MIT Sloan School of Management

MIT Sloan School Working Paper 4673-07
12/01/2007

Enabling Global Price Comparison through Semantic Integration of Web Data

Hongwei Zhu, Stuart Madnick, Michael Siegel

Enabling Global Price Comparison through Semantic Integration of Web Data

Hongwei Zhu
Stuart Madnick
Michael Siegel

Working Paper CISL# 2007-12

December 2007

Composite Information Systems Laboratory (CISL)
Sloan School of Management, Room E53-320
Massachusetts Institute of Technology
Cambridge, MA 02142

Enabling Global Price Comparison through Semantic Integration of Web Data

Hongwei Zhu

College of Business and Public Administration
Old Dominion University
Norfolk, VA 23529, USA
Fax: +1-757-683-5639   Email: hzhu@odu.edu

Stuart E. Madnick*

Sloan School of Management
Massachusetts Institute of Technology
Cambridge, MA 02139, USA
Fax: +1-617-253-3321   Email: smadnick@mit.edu
*Corresponding author

Michael D. Siegel

Sloan School of Management
Massachusetts Institute of Technology
Cambridge, MA 02139, USA
Fax: +1-617-253-7579   Email: msiegel@mit.edu

Abstract: "Sell Globally" and "Shop Globally" have been seen as a potential benefit of web-enabled electronic business. One important step toward realizing this benefit is to know how things are selling in various parts of the world. A global price comparison service would address this need. But there have not been many such services. In this paper, we use a case study of global price dispersion to illustrate the need and the value of a global price comparison service. Then we identify and discuss several technology challenges, including semantic heterogeneity, in providing a global price comparison service. We propose a mediation architecture to address the semantic heterogeneity problem, and demonstrate the feasibility of the proposed architecture by implementing a prototype that enables global price comparison using data from web sources in several countries.

Biographical notes: Hongwei Zhu is an Assistant Professor of Information Technology at the College of Business and Public Administration, Old Dominion University. He holds a Ph.D. in Technology, Management and Policy from MIT. His research interests include data integration and reuse

technologies, data quality management, data mining, information policy analysis, and information economics.

Stuart Madnick is the John Norris Maguire Professor of Information Technology, Sloan School of Management and Professor of Engineering Systems, School of Engineering at the Massachusetts Institute of Technology. He has been a faculty member at MIT since 1972. He has served as the head of MIT's Information Technologies Group for more than twenty years. He has also been a member of MIT's Laboratory for Computer Science, International Financial Services Research Center, and Center for Information Systems Research. Dr. Madnick is the author or co-author of over 250 books, articles, or reports including the classic textbook, *Operating Systems*, and the book, *The Dynamics of Software Development*. His current research interests include connectivity among disparate distributed information systems, database technology, software project management, and the strategic use of information technology. He is presently co-Director of the PROductivity From Information Technology Initiative and co-Heads the Total Data Quality Management research program. He has been active in industry, as a key designer and developer of projects such as IBM's VM/370 operating system and Lockheed's DIALOG information retrieval system. He has served as a consultant to corporations, such as IBM, AT&T, and Citicorp. He has also been the founder or co-founder of high-tech firms, including Intercomp, Mitrol, and Cambridge Institute for Information Systems, iAggregate.com and currently operates a hotel in the 14th century Langley Castle in England. Dr. Madnick has degrees in Electrical Engineering (B.S. and M.S.), Management (M.S.), and Computer Science (Ph.D.) from MIT. He has been a Visiting Professor at Harvard University, Nanyang Technological University (Singapore), University of Newcastle (England), Technion (Israel), and Victoria University (New Zealand).

Michael Siegel is a Principal Research Scientist at the MIT Sloan School of Management. He is currently the Director of the Financial Services Special Interest Group at the MIT Center For eBusiness. Dr. Siegel's research interests include the use of information technology in financial risk management and global financial systems, eBusiness and financial services, global ebusiness opportunities, financial account aggregation, ROI analysis for online financial applications, heterogeneous database systems, managing data semantics, query optimization, intelligent database systems, and learning in database systems. He has taught a range of courses including Database Systems and Information Technology for Financial Services. He currently leads a research team looking at issues in strategy, technology and application for eBusiness in Financial Services.

1. Introduction

With its increasing connectivity and capability, the World Wide Web has become an important platform for e-business. Many innovative services have emerged to facilitate business transactions on the new platform. One of the increasingly useful services is the price comparison service, which is often used by consumers to find the best deals online and by vendors to make informed decisions on their pricing strategies. *Comparison service providers* are also known as *comparison aggregators* for their capability of transparently aggregating information from multiple web sources (Madnick and Siegel,

2000). They are also called *price comparison agents*, *shopping agents*, and *shopping robots* (or *shopbots* for short) elsewhere; we will use these terms interchangeably in the rest of the paper.

The global connectivity of the Web has yet to be fully exploited. In this paper, we investigate the technology challenges of taking price comparison from a regional level to a global level and develop a solution architecture to address the challenges identified.

1.1. Motivation

Most current price comparison services still only offer regional price comparison, where regional (as opposed to global) information sources are used (Zhu, 2002). We are aware of only one global comparison service, *AddAll.com*, that compares prices of books sold in different countries.

What if a price comparison service is offered on a global basis? Imagine for the moment you are from Sweden and interested in buying a pocket sized digital camcorder. After some research on the Web you decide to buy a SONY MICROMV DCR-IP5, which records video in MPEG format for easy editing on computers and weighs only 0.336 kilograms (i.e., 12 ounces). You use your favorite comparison service *Kelkoo* (at www.kelkoo.se) to find the best deals and it returns information as shown in Figure 1.

<< Figure 1 >>

Among the five vendors, 18,082 Swedish Krona (SEK) is the lowest total price (see data in Totalpris column in Figure 1). Is this the best deal, or is there a substantially better deal, on a global basis? A systematic and extensive (if not exhausting) search is needed to answer this question. Without a global aggregator, this can only be done manually by visiting numerous regional comparison aggregators available in other countries. Our manual exercise found one vendor in the U.S. who ships the product to worldwide destinations at a total price of $1,099.99 ($999.99 plus $100 international shipping charge), as shown in Figure 2.

<< Figure 2 >>

Between 18,082 SEK and $1,099.99, which is the better deal? Information, such as knowing that 1 US dollar is around 10 SEK (at the time), will be useful in answering the question, but such information usually is not readily available from a regional aggregator. After this information is obtained, for example by querying a currency conversion service, such as www.oanda.com, we can use it to convert the prices into the same currency. Only after all these steps have been done do we know that the Swedish offer is 64% more expensive than the U.S. offer.

As we have illustrated so far, there are certain "inconveniences" to be overcome to compare prices globally: a user has to visit and collect data from multiple sites, determine if data needs to be converted, and perform the conversions to make the comparison meaningful. This process is time consuming and error prone. A global comparison service could alleviate the user from these tedious tasks. Such a service provider would need to ensure that the information is properly processed so that data coming from different parts of the world can be correctly interpreted by users who are also geographically dispersed around the world. The users of such services can be consumers (looking for the best deals), economists (studying the global markets), vendors (developing competitive pricing strategies), manufacturers (monitoring vendor pricing behaviors), or arbitragers (buying low at one place and selling high at another).

