WEAVE: AN AUTOMATED SYSTEM FOR COLLATING UNSTRUCTURED DATA FROM WEB AND LEGACY SOURCES TO ENHANCE THE MRO SUPPLY CHAIN

L. Robert Pokorny
XSB, Inc.
pokorny@xsb.com

Harpreet Singh
XSB, Inc.
h.singh@xsb.com

Abstract: Gleaning consistent and complete data from multiple sources of unstructured information is often a difficult and time consuming process. In this paper we outline the WEAVE® system which automates the structuring and collating of unstructured data from multiple on-line Websites. WEAVE® is presented in the context of the maintenance, repair, and operations supply chain. The underlying knowledge representation for WEAVE® is an MRO product ontology. This ontology drives classification of product descriptions harvested from Websites and attribute value extraction from the descriptions. The system uses logic programming to manage the ontology driven classification and extraction and the Java 2 Enterprise Edition platform and Open Business Engine workflow engine to continuously harvest and collate data from multiple MRO catalog Websites. It uses this coherent view of MRO data to allow a user to quickly locate and compare MRO products.

Key Words: Mining Unstructured Data, Data Standardization, Data Extraction, Ontologies for Knowledge Representation

INTRODUCTION

Data that is inconsistent and incomplete is of little value in supporting decision making about the items represented in the data. Much of relational database technology is devoted to maintaining data consistency and completeness. However, even a well designed relational database requires constant vigilance to avoid problems with inconsistency and incompleteness. When the data of interest is unstructured such as text descriptions or memo fields in relational database tables, the problem of inconsistency and incompleteness is greatly compounded. Gleaning useful information from multiple sources of unstructured data is often a difficult and time consuming manual process. An example of a data domain where this problem manifests itself is the maintenance, repair, and operations, (MRO), supply chain.

Manufacturing companies spend approximately 30% of their material purchasing on items for MRO. Unlike materials bought as raw input to the manufacturing process, these purchases are often unplanned and unpredictable. In this environment, it is often impossible to set up long term purchase agreements or blanket purchase orders to ensure most favorable pricing when neither the type nor the quantity of items being purchased is known with any certainty. Therefore, having ready information on the market suppliers for MRO items is critical for making wise MRO purchasing decisions.

There are many sources for such information. Internal data the company has about past MRO purchases and stock room inventory provide reference about the types of MRO items being purchased. On-line catalogs from industrial distributors and OEM manufacturers contain descriptive and pricing data about MRO items available in the marketplace. However, this information is often unstructured. Internal databases of purchasing and inventory data have structured fields for supplier catalog numbers and inventory stock numbers, but if the supplier is a distributor, the manufacturer's part number for the item, if available at all, is included in the part description. The database fields containing the descriptions of items are usually unstructured text. Likewise, item descriptions in online catalogs and manufacturer Websites are usually embedded text in html documents. The key to utilizing this information is to organize it in a structure that classifies items to a taxonomy and associates each item to a set of descriptive attributes. This structuring facilitates easy comparison of items within the company's own inventory and across the online marketplace.

We outline here the WEAVE® system which automatically gathers, extracts, and structures MRO item information from targeted Websites and a company's legacy data. WEAVE® also provides a user interface to search for and compare items across this data landscape so that similar items can be quickly identified and pricing from various sources can be easily referenced. WEAVE® is composed of a number of technologies, all of which rely on an ontological representation of knowledge. It describes MRO product using an ontology based the Federal Cataloging System implemented by the Department of Defense and NATO. Textual descriptions of items are gathered from a company's legacy databases and, using web agents, are also harvested from targeted Websites. Ontology Directed Classification (ODC) associates each description to a product node in the MRO ontology. Ontology Directed Extraction (ODE) parses the description in the context of its classification to extract values of appropriate descriptive attributes. Finally, Ontology Directed Matching (ODM) compares extracted attributes to identify similar items. Data is updated using an automated scheduling workflow so that a company always has a current structured picture of its MRO inventory and the MRO supply chain.

