JITTA

JOURNAL OF INFORMATION TECHNOLOGY THEORY AND APPLICATION

A Publication of the Association for Information Systems

The Semantic Web in Federated Information Systems: A Space Physics Case Study

Tom Narock
Department of Information Systems
University of Maryland, Baltimore County
tnarock@umbc.edu

Victoria Yoon
Department of Information Systems
University of Maryland, Baltimore County
yoon@umbc.edu

Jan Merka
Goddard Earth Sciences and Technology Center
University of Maryland, Baltimore County
Jan.Merka@nasa.gov

Adam Szabo
Heliospheric Physics Laboratory
NASA/Goddard Space Flight Center
Adam.Szabo@nasa.gov

Abstract:

This paper presents an innovative framework for integrating Semantic Web technology in FIS and a case study of the framework for the search and retrieval of disparate data sources in NASA's space physics domain. Our case study involves utilizing the Semantic Web within a community that has little knowledge of the technology. As a result, our case study uses and evaluates the proposed framework for formal ontologies in FIS that shields participants and users from the details of this technology. The framework is a middle-of-the-road approach for utilizing semantics in FIS. Our work also evaluates the Semantic Web under real-world conditions providing empirical results of efficiency and efficacy for practitioners. Moreover, our case study compares two identical systems, one with semantics and one without, which to the best of our knowledge, is the first study on the side-by-side comparison of the Semantic Web with existing relational database technology. This comparison case study will benefit researchers and practitioners as many organizations begin augmenting their relational databases with Semantic Web technology.

Keywords: Semantic Web, Federated Information Systems, ontologies, knowledge management

Volume 11, Issue 3, pp. 25-41, September 2010

Balaji Rajagopalan was the Senior Editor for this paper.

JOURNAL OF INFORMATION TECHNOLOGY THEORY AND APPLICATION

INTRODUCTION

Today's world of inexpensive processing and storage capabilities has enabled organizations to easily create, store, and make accessible vast amounts of data. Cho and Giustini's (2008) review of the medical domain concludes that, "the information age has simply become too efficient in creating vast amounts of new medical knowledge." In addition to vast quantities of data, an organization's diverse operations have led to heterogeneous data sets, and the integration of such diverse data within and across organizations has become a major challenge (Zhao and Ram 2007). This challenge is by no means restricted to any single domain but exists across the spectrum from health care (Bell and Sethi 2001), to military information systems (Clifton et al. 1997), and to the physical sciences (Dalton 2007). Effectively utilizing heterogeneous data requires users to have knowledge of relationships and hierarchies between the constituent data. Broader, cross-organization, integration involves the assimilation of schema and the alignment of synonymous concepts, attributes, and data instances. These tasks require the effective capture and use of domain knowledge and semantics on large heterogeneous data sets.

Several methods have been proposed to meet these integration challenges. Zhao and Ram (2007) have developed a technique that integrates disparate data sources into a single unified data source. The technique integrates at both the schema and instance levels. Alternatively, Busse and colleagues (1999) have developed the concepts of Federated Information Systems (FIS) to provide integrated access to a finite, predefined set of autonomous and heterogeneous databases. FIS is characterized by the presence of a federation layer that accomplishes interoperability among the underlying heterogeneous databases while retaining their autonomy and heterogeneity (Sheth and Larson 1990; Busse et al. 1999). One of the prime research areas in FIS is semantic integration (Hasselbring et al. 2000). Some researchers have explored the Semantic Web approach (Berners-Lee et al. 2001) to integrate the semantics of heterogeneous data sources in the federation layer (Thomas et al. 2007; Vdovjak et al. 2003). However, more research is necessary to effectively incorporate Semantic Web technology in FIS.

The objectives of this study are to propose a conceptual framework for integrating Semantic Web technology in FIS and to present a case study of the framework for the search and retrieval of disparate data sources in NASA's space physics domain. Our conceptual framework presents **a unique middle of the road approach,** which is designed to overcome the limitations of two dominant methods used in integrating semantics in FIS. Such a framework is very effective for domains, like our space physics domain, whose users have limited Semantic Web experience. The study examines how the Semantic Web can be utilized to enhance integration and search of the heterogeneous data sources in FIS. Specifically, we explore using formal ontologies to capture and work with domain semantics. A prototype search implementation, relying solely on relational database technology, is already in existence in our chosen domain. This allows us to obtain a quantitative measure of the effectiveness of semantic technologies when directly compared to relational database technologies. We provide empirical evidence comparing the semantic interface to queries over a relational database approach from a space physics data integration problem to illustrate the utility and effectiveness of Semantic Web technologies.

The remainder of the paper is organized as follows. The next section discusses related work in using the Semantic Web to integrate heterogeneous data sets. The paper then presents our conceptual framework that incorporates the Semantic Web approach in the federation layer to integrate heterogeneous data. This is followed by our case study of a real world NASA data integration problem in space physics. Finally, we conclude with a discussion of the efficacy of the Semantic Web as a data integration technology and discuss its broader implications.

CONTRIBUTION

This paper is an application case study. It applies the emerging Semantic Web within NASA's space physics environment, a community that is mostly unfamiliar with this technology. As such, this work makes two important contributions. First, it presents a new theoretical framework for applying the semantic web in communities with limited understanding of the technology. The case study is then used to evaluate the new framework. An additional contribution comes in the form of a comparison of semantic Web technologies versus traditional relational database technologies. Within our case study, a system relying on formal ontologies is compared against a system with the same functionality, but built using relational database technology. The advantages and disadvantages of each technology are compared through real-world uses of the systems.

RELATED WORK

The need for the Semantic Web and its early adoption has been well documented. For example, the medical community (Slaughter et al. 2006) has discussed the need for formal ontologies, and information science professionals have begun to educate their communities (Cho and Giustini 2008) about this technology. Along the way to implementing the Semantic Web, researchers have discussed the trials and tribulations associated with this emerging technology. Brunner and colleagues (2007) have discussed the benefits and difficulties of the Semantic Web in representing business product information. In the field of space physics, McGuinness and colleagues (2007) have developed a data search and retrieval system based on Semantic Web technologies. They have constructed a domain ontology and captured domain relationships to assist users in finding available data. McGuinness' experiences (2007) highlight the effectiveness of various Semantic Web technologies in the steps leading up to the completion of their information system.

Our particular application looks at applying semantics to FIS. A FIS consists of a three-tier approach (Busse et al. 1999) with independent participants forming the foundation layer. These participants are then unified via a common schema and metadata in the federation layer. Users interact with the unified information through a presentation layer. As noted by Hasselbring and colleagues (2000), semantic integration, especially the role of formal ontologies, is one of the prime research areas of FIS. Some choose to implement a large and comprehensive ontology while others choose multiple smaller ontologies. Each approach has its advantages and disadvantages. The former creates a comprehensive model of the domain that contains the global schema and all associated terminology from the foundation layer, as shown in Figure 1.a. Such ontologies can be exceedingly large and unwieldy to work with; thus, it can be difficult to use in practice. Consider the OpenCyc[1] project. The foundational Cyc ontology has the goal of providing high-level concepts that can be used to map low-level FIS concepts. Version 0.7.8b of Cyc contains over 60,000 concepts, and its resulting OWL file is larger than 700 MB (Bao and Honavar 2006) taking roughly nine hours to be loaded into the Protégé ontology-editing tool (posted 4 June 2004 on the OpenCyc website and quoted in Bao and Honavar 2006).