1.2. Summary of Contributions

There are several technical obstacles that need to be overcome to provide global price comparison services. This paper focuses on identifying these obstacles and presenting a solution. This paper makes the following specific contributions:

- It makes a useful observation about the lack of global price comparison services and illustrates the needs for such services;

- It identifies three semantic heterogeneity problems to be resolved to enable global price comparison. The problems are related to the **D**efinition, **I**dentification, and **R**epresentation (**DIR**) of concepts, entities, and their attribute values;

- It presents a solution architecture to address the DIR problems;

- It demonstrates the feasibility of the solution architecture using a prototype application; and

- It discusses the scalability issue and show that the implementation used in the prototype is scalable in the sense that it requires a significantly smaller number of data conversion programs than that of a more traditional brute-force pair-wise conversion strategy.

1.3. Organization of the paper

The rest of the paper is organized as follows. In section 2, we present a case study showing worldwide price dispersion for the Sony camcorder mentioned earlier. The data for the case study was first assembled by manually integrating data from nearly two dozen regional price comparison aggregators. This manual exercise allowed us to identify the issues to be addressed by a global price comparison aggregator. In section 3, we discuss deficiencies of existing comparison aggregators and technological challenges to providing global price comparison services. In section 4, we present an architecture for dealing with semantic heterogeneity that exists in data from global sources and demonstrate its feasibility using a prototype that can meaningfully compare price data from sources in different countries. Scalability of the implementation of the prototype is also discussed in this section. In section 5, we discuss related work. We present some brief conclusions in Section 6.

2. Case study – global price dispersion for Sony DCR-IP5

The same product can be sold at different prices at different places (or at different times). This phenomenon is known as price dispersion and has been studied by economists because price dispersion is an indicator of market efficiency, or the lack thereof. For readers not familiar with price dispersion, below we provide some background information before presenting our case study.

One of the expectations of the European Union (EU) is to have an efficient integrated market with small price differences among member countries. A survey in the EU (EU Economic Reform, 2001) shows that in the fresh food market "high price countries are often two times more expensive than countries with minimum prices"; even in the

consumer electronics market, one country could be over 50% more expensive than another for a particular product. Data for that study was collected by three consultants who sampled various products in different stores across the EU.

Because comparison aggregation is a great tool for collecting price information, it has been used in a number of price dispersion studies in the U.S. for products such as books, CDs, and consumer electronics. Inter-store price differences were found to be 25-40% (Brynjolfsson and Smith, 2000; Clay et al., 2001; Baye et al., 2004). There are few studies on price dispersion in the global online market. Clay and Tay (2001) studied book prices in the U.S. and the U.K., and showed that a U.S. buyer could save 42% for a particular textbook by purchasing it from the U.K. instead of from the U.S. Maier (2005) found significant price differentials in eBay auction prices among European countries.

As there have been few studies on global price dispersion of the online market, we conducted an empirical study on the SONY digital camcorder mentioned earlier: MICROMV DCR-IP5, which was introduced into the consumer electronics market in early 2002. Market prices for such a new product are volatile; we took a snapshot of global prices by collecting data within 24 hours between March 8 and 9, 2002.

We used a number of regional comparison aggregators to retrieve the prices of the product. These aggregators include *BizRate*, *mySimon*, *Dealtime*, *Shopper*, *PriceRunner*, *PriceGrabber*, *Kelkoo*, and *Kakaku*. Some of the aggregators have country-specific sites to compare prices in those countries. For example, *Kelkoo* currently operates in 11 countries; each *Kelkoo* country-specific site is considered as a regional aggregator. We report our analysis on the unique vendor/price basis within a country. That is, if multiple aggregators in a country report on the same vendor, we treat them as one observation if the prices are the same or within $1 difference. All prices are listing prices not including shipping charges.

We collected 172 observations covering US, Mexico, Brazil, Japan, and nine European countries. Because the product is not "officially" available in Mexico or Brazil, prices found in the two countries are offered by US vendors who charge 52.8% and 61.8% import taxes, respectively. Figure 3 shows the histogram of prices excluding Mexico and Brazil. It is obvious that prices are highly dispersed. Most prices are within the range of $1200-2000 and they are nearly evenly distributed in this range. Prices outside this range exist at both ends.

<< Figure 3 >>

Figure 4 shows the price distribution for all 13 countries, with the number of observations at the bottom. This is a box plot with each box representing 50% of price observations (i.e., the 25% and 75% quartiles) and the line within the box being the median. The ends of the vertical lines indicate the lowest and the highest prices, unless outliers are present in which case the lines extend to a maximum of 1.5 times the inter-quartile range (the height of the box), and the outliers are marked as solid circles. Clearly, prices are different between countries (e.g., prices for the Sony camcorder in the U.S. tend to be lower than those in most of the European countries; this is in contrast to the international book price study (Clay and Tay, 2001) which showed that the UK had lower prices for books.)

<< Figure 4 >>

Prices are also dispersed within a country. Let us look at US prices in more detail, shown in Figure 5. These 53 unique price observations do not include SonyStyle US, Sony's online store in the U.S., and major consumer electronics vendors like BestBuy and CircuitCity, which offer the product at the same "official" price: $1299.99. We can

see from the figure most prices are at or below this price level. The average price is $1203, which is 7.7% below the "official" price. The U.S. average price is 26.3% lower than the worldwide average.

<< Figure 5 >>

Although the global price dispersion found in the case study is interesting, the main purpose of the case study is to "experience the inconvenience" of comparing prices globally using existing tools, and thereby identify the challenges to be addressed by a global aggregation service. In addition to currency differences in the data reported by different regional aggregators, there are other differences that need to be reconciled to make the comparison meaningful. With regional aggregators, obtaining "raw" price data from different countries was indeed easy, which took us only *a few hours* for the case study. But understanding the data, identifying subtle differences in the meaning of the data and subsequently reconciling these differences were much more tedious and took us *more than a week*. That is, it can take more than a week to normalize the price data of a particular product sold in various parts of the world, even though the raw data for answering the question is available "at our fingertips".

Next, we will examine the deficiencies of existing comparison aggregators and identify technological challenges to advancing from regional price comparison to global price comparison. Then we discuss an enabling technology that can make global price comparison much easier.

3. Technology challenges to global price comparison

3.1. Characteristics and deficiencies of existing price comparison

Many price comparison service providers have emerged since the introduction of the first online price comparison agent, which was implemented as an internal project of Anderson Consulting in 1995 (Smith, 2002). We listed a few such service providers in the case study discussed above. As noted earlier, except for AddAll.com, all the other service providers only compare prices within a country. This is the case even for those that have an international presence. This situation makes global price comparison difficult because manual data integration has to be performed.