Data quality is addressed in WEAVE® by presenting a Coherent View of product data gathered from multiple sources. This Coherent View shows composite product attributes with traceability back to the original data source. If inventory data describes an electric motor as having a horsepower rating of 2hp but the motor manufacturer's website says the horsepower is 20hp, both values are presented to the user with their reference sources. The user can then make an informed decision as to the correct value or decide that further investigation is necessary. Likewise, if five different data sources claim that a motor is 3 Phase 230 Volt, the user has increased confidence that this is indeed the case.

In the following sections we first discuss background and rationale for using an ontological framework in WEAVE®. We then describe the architecture of the WEAVE® system. We continue with examples of using WEAVE® to compare MRO items. Finally, we conclude with a discussion of the usefulness and limitations of WEAVE® and future improvements to the WEAVE® system.

BACKGROUND AND RATIONALE: USING ONTOLOGIES TO REPRESENT PRODUCT KNOWLEDGE

Over the last few years we have been actively seeking ways to take advantage of unstructured knowledge about products to enhance supply chain management. Much of this effort has been done for and sponsored by the U. S. Defense Logistics Agency. The Defense Logistics Agency and its subordinate agencies manage logistics and acquisition information required to support DOD in times of Peace and War. Each year, the Defense Logistics Agency conducts more than 23 million transactions over 4 million

stock numbers from over a quarter of a million vendors. The broad DLA mandate dictates that the agency work across a wide spectrum of domains – clothing, textiles, medical, fuel, general industrial and construction supplies, as well as military-specific items. A critical issue in fulfilling this mandate is locating alternate sources of supply when current sources become unreliable. The DLA has a large body of legacy data about product characteristics and requirements from which to draw when making sourcing decisions. However, most of this data is unstructured text in buyers' notes and memo fields in databases. Our work with DLA has focused on creating a structured coherent view of this data to support sourcing decisions.

Early on, we realized that products are best described by a set of critical attributes, each of which can have certain defined values. Products described in this way can easily be compared and contrasted based on their attribute values. We initially worked with aircraft maintenance parts where the part material, part manufacturing process, and aircraft platform were critical attributes in determining suppliers that could make the part. We developed parsers that that used preference logic grammars [9] that could extract values for these attributes from unstructured text descriptions. This was done in XSB, an open source tabled logic programming language [11]. As we expanded this effort to other product domains we realized that different products required different attributes. Because of this we decided to represent product knowledge in an ontology.

Ontologies provide a taxonomic structure to classify objects but extend taxonomies by allowing objects to have arbitrary attributes. These attributes can represent any complex property. We built the Conceptual Data Framework, (CDF), ontology management system to allow us to use ontological knowledge representation in a logic programming environment. CDF is a prolog based ontology description language that can efficiently manage and reason about large ontologies with thousands of classes and multiple thousands of objects in those classes. CDF embodies class expressions from description logic for providing rule based inference rules. It is available in open source as a package and documentation bundled with the XSB Prolog Language and its theoretical foundations are described in [13]. Using CDF, We created an MRO product ontology from the Federal Cataloging System, (FCS). DLA uses the FCS to describe parts. The basic entity in this system is the Item Name Code, (INC). There are approximately 40,000 approved INCs and critical attributes are defined by Master Requirement Codes, (MRC) for each INC. Each INC is associated to a Federal Supply Class, (FSC) and Federal Supply Classes are combined in Federal Supply Groups, (FSG). Figure 1 shows an example of an aircraft bell crank described in this system. The FCS has the structure of an ontology in that there is a three level classification taxonomy of FSG, FSC, and INC and each INC has specific MRCs to define the attributes that describe it. We adapted this ontology to the CDF system to define a product knowledge base. By taking advantage of the preexisting product ontology implicit in the FCS we were able to avoid the daunting task of developing an MRO product ontology from scratch.

