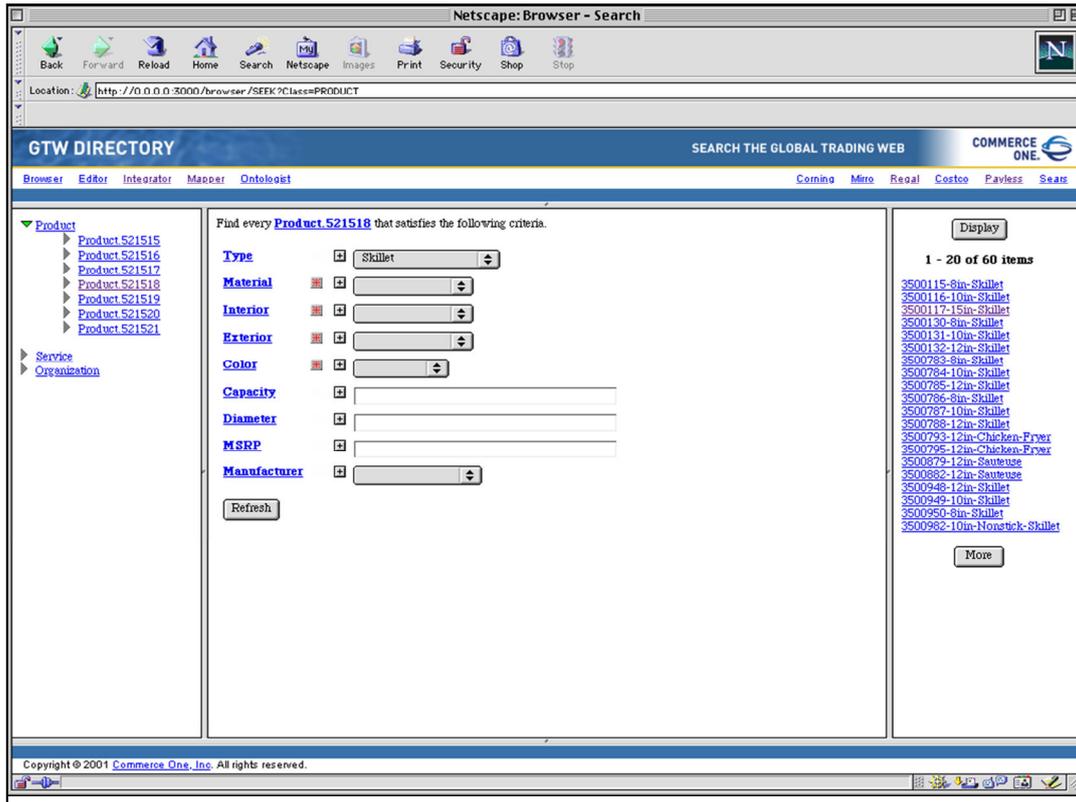
The opening page provides the user with two type of search. The type-in box allows the user to search to search by keyword. The high-level categories allow the user to search by category.



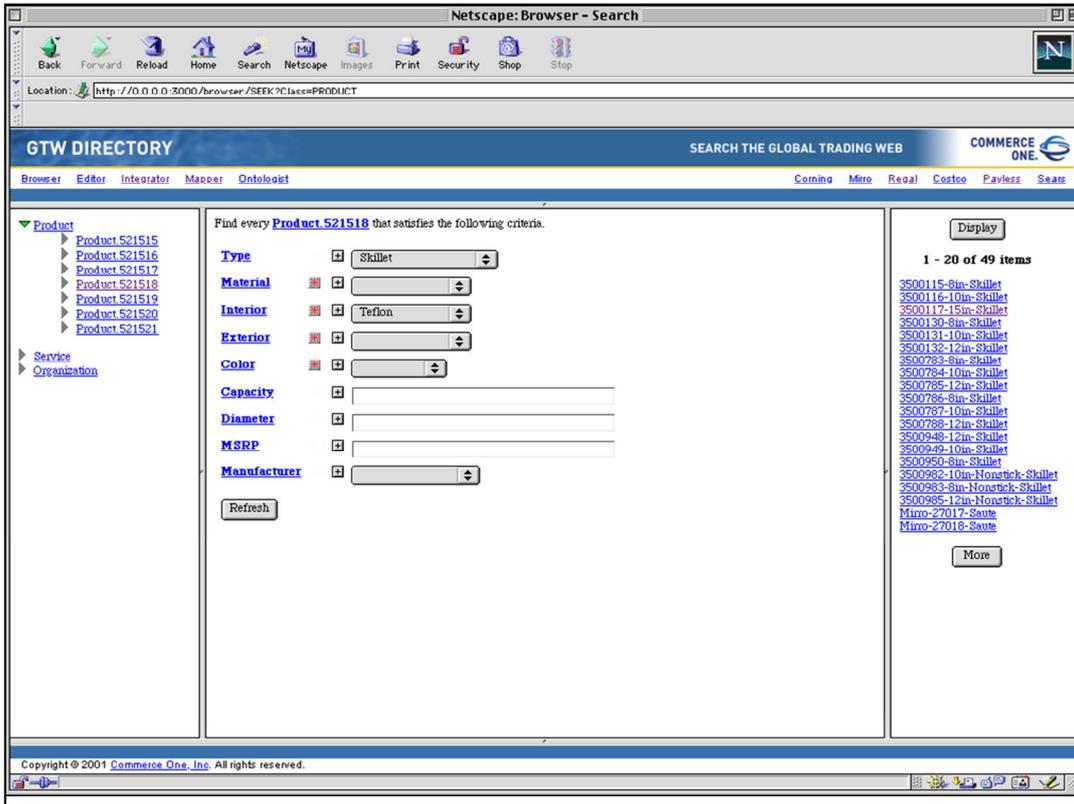
Clicking on the triangle next to a category name expands that category to show its subcategories. This process can be repeated until one finds a sufficiently narrow category. In this case, we have expanded the Product category three times to get Domestic cookware.



