Computers & Geosciences ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at SciVerse ScienceDirect

Computers & Geosciences

journal homepage: www.elsevier.com/locate/cageo

Review

From science to e-Science to Semantic e-Science: A Heliophysics case study

Thomas Narock [a,b,*], Peter Fox [c]

[a] Adnet Systems, Inc, Rockville, MD, USA
[b] NASA/Goddard Space Flight Center, Heliospheric Physics Laboratory, USA
[c] Rensselaer Polytechnic Institute, Tetherless World Research Constellation, USA

ARTICLE INFO

Article history:
Received 30 November 2010
Received in revised form
15 November 2011
Accepted 16 November 2011

Keywords:
Semantic e-Science
Semantic Web
Ontology
Provenance

ABSTRACT

The past few years have witnessed unparalleled efforts to make scientific data web accessible. The Semantic Web has proven invaluable in this effort; however, much of the literature is devoted to system design, ontology creation, and trials and tribulations of current technologies. In order to fully develop the nascent field of Semantic e-Science we must also evaluate systems in real-world settings. We describe a case study within the field of Heliophysics and provide a comparison of the evolutionary stages of data discovery, from manual to semantically enable. We describe the socio-technical implications of moving toward automated and intelligent data discovery. In doing so, we highlight how this process enhances what is currently being done manually in various scientific disciplines. Our case study illustrates that Semantic e-Science is more than just semantic search. The integration of search with web services, relational databases, and other cyberinfrastructure is a central tenet of our case study and one that we believe has applicability as a generalized research area within Semantic e-Science. This case study illustrates a specific example of the benefits, and limitations, of semantically replicating data discovery. We show examples of significant reductions in time and effort enable by Semantic e-Science; yet, we argue that a "complete" solution requires integrating semantic search with other research areas such as data provenance and web services.

Contents

1. Introduction

Many organizations are leveraging the Information Age (Hey and Trefethen, 2005) to move online to find new avenues of reaching their consumers. The past few years have witnessed unparalleled efforts (Dalton, 2007) to make scientific data web accessible especially in the physical sciences. The Semantic Web (Berners-Lee et al., 2001) has proven invaluable in this effort. The Semantic Web aids in data discovery and integration and several implementations can be found in the literature (Madin et al., 2007; Neumann, 2005; Fox et al., 2007). Specifically, within the field of heliophysics three Semantic Web enabled search and integration systems are available (McGuinness et al., 2007a; Narock et al., in press).

However, much of the literature is devoted to system design, ontology creation, and trials and tribulations of current technologies (e.g. Madin et al., 2007; McGuinness et al., 2007a; Narock et al., in press). Moreover, Semantic e-Science is emerging as its own discipline (Fox and Hendler, 2009a; McGuinness et al., 2009) in which intelligent applications automate scientific research tasks. The aforementioned implementation and deployment studies serve

* Corresponding author at: NASA/Goddard Space Flight Center, Code 672, Greenbelt MD 20771, USA. Tel.: +1 301 286 1086.
E-mail address: Thomas.W.Narock@nasa.gov (T. Narock).

0098-3004/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.cageo.2011.11.018

as valuable first steps; yet, we need to complement these technology demonstrations in order to move Semantic e-Science mainstream. We need to investigate the socio-technical implications of semantic search and the effects on end users. Technology research and user behavior are not dichotomous, they are inseparable (Lee, 2000). In order to fully develop Semantic e-Science we must take a design-science approach (Hevner et al., 2004) in which we investigate not only how a system works, but also why it works. Evaluation of systems in real-world settings is invaluable to this process.

A recent National Science Foundation report (Cummings et al., 2008) concluded that "few studies integrate the distinct social, organizational, and infrastructure dimensions of dynamic distributed collaborations". The Cummings report (2008) also touches on the physical sciences as a key application area for this type of study. In regards to semantic technologies, the physical sciences are amongst the early adopters of semantic technologies (Finin and Sachs, 2004) and serve to benefit greatly from it.

In this work we describe a case study within the field of Heliophysics and provide a comparison of the evolutionary stages of scientific research. Specifically, we start with a recent Heliophysics study conducted under traditional means—manual inspection of large quantities of data and personal interaction with data producers and other domain experts. We label this effort ***science*** as it describes the norm in heliophysics over the past few decades. The data discovery portion of this Heliophysics study is then reproduced using multiple online search and retrieval systems. This effort is termed ***e-Science*** and describes the recent trend to automate data discovery using web services, databases, and other Internet technologies. Finally, we reproduce the data discovery a second time using a new semantic search and retrieval system. We label this scenario ***Semantic e-Science***.

The ***science*** scenario is used as a baseline. It provides us with an estimate of the status quo. It tells us the processes by which the heliophysicists conducted their research and the results that they achieved. Comparing the ***Semantic e-Science*** results with ***science*** and ***e-Science*** allows us to see how well semantic search enhances the scientific process. In other words, we are primarily focused on how well semantic search enhances traditional scientific data discovery. We are interested in the socio-technical implications of moving toward automated intelligent discovery and we seek to understand how well this process enhances what is currently being done manually in various scientific disciplines. The novelty of this research is that we begin to understand the broader impacts of moving Semantic e-Science mainstream and in doing so we are able to come full-circle and report on the relationship between semantic technologies and their efficacy in scientific research. We also provide empirical evidence in support of several founding propositions of Semantic e-Science.