Three separate data processing processes

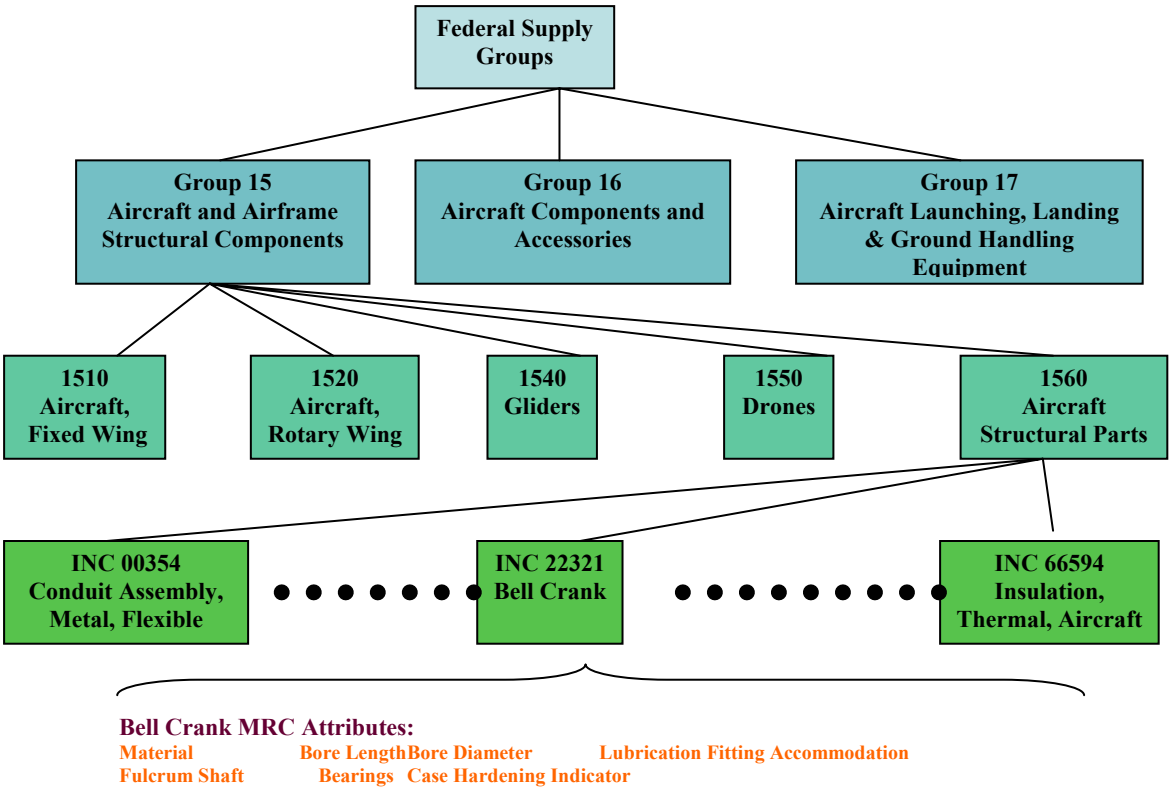

Figure 1: Federal Cataloging System

Having established this ontological view of product knowledge, we developed tools to classify descriptions to ontology classes and use the ontology information to drive parsing of the descriptions to extract attribute values. These processes, Ontology Directed Classification (ODC), Ontology Directed Extraction (ODE), and Ontology Directed Matching (ODM) are presented in [10]. ODC is used to classify short text descriptions to the MRO product ontology. It is based on matching the words in a description to the words describing ontology classes. The weight of a word match is based on both the frequency of the word in ontology class descriptions and the distribution of the word across class descriptions. ODE takes over when a description is classified to a particular ontology class. It automatically parses a product description using a grammar based on the structured knowledge about the class's attributes and their allowable values. Finally ODM is used to compare to items based on the similarity of their attribute values.