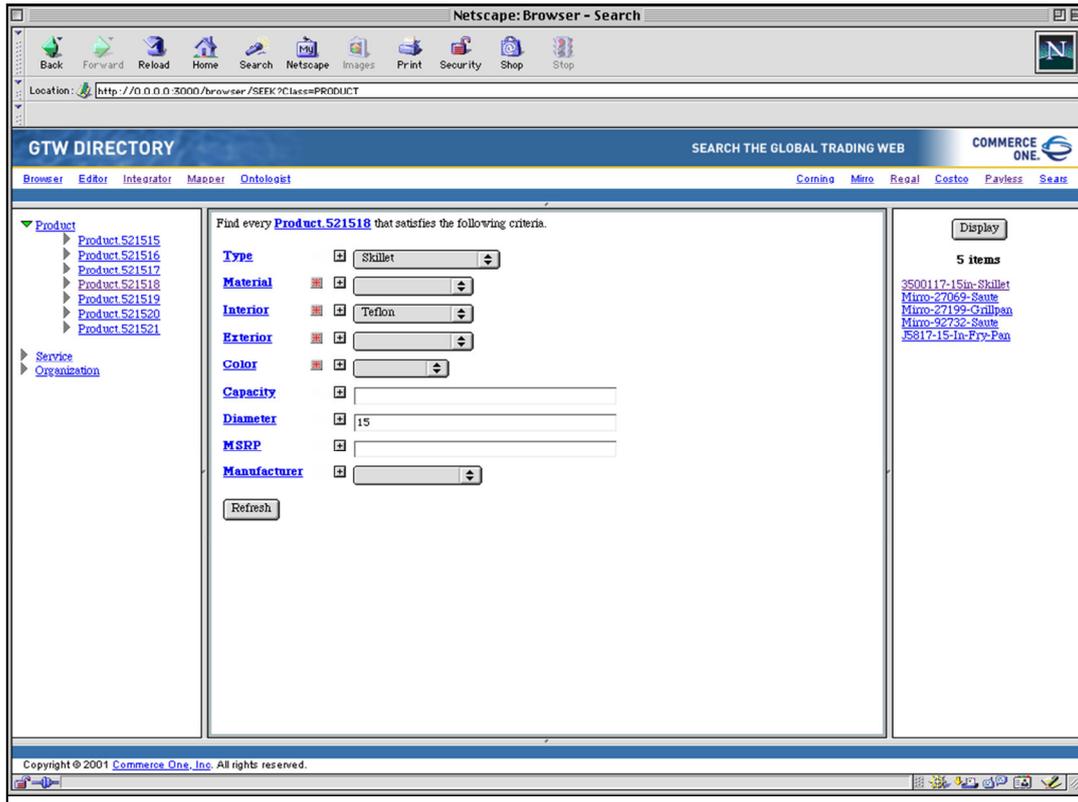
Clicking on a category name brings up a parametric search page. The left pane shows the category hierarchy. The middle pane shows attributes appropriate to the selected category, each with appropriate values. The right pane shows entries that match the search criteria specified. In this case, no attribute values are specified; and the system shows all 260 entries in the Cookware category.



By selecting values for attributes, the set of possible answers is pared down. As each choice is made, the answers in the right pane are automatically refreshed. For example, selecting skillet decreases the possibilities to 60.



Selecting Teflon for interior decreases the list to 49 items.



Entering 15 inches for diameter narrows the set of solutions to just 5 products.

Netscape: Browser - Display

Location: <http://0.0.0.0:3000/browser/ENHMERATE?>

GTW DIRECTORY SEARCH THE GLOBAL TRADING WEB **COMMERCE ONE**

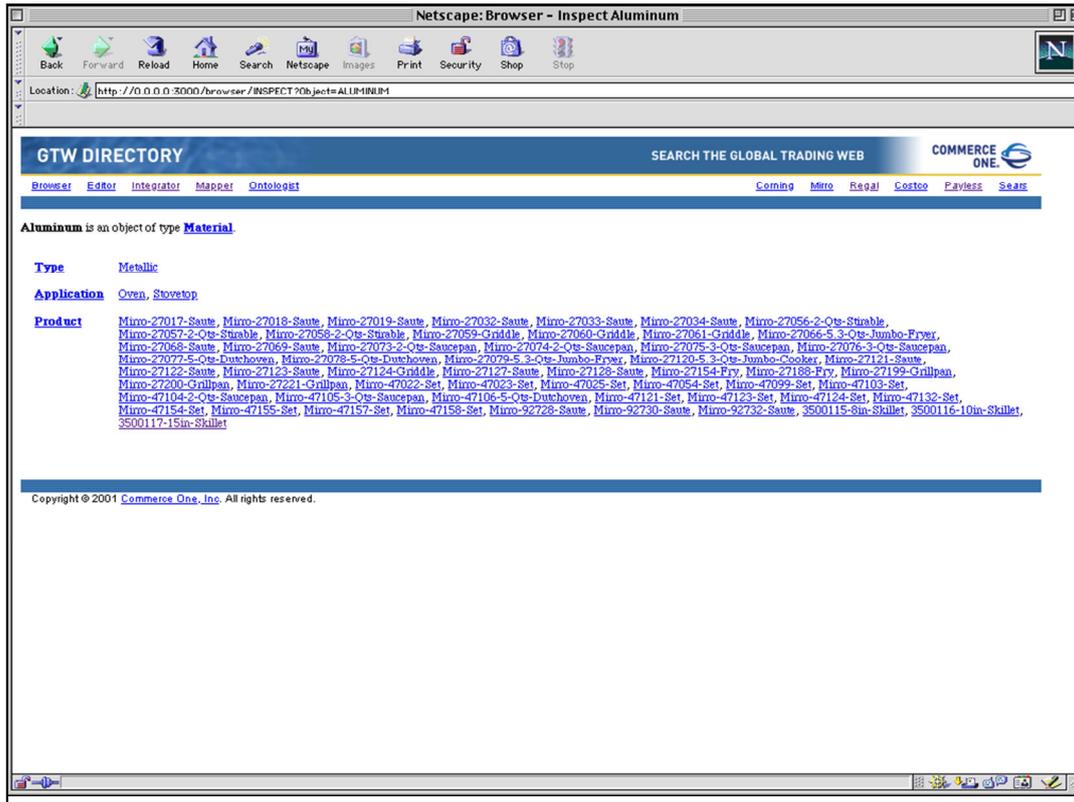
[Browse](#) [Editor](#) [Integrator](#) [Mapper](#) [Ontologist](#) [Coming](#) [Miro](#) [Regal](#) [Costco](#) [Favies](#) [Seas](#)