Specifically we aim to address the following questions: How efficiently and effectively does semantic search allow us to facilitate ingrained data discovery techniques in one physical science area? Does semantic search arrive at the same (or better) conclusions as researchers using manual data discovery means? Is semantic search by itself sufficient for data discovery in scientific research? What affect does semantic search have on the research process? What general statements of applicability can be made to other science areas?

2. Setting: NASA Heliophysics data environment

The NASA Heliophysics data environment[1,2] is broken into several sub-disciplines each with their specialized instrumentation and data discovery methods. Following the recent trend of online data discovery, and wanting to retain the uniqueness of each of these sub-disciplines, NASA, in 2007, commissioned several federated information systems. Five information systems were funded serving the solar, heliospheric,[3] magnetospheric,[4] Earth's radiation belts, and Earth's upper atmosphere communities. Each of these systems is responsible for providing uniform access to their respective underlying sources of heterogeneous and distributed data. These systems are broadly known as Virtual Observatories, a paradigm (Szalay and Gray, 2001) that began in astronomy and quickly spread to heliophysics, oceanography, volcanology, and other diverse scientific communities. Specifically, the Virtual Observatory paradigm unites large quantities of disparate and heterogeneous data usually under one web-based portal. The underlying data remain heterogeneous and distributed, yet common metadata, access protocols, and terminology provide transparent access to users. Within the NASA Heliophysics domain the five Virtual Observatories are supported by numerous data analysis and visualization capabilities,[5,6] that enable the provision of a diverse Heliophysics data environment.

The systems within NASA's Heliophysics Data Environment implement search capabilities relevant to their domain. For example, the Virtual Solar Observatory,[7] dealing primarily with solar images, focuses on optical search parameters such as wavelength and intensity. The Virtual Heliospheric Observatory,[8] by contrast, deals primarily with in situ time series data. In a similar manner the remaining NASA Virtual Observatories implement search capabilities analogous to the types of data they contain.

The diversity of scientific data and numerous methods by which to relate it makes for a challenging search problem. Previous research (Merka et al., 2008a; King et al., 2008) into the applicability of traditional information retrieval techniques implemented in contemporary internet search engines (e.g. Google, Yahoo, etc.) has shown that relevance scoring, such as PageRank (Brin and Page, 1998), while successful in web search, fails in scientific search scenarios as there are few predefined connections between data. Any connections, if they exist, are determined during the research process by domain experts.

The recently emerged microformats[9] are also impractical for scientific search. Specifically, microformats offer a means to augment XHTML with well-defined semantics. Microformats have seen extensive usage lately as the number of web sites using them has been estimated to be in the hundreds of millions.[10] However, the use of microformats implies search over web pages, which is not how search is conducted within the sciences (Merka et al., 2008a, King et al., 2008). The aforementioned lack of predefined connections between data and the fact that scientific data, and associated metadata, are often not stored as web pages precludes the use of microformats. In a similar manner, popular techniques such social tagging and collaborative filters are also not relevant for many scientific search and retrieval systems. Tags and recommendations can be useful in scientific collaborations; however, they are primarily useful for initial discovery and latter evaluation of fitness for use, and not the primary means of data retrieval. Rather, users are supplying a set of constraints on the data and the system is identifying time periods and data sets in which those constraints are met. Thus, solutions involving domain

[1] http://dx.doi.org/10.1029/2009EO470001.
[2] http://hpde.gsfc.nasa.gov/.
[3] The region of the solar system affected by the Sun.
[4] The region of space encompassed by the Earth's magnetic field.
[5] http://hpde.gsfc.nasa.gov/hpde_data_access.html.
[6] http://spdf.gsfc.nasa.gov/.
[7] http://virtualsolar.org.
[8] http://vho.nasa.gov.
[9] http://microformats.org.
[10] http://microformats.org/blog/2007/06/21/microformatsorg-turns-2/.

Please cite this article as: Narock, T., Fox, P., From science to e-Science to Semantic e-Science: A Heliophysics case study. Computers & Geosciences (2011), doi:10.1016/j.cageo.2011.11.018

semantics and the use of logical inference are the ideal solution for most scientific search use cases.

Combining the diversity of data, the diversity of search types, and the need to provide uniform online access we arrive at a scenario ripe for semantic search. Semantic technologies, such as the Web Ontology Language (OWL) (McGuinness and van Harmelen, 2004b) utilize the meaning of the data as well as relationships between constituent data. The ability of semantic search to enable subsumption, inheritance, and inferencing capabilities helps overcome data heterogeneity and terminology differences and allows encoding of previously unconnected concepts. The lack of these capabilities in syntactic search makes its utilization in scientific data discovery limited.