Each of these processes is in essence a data transformation. Each transformation is trained by manually providing examples of correct transformations. In classification, descriptions of items that are known to represent a particular ontology class are added as alternate descriptions of that class. In extraction, alternate forms of attribute values are added as abbreviations for these values. In attribute based matching, scoring is adjusted by changing relative weights for matching each attribute.

The quality of these data transformation processes is measured using statistical sampling of batches of real data transformed by the process. This validation procedure, described in [12], assigns a quality measure to the data transformation process. Alternating training and validation are iterated until an acceptable quality level for the data transformation is achieved.

With this background in place, we received an NSF SBIR grant to develop the WEAVE® system. WEAVE® was originally envisioned as a tool to use the WEB to create and manage virtual enterprises, but we quickly realized that applying the technology to on-line product sourcing would produce great benefits. We chose the MRO marketplace as a test bed for this development. To establish an ontology that would provide the knowledge base for a WEAVE® MRO sourcing system, we collaborated with a national manufacturer. We built a taxonomy of MRO items that were a subset of the FCS and classified the product descriptions in the manufacturer's MRO purchasing data to this taxonomy. This allowed us to determine the classes of MRO items that represented the universe of items the manufacturer purchased. An analysis of the classification results identified a set of items that accounted for a significant portion of their MRO spend. We then built ODE attribute extractors for the items identified. These were used to extract attribute values from the text item descriptions. We also built Web agents that can harvest product descriptions for these items from six manufacturer and distributor Websites. The MRO ontology, classifier, extractors, and Web agents provided the tools to collate data from the MRO marketplace. The WEAVE® system uses these tools to provide an integrated interface for exploring that marketplace. The architecture of WEAVE® follows below.

METHOD: THE WEAVE ARCHITECTURE

The WEAVE® architecture links on-line MRO vendor catalog data sources and the company's legacy MRO inventory data. Each data source describes parts using different terminology, standards and abbreviations. The main purpose of the WEAVE® system is to create a master catalog which is continuously in sync with these various vendor and legacy catalogs. The system doesn't just copy information from the various sources, but also standardizes it. The standardization is two fold; the first pass maps the parts in the catalogs to classes in the MRO ontology. In the second pass the part descriptions are converted to attribute-value pairs using our extraction technology. This second pass also removes any catalog specific terminology or abbreviations. Thus, data standardization is driven by the MRO ontology. The standardized description of a product is represented by its canonical ontology class description.

Having all the information from the different catalogs mapped to a single ontology, (the MRO coherent view), allows for easier searching and comparison. To accomplish all of this the WEAVE® system uses a server client application model. The server is responsible for harvesting information and creating the coherent view. The client application allows the user to query and view the catalog generated by the server.

The WEAVE® server is implemented using Java 2 Enterprise Edition, (J2EE), and uses the Open Business Engine, (OBE), workflow engine to manage its complex set operations. The primary workflow on the server side is responsible for maintaining an up to date account of all parts available in the different vendor and customer warehouses. This is done by using our proprietary automated agent technology to mine product information, including price and availability information, from the six vendor website. These Web agents use a round robin schedule so as to not overwhelm any of the websites. Because of this, a complete refresh of all the underlying data can take a couple of weeks. To overcome this, a secondary workflow process can be executed by the client. The secondary workflow allows the client to request up to date information about a part or set of parts in the catalog. A Web query agent is run to get the most current information on that part from the Web data source. This ensures that the data in which user is most interested is accurate and timely.

As each catalog harvest is completed it is compared to the last harvest to determine which parts have been

added, removed or updated. Once the database has been updated with the changes, an ODC is used to classify the new and updated parts to the MRO ontology. The ODC is necessary to set the context for attribute value extraction. The ODE process then extracts structured attribute-value pairs from the harvested product description. The attributes are dependent on the location of the part in the MRO ontology. This phase creates a single standardized view for all the parts in a domain across the different catalogs. The ODE also normalizes the part numbers and manufacturer names. The ODC and ODE are implemented as XSB logic programs which are called as subprocesses from the OBE workflow.