Most existing regional price comparison services are implemented using information extraction techniques (Chang et al., 2006) to extract data from web sources. These techniques often rely on the syntactic structures of web pages (Fasli, 2006) to obtain the raw data from multiple web sites, including those non-cooperative sites. When these techniques are used for regional comparison services, the direct use of raw data from multiple sites usually does not create a problem of understanding or comparing the data because the data sources usually report data using the same convention that is understood by the customers in the region. For example, most online vendors in the U.S. list prices in USD, excluding taxes and shipping charges. But when the service is provided on the global basis, there will be significant difficulties of understanding the data because the sources from different parts of the world use different conventions to report data, whereas the users of such a global system are familiar with their local conventions only. To make the information from diverse sources meaningful to equally diverse customers, a

comparison service provider must address the data semantics issues that we will discuss next.

3.2. *Semantic heterogeneity of global web sources*

We have seen in the motivational example that currency needs to be converted to make sensible comparison for the Swedish user. Many other such issues also exist. Let us illustrate these issues with an example of information about laptop computers from several web sites of Sony, summarized in Table 1. We will ignore most language differences in the following discussion.

<< Table 1 >>

First to note is that not all information is available at a single source. In this case the thickness information is not immediately available from the U.K. sources (it is buried in a PDF document). If an aggregator takes the information from the U.S. source and directly reports to its German users, 1.09" probably would not be helpful to users who are only familiar with the metric system for measurement. In addition to different units being used (lbs vs. kilograms, inches vs. millimeters, US dollars vs. British Pounds, etc.) there are other representational differences, such as symbols for thousands separator and decimal point. These differences have to be reconciled for the users.

There is a more complicated problem in the data shown in Table 1. The last row shows pricing information for the product. Aside from representational differences, we notice that different sources use different definitions for what appears to be same notion. In the case of price, different definitions will cause differences in what components are included in the price.

Price, however simple as it appears, is in fact a complicated concept that has different meanings from different perspectives: nominal price, tax-included price, price indicating cost of acquiring the item, to name only a few. How much an item costs for someone to acquire is often different from the listed price because the listed price may not include other costs associated with this transaction, such as taxes, duties if it involves international trade, shipping and handling, etc. An accurate calculation for price in the sense of "cost to acquire" could be very complicated in the context of global e-business. Calculation of VAT alone requires additional information because VAT varies depending on the type of product, origination, destination, and special treaties between regions. The variations range from 0 to 25% of the listing price in European countries. This makes aggregation and meaningful comparison difficult. A classic example of this problem is given in McCarthy and Buvac (1994), where different prices of the same GE aircraft engine are perceived by different organizations, such as the U.S. Air Force and U.S. Navy depending on whether the price includes spare parts, warranty, etc.

Another problem not explicitly shown in Table 1 is how the aggregators identify the same product from different regions. In the process of manually composing the Table, we noticed that the model numbers are different between laptops in the U.S. and those in Europe. We recognize their similarity (in this case identity except for the model numbers) by examining the configurations (e.g., CPU speed, hard disk capacity, weight, etc.). The fact that manufacturers often market the same product with different names in different regions (Bergan, et al., 1996) makes it difficult for the aggregator to recognize their identity. This problem is best described in the following Camera example from Focuscamera.com:

> "... a USA Minolta Maxxum is a Minolta Dynax overseas, the USA Canon
> EOS Rebel 2000 is an EOS 300 overseas, Pentax IQ Zooms are Pentax Espios
> overseas, etc."

Conversely, when models with different features are named the same or slightly differently in different regions, aggregators sometimes cannot recognize the distinction. In the Sony DCR-IP5 case study we found that some vendors label the product as DCR-IP5E to indicate that it is an international model compatible with the PAL standard rather than the NTSC standard in the U.S. What makes it worse is that most vendors use DCR-IP5 for both the NTSC model and the PAL model. Although this does not cause big problems because of its common MPEG recording format, for other types of products this could be an issue.

The preceding discussions can be summarized into three **DIR** issues related to semantic heterogeneity:

- **D**efinition – a general concept having different definitions, e.g., base price vs. tax-included price

- **I**dentification – same entity being named or described differently, e.g., Minolta Maxxum vs. Minolta Dynax.

- **R**epresentation – the same information being represented differently, e.g., using different currencies and different thousands delimiters, such as 1.250 vs. 1,250.

A global comparison service provider has to reconcile the DIR differences between the data sources and the data users. The provision of the reconciliation is known as an n^2 problem because when n different parties attempt to exchange data, $n(n-1)$, which is $O(n^2)$, conversion programs need to be implemented to ensure the correct interpretation of data by the receiving party. This is not scalable because the number of conversions to be provided and maintained grows quickly as n becomes large.

Next, we will propose an architecture that aims to address these issues.

4. Enabling technology for global comparison services

The adoption of XML data standards (Madnick, 2001) and the emergence of Web services (Curbera, et al., 2003) will make it much easier to obtain the raw data from various sources. But semantic heterogeneity will continue to exist because of cultural diversity and different user preferences (e.g., prices will be quoted in local currencies until a global currency is adopted, which will not happen in any foreseeable future). The Semantic Web (Berners-Lee, et al., 2001) initiative aims to develop an architecture of the future Web and a set of technologies to represent and reason with data semantics. But the development and wide deployment of Semantic Web technologies will take some time. Furthermore, the Semantic Web does not directly address the n^2 problem. Thus we need a near-term solution to address the issues identified in the previous section.

4.1. Mediation architecture

The concept of a mediation architecture was proposed more than a decade ago (Wiederhold, 1992) to enable applications to use heterogeneous data sources. This

architecture is well suited for enabling global price comparison. Below we present a generic mediation architecture and discuss how it can be used for global comparison.

<< Figure 6 >>

We will use Figure 6 to illustrate the architecture. A software system known as a mediator exists between the data sources and the users (or user applications) to provide three services with which the users can obtain information from data sources. Each service may or may not use a certain data model to facilitate its task. Details of each service are described below.

Wrappers. Wrappers are software components that provide the other mediation services with a uniform access to heterogeneous data sources and are responsible for extracting data from sources using source-specific protocols and extraction rules. A data model, usually in the form a data schema (e.g., a relational schema), is used to provide a uniform view of the data in various sources. A schema is needed because sources can have different schemas, or they may not have an explicit schema, such as in the case of semi-structured or unstructured sources (e.g., web pages). The uniform schema is usually application specific. Source-specific extraction rules can be provided manually as in Cameleon (Firat et al., 2000) and TSIMIS (Garcia-Molina et al., 1995), semi-automatically as in STALKER (Muslea et al., 2001), or automatically as in RoadRunner (Crescenzi et al., 2001). Recent research has been focusing on how to automatically and reliably generate data extraction rules for semi- or un-structured data sources.

Entity resolution/record linkage. This service addresses the identification issue. Information about the same entity (e.g., a person) or the same kind of entity (e.g., a SONY DCR-IP5 camcorder) often appears in different forms in disparate sources, making it difficult to identify and link them. The service of entity resolution, also known as record linkage or inter-database instance identification (Madnick and Wang, 1989), uses a software component to identify data about the same entity in disparate sources. Many entity resolution techniques measure a certain similarity (or conversely a certain distance) of different records referring to the same entity (Winkler, 2006; Birzan and Tansel, 2006). Recent research has explored the use of an entity ontology (a collection of known entities and their attribute values) to identify these records (Michelson, 2005; Michelson and Knoblock, 2007). For purpose of comparison shopping, it is desirable to identify the same kind of entity (e.g., any SONY DCR-IP5) instead of the same entity (e.g., a particular SONY DCR-IP5 identifiable perhaps by a serial number). In this paper, we do not distinguish between "same entity" and "same kind of entity", and loosely use "same entity" to refer to both cases.