The latter approach of multiple ontologies simplifies implementation and offers the benefits of having small, easy to work with, modular ontologies. However, this approach requires ontology mapping, as shown in Figure 1.b, which has been shown to be a difficult problem (Ding and Foo 2002). Vdovjak and colleagues' (2003) work illustrates the difficulties of autonomously mapping concepts between ontologies, which frequently yielded imprecise mappings propagating throughout their system. In order to overcome these difficulties, Vdovjak et al. allowed the domain experts to specifically align their terminology with the global schema of the federation layer. However, such ontology mapping can be very tedious and time-consuming. Thomas et al. (2007) also use the multiple ontology approach and define a framework in which intelligent software agents are overlaid on a loosely coupled FIS. These software agents, in conjunction with ontologies at the data provider sites, allow for quick and efficient integration of new data sources into the FIS. Unfortunately, their work has strong reliance on ontological documents that may be infeasible in many domains. Managers of various data sources in the foundation layer are required to create OWL documents, and sophisticated web services must be capable of interacting with the agent environment. While interesting and applicable in some areas, this framework will not effectively service user communities with limited Semantic Web experience.

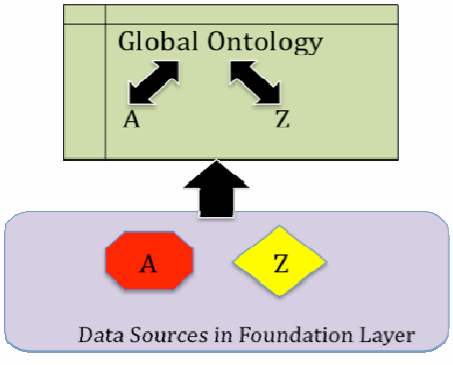


Figure 1.a. Terminologies from diverse date sources are integrated by mapping to a global ontology.

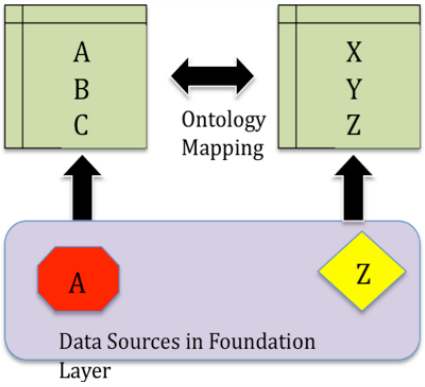


Figure 1.b. Modular ontologies are created for each distinct data source.

[1] http://opencyc.org/

PROPOSED FRAMEWORK

In order to overcome the drawbacks of the two aforementioned approaches, we propose a middle-of-the-road approach. As shown in Figure 2, instead of having many modular ontologies, our framework utilizes a single domain ontology. However, this domain ontology is not a monolithic ontology (Figure 1.a) that incorporates all terminology from the domain. It is a coarse domain ontology that resides at the federation layer unbeknown to the members of the foundation layer. Participants in this FIS are not required to be familiar with the Semantic Web; however, they garner all the benefits of semantic data integration. Members of the foundation layer deal exclusively in XML while users in the presentation layer utilize a user interface that hides the complexities of the Semantic Web. Thus, this approach is appropriate for user communities, like our space physics domain, which have limited Semantic Web knowledge.

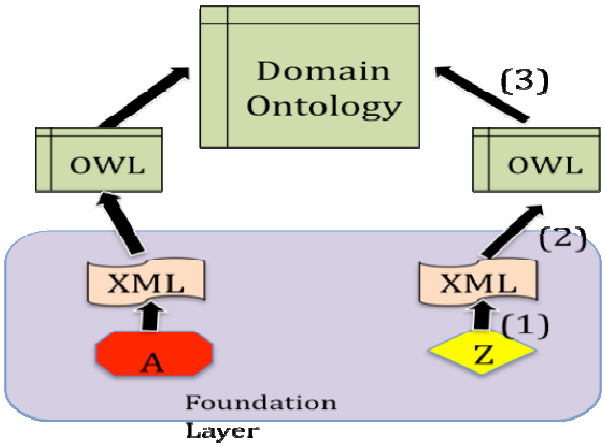


Figure 2. Our proposed middle-of-the-road approach.

Our framework utilizes a three-step process that is illustrated in Figure 2: XML generation for each data source, XML to OWL conversions, and integration of OWL instances into a domain ontology. In the first step, data providers in the foundation layer produce XML documents, conforming to a provided schema, that completely describe their underlying data. This can be accomplished with minimal effort and without regard for ontologies or semantic documents. These XML documents are then forwarded to the federation layer and periodically harvested for updates thereafter. The XML schema is a flattened and less expressive version of the ontology. The advantage here is that XML is a widely accepted industry standard.

Once these XML documents are at the federation layer, software tools are used to convert the XML into instances of our ontology. There is not a direct correspondence between XML elements and domain ontology concepts. In other words, each XML element does not equate to an OWL class. Rather, various combinations of XML elements form the properties of OWL classes and are used to instantiate these classes, enabling us to minimize mapping issues between the foundation layer schema and our domain ontology. A foundation layer manager may feel that allowed values of a particular XML element do not map exactly to his/her schema. However, because the OWL instances are combinations of XML elements, we can obtain very good semantic integration and have found no cases where the resulting OWL instance does not accurately represent the foundation layer data. The third step integrates the OWL instances into the domain ontology, which is used to semantically integrate the heterogeneous foundation layer data.

Our framework utilizes a semi-autonomous approach based on the On-To-Knowledge methodology (Staab et al. 2001). On-To-Knowledge utilizes (Gomez-Perez et al. 2004) an incremental lifecycle (McCracken and Jackson 1982) in combination with evolving prototypes (Kendall and Kendall 1985). As such, the resulting ontology is application dependent and is continually, and routinely, evolved as application requirements evolve. In this regard, a team of domain experts evolves the domain ontology (described in section 4.3) and associated XML to OWL mappings as application requirements mandate. However, once an incremental release is approved, the system is able to autonomously populate the ontology with instance data. As our framework autonomously converts XML to OWL, it creates only instances of classes. The predefined mappings, as well as adherence to the XML schema, assure that conflicts and ontology inconsistency do not occur. Possible inferential inconsistencies may occur; however, these can be easily addressed and are detailed in subsequent sections. Storing and reasoning with

millions of instances is currently an active Semantic Web research area. Our case study data did not approach such limitations (~10,000 instances), however; practitioners should be aware of potential scalability issues.

Evaluating an ontology creation methodology is a nontrivial task (de Hoog 1998; Gomez-Perez et al. 2004, Cristani and Cuel 2005). Numerous methodologies, and evaluation criteria, exist and the choice is inherently subjective. While our framework favors the flexibility of the On-To-Knowledge methodology, it is not strictly coupled to it and other methodologies may prove suitable within other domains.

CASE STUDY

NASA's Heliophysics Data Environment

In a recent strategic shift NASA adopted the term *heliophysics* to describe its current and future research emphasis (NASA Recommended Roadmap 2005). From the Greek term *helio,* which means "relating to the sun," heliophysics is meant to encompass the Sun-Solar System connection and reflect NASA's increasing priority to study this system as a collective whole. The vast amount of spacecraft now available provides unprecedented abilities to study the three-dimensional structure and related phenomena of our universe. However, utilizing this collection of data presents a formidable challenge. Each spacecraft mission is independently managed and years of operation have led to disparate data storage formats and a multitude of variations on parameter names. Integrating these data, along with recently digitized legacy data, have caused NASA to commit to facing data interoperability challenges head on.

NASA first combated interoperability challenges in the astronomical community (Szalay and Gray 2001). Within the astronomical community NASA commissioned the development of a so-called "Virtual Observatory." The community bestowed this name to designate the FIS that now served as a single point of entry to distributed and heterogeneous data. This system provided transparent online access to the brick and mortar observatories familiar to astronomers. No longer did astronomers have to search each observatories holdings individually. The Virtual Observatory allowed for unified search over the underlying heterogeneous data.