Product 521518	Type	Material	Interior	Exterior	Color	Capacity	Diameter	MSRP	Manufacturer	UPC	Image
3500117-15in-Skillet	Skillet	Aluminum	Teflon	Aluminum	Silver		15	33	Coming Inc	050035001176	
Miro-27069-Saute	Skillet	Aluminum	Teflon	Porcelain	Black		15	49	Miro Inc	071108270695	
Miro-27199-Grillpan	Skillet	Aluminum	Teflon	Porcelain	Blue		15	48	Miro Inc	071108271999	
Miro-92732-Saute	Skillet	Aluminum	Teflon	Porcelain	Black		15	44	Miro Inc	071108927322	
J5817-15-In-Fry-Pan	Skillet		Teflon	Teflon	Grey	4	15	92	Regal Inc	078008018563	

Display answers 1 through 5

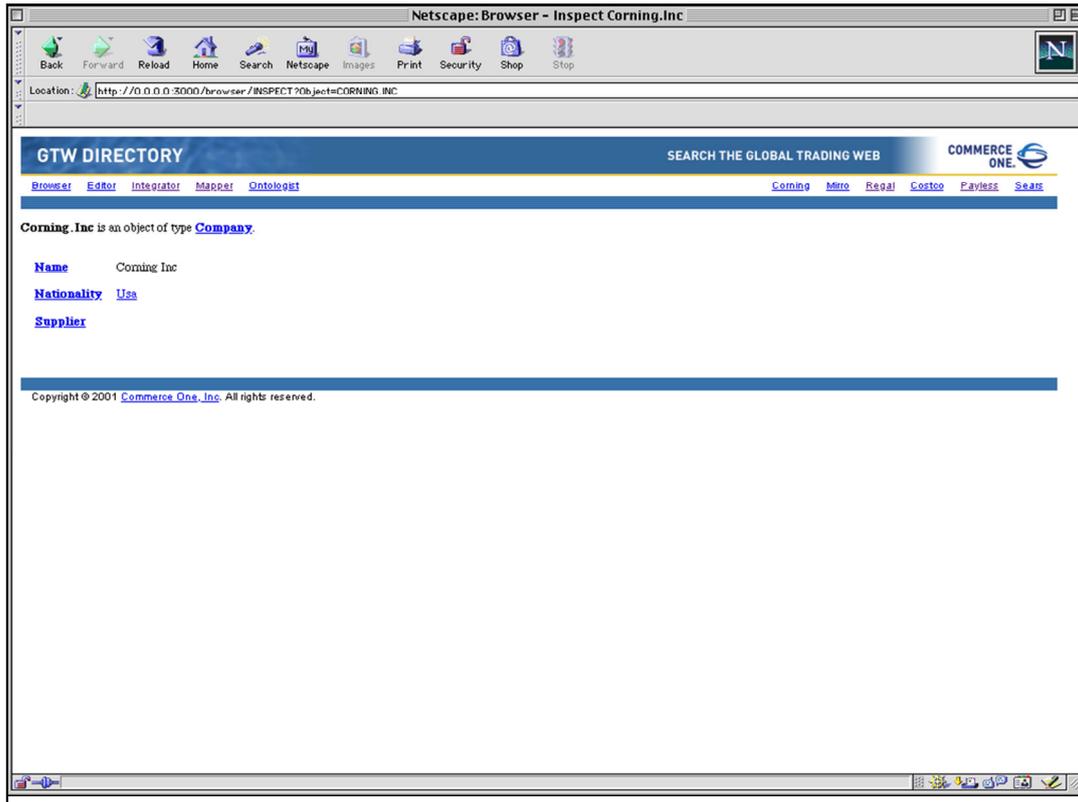
Copyright © 2001 [Commerce One, Inc.](#) All rights reserved.

Pressing the Display button brings up a table showing further information about the selected products.