Context mapping. This service addresses the definition and representation issues. Both the sources and the users often make different semantic assumptions that affect the interpretation of data, in which case we say the sources and the users are in different *contexts*. For example, in one context, the price may be reported in USD, using dot (.) as the decimal point, and not including taxes or shipping charges, yet in another context the price may be reported in Euros, using comma (,) as the decimal point, and including 15% taxes but not including shipping charges. The context mapping service takes the semantic assumptions into account to appropriately transform data from source contexts to the user context.

Figure 7 uses an example to illustrates how this mediation architecture can be applied to implementing global price comparison services.

<< Figure 7 >>

Users and sources. Different users, shown in the top portion of Figure 7, have different needs. For illustration purposes, we show two users: one wants prices, in USD, of a product represented using a black dot, the other wants prices, in Japanese Yen, of a product represented using a hollow circle. The sources, shown in the bottom portion of Figure 7, can be from anywhere in the world and in different forms (e.g., HTML web pages, web services, or relational databases, etc.). The users and the sources may share the same context or be in different contexts.

Wrappers. Within the mediator, a wrapper is created for each source. The wrappers communicate with the sources using source-specific protocols. The output of the wrappers uses a uniform data schema (e.g., a relation of three attributes: <product, vendor, price>, as alluded to in Figure 7). The schema is instantiated with data extracted from each source. The data extracted, although organized according to the uniform schema, is still its original form without any syntactic or semantic transformation (e.g., if the source reports prices in Euros, the extracted prices are in Euros regardless of user preference).

Entity resolution. In Figure 7, we use different symbols to represent different entities. The actual representations of an entity can be different across sources (e.g., using "Sony DCR-IP5", "DCR-IP5", or "MicroMV IP5" to represent a SONY MICROMV DCR-IP5). The entity resolution service identifies these different representations for the same entity and establishes the linkage of these records.

Contexts and context mapping. To simplify explication, we only show different currencies used for price in Figure 7. Later we will introduce more context differences. Different contexts need to be recorded only once before using the service or if there is a context change. When the mediator receives a user request, it compares source and user contexts; if there is a mismatch, it invokes a conversion to reconcile the difference (e.g., using a currency conversion to convert prices in Euros to USD or Japanese Yen). This mediator service ensures that users always receive data that can be correctly interpreted in their contexts, without the need to manually consult other data sources and perform manual data conversions. In the example, the user in the USA receives price data in USD (thanks to the context mapping service) for the desired product (thanks to entity resolution service), similarly for the user in Japan.

4.2. Prototype demonstration

We have developed a prototype to demonstrate the feasibility of the proposed architecture. The prototype uses the COntext INterchange (COIN) technology (Goh et al., 1999; Firat, 2003) to implement the context mapping service, and the Cameleon web wrapper engine (Firat et al., 2000) to allow web sources to be queried as if they were relational databases. Data sources used in the prototype are regional price comparison aggregators, such as those mentioned in the case study. To a large extent, regional aggregators have performed entity resolution tasks for various products offered by different vendors. Therefore, we did not include an entity resolution component in the prototype.

In this demonstration, we focus on price data, which may have different contexts in such aspects as domestic and international taxation, shipping charges, and currency. Example contexts are given in Table 2. In addition, conversion functions are provided to deal with potential differences in each aspect.

<< Table 2 >>

We will use an example query to show how the prototype system can help users such as the Swedish buyer mentioned earlier to perform global price comparison.

The Swedish buyer is interested in knowing the total cost of the camcorder from worldwide vendors. For illustration purposes, we simplify the data schema of the wrappers to only include <seller, price> for the SONY DCR-IP5 camcorder. The buyer can issue a query to compare prices of vendors in multiple countries reported by regional price comparison aggregators (such as kelkoo, pricerunner, etc.) using a predefined SQL, compare_all:

Select seller, price from kelkoofrance union	//French source
Select seller, price from pricerunnersweden union	//Swedish source
Select seller, price from pricerunneruk union	//UK source
Select seller, price from cnetshopper union	//US source
... //etc.	

As illustrated in the sample contexts, differences exist between the sources and the user. The COIN reasoner is designed to determine these differences (from the context definitions) and reconcile them by revising the original query to incorporate necessary conversion functions. This process generates mediated queries that perform all the necessary conversions from source context to user context. Some of the conversions that the system automatically generates are given in Table 3.

<< Table 3 >>

Within the COIN system, the original input SQL query is translated into a DATALOG (Ceri et al., 1989) query representation which the reasoner operates upon to generate the mediated query, also in DATALOG. The mediated DATALOG query can be displayed as a SQL query for inspection. Assuming that multiples sources are to be used, the mediated query is divided into sub-queries, one for each source, which, when possible, are executed in parallel by the executioner. The following shows the final mediated query, in SQL format, generated by the system to answer the user's initial query. We hope that readers can examine this and be convinced that all necessary conversions are indeed performed by the following query. In order to accomplish the conversions, sometimes auxiliary data sources are used. In this example, olsen (see www.oanda.com) is an auxiliary online source that provides current and historical currency exchange rates; the system uses current date (*i.e.*, date when the query is issued).

```
//French source. Deduct 19.6% French tax; add 25% Swedish tax;
//add €80 int'l shipping; convert Euros to Krona
select kelkoofrance.seller,
((((kelkoofrance.price/1.196)+((kelkoofrance.price/1.196)*0.25))+80)*olsen.rate)
from (select seller, price
   from   kelkoofrance) kelkoofrance,
//find exchange rate using auxiliary source
(select 'EUR','SEK',rate,'11/01/02' from olsen
     where  exchanged='EUR'
     and    expressed='SEK'
     and    date='11/01/02') olsen
union
//Swedish source. Add 20 Krona domestic shipping
```

```
select pricerunnersweden.seller, (pricerunnersweden.price+20)
from (select seller, price
    from  pricerunnersweden) pricerunnersweden
union
//UK source. Deduct 17.5% UK tax; add 25% Swedish tax;
//add £35 int'l shipping; convert Pounds to Krona
Select pricerunneruk.seller,
((((pricerunneruk.price/1.175)+((pricerunneruk.price/1.175)*0.25))+35)*olsen.rate)
from (select seller, price
    from  pricerunneruk) pricerunneruk,
//find exchange rate using auxiliary source
(select 'GBP','SEK',rate,'11/01/02' from olsen
     where  exchanged='GBP'
     and    expressed='SEK'
     and    date='11/01/02') olsen
union
//US source. Add 25% Swedish tax; add $100 int'l shipping;
// convert USD to Krona
select cnetshopper.seller,
(((cnetshopper.price+(cnetshopper.price*0.25))+100)*olsen.rate)
from (select seller, price
    from  cnetshopper) cnetshopper,
//find exchange rate using auxiliary source
(select 'USD','SEK' rate '11/01/02' from  olsen
     where  exchanged='USD'
     and    expressed='SEK'
     and    date='11/01/02') olsen
union
...
```

An excerpt of the query execution results is shown in Table 4 (reformatted from prototype output). All prices have been translated into the context of the Swedish user, who can easily compare them on the same basis. Finding the best deal globally is now as simple as clicking the predefined query with the help of this prototype of global comparison aggregation services.