The "Virtual Observatory" contains the three-tiered approach common to FIS. Since the system deals with search and retrieval, it also offers connections to web services at the presentation layer. Users have the option of visualizing or applying analysis techniques to the data prior to retrieval (for an example and overview see Zhizhin et al. 2008). Figure 3 illustrates the paradigm of a Virtual Observatory (and FIS in general).

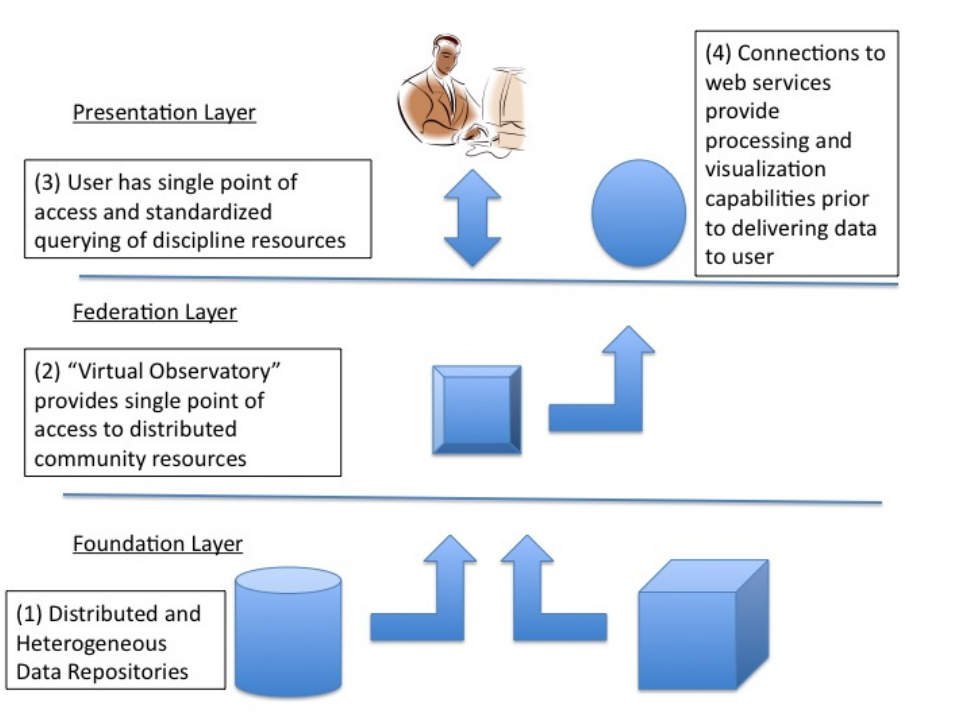


Figure 3. A graphical depiction of the Virtual Observatory concept. Such concepts are common in the geosciences for data search and retrieval.

Disparate community data resources (1) are described using a domain metadata schema. The schema allows for the description of data products at both a high and low level. That is, a high-level overview of the data product is made available in addition to low-level descriptions of individual data files. These metadata descriptions are aggregated (2) to a centralized repository. The repository continually checks for new or modified metadata, and procedures are in place for the addition of new data sources. Users can access (3) the metadata repository, which offers a standardized view of the disparate resources, to search for and retrieve links to data files of interest. Users also have the ability to request data analysis and/or visualization (4) through accompanying web services. The Virtual Observatory concept includes the FIS and the human component charged with the governance of uniting the data resources of their community.

This paradigm quickly spread through the geosciences and is now the basis for unifying the NASA heliophysics data environment. However, heliophysics is a diverse domain with many unique sub-domains, each with unique data and data requirements. Thus, sub-domain experts should be tasked with identifying, aggregating, and providing the appropriate search mechanisms for their data. As such, NASA took a multi-staged unification approach within heliophysics. First, several Virtual Observatories were commissioned and each sub-discipline has its disparate data unified through a Virtual Observatory. Figure 4 shows a depiction of the five Virtual Observatories in the time dependent Sun-Earth environment. The Sun on the left edge influences the region between the Sun and Earth and ultimately affects Earth's upper atmosphere and surrounding environment, which are depicted on the right-hand portion. In order to gain a complete understanding of the system, researchers need to couple data from all of the regions. Thus, the second unification stage aims to integrate these five Virtual Observatories in order to present comprehensive access to the complete collection of space physics data resources.

However, as will be shown, the lack of formal semantics in the current design requires significant complexity in the federation layer and also in the presentation layer interface. The use of a formal ontology removes much of this complexity from the system and assists in making great strides toward the second stage goal of unification.

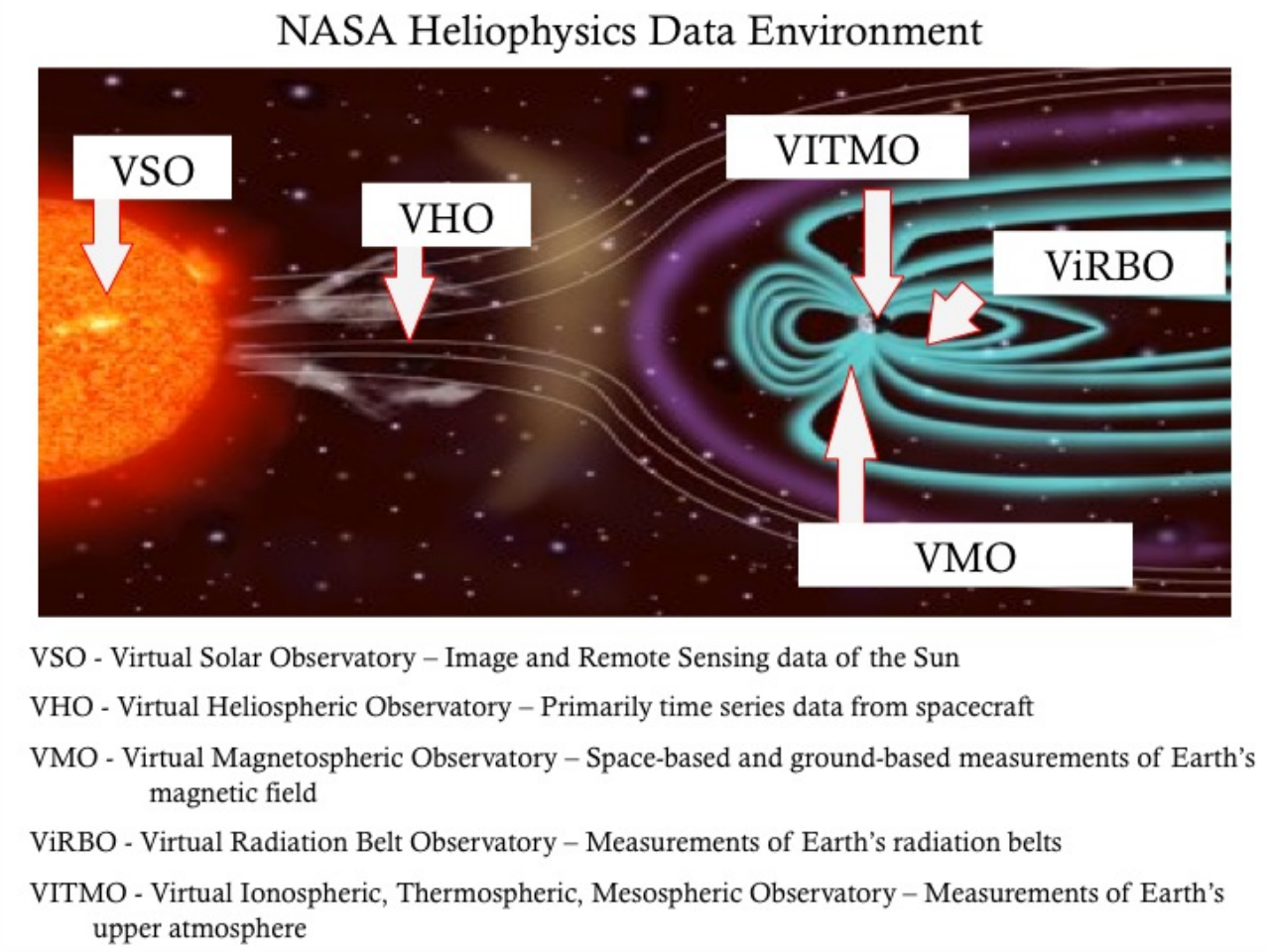


Figure 4. An illustration of the NASA heliophysics data environment (each Virtual Observatory represents a sub-discipline specific information system).