As the primary workflow continuously runs in the background, the coherent view, or master catalog, is continually updated to reflect the changes and additions to the underlying database. The master catalog is comprised of virtual catalog objects. Since a part can be sold in multiple catalogs with each catalog describing the part differently, virtual catalog objects are needed. Virtual catalog objects are a view on top of regular catalog objects and each represents a single standardized part. The attributes of virtual catalog object are the combination of all the attributes harvested from the underlying harvested catalog objects they represent. It is possible to create the virtual objects because all catalog objects have a standardized manufacturer name and part number, allowing the workflow to join across catalogs.

Technically, the WEAVE® system needs to combine workflow management for data updates and GUIs for user interaction written in JAVA and reasoning provided by CDF running on XSB prolog. This has been accomplished through a proprietary system architecture called XJ. XJ allows the definition of JAVA GUI interfaces declaratively in XSB Prolog. It is built on top of interprolog, a JAVA-XSB interface that is available as an open-source package included in the XSB Prolog system.

The WEAVE® client has traditionally been a Java based desktop application, although recently we have experimented with a web based application. The client allows the user to search the master catalog using a variety of different methodologies. These include exact manufacturer and part number search, search by product description, and search for similar parts. Search by product description uses the ODC and ODE classify the user description to a class in the MRO ontology and extract attribute-value pairs to describe the users object. The Ontology Directed Matcher (ODM) compares the specifications of the user object with the virtual catalog objects in the master catalog on an attribute by attribute basis. The user can specify the importance of different attributes before running the search. The ODM returns a list of virtual catalog objects ranked by their similarity to the user's object. Once the user is ready to purchase the part, the underlying catalog objects are used to display sources and pricing information.

	LEGACY DATA	Baldor Motor (www.baldor.com)	G.E. Motor (www.grainger.com)
DESCRIPTION	MOTOR 10HP 1760 EM3774T BALDOR	10HP | 1760RPM | 3PH | TEFC | 215T NEMA …	3 Phase Totally Enclosed Fan Cooled Motor, HP 10, RPM 1755, Voltage 230/460 V, NEMA Frame 215T, Service Factor 1.15, Frequency 60&50 Hz, Efficiency 89.5 …
INCLOSURE FEATURE		TOTALLY ENCLOSED	TOTALLY ENCLOSED
COOLING METHOD		FAN COOLED	FAN COOLED
HORSEPOWER	10	10	10
TEMP RATING (° C)		105	105
ROTOR SPEED (RPM)	1760	1760	1755
FREQUENCY (Hz)		60	60
WEIGHT (Pounds)		202	213
VOLTAGE (Volts)		230	230/460
FRAME	215T	215T	215T
CURRENT RATING (Amps)		25	23.5
PRICE ($)		929.00	633.00 (882.00)

Figure 2: Enrichment of Legacy Data

RESULTS

The WEAVE® systems attribute extraction/enrichment leads to better product equivalence and better spend decision making. Cleaning up the legacy data reduces redundancy and improves accuracy in buying process. Use of agents and extraction technologies provide a coherent view of the market for determining price realism.

Data quality is enhanced in WEAVE® through data enrichment. The concept of enrichment is based on two interrelated data transformations. The first is classification to an ontology using ODC. The second is extraction of item attributes using ODE. These processes take textual product descriptions and standardize them to canonical objects in an ontology class with specific attribute values. The quality of the transformation processes is validated using a statistical validation process.

An example of enrichment of legacy data is shown in figure 2: The figure show information about an electric motor. The first column contains information available in the company's inventory system about this motor. As can be seen, only 3 attributes are extracted from the company's description of this motor, horsepower, rotor speed and frame size. When this information is fed into the WEAVE® system we are able to increase our knowledge about the part by combining the information in the legacy data with information from the Baldor Motors Website shown in the second column. Finally, the third column shows information about a General Electric motor available from the Grainger on-line catalog which might be considered as an equivalent for this part. This merging of information is done automatically as described earlier in the coherent view generation phase of the server workflow. Since we now have more attributes and values to compare we are able to better target our searches and find more plausible replacements for the motor.