    << Table 4 >>

    We demonstrated the case showing how the application works for one particular receiver context. The application also works for any user whose context is one of the example contexts (both source contexts and receiver contexts) in Table 2. Due to space limitation, we do not show the other cases here.

4.3. Prototype implementation: COIN as a solution to the n^2 problem

In this section, we provide an overview of the COIN technology and show how is used to implement the global comparison service demonstrated earlier. We also discuss how it solves the n^2 problem.

The COIN mediator consists of the application specific as well as generic components, as shown in Figure 8. The application-specific components are created only once during initial configuration of an application, or when a new user/source is added or context of a user/source has changed. After this has been done, a user can query the sources as if all sources were in the user contexts. The generic components intercept user queries, compare the user contexts with the contexts of the sources involved; if there is any difference, they rewrite user queries to generate mediated queries that incorporate necessary conversions to reconcile context differences, and execute the mediated queries to extract data from the sources and transform the data into the user context. The generic components can be used for other applications when it is supplied with the corresponding application-specific components. Below we briefly describe each component.

<< Figure 8 >>

Shared ontology. It models the application domain using a collection of semantic types (which corresponds to the high level concepts in the domain) and their relationships. It uses a "lightweight ontology" approach (Zhu and Madnick, 2006a), which we explain later. Figure 9 gives a graphical representation of the ontology used in the prototype. A type can be related to another in three ways: 1) as a subtype or supertype (e.g., *price* is a subtype of *monetaryValue*); 2) as a named attribute (e.g., *price* is the *prodPrice* attribute of *product*); and 3) as a modifier or contextual attribute, whose value is specified in context descriptions and can functionally determine the interpretation of instances of the type that has this modifier (e.g., *monetaryValue* type has a *currency* modifier). There is a modifier-free type *basic* that serves as the supertype of all the other types in the ontology. A subtype recursively inherits the attributes and modifiers of its supertypes.

<< Figure 9 >>

Context descriptions. They are the declarative descriptions of the representational and definitional variants of the high level ontological concepts (e.g., *price* can be quoted in different currencies). The descriptions can also include implicit assumptions (e.g., domestic tax rates) made by the data sources and the data receivers (which can be users or user applications). As mentioned earlier, a context is described by specifying values for the modifiers. Two contexts are different if there is at least one modifier that has different values. Labels (e.g., C1-C4 in Figure 8) can be used to identify different contexts. In addition, the context descriptions also include conversion rules for each modifier between different modifier values (e.g., rules that specify how to convert price from one currency to another). We call such conversions *component conversions*.

The correspondence between the data elements (i.e., the table fields) in the data sources and the ontological concepts are established via declarations. Each data element is also associated with a context via declarations. These declarations as well as the ontology and the context descriptions are expressed using a logic formalism of the F-logic (Kifer, et al., 1995) family.

Wrappers. The wrapper specifications consist of source-specific schema declarations and data extraction rules for each data element. Details of wrapper specification can be found in Firat, et al. (2000). The actual wrappers are produced by the Cameleon engine (discussed later) using the wrapper declarations.

COIN reasoner. This is a query rewriting engine that allows users to issue queries against sources without concerning context differences. The query it generates, called the mediated query, incorporates necessary conversions to reconcile context differences between the sources involved and the user. Context differences are determined by

comparing the modifier assignments between the source context and the user context for each concept involved.

Planner/Optimizer/Executioner (*POE*). It takes the mediated query as the input, generates a query execution plan that considers source capability constraints, optimizes the plan by imposing an execution order of sub-queries and employing parallel execution when possible, and subsequently executes the plan to produce the dataset interpretable in the user context. Details of POE can be found in Alatovic (2002). The parallelism significantly reduces the execution time. In the prototype, the execution time is determined by the regional aggregator with the longest response time, and it is not dependent on the number of regional aggregators used.

Cameleon engine. It uses the wrapper specifications to produce the source-specific wrappers and provide an SQL interface to other components that use the wrappers.

These components work together to underpin the global comparison prototype.

The COIN approach solves the n^2 problem by using a lightweight ontology and a reasoner that can dynamically compose composite conversions using a small number of component conversions. We briefly describe these mechanisms here, the detailed of which can be found in Zhu and Madnick (2006a).

A lightweight ontology includes only high level concepts (e.g., *price*) as opposed to their well-specified variants (e.g, "*price in USD not including taxes or shipping charges*", "*price in Euros including 15% taxes but not including shipping charges*", etc.). Imagine how big the ontology would be had we included all possible variants of *price* in Figure 9. As an artifact, a lightweight ontology is much easier to create than a well-specified ontology. The ambiguity of the high level concepts is removed by the context descriptions. In other words, the well-specified ontology can be derived from the lightweight ontology and the context descriptions.

The lightweight ontology provides the vocabulary and a structure of describing contexts. The reasoner exploits the structure to dynamically compose conversions. The conversions specified for each modifier between different modifier values are called *component conversions*. An ontological concept may have multiple modifiers. A *composite conversion* is composed by the reasoner to incorporate the component conversions, if necessary, of all modifiers of a concept. A simplification of the algorithm is described in Figure 10 and is implemented using abductive constraint logic programming (ACLP) (Kakas, et al., 2000) and constraint handling rules (CHR) (Frühwirth, 1998).

In the worse case scenario, the number of predefined component conversions required by the COIN approach is:

$$\sum_{i=1}^{m} n_i (n_i - 1)$$

where n_i is the number of unique values that the ith modifier has in all contexts, m is the number of modifiers in the lightweight ontology.

While the formula appears to be n^2, it is fundamentally different from the approach that requires the manual creation of *comprehensive conversions* between each pair of the parties engaging in data exchange. The supplied conversions needed in COIN are only the component conversions, which are much simpler than the comprehensive conversions that consider the differences of all data elements in all aspects between two parties.

Furthermore, the number of component conversions is significantly smaller than the number of pair-wise comprehensive conversions. This can be illustrated by considering a particular scenario we have studied where there are 50 data sources, each having its own context, and the users share contexts with their preferred data source. In the lightweight ontology in Figure 9, the *price* concept has three modifiers. Concept *country* also has two modifiers, but they are used to describe the tax rates used by the component conversions of *location* modifier. No component conversions for these two modifiers are needed in the prototype. Assume the 50 sources are located in 10 countries with different currencies, and there are four *types* of prices (e.g., base price, price with domestic taxes included, price with domestic taxes and shipping charges, and price with all taxes and shipping charges). The pair-wise approach requires 2450 (i.e., 50*49) predefined comprehensive conversions, the COIN approach requires only 192 (i.e., 10*9+10*9+4*3) component conversions. All the 2450 comprehensive conversions can be composed dynamically, as needed, using the 192 component conversions by the COIN reasoner. As explained in (Zhu and Madnick, 2006a), the actual number of component conversions can be considerably less if they can be parameterized – for example, the olsen web source can be used to create a general conversion between any two currency values.