VHO—Architecture and Challenges

This case study focuses on the integration effort of the Virtual Heliospheric Observatory (VHO) that encompasses the Sun to outer solar system environment. The nature of space physics data sets is hierarchical and tree-like. Each spacecraft contains multiple instruments, each instrument can operate in multiple modes producing several types of data, and each type of data contains numerous data files corresponding to the time periods they are captured. Moreover, several space physics data sets are combinations of other existing data products. Unlike some other fields, where data is collected as needed, NASA spacecraft are continually collecting and transmitting data to Earth. As a result, large pools of data exist for each spacecraft, and exploring such extensive data is very time consuming. Moreover, users want to enter the tree at various stages and search for data in various ways. For example, some users are interested in all known instruments of a certain type while others may be interested in all data available for a specific time period regardless of which spacecraft or instrument it came from. Supporting such flexible query processing is an intractable task.

Initially the VHO created a relational database to manage information and execute user queries. Users interact with a web-based form in order to construct queries. These queries are then passed to a software layer that transforms the web input into SQL. Next, the SQL query is executed in the relational database. Finally, the results are formatted and processed for presentation and displayed to the user on a results page.

Domain Ontology Development for VHO

In order to evaluate Semantic Web technologies and assist the NASA space physics community, who has limited knowledge of semantic technologies, we have adopted the middle-of-the-road approach that was presented in Section 3. Our proposed middle of the road approach in Figure 2 takes the XML document for each data source in the foundation and converts it to OWL. For this process, we use the XML-based schema developed by the Space Physics Archive Search and Extract (SPASE) consortium (Harvey, et al. 2008). The SPASE consortium, which consists of space physics researchers, software developers, and data providers, was founded to aid in the integration of space physics data. The original intent of SPASE was to create a metadata schema that would drive the federation layers of NASA's Virtual Observatories and lead to standardized descriptions of data resources. The SPASE effort follows the On-To-Knowledge methodology (Staab et al. 2001) where application requirements (creation of a comprehensive federation layer) drive the evolution of the schema. Consistent with the incremental life cycle and evolving prototypes of On-To-Knowledge, SPASE members conduct periodic meetings to access current status and discuss future directions. It should be noted that On-To-Knowledge was developed as an ontology creation methodology while the SPASE effort's primary goal is a more general metadata schema. Our work parallels this primary goal with concurrent expressions of the metadata schema in OWL.

A conversion to OWL was initially begun in earlier work (Narock et al. 2009), but has been significantly updated and expanded for this case study. The resulting OWL-DL ontology consists of thirty-two classes, thirty-eight data type properties, and thirteen object properties. This ontology, and the XML schema on which it is based, addresses the entire domain of heliophysics and each sub-domain instantiates varying parts. Currently we have 9267 individuals for VHO. As the VHO is an evolving project these individuals are expected to increase by an order of magnitude in the coming years.

Figure 5 shows selected classes in our domain ontology with associated object properties illustrated by dashed arrows. The classes shown are those primarily instantiated by the VHO and the image was created via the Jambalaya plug-in to the Protégé ontology editor. Figure 6 focuses on the *Data* class and its relationships to the *Parameter* and *Observatory* classes, which is a specific branch of our ontology, providing insight into its conceptual layout. In a search and retrieval application, such as ours, the *Data* class is fundamental. The various types of instruments and flavors of data they produce are engrained in the minds of heliophysics researchers. However, this semantic information is traditionally lost on information systems. The ontology captures the lineage of data in our domain as well as other pertinent information. For example, the *hasAccessInformation* object property (dashed arrow between *Data* and *AccessInformation* in Figure 5) links the *Data* class to the *AccessInformation* class. In many ways Data is the central point of our ontology. The inner workings of the data are modeled through datatype properties (attributes) such as *MeasurementType* (one or more cardinality), *Name*, *ReleaseDate* and *Format*. The ontology uses the formal semantics of OWL to represent relationships familiar to heliophysics researchers, as well as data lineage and access relationships. Once the data lineage is properly described the data product is then linked, Figure 6, to the spacecraft from which it came as well as to instantiations of the parameters it contains.

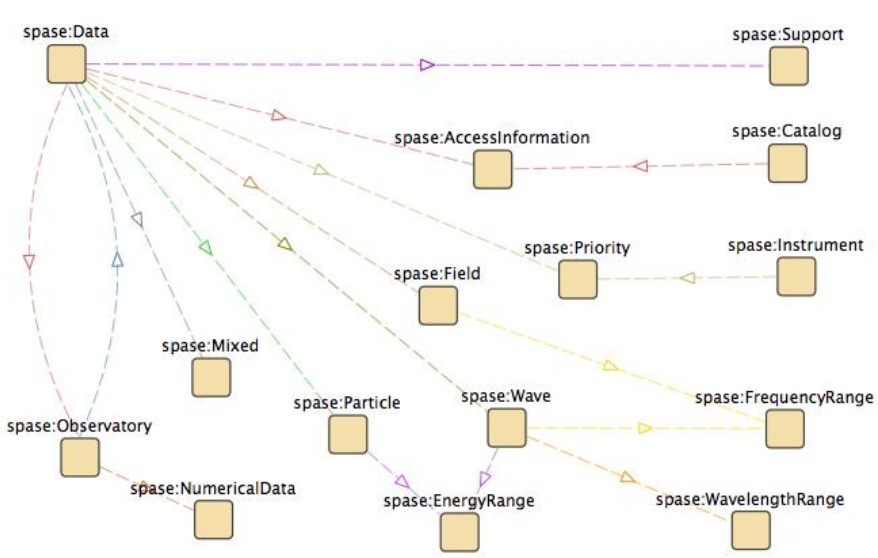


Figure 5. A depiction of selected classes in our SPASE ontology.

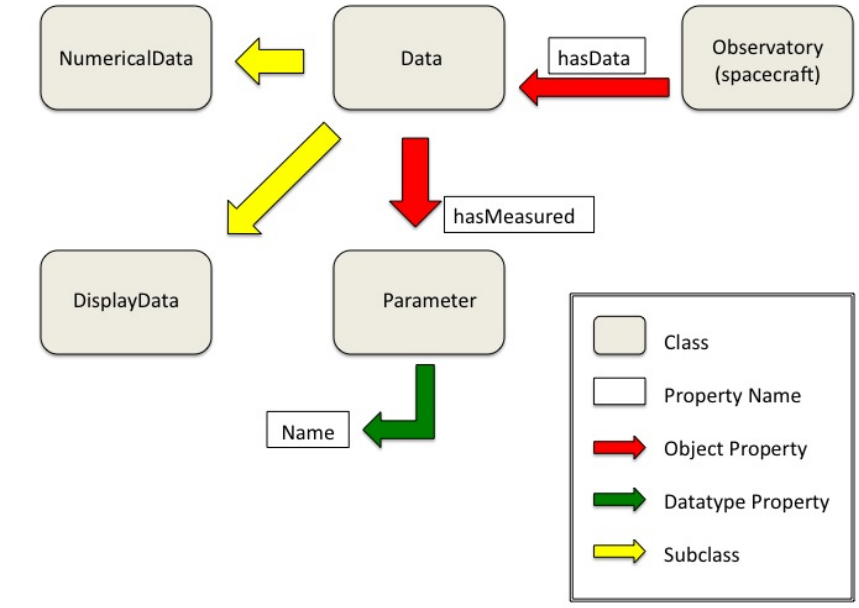


Figure 6. A branch of our ontology on Data, Parameter and Observation.