The use of agents and automated workflow processes to continuously update and standardize data reduces the amount of labor needed to find parts. One experiment set the amount of time needed by a single person to find a replacement v-belt to 1.5 hours. Most of this time is spent going to different vendor websites, filling out search forms, and looking through product data sheets. The same process while using the WEAVE® application should not take more than 10 minutes. WEAVE® lowers the cost barrier to finding the best price for low value, high volume items.

DISCUSSION

In 2002, the Aberdeen Group, Inc [4] estimated that inadequate spending analysis costs business $260 billion in missed savings each year. While perhaps not the corporate epidemic that Aberdeen predicts, further research indicates opportunities exist to reduce total spending for MRO by 12% or more. Realizing this opportunity requires comparison of unstructured product descriptions from multiple data sources. The cost of doing this comparison manually often negates any savings achieved by finding the best cost vendor.

WEAVE® automates the process of collating the information necessary to make wise buying decisions. There are three areas in which this is particularly evident.

1. Sourcing of low value, high volume items – These are items that are purchased often but in a wide variety of styles of sizes. No one size dominates the aggregated purchases of the item. These items also have small price differences from different vendors. This was the case with the v-belt mentioned above. The cost savings buying from the low cost vendor can only be realized if the low cost vendor can be rapidly identified
2. Sourcing of high value items – This is the case illustrated by the motor in figure 2. Here, significant savings can be saved by buying the equivalent product from the lowest cost vendor. However, equivalence can only be determined when the product description is enriched with all the critical attributes on which a decision about equivalence is based
3. Finding equivalent items in company inventory – This utilizes attribute based equivalence matching to indicate when one item in inventory can be substituted for another item that is out of stock. Here again, the decision about replacement can only be made if the items are fully described based on their attributes.

The automated process for collating and standardizing parts descriptions from multiple data sources embodied in WEAVE® can help realize the potential savings in MRO purchasing highlighted in the Aberdeen Group study.

WEAVE® can be compared to other systems that strive to coordinate data from multiple Web sources. There are a number of systems based on Web Services and the Semantic Web [3]. BPEL4WS [1] provides a structure for interconnecting Web Services. Universal description, discovery and integration protocol (UDDI) [2] provides a directory service giving semantics for locating Web Services. Both these systems are concerned with locating and connecting sources of information but say nothing about the content of those sources. A system much closer to WEAVE® is SEWASIE [7,8]. SEWASIE organizes knowledge based on multiple Web sources using an ontological representation. However, SEWASIE is based on building multiple ontologies to represent different nodes of web knowledge characterized by similar websites. SEWASIE then builds a peer-to-peer network of these nodes managed by broker agents that match ontologies to user queries. This model requires a good deal of manual effort to build and maintain ontologies for the different knowledge nodes. SEWASIE uses MOMIS [6] as an ontology language for managing this process. WEAVE® avoids this by bootstrapping the FCS for a standard

product ontology. Also WEAVE® provides quality measures for the classification, extraction, and matching processes that develop standard product content based on this ontology. Finally, Lixto [5] is one of a number of systems to automate harvesting of structured data from HTML documents. This is similar to the harvest agents in WEAVE® but only represents a component of the WEAVE® system.

LIMITATIONS

While WEAVE® provides structure to the view a company has of the MRO supply chain, there are limitations to what it can do. The ODE process of extracting attribute values from an item description needs to be tuned for each item class in the ontology. This is primarily a manual process of indicating the types of values expected for each attribute and standardizing the abbreviations and jargon used to represent these values in each data source. This requires additional work whenever a new item is added or a new data source is added for an existing item. There is a breakeven point where the effort to cover additional items or add additional data sources is more that the value of information gained by making the addition.