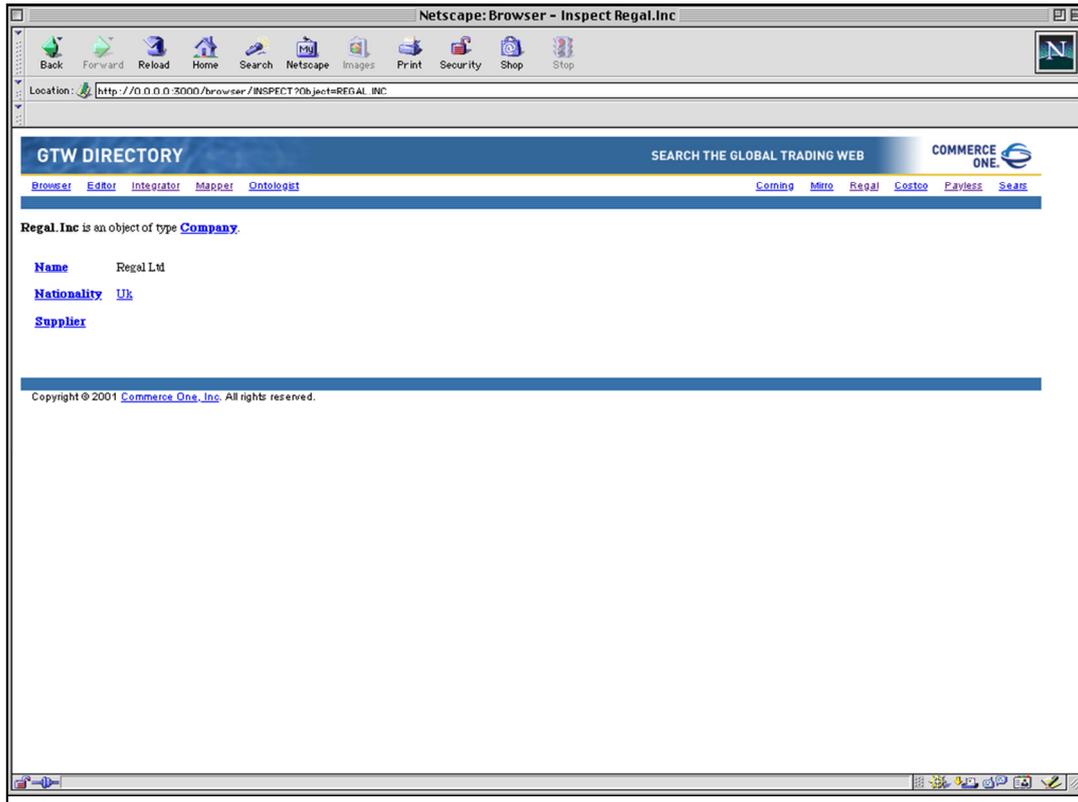


Clicking any link bring up an “Inspect” page containing of information about the associated concept. These pages are constructed on the fly from the databases available at the time. Each page contains information about a single concept and contains all information about that concept.

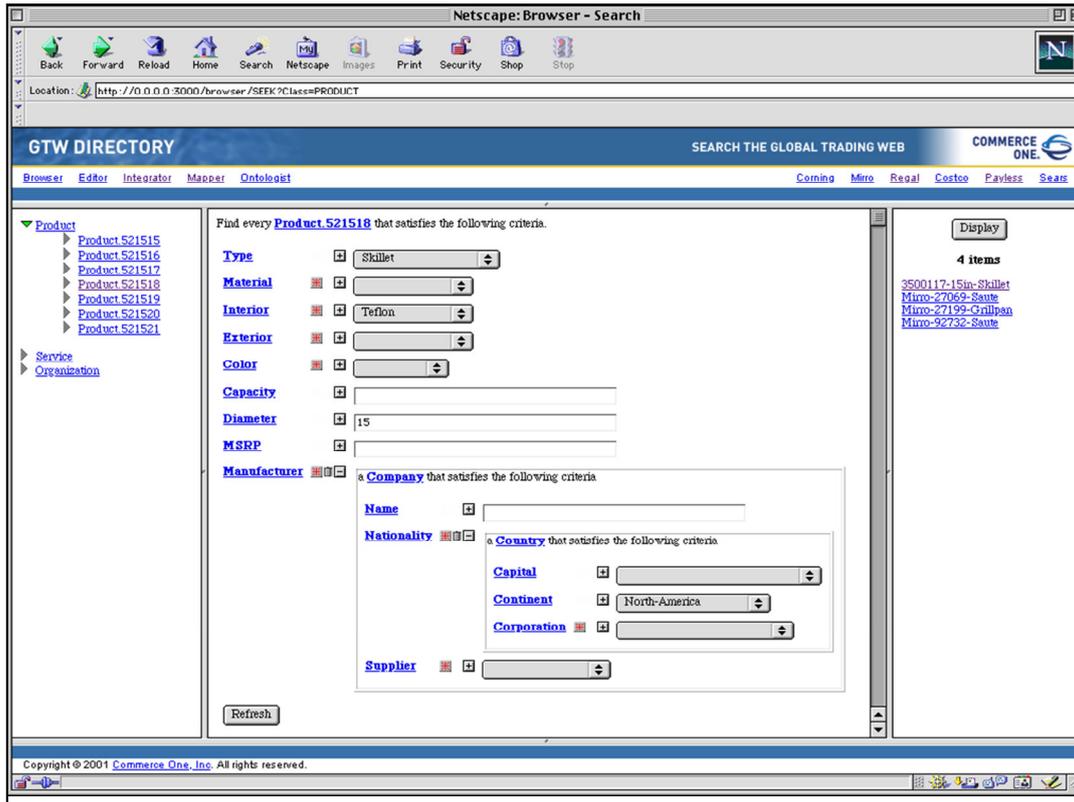
For example, clicking on Aluminum shows information about the material aluminum, such as its type (metallic), its possible uses (on the stove top and in the oven but not in the microwave), and the various products that contain aluminum.



Similarly, clicking on the name of a company brings up information about that company. For example, here we see that Corning is a US company. There is also a place for suppliers, but none are known.