To create our ontology and associated individuals we utilized two freely available tools. First, we used Protégé to manually create the ontology classes and properties from the SPASE schema and documentation. Next, we utilized the XML2OWL (Bohring and Auer 2005) tool to map available XML metadata to OWL. XML2OWL provides a graphical user interface to assist the user in mapping XML elements to OWL classes and properties. Once this is done, the tool produces an XSLT that can be applied to subsequent XML files. Thus, the creation of the mapping is a brief one-time effort and all remaining XML files can be mapped autonomously. One main benefit of this approach is that any future changes to the XML can quickly and easily be expressed in the mappings.

System Architecture

Figure 7 shows the current system architecture, which is enhanced with the domain ontology for VHO. A user first submits a query using the form-based User Interface (UI) on the Web. The UI converts the query into the SPARQL query language (Prud'hommeaux and Seaborne 2008) and submits it to the domain ontology. The ontology then retrieves relevant classes, properties, and/or instances, and returns the retrieved information to the UI. The UI then generates an SQL query using the retrieved information; relevant classes, properties, and/or instances become tables, rows, and/or columns in the SQL query. The UI submits the SQL query to the relational database, and the database finally returns the query result to the UI, which then presents the results to the user.

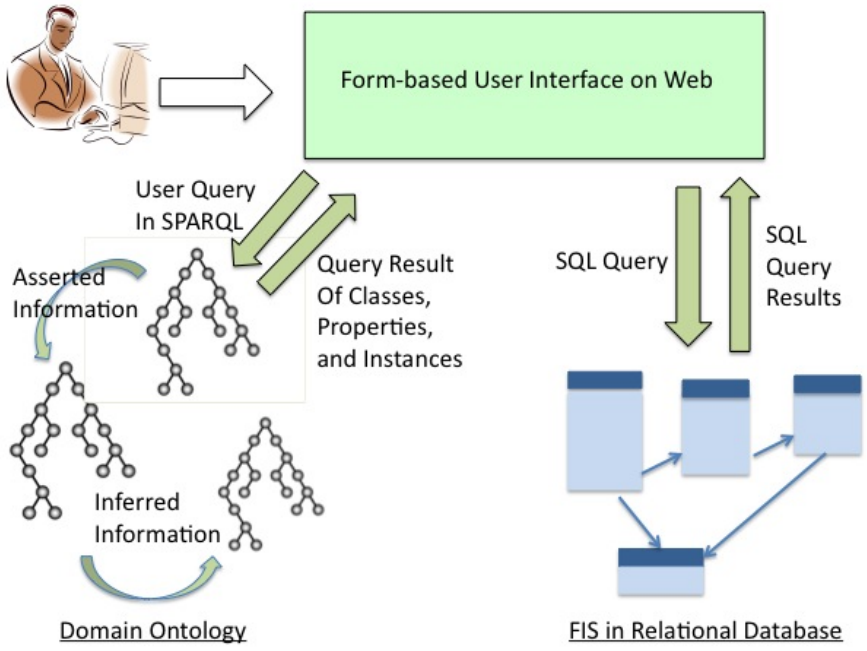


Figure 7. The architecture of the Semantic Web approach.

System Maintenance

The VHO data holdings are evolving; thus, both the ontology and relational database will periodically need to be updated. As new information is added and the ontology changes, previously inferred information may no longer be valid and new inferences are needed. In the case of the relational database, new indices will have to be calculated in order to optimize search over large tables. In order to accomplish these tasks, the system in our case study schedules periodic maintenance operations. The system is taken offline for a short period of time while indices and inferences are calculated and then brought back online for user interaction. The inferred ontology is stored in the memory on the VHO server. Thus, inferencing only needs to occur once during "maintenance mode," and the inferred data is then readily available for all future users. The entire operation of creating indices and inferring new information can be completed on the order of minutes. However, the process of ingesting new information into the relational database can be tedious. The advantages of the Semantic Web approach to information ingestion are discussed in forthcoming sections.

EVALUATION

Query Selection

In order to evaluate the utility of our proposed framework we sought to validate it against a number of common heliophysics queries. Through informal discussion with domain experts we arrived at a set of queries that generalizes the types of questions being asked by today's space physics researchers. We felt that these queries are representative of the community; however, we sought independent verification through an examination of the space physics literature. Specifically, we examined a leading space physics journal (*Journal of Geophysical Research*—impact factor 3.147 (Thomson Reuters 2009)) for papers published during the second half of 2009 (July through December). The six months of space physics articles amounted to forty-three published papers. We used a coding scheme described in Miles and Huberman (1994) and coded these forty-three articles into categories based on the type(s) of data used and how that data was obtained. Figure 8 illustrates the results of our coding.

Of the forty-three published articles twenty-six (60 percent) were applicable to our case study. The excluded works were review papers and computational/theoretical studies that did not require the search and integration of existing data. The remaining papers gave an indication of the data-intensive research questions being asked by space physicists. For example, in 35 percent of the papers, the authors were interested in characteristics of the data. They did not mandate any specific data source, but were open to any results as long as those results contained the

specified characteristics. Contrary to that, 26 percent of the researchers knew precisely which spacecraft and time period they were interested in. The examples found in the literature were in good agreement with the results obtained from discussions with domain experts. As a result, we constructed the first two query categories in Table 1. We refer to these as *query categories* because they contain variables such as "Observatory O." During the evaluation process, these query categories are expanded into a longer list of actual queries. These query types mimic the predominant methods that are currently used to find space physics data.

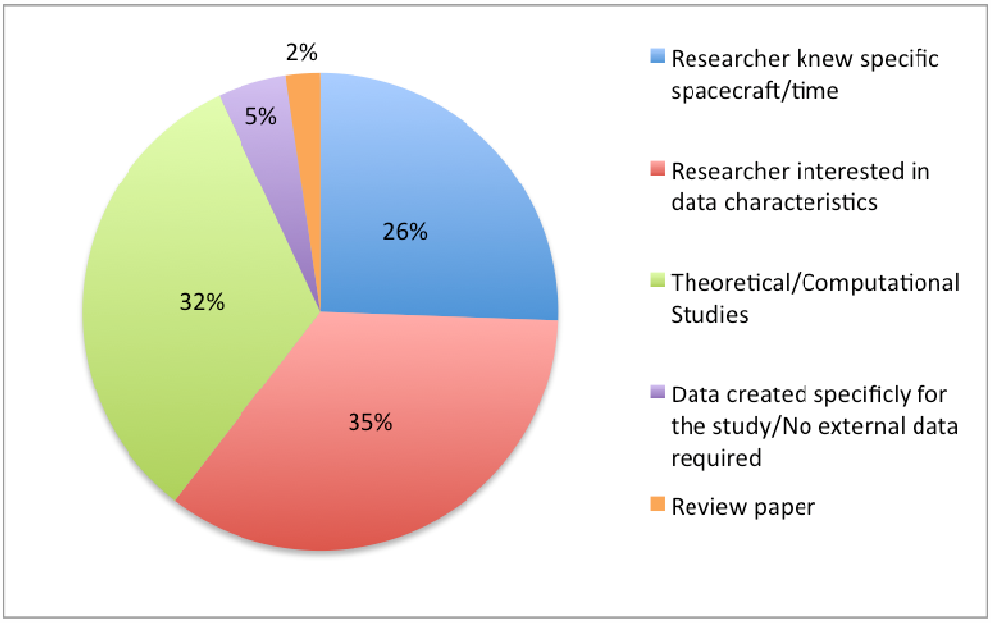


Figure 8. The distribution of information sources used by researchers from July 2009 to December 2009.