5. Related Work and Remarks

As discussed earlier, most existing commercial price comparison services are provided only regionally and regional sources usually do not have representational and definitional semantic problems that are pervasive in global data sources. Most commercial comparison service providers use a product catalog, which can be considered as an entity ontology, to address the identification problem. We have mentioned a few commercial providers and used them as the data sources in the global comparison prototype. Brief reviews of other commercial providers can be found Smith (2002) and Fosli (2006).

There have been a few research prototypes of price comparison. As in the commercial case, the research prototypes that we are aware of do not address the semantic heterogeneity issues discussed here. The PriceBot agent (Doorenbos, et al., 1997) uses inductive learning techniques to learn how to extract product and price data from web sites about a certain category of products. The WhereBuy shopping broker (Santos, et al., 2001) uses product catalogs to identify the products offered by different vendors; it subsequently extract the prices (the raw data) for price comparison. The IPIS (Intelligent Product Information Search) system (Kim, et al., 2005) focuses on the product identification issue. It uses ontology matching techniques to rewrite queries to match products of vendors that may use different category schemes (e.g., Television vs. TV&HDTV) and different attribute names (size vs. diagonal).

We have focused on price comparison in this paper. But comparison can involve other dimensions to serve different purposes (e.g., compare other features, such as size and weight, compare vendor reputation, or assist buyers to make decisions). A classification scheme of different comparison agents can be found in Wan, et al., (2003). The architecture presented here can address DIR issues that arise when these various agents extract and integrate data from multiple data sources.

In addition to the technical issues discussed here, there are other issues that concern various stakeholders of price comparison. There issues, some of which are summarized in Smith (2002), include the strategy of comparison service providers (Madnick and Siegel,

2002), pricing strategies (Koças, 2005), data reuse regulations (Zhu and Madnick, 2006b).

6.   Conclusions

Despite the increasing presence of comparison aggregation, most of these services are offered regionally, not globally. There are substantial price differences in the global market. Our price dispersion case study shows that the worldwide prices for DCR-IP5, a Sony digital camcorder, can differ by nearly three times. A global aggregator can help us to better understand this dispersion.

With this motivation, we present a mediation architecture to address data semantics issues in global aggregation. A prototype global aggregator has been developed to validate the architecture. The technologies used here show promising signs for enabling global comparison aggregation services. These new services will benefit a variety of users. They can help consumers find the best deals around the world; they can also assist researchers and policy makers to systematically and efficiently collect global market data at low cost (recall that the E.U. price dispersion survey mentioned in section 2 relied on three consultants who visited stores to manually collect retail prices); manufacturers can also use the services to find out the actual retail prices of their products around world, with which they can better assess demand and set appropriate wholesale and suggested retail prices. The emergence and the wide usage of global aggregation services will make the web the truly efficient platform for e-business.

References

Alatovic, T. (2002) 'Capabilities Aware Planner/Optimizer/Executioner for Context Interchange Project', M.S. thesis, MIT.

Bailey, J.P. (1998) Intermediation and Electronic Markets: Aggregation and Pricing in Internet Commerce, Ph.D. thesis, MIT.

Baye, M.R., Morgan, J. and Scholten, P. (2004) 'Price Dispersion in the Small and in the Large: Evidence from an Internet Price Comparison Site', *Journal of Industrial Economics*, Vol. 52, No. 4, pp. 463-496.

Berners-Lee, T., Hendler, J. and Lassila, O. (2001) 'The Semantic Web', *Scientific American*, Vol. 284, No. 5, pp.34-43.

Bergan, M., Dutta, S. and Shugan, S.M. (1996) 'Branded Variants: A Retail Perspective', *Journal of Marketing Research*, Vol. 33, No. 1, pp. 9-19.

Birzan, D.G., Tansel, A.U. (2006) 'A Survey of Entity Resolution and Record Linkage Methodologies', *Communications of the IIMA*, Vol. 6, No. 3, pp. 41-50.

Brynjolfsson, E. and Smith, M.D. (2000) 'Frictionless Commerce? A comparison of the Internet and Conventional Retailers', *Management Science*, Vol. 46, No. 4, pp. 563-585.

Ceri, S., Gottlob, G. and Tanca, L. (1989) 'What You Always Wanted to Know About Datalog (And Never Dared to Ask)', *IEEE Transactions on Knowledge and Data Engineering*, Vol. 1, No. 1, pp. 146-166.

Chang, C.H., Kaye, M., Girgis, M.R. and Shaalan, K.F. (2006) 'A Survey of Web Information Extraction System', *IEEE Transactions on Knowledge and Data Engineering*, Vol. 18, No. 10, pp. 1411-1428.

Clay, K., Krishnan, R. and Wolff, E. (2001) 'Prices and Price Dispersion on the Web: Evidence from the Online Book Industry', *Journal of Industrial Economics*, Vol. 49, No. 4, pp. 521-539.

Clay, K. and Tay, C.H. (2001) 'Cross-Country Price Differentials in the Online Textbook Market', Working Paper, Carnegie Mellon University.

Crescenzi, V., Mecca G. and Merialdo, P. (2001) 'RoadRunner: Towards Automatic Data Extraction from Large Web Sites', *Proceedings of the 27th International Conference on Very Large Databases* (*VLDB 2001*), pp. 109-118.

Curbera, F., Khalaf, R., Mukhi, N., Tai, S. and Weerawarana, S. (2003) 'The Next Step in Web Services', *Communications of the ACM*, Vol. 46, No. 10, pp. 29-34.

Doorenbos, R.B., Etzioni, O. and Weld, D.S. (1997) 'A Scalable Comparison-Shopping Agent for the World-Wide Web', *Proceedings of the First International Conference on Autonomous Agents* (*Agents'97*), pp. 39-48.

EU Economic Reform (2001) 'Price Dispersion in the Internal Market', available at http://ec.europa.eu/internal_market/economic-reports/docs/pricestudy_en.pdf.

Fasli, M. (2006) 'Shopbots: A Syntactic Present, A Semantic Future', *IEEE Internet Computing*, Vol. 10, No. 6, pp. 69-75.

Firat, A., Madnick, S. and Siegel, M. (2000) 'The Cameleon Web Wrapper Engine', *Proceedings of the Workshop on Technologies for E-Services*, September 14-15, 2000, Cairo, Egypt.

Firat, A. (2003) 'Information Integration using Contextual Knowledge and Ontology Merging', Ph.D. thesis, Sloan School of Management, MIT.

Frühwirth, T. (1998) 'Theory and Practice of Constraint Handling Rules', *Journal of Logic Programming*, Vol. 37, No. 1-3, pp. 95-138.

Garcia-Molina, H., Hammer, J., Ireland, K., Papakonstantinou, Y., Ullman, J. and Widom., J. (1995) 'Integrating and Accessing Heterogeneous Information Sources in TSIMMIS', AAAI Symposium on Information Gathering, Stanford, California, 61-64.