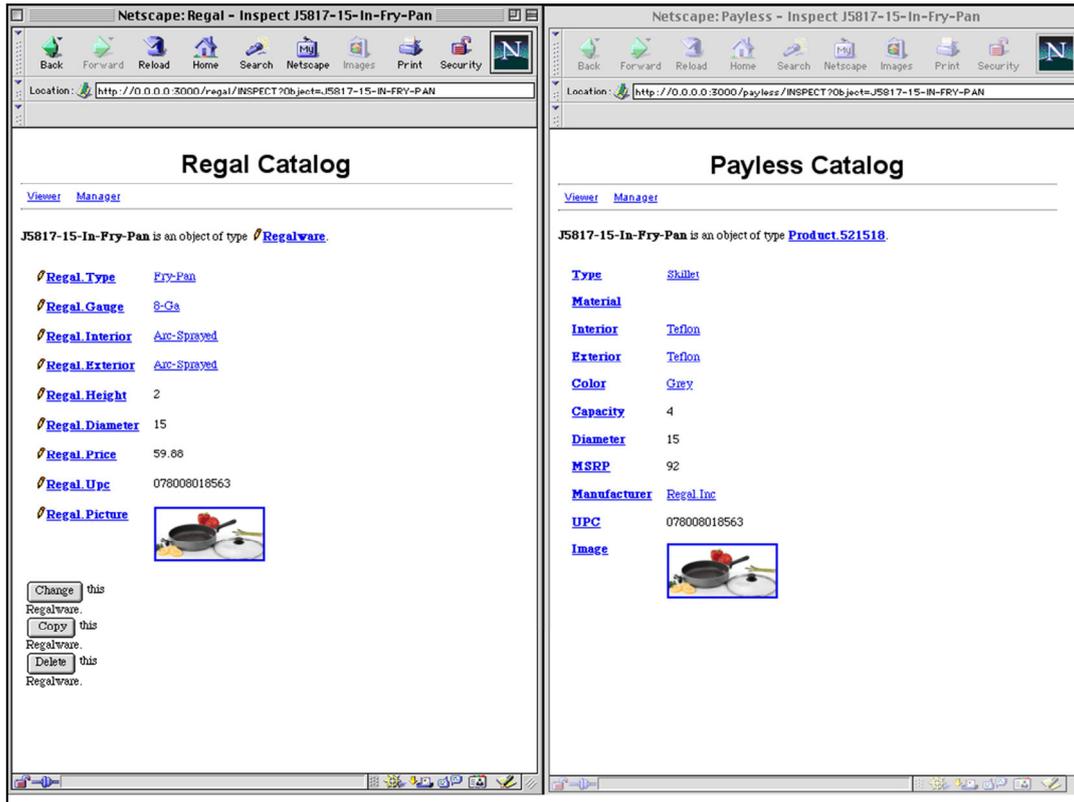


Clicking on Regal brings up analogous information. In this case, we see that Regal is a company in the UK. (This is not actually true, but this is just a demo.)



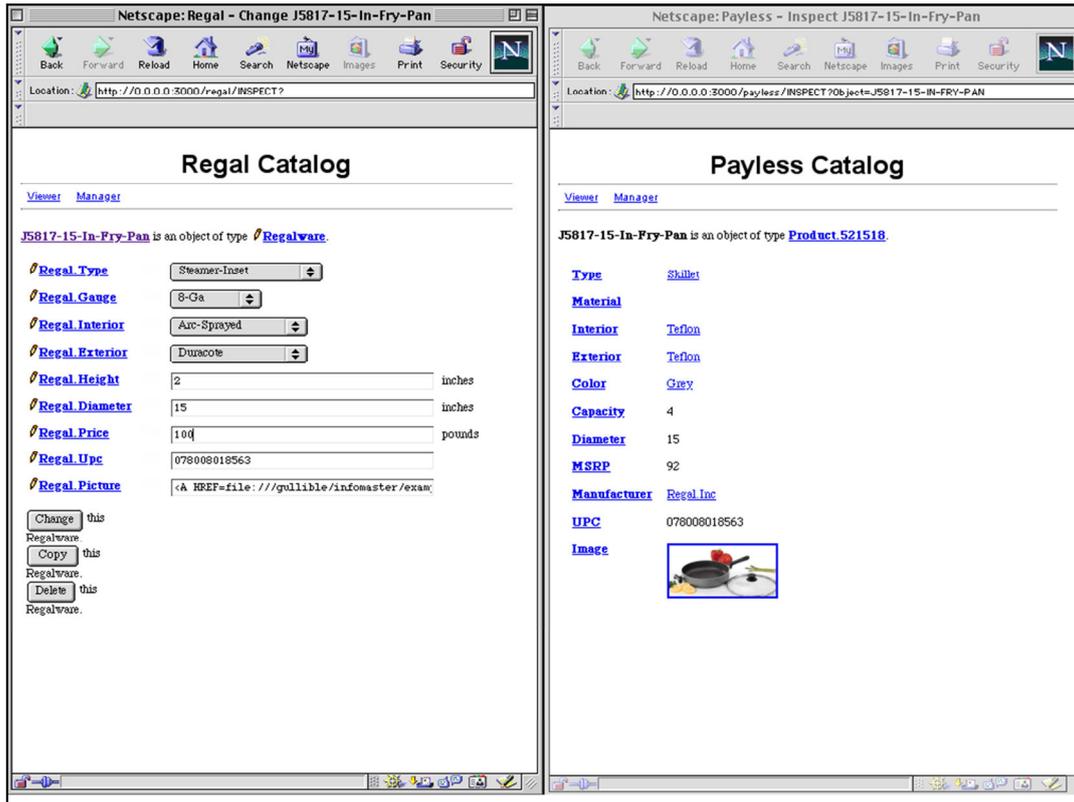
Browsing works well when there are just a few entries, as in this case. However, when there are thousands of entries, it can be tedious. Parametric search works well when the features of interest are attributes of the type of item being sought. In some cases, the features are indirect, i.e. they are attributes of the values of attributes. Since the search in such cases involves items of multiple categories, it is often called “cross-category” search. In order to do cross-category searches in Infomaster, the user clicks on the small icon beside the Attribute name. This opens up a sub-search that allows the user to enter features of the value of that attribute.

In this case, we have opened up the Manufacturer attribute to allow us to specify properties of companies. And we have gone further and opened up the Nationality attribute of companies to allow us to specify properties on countries. Clicking on North America here causes the system to display products with the specified features made by companies incorporated in countries on the North American continent. Note that there are only 4 answers. The Regal product is no longer an answer, as its manufacturer is incorporated in the UK.