As previously mentioned, the VHO addresses a specific sub-domain of space physics. This is due to a NASA programmatic decision and not related to the scientific questions that researchers would like to ask. Through our discussions with domain experts, and also through the literature coding, we found that users often had queries that spanned multiple sub-domains. As a result, users would have to artificially decompose their questions to match the specific capabilities of each system (see Figure 4). This led us to the third query type in Table 1. We sought to explore the Semantic Web's reasoning capabilities and autonomously determine which portions of a query were relevant and which portions needed to be forwarded on to another information system.

Research is a dynamic and fluid process and as a result it is impossible to predict the long-term directions of space physics research. However, based on our domain expert interviews and their subsequent correlation with current research, we believe we have a representative set of queries for the foreseeable future. Table 1 highlights the query categories used in our evaluation.

Table 1: Query Categories Used For Evaluations

Query Category	Query Description
Q1	Find data files associated with observatory *O* over time period *T*
Q2	During time period *T* find times when the parameter Velocity was exceptionally high (V > 1000) and find data during those times from all instruments capable of measuring Velocity.
Q3	Find data from another sub-domain during the time of known events in my sub-domain

Query Evaluation

Using the aforementioned query categories, we evaluated the performance of our system first without the domain ontology, which is the pure relational database. For the relational database, we show complexities in retrieving the required information using a purely relational model. We then assess the utility of the system with the Semantic Web domain ontology and discuss the reasoning aspects of answering the query. In the end we present a complete overview of both systems under real-world use cases and offer empirical evidence regarding the efficacy of our proposed solution. All query categories were carried out on the VHO production server that contains dual 2.6 GHz processors and 4 GB of RAM.

Evaluation of Query Category 1

Query Category 1 (Q1) follows from our coding data in which a number of researchers knew the specific spacecraft and time period they were interested in. They subsequently wanted to find all available data from that spacecraft. Our ontology contains the *hasData* object property that links observatories to the data that they produce. When reasoning occurs the ontology is able to infer all data sets related to a user input observatory. Conversely, the relational database utilizes three tables—OBSERVATORIES, INSTRUMENTS, and PRODUCTS. Instrument-Observatory relationships are found manually prior to database ingestion and made explicit through a foreign key relationship. During query execution of the relational database system the OBSERVATORIES and INSTRUMENTS tables are joined using the observatory_id column.

We substituted various values for "Observatory O" in Q1 and executed each query 100 times in both the purely relational and semantic web scenarios. The average and standard deviation that result from 100 executions allows us to provide an accurate comparison of query execution time. By examining a distribution of execution times, we are able to account for variations in how the relational database query planner implements a given query on subsequent executions. Figure 9 shows the results of these executions, with open circles representing queries using the ontology and filled circles represent SQL only queries and the standard deviation is shown as error bars.

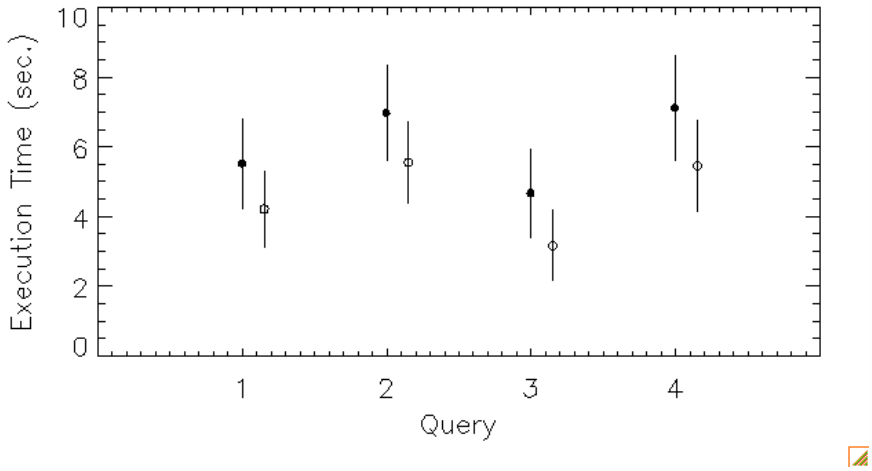


Figure 9. Execution time for queries within Query Category 1 (open circles represent queries using the ontology while filled circles represent SQL only queries. The standard deviation of 100 queries is shown as error bars).

Over 100 iterations the Semantic Seb implementation performed as well, or slightly better, than the database system. Performance does not suffer by having to first query the ontology and then the database. Further, there is an implicit benefit to the Semantic Web approach. The relational model cannot function without explicit OBSERVATORY–INSTRUMENTS–PRODUCTS relationships being defined. The dependence on a database administrator affects the rate at which new information enters the systems and leads to a bottleneck in new data being available to users—thus slowing the rate of scientific progress. Conversely, the reasoning capabilities of the semantic web allow data to be ingested nearly autonomously with the relationships inferred by the system.

Evaluation of Query Category 2

A common scenario that emerged from our analysis of space physics research is the need to find data that satisfied certain criteria regardless of which spacecraft it came from. Query Category 2 (Q2) models this scenario and looks at its implications.

In the purely relational database we need to implement a sub-query and several JOIN operations in order to determine data sets of interest. This is because we know only characteristics of the data sets and need to obtain the specific names of matching data sets. Conversely, the Semantic Web approach offers us a short cut by utilizing domain semantics. Through reasoning we are able to infer which data sets should be of interest. By placing this knowledge within the formal ontology the UI can be less complex. That is, the UI need not know how to formulate the complex SQL query needed in the purely relational scenario. The design and maintenance of the federation layer thus becomes less cumbersome. A quantitative measure of this affect is discussed in section 5.3. Figure 10 shows the results of queries within Query Category 2.

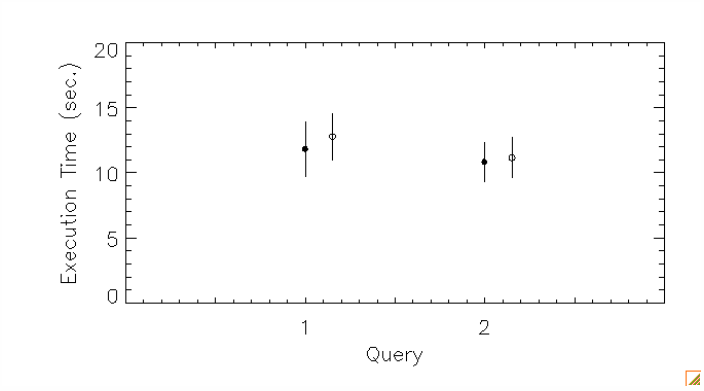


Figure 10. Execution time for queries within Query Category 2 (open circles represent queries using the ontology while filled circles represent SQL only queries. The standard deviation of 100 queries is shown as error bars).