A second issue arises from the automated Web agents we use to gather on-line product information. The proprietary technology we use to build these agents allows us to create an agent fairly quickly. The agents are also robust to changes in the Website from which they harvest information. However, some Website changes can cause the agents to fail in collecting data. Therefore, it is necessary to monitor the harvesting workflow for indications of agent failure and modify the agent software when this happens. This is manageable in the current WEAVE® system where we have six on-line sources. It becomes problematic when WEAVE® is scaled to hundreds of sources. Scalability is an area where we need to do more investigation.

CONCLUSION

There is a large body of information available to a company to aid in managing its MRO supply chain. Internal company data can be enhanced with on-line data from potential distributors and OEM equipment manufacturers. However, most of this information is unstructured text. Standardizing, structuring, and collating this information is necessary to reap the advantages of having access to vast numbers of information sources.

The WEAVE® system presents an approach to automating the collection, classification and structuring of information from multiple data sources. An up front effort to review the data landscape for MRO products and generate an MRO ontology was required to create the knowledge base necessary to support the WEAVE® MRO system. This effort has allowed us to implement a system that can supply continual up to date structured data for the MRO marketplace.

The WEAVE® methodology can be extended to other product domains. We are currently looking an office supplies and medical products. The key ingredients necessary for a WEAVE® system to be successful are an ontology that organizes knowledge for a data domain, and a rich, albeit unstructured collection of data sources for the domain. Any data domain that has these properties is a potential application for our methodology.

REFERENCES

[1] Business process execution language for web sevices (BPEL4WS). http://www.ibm.com/developerworks/library/ws-bpel/.

[2] Universal description, discovery and integration protocol. http://www.uddi.org.

[3] W3c semantic web. http://www.w3.org/2001/sw/.

[4] Aberdeen Group, Inc. The Spending Analysis Benchmark Report – Dissecting a Corporate Epidemic; 2002 Aberdeen Group, Inc and Penton Media Inc

[5] R. Baumgartner, S. Flesca, G. Gottlob: Visual Web Information Extraction with Lixto, in proceedings of the 27th VLDB Conference, Rome, Italy, 2001

[6] R. Benassi, D. Beneventano, S. Bergamaschi, F. Guerra, M. Vincini: "Synthesizing an Integrated Ontology with MOMIS", International Conference on Knowledge Engineering and Decision Support (ICKEDS). Porto, Portugal, 21-23 July 2004 Paper (pdf)

[7] D. Beneventano, S. Bergamaschi, F. Guerra, M. Vincini: "Building an integrated Ontology within SEWASIE system", in proceedings of the First International Workshop on Semantic Web and Databases (SWDB), Co-located with VLDB 2003 Berlin, Germany, September 7-8, 2003. Paper (pdf)

[8] S. Bergamaschi, F. Guerra, M. Vincini: "A peer-to-peer information system for the semantic web", in proceedings of the International Workshop on Agents and Peer-to-Peer Computing (AP2PC03), held in AAMAS 2003 International Conference on Autonomous Agents and MultiAgent SystemsMelbourne, Australia, July 14, 2003 Paper (pdf).

[9] Cui B. and T. Swift. Preference Logic Grammars: Fixed-Point Semantics and Applications to Data Standardization. *Artificial Intelligence*. To appear

[10] Davulcu H., Jones J., Pokorny L.R., Rued C, Swift T., Vidrevich T, and Warren D.S. Ensuring the Consistancy of Self-Reported Data: A Case Study. in proceedings of the 7th International Conference on Intormation Quality ICIQ02, Cambridge, Mass. 2002

[11] http://xsb.sourceforge.net

[12] Pokorny L. R. Assigning a Quality Measurement to Matching Records from Heterogeneous Legacy Databases:A Practical Experience, in proceedings of the 5th International Conference on Information Quality ICIQ00, Cambridge, Mass. 2000

[13] Swift, T. **D**eduction in Ontologies via Answer Set Programming. In proceedings of the Conference on Logic Programming and Non-Monotonic Reasoning, Springer Lecture Notes in Artificial Intellegence #2923 pg 275-289, 2004