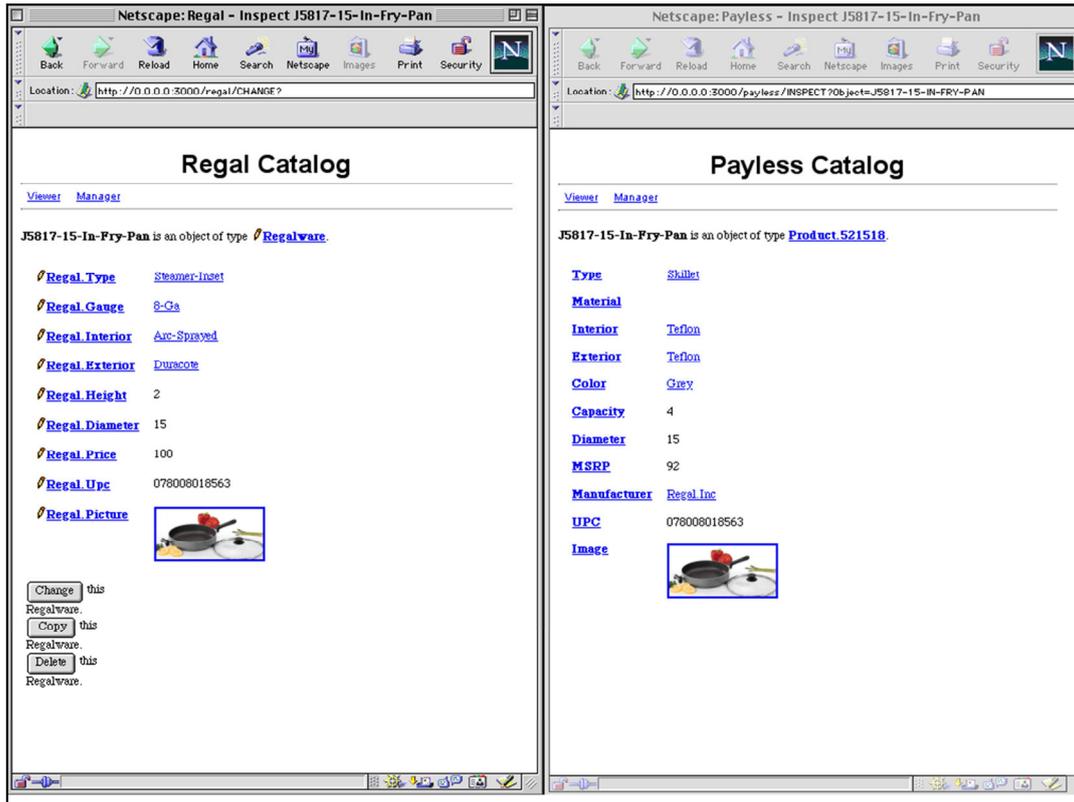


In order to illustrate Infomaster's support for conceptual heterogeneity, let's compare the source data for a product from one of the suppliers with the view available to one of the buyers. Here we have the supplier's view in the window on the left and the buyer's view in the window on the right.

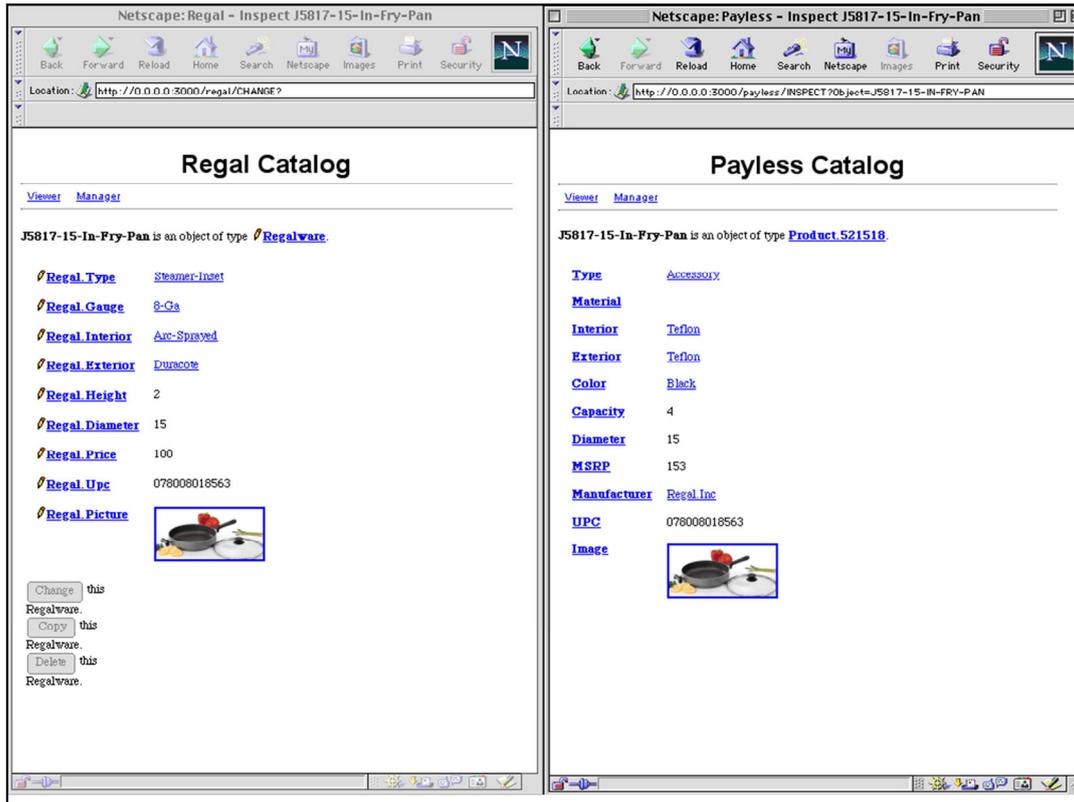
In the Regal catalog, the product is listed as a frying pan, whereas, in the Payless catalog, it is a skillet. In the Regal catalog, the interior and exterior are arc-sprayed, whereas in the Payless catalog, the surface is listed as Teflon. The Regal catalog contains diameter and height, both in inches, whereas the Payless catalog has capacity in quarts and diameter in inches. The Regal catalog lists Price in Pounds Sterling, whereas the Payless catalog lists the price in US dollars.