Evaluation of Query Category 3

Cross-disciplinary integration is the second-stage goal of the NASA heliophysics data environment. In attempting to accomplish this, it is easy to overlook a primary requirement. Each Virtual Observatory must know the capabilities, content, and scope of all other Virtual Observatories. For domain scientists this is a straightforward task and something that comes naturally from working in the field. For information systems this is a formidable challenge, especially with the lack of inference capability in relational databases. In order for relational databases to recognize that parts of a query are potentially relevant to another discipline, they would need to deploy various ad hoc methods. Attempting to match various keywords, using lookup tables, or other heuristics would have to be employed. Such mechanisms do not guarantee accuracy, could vary in implementation from system to system, and are a direct result of not having the ability to utilize domain semantics. On the other hand, our Semantic Web approach takes advantage of the necessary semantic information. The domain ontology covers all of heliophysics relevant to our problem. With this approach the VHO is capable of reasoning about concepts within a query and we were able to infer to which part of the heliophysics domain they belong.

The lack of inference capabilities means that Query Category 3, a significant domain and user requirement, is not executable within the purely relational model. Our interviews with domain experts revealed that the lack of this capability lead to increased time and effort on their part to conduct their research. Researchers were required to artificially decompose their questions in order to match the capabilities of the various information systems (Figure 4). This was generally not a trivial process for users.

However, within the Semantic Web system, we were able to execute this query. Utilizing the domain ontology allowed us to offer more choices to the user. Users did not need to artificially decompose their query to match the capabilities of multiple information systems. Rather, the Semantic Web system was capable of accepting complex queries and, without user intervention, infer what it could answer and forward the remaining portions onto other information systems within the domain.

Evaluation of Query Complexity

Databases are prevalent in today's information systems, and the relationship of dynamically created SQL statements and software faults is an emerging topic within software engineering. Quah and colleagues (Quah et al. 2004; Quah et al. 2006) have investigated this relationship and empirically derived weights for various SQL commands. Their results show that the number of SELECT statements, number of sub-queries, number of conditions after the WHERE clause, number of GROUP By statements, and number of insertions/deletions are the leading causes of software faults (Quah et al. 2004; Quah et al. 2006). Moreover, Quah provides a weighting for each of these statements from which a query complexity can be derived and the number of expected faults can be predicted. Unfortunately, Quah does not provide the functional form of the complexity measure. We assume a linear combination of components, and we utilize such a combination to estimate the complexity of the UI generated queries with and without the ontology. Figure 11 illustrates the complexity of the SQL queries Q1 and Q2 assuming a linear combination of components. The solid circles indicate complexity in the purely relational model where the UI must create SQL statements directly from user input. The open circles in Figure 11 show the complexity the same SQL statement after the UI is able to consult the ontology and obtain more domain information. The numbers in parenthesis are the difference in complexity from relational model to Semantic Web model. Figure 11 shows that the Semantic Web model reduces the complexity of the UI software and will lead to fewer software faults within a production environment.

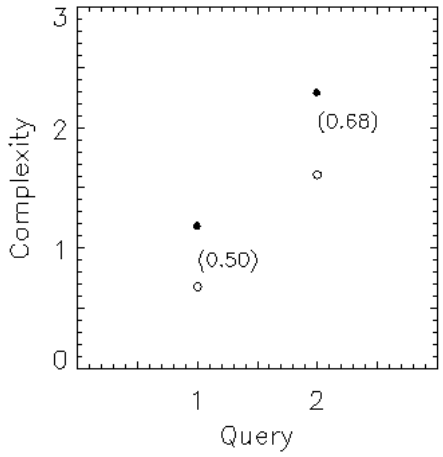


Figure 11. The complexity of SQL queries from direct UI-database interaction (solid circles) and UI-ontology-database interaction (open circles) (the numbers in parenthesis indicate the differences in the two approaches).

The Semantic Web approach also led to a complexity reduction within the relational database. Thirty percent of the relational database was found to be tables and procedures that simulated reasoning. The creation of the ontology, and subsequent removal of these tables, allows for easier maintenance of the database. This reduced maintenance translates into fewer person hours and lower costs.

Discussion and Conclusion

This paper presents a novel middle of the road approach to apply semantics to the federation layer in FIS. This approach overcomes many of the challenges inherent in the two current dominant approaches. Moreover, we have applied our proposed framework to a real-world data integration problem in space physics and have shown how it can enhance the capabilities of existing relational database technologies. Our application in space physics has shown that the Semantic Web is a powerful tool for data integration; yet, its infancy still leaves several shortcomings.

On the positive side, our framework provides a more straightforward, and often quicker, means of executing queries. Complex SQL statements can often be expressed much easier, and more intuitively, as ontology queries. This further results in a simpler and easier to maintain software interface. For example, the initial UI needed to be sophisticated enough to create, and validate, complex SQL queries owing to the multitude of ways space physicists want to query data. With the ontology handling reasoning and inference, the size and complexity of the UI software decreases dramatically. Additionally, many relational database tables were found to simulated reasoning. The creation of the ontology, and subsequent removal of these tables, allows for easier maintenance of the database. Moreover, the inference capabilities of the Semantic Web remove the bottleneck of explicitly stating relationships and getting new information into the system.

In addition to easier query writing the Semantic Web provided us with capabilities not found in the relational database implementation. During the course of the case study something as simple as a cardinality change provided significant problems for the relational database. A change from one-to-one to one-to-many meant a complete redesign of the database and a new Entity-Relationship Model (Chen 1976). However, such a change could be implemented quickly and easily in the ontology. It was a matter of changing a property value on one of the classes, was accomplished in seconds, and did not affect the rest of the system. Additional capabilities come in the form of reasoning that cannot be matched in the relational database system. As mentioned previously, the Virtual Observatory paradigm consists of web services in the presentation layer with the intention of providing visualization and data analysis. These services often operate on broad categories of data. For example, a visualization service may be capable of plotting data of a certain measurement type. The classification and reasoning capabilities of the Semantic Web provide us a convenient means of determining which services apply to our results data. Once we've returned results from the relational database system the semantics are lost. The system no longer knows what type of data it is or from where it originated. It would involve a complex SQL query at this point to retrieve relevant web services. On the contrary, the semantics of the ontology can easily recognize the type of data and determine appropriate services. However, the tool support often required a significant knowledge of OWL and XSLT, thus it was mandatory to have an expert on hand. Heliophysics researchers would have faced a steep learning curve and would have been limited in their efforts had they attempted it alone.

Mixed within this case study are a number of positives regarding the Semantic Web as well as a few negatives. Our proposed framework simplified the data integration task and made system maintenance easier. It has also provided an ideal way of achieving NASA's second stage goal of cross-disciplinary integration. Despite the short-term limitations of the Semantic Web, we see the technology as a key component in data discovery and integration both within space physics and beyond.

REFERENCES

Bao, J. and V. Honavar, "Divide and conquer Semantic Web with modular ontologies—a brief review of modular ontology language formalisms," *Proceedings of the ISWC'06 International Workshop on Modular Ontologies (WoMo '06),* 2006, Atlanta, GA.

Bell, G.B. and A. Sethi, "Matching records in a national medical patient index," *Communications of the ACM*, 2001, 44:8, pp. 83–88.

Bohring, H. and S. Auer, "Mapping XML to OWL ontologies," *Proceedings of 13th Leipziger Informatik-Tage (LIT 2005)*, 21–23 September 2005, Lecture Notes in Informatics (LNI).

Brunner, J-S., L. Ma, C. Wang, L. Zhang, D. Wolfson, Y. Pan, and K. Srinivas, "Explorations in the use of Semantic Web technologies for product information management," *Proceedings of the International World Wide Web Conference 2007*, 8–12 May 2007, Banff, Alberta, Canada.

Burners-Lee, T., J. Hendler, and O. Lassila, "The Semantic Web," *Scientific American*, 17 May 2001.