Goh, C.H., Bressan, S., Madnick, S. and Siegel, S. (1999) 'Context Interchange: New Features and Formalisms for the Intelligent Integration of Information', *ACM Transactions on Information Systems*, Vol. 17, No. 3, pp. 270-293

Madnick, S.E. (2001), 'The Misguided Silver Bullet: What XML will and will NOT do to help Information Integration,' *Proceedings of the Third International Conference on Information Integration and Web-based Applications and Services* (IIWAS2001), September 2001, Linz, Austria, pp. 61-72.

Kiffer, M., Laussen, G. and Wu, J. (1995) 'Logic Foundations of Object-Oriented and Frame-based Languages', *Journal of the ACM*, Vol., 42 No. 4, pp. 741-843.

Kim, W., Choi, D. and Park, S. (2005) 'Product Information Meta-search Framework for Electronic Commerce Through Ontology Mapping', *The Semantic Web: Research and Applications*, LNCS 3532, pp. 408-422

Koças, C. (2005) 'A Model of Internet Pricing Under Price-Comparison Shopping', *International Journal of Electronic Commerce*, Vol. 10, No. 1, pp. 111-134.

Madnick, S.E. and Siegel, M.D. (2002) 'Seizing the Opportunity: Exploiting Web Aggregation', *MISQ Executive*, Vol. 1, No. 1, pp. 35-46.

Madnick, S.E. and Wang Y.R. (1989), 'The Inter-Database Instance Identification Problem in Integrating Autonomous Systems', *Proceedings of the Fifth International Data Engineering Conference*, February 1989, Los Angeles, CA.

Maier, P. (2005) 'A "Global Village" without borders? International price differentials at eBay', Netherlands Central Bank Working Paper, #044.

McCarthy, J. and Buvac S. (1994) 'Formalizing Context (Expanded Notes)', Stanford University.

Michelson, M.J. (2005) 'Building Queryable Datasets from Ungrammatical and Unstructured Sources', M.S. thesis, University of Southern California.

Michelson, M.J. and Knoblock, C.A. (2007) 'An Automatic Approach to Semantic Annotation of Unstructured, Ungrammatical Sources: A First Look', *IJCAI'07 Workshop on Analytics for Noisy Unstructured Text Data*, January 8, Hyderabad, India, pp. 123-130.

Muslea, I., Minton, S. and Knoblock, C. (2001) 'Hierarchical Wrapper Induction for Semistructured Information Source', *Journal of Autonomous Agents and Multi-Agent Systems*, Vol. 4, No. 1-2, pp. 93-114.

Santos, S. C., Angelim, S. and Meira, S. R. (2001) 'Building Comparison-Shopping Brokers on the Web', *Proceedings of the Second international Workshop on Electronic Commerce* November 16-17, L. Fiege, G. Mühl, and U. G. Wilhelm, Eds. LNCS 2232, pp. 26-38.

Smith, M.D. (2002) 'The Impact of Shopbots on Electronic Markets', *Journal of the Academy of Marketing Science*, Vol. 30, No. 4, pp. 446-454.

Wan, Y., Menon, S. and Ramaprasad, A. (2003) 'A Classification of Product Comparison Agents', *Proceedings of International Conference on Electronic Commerce* (*ICEC'03*), Sept. 30 – Oct. 3, Pittsburgh, PA.

Wiederhold, G. (1992). 'Mediators in the Architecture of Future Information Systems.' *Computer*, Vol. 25, No. 3, pp. 38-49.

Winkler, W.E. (2006) Overview of Record Linkage and Current Research Directions, Research Report, Statistics #2006-2, US Census Bureau, available at http://www.census.gov/srd/papers/pdf/rrs2006-02.pdf.

Zhu, H. (2002) 'A Technology and Policy Analysis for Global E-Business', M.S. thesis, Massachusetts Institute of Technology.

Zhu, H. and Madnick, S. (2006a) 'A Lightweight Ontology Approach to Scalable Interoperability', *VLDB Workshop on Ontologies-based techniques or DataBases and Information Systems* (*ODBIS'06*), September 11, 2006, Seoul, Korea.

Zhu, H. and Madnick, S.E., (2006b) 'Reutilization and Legal Protection of Non-Copyrightable Database Contents', *Proceedings of the Fourth IASTED International Conference on Law and Technology* (*LawTech'06*), October 9-11, Cambridge, MA, USA.

Table 1 Information from multiple sources

	U.S.	*U.K. (in English)*	*U.K. (in German)*
Weight	2.76 lbs	1.26 kg	1,26 kg
Thickness	1.09"	NA	NA
Price	$2,029 plus $25 shipping	1,699.00 GBP incl. VAT	1.699,00 GBP inkl. MwSt.

Table 2 Sample contexts for price

		Currency	*Tax*	*Shipping*[+]
Source contexts	*France*	Euro	Included, 19.5%	Domestic: 15 Int'l: 80
	Sweden	Krona	Included, 25%	Domestic: 20 Int'l: 800
	UK	Pound	Included, 17.5%	Domestic: 10 Int'l: 35
	US	USD	Excluded	Domestic: 50 Int'l: 100
Receiver contexts	*US, Base*	USD	Excluded	Excluded
	US, Cost	USD	If domestic vendor, no tax; otherwise, add 3% import tax	Include domestic or int'l shipping accordingly
	Sweden, Cost	Krona	Include 25% tax regardless	int'l shipping accordingly

[+]: Assume vendors only distinguish between domestic and interchange shipping charges. This can be refined to use online shipping inquiry services to calculate shipping costs by supplying product's dimensions and weight.

Table 3 Appropriate conversions for reconciliation of context differences

Source	*Conversion*
France	Deduct 19.5% French tax, add 25% Swedish tax, add €80 international shipping, convert Euros to Krona
Sweden	Add 20 Krona domestic shipping
US	Add 25% Swedish tax, add $100 international shipping, convert USD to Krona
UK	Deduct 17.5% UK tax, add 25% Swedish tax, add £35

Table 4   Excerpt of results in user's context

Source	Seller	Price (i.e. total cost in Krona)
Sweden	Foto & Elektronik AB	15815
	Expert Citybutiken/Konserthuset	16015
	…	…
	Click ontime	23470
…	…	
US	Bridgeviewphoto.com	10255
	PC-Video Online	10594
	…	…
	Circuit City	14933

Figure 1   Prices for DCR-IP5 in Sweden

Figure 2   An offer for DCR-IP5 from the U.S.