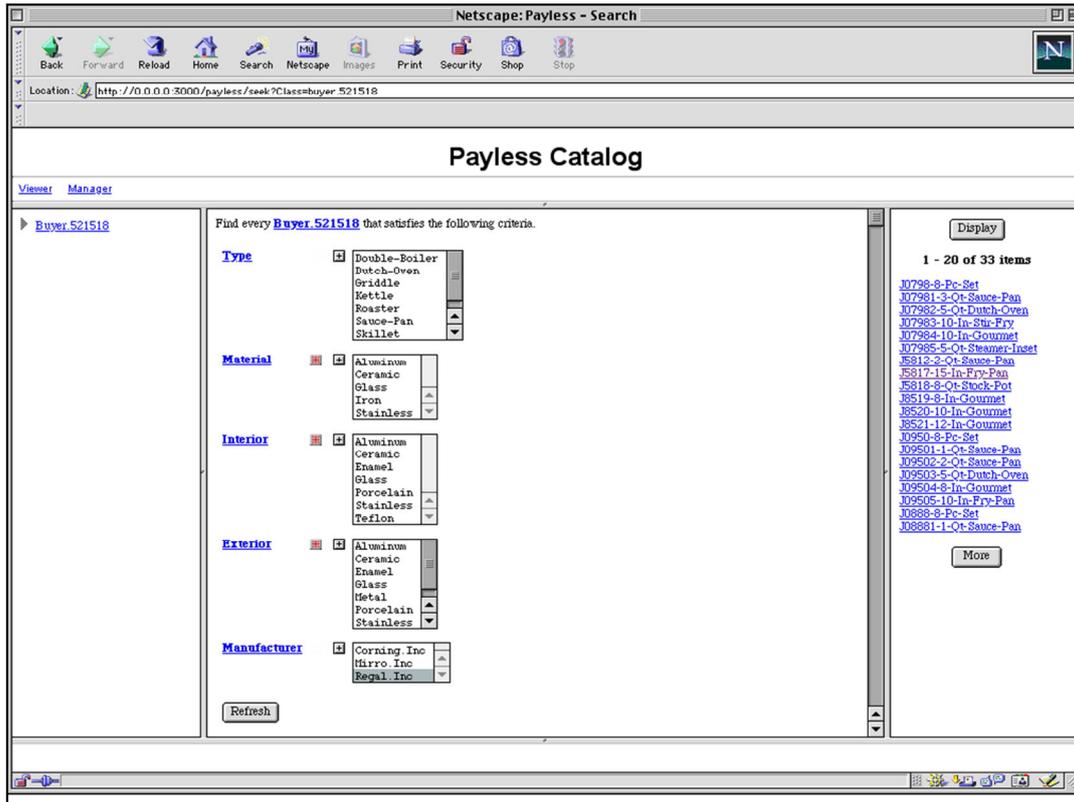
Now, let's change some of the information about this product. We click on the Change button to convert the Inspect page into a Change page. Then we can make modifications, e.g. switching the product to be a steamer insert, changing the exterior to be Duracote, changing the price to 100 pounds.



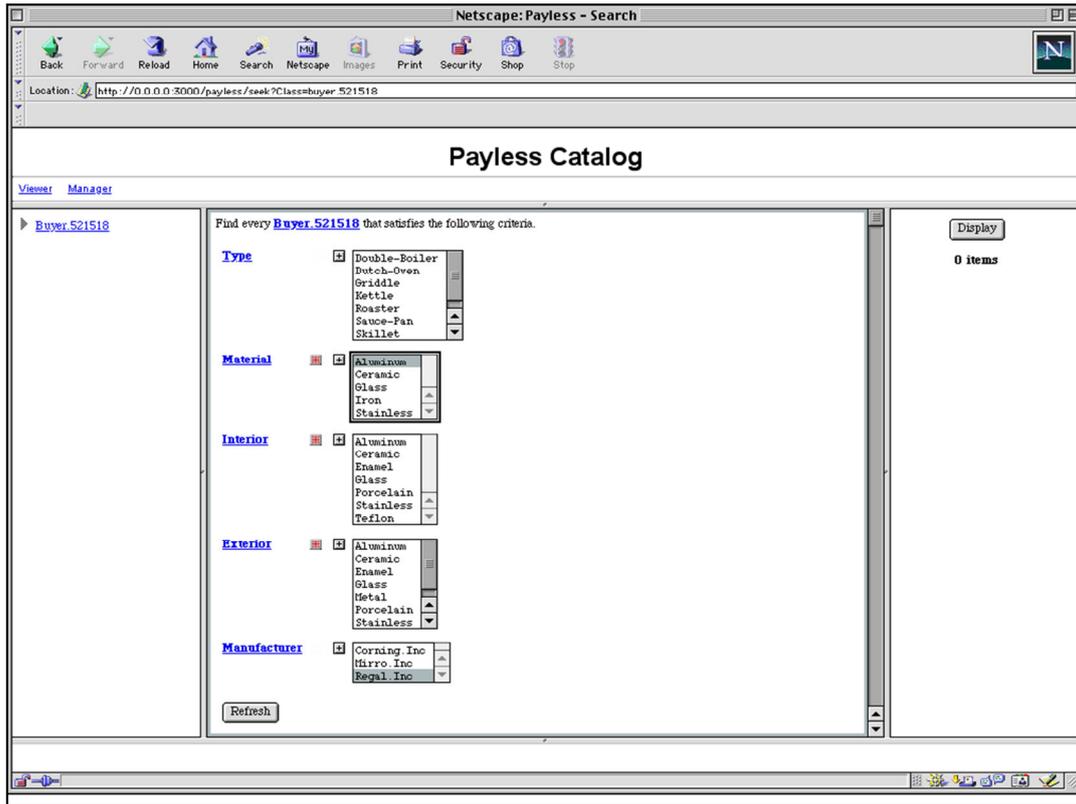
We then click on Change to send these changes to Infomaster. The Inspect page shows that the changes have been accepted. Of course, the Payless page is the same, since it has not been refreshed.