Busse, S., R. Kutsche, U. Leser, and H. Weber, "Federated Information Systems: Concepts, terminology and architectures," Forschungsbericht Nr. 99–9, Fachbereich Informatik, Technical Report, April 1999, Berlin, Germany.

Chen, P., "The Entity-Relationship Model—toward a unified view of data," *ACM Transactions on Database Systems*, 1976, 1:1, pp. 9–36.

Cho, A. and D. Giustini, "Web 3.0 and health librarians: An introduction," *Journal of the Canadian Health Libraries Association*, 2008, 29, pp. 13–18.

Clifton, C., E. Housman, and A. Rosenthal, "Experience with a combined approach to attribute-matching across heterogeneous databases," *Proceedings 7. IFIP 2.6 Working Conference on Database Semantics,* 1997*,* pp. 429–451.

Cristani, M. and R. Cruel, "A survey on ontology creation methodologies," *International Journal on Semantic Web and Information Systems,* 1:2, April–June 2005, pp. 49–69.

Dalton, R.; "Geophysicists combine forces," *Nature,* 2007, 447:7148, pp. 10–37.

de Hoog, R., "Methodologies for building knowledge based systems: Achievments and prospects," In *Handbook of Expert Systems*, Liebowitz, J. (ed.),. CRC Press, Boca Raton, FL, 1998, chapter 1.

Ding, Y. and S. Foo, "Ontology research and development, part 2—a review of ontology mapping and evolving," *Journal of Information Science*, 2002, 28:5, pp. 375–388.

Gomez-Perez, A., M. Fernandez-Lopez, and O. Corcho, *Ontological Engineering*, *3rd edition,* London: Springer-Verlag, 2004.

Harvey, C., M. Gangloff, T. King, C. Perry, D. Roberts, and J. Thieman, "Virtual observatories for space and solar physics research," *Earth Science Informatics*, April 2008, 1:1, pp. 5–13.

Hasselbring, W., G.-J. Houbeny, B. Riegerz, M. Roantreex, and K. Subieta, "Research issues in Federated Information Systems: Report of the EFIS International Workshop," In *Engineering Federated Information Systems, Proceedings of the 3rd Workshop EFIS 2000,* Roantree, M., W. Hasselbring, and S. Conrad (eds.), June 2000, Dublin, Ireland, pp. 19–20.

Kendall, K.E. and J.E. Kendall, *Systems Analysis and Design*, *3rd edition,* Upper Saddle River, NJ: Prentice Hall, 1995.

McCracken, D.D. and M.A. Jackson, "Life cycle concept considered harmful," ACM Software Engineering notes, 1982, 7, pp. 229–232.

McGuinness D., P. Fox, L. Cinquini, P. West, J. Benedict, and J. Garcia, "Current and future uses of OWL for scientific data frameworks: Successes and limitations," *OWL Experiences and Directions 2007, CEUR Workshop Proceedings*, 2007.

Miles, M.B. and A.M. Huberman, *Qualitative Data Analysis: An Expanded Sourcebook, 2nd edition,* Newbury Park, CA: Sage Publications, 1994.

Narock, T.W., A. Szabo, and J. Merka, "Using semantics to extend the space physics data environment," *Computers & Geosciences*, April 2009, 35:4, pp. 791–797.

NASA Recommended Roadmap for Science and Technology 2005–2035, "The New Science of the Sun-Solar System Connection," 2005. Available at http://sec.gsfc.nasa.gov/sec_roadmap.htm, last accessed 30 August 2009.

Prud'hommeaux , E. and A. Seaborne, "SPARQL Query Language for RDF," W3C Recommendation, 15 January 2008. Available at: http://www.w3.org/TR/2008/REC-rdf-sparql-query-20080115/, last accessed 30 August 2009.

Thomas, M.A., V.Y. Yoon, and R. Redmond, "Extending loosely coupled Federated Information Systems using agent technology," *International Journal of Intelligent Information Technologies*, 2007, 3:3, pp. 1–20.

Tong-Seng Quah, "Mie Mie Thet Thwin: Prediction of software development faults in PL/SQL files using neural network models," *Information & Software Technology*, 46:8, 2004, pp. 519–523.

Tong-Seng Quah, "Mie Mie Thet Thwin: Utilizing computational intelligence in estimating software readiness," IJCNN, 2006, pp. 2999–3006.

Slaughter, L., D. Soergel, and T. Rindflesch, "Semantic representation of consumer questions and physician answers," *International Journal of Medical Informatics*, 2006, 75, pp. 513–529.

Staab, S., H.P. Schnurr, R. Studer, and Y. Sure, "Knowledge processes and ontologies," IEEE Intelligent Systems, 2001, 16:1, pp. 26–34.

Szalay, A., J. Gray, "The world-wide telescope," *Science,* 293, 14 September 2001, pp. 2037–2038.

Thomson Reuters, "Introducing the Impact Factor." Available at http://thomsonreuters.com/products_services/science/academic/impact_factor/, last accessed 5 February 2009.

Vdovjak, R., P. Barna, and G.-J. Houben, "Designing a Federated Multimedia Information System on the Semantic Web," *Advanced Information Systems Engineering, Lecture Notes in Computer Science*, 2003, 2681, Springer Berlin/Heidelberg.

Zhao, H. and S. Ram, "Combining schema and instance information for integrating heterogeneous data sources," *Data and Knowledge Engineering*, 2007, 61, pp. 281–303.

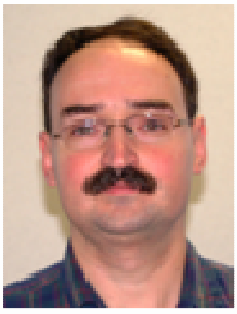
Zhizhin, M., E. Kihn, R. Redmon, D. Medvedev, and D. Mishin, "Space physics interactive data resource—SPIDR," *Earth Science Informatics*, 2008, 1:2, pp. 49–103.

ABOUT THE AUTHORS

Tom Narock is a doctoral student in the Department of Information Systems at the University of Maryland, Baltimore County. He received his B.S. from the University of Maryland, College Park, and his M.S. from Johns Hopkins University. His research interests are in subjectivity and context and the role they play in knowledge management and knowledge discovery. Recently he has focused on the emerging Semantic Web and the role it can play in knowledge management.

Victoria Yoon is a Professor in the Department of Information Systems at the University of Maryland, Baltimore County. She received her M.S. from the University of Pittsburgh and her Ph.D. from the University of Texas at Arlington. She has published articles in such leading journals as *MIS Quarterly*, *Decision Support Systems, Communications of the ACM*, and *Journal of Management Information Systems*. Her primary research area has been the application of Artificial Intelligence to business decision-making in organizations and technical and social issues surrounding such applications.

Jan Merka has been an Associate Research Scientist at the University of Maryland, Baltimore County, since 2006. Located at NASA/Goddard Space Flight Center since 2001, he works on several space physics research projects. He is the Principal Investigator of the Virtual Magnetospheric Observatory. Dr. Merka recieved his Ph.D. in plasma physics from the Charles University, Prague, Czech Republic.



Adam Szabo is Director of the Heliospheric Physics Laboratory at NASA's Goddard Space Flight Center. He has authored or co-authored over ninety scientific papers. Dr. Szabo is the data manager for the magnetic field investigation of the WIND spacecraft and he serves as the Principal Investigator of the Virtual Heliospheric Observatory. He received his Ph.D. in Physics from the Massachusetts Institute of Technology.

JITTA

JOURNAL OF INFORMATION TECHNOLOGY THEORY AND APPLICATION

JITTA is a Publication of the Association for Information Systems
ISSN: 1532-3416