Figure 3   Price histogram excluding Mexico and Brazil (N=128)

Min	783
Max	2254
Median	1569
Mean	1524
Standard Deviation	335
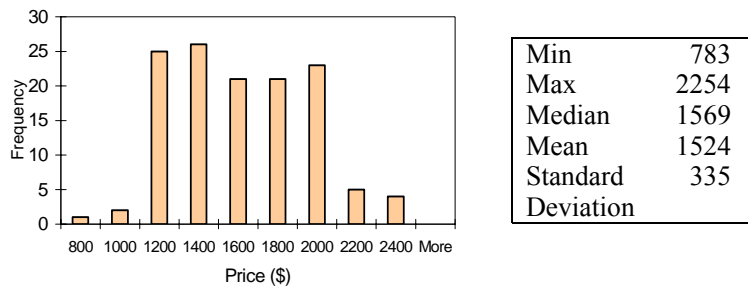
Figure 4   Price distribution in different countries
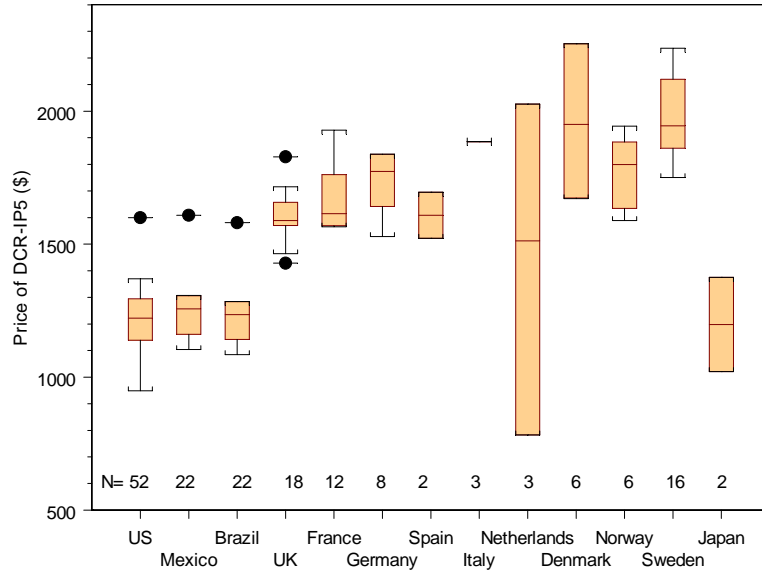
Figure 5   Price distribution in the U.S. (N=52)
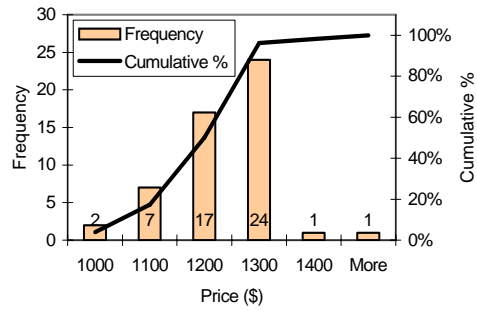
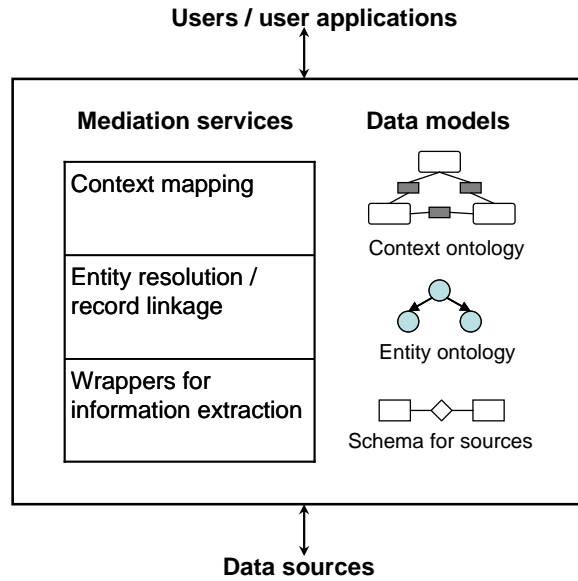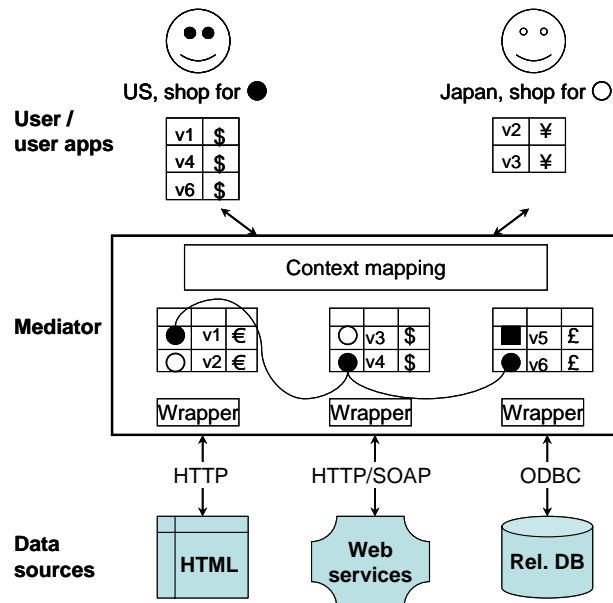**Figure 6** Mediation architecture for effective use of heterogeneous data

Figure 7 Mediation architecture for global price comparison

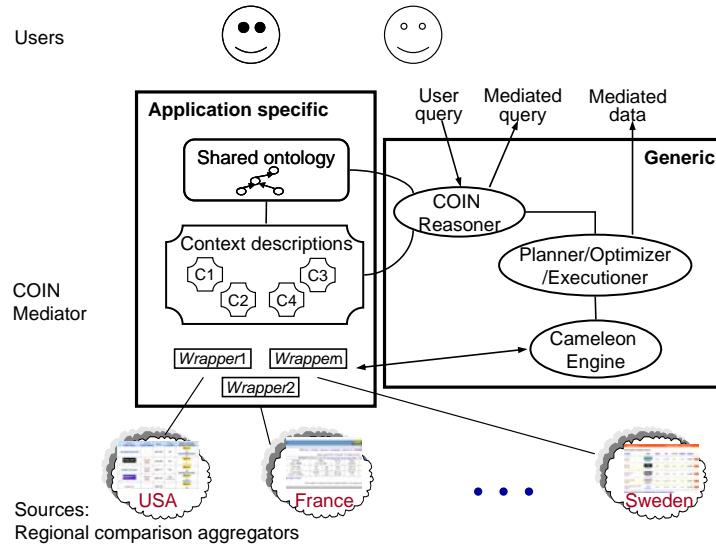**Figure 8**   Context mediation prototype for global price comparison

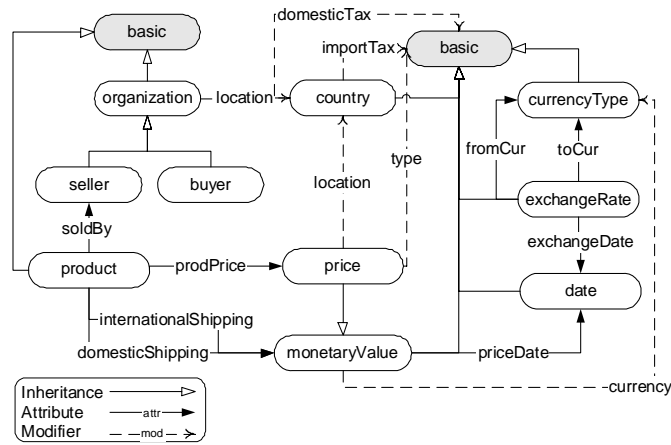**Figure 9**   Lightweight ontology for global price comparison

Figure 10 Algorithm for composing composite conversion using component conversions

```
Input: data value V, corresponding concept C in ontology,
       source context C1, target context C2
Output: data value V (interpretable in context C2)

Find all modifiers of C
  For each modifier mi
       Find and compare mi's values in C1 and C2
       If different: V=cvt_mi(V); else, V=V
Return V
```