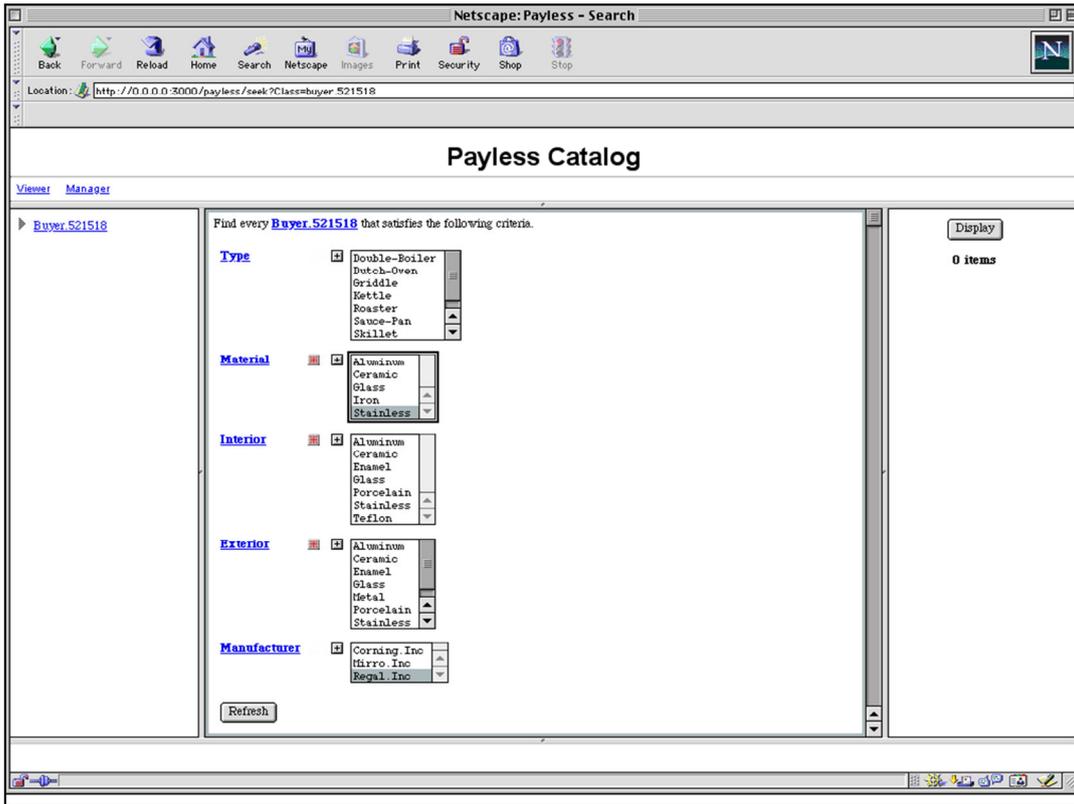
However, clicking the Refresh button causes the new values to be displayed. The product is now listed as an accessory in the Payless terminology. Its exterior is still Teflon. However, the color, computed from the actual exterior material, is listed as black. Finally, the price has changed proportionally.



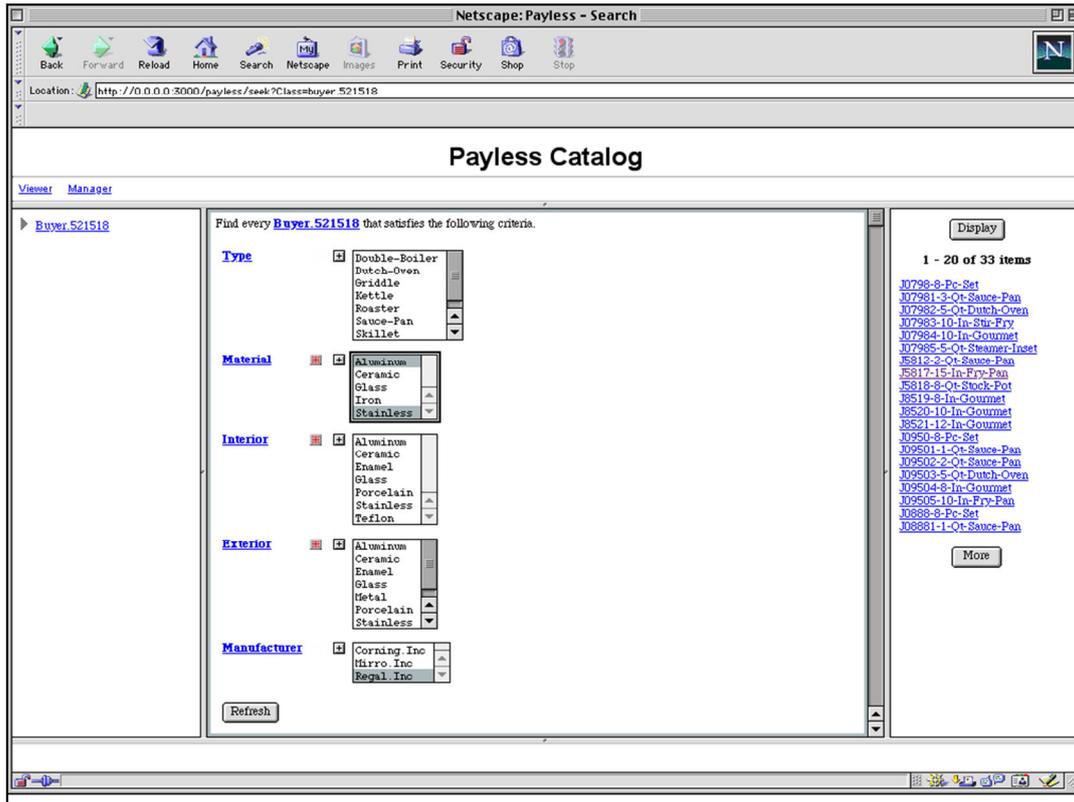
Finally, let's take a look at Infomaster's ability to deal with incomplete information. One thing we know about Regal is that it makes its products out of just two materials, aluminum and stainless. Unfortunately, the Regal catalog does list any material for its products. Consequently, Infomaster cannot fill in this field. On the other hand, things are not so bleak. Let's consider what happens when we search for Regal products. Here we see the Payless search page. Selecting Regal as manufacturer gives 33 products.



Naturally, clicking on Aluminum as material leads to zero hits, as none of the Regal products are known for sure to be made of aluminum.



Similarly, clicking on stainless leads to zero hits.



However, when we click both aluminum *and* stainless, all of the products reappear. Although Infomaster does not know exactly which material is in a product, it knows that in either case it satisfies the user's request.

Note that this is a very different answer than one would get from a traditional database system, where a disjunctive query like this one is handled by forming the union of the answers to queries formed from each disjunct. In this case, the result would be zero hits, despite the fact that all Regal products satisfy the user's request.

Why is it that database systems do not provide the correct answer in this case? The fact is that most Database systems were designed for use in corporate settings, where the data schema could be designed to ensure that no information is missing. Sadly, in the Internet setting, there is no overarching authority requiring that all fields be present. Yet we want to get as much information from the available databases as possible. The logical reasoning in Infomaster ensures that this criterion is met.