In particular, two of the NASA systems (the Virtual Heliospheric Observatory (VHO) and the Virtual Magnetospheric Observatory (VMO)) now use formal ontologies to discover and integrate data (Narock et al., in press). Moreover, these two systems were specifically designed to share technology and infrastructure (Merka et al., 2008a) in order to reduce costs and shorten development time. As a result, operations of each of these systems are completely driven by the ontologies. In other words, there is one underlying software infrastructure with the two domain semantics completely captured in independent formal ontologies. Simply pointing the software to one of the ontologies changes the domain that is searched. McGuinness et al. previously developed (2007a) a similar framework, which may hint at a convergence of semantic deployment methodologies. However, the questions posed in Section 1 regarding the efficacy of these systems still remain largely unanswered.

3. Case study

3.1. Science

The Slavin et al. (2002) study that is the basis of our case study represents research questions that are common within the field of Heliophysics. Moreover, the Slavin (2002) study requires multiple types of data to be available during a stringent set of spatio-temporal conditions. Specifically, the Slavin study specifies a spatial configuration of four spacecraft while specific instruments onboard these spacecraft are operating and collecting data. Any four spacecraft are acceptable as long as they meet the spatio-temporal restrictions listed in Table 1. All restrictions must occur simultaneously. The spatial restrictions are listed using the Geocentric Solar Magnetospheric (GSM) coordinate system—an Earth-centered system common in Heliophysics. Such a configuration of spacecraft is rare and heliophysicists do not know its occurrence intervals a priori.

As a result, Slavin and colleagues (2002) selected the most likely subset of spacecraft that might fit their needs. While this selection was based upon extensive experience within the domain it nevertheless highlights the gap between data sources domain experts are familiar with and data sources that are regularly emerging and appearing online. In recent years this gap has been widening (Gray et al., 2005) to the point of being acknowledged in the popular media (Hotz, 2009). With more, and increasingly voluminous, data constantly coming online, knowing what to search is becoming intractable. This is increasing the need for search in general and semantic search in particular.

Slavin's team manually investigated 16 months of an initial data source to undercover 43 possible events. Gradually, these events were whittled away as the remaining spatio-temporal restraints were added. All told, 2 useable events where identified after ~100 h of manual labor.[11]

[11] Private communication with the study's first author Dr. James Slavin.

Table 1
Spatio-temporal restrictions used by Slavin et al. (2002) and Merka et al. (2008b). The restrictions for the queries used in our case study.

1.	Measurements of type magnetic field occurring in the range $-150 <= \text{GSM X (Re)} <= 15$
2.	Measurements of type thermal plasma occurring in the range $-15 <= \text{GSM X (Re)} <= -5$
3.	Measurements of type magnetic field occurring in the range $-5 <= \text{GSM X (Re)} <= -2.5$
4.	Magnetic field and thermal plasma measurements occurring in the range $20 <= \text{GSM X (Re)} <= 50$
5.	Restrictions 1., 2., and 3. must also be restricted to the range $-10 <= \text{GSM Y (Re)} <= 10$
6.	Restriction 4. must also be restricted to the range $-40 <= \text{GSM Y (Re)} <= 40$

3.2. e-Science

While automated web technologies can reduce much of the aforementioned workload the question remains as to the efficacy of such a system and technology's ability to enhance the search process. To examine this in a controlled setting we first replicated (Merka et al., 2008b) the Slavin (2002) study using the initial versions of VHO and VMO. The initial systems used purely relational database technology converting web form input into Structured Query Language (SQL) queries. While these systems had many advantages, reducing the search time to ~2 h, they also highlighted several socio-technical limitations.

The VHO and VMO integrate data within two distinct heliophysics sub-domains. However, heliophysics research questions do not always fall into one sub-domain or the other. Rather, in today's world of cross-disciplinary and system science, research questions often encompass resources from multiple domains. Thus, researchers think in terms of complex interrelated systems while managers are often forced to deploy multiple component-based IT systems that are more manageable. Moreover, despite best efforts, users are often unaware of such divisions. In our particular case, our spatio-temporal requirements span the sub-domains covered by VHO and VMO. A standardized messaging framework (Narock and King, 2008) allows results from one system to be saved and used as a query in the other system. However, users must first be aware of the existence and capabilities of both systems. Subsequently, the researcher must artificially decompose their query into multiple questions of the respective systems—a process that is foreign to most researchers.

Other limitations come behind the scenes of the information systems. In a relational search system all relationships must be stated explicitly. This simple fact has two far-reaching consequences. First, it limits the discovery of new knowledge. Users may find relationships unbeknownst to them; however, they will not find a relationship that was not pre-defined by system maintainers. Semantic search alleviates this limitation and offers the potential for knowledge discovery—a capability essential to scientific research. Second, the system maintainer who defines, approves, and commits these pre-defined relationships serves as a bottleneck to the system. This person slows down the addition of new data, which in turn slows down the scientific process.

3.3. Semantic e-Science

The data search portion of the Slavin (2002) study was once again reproduced. By replacing the e-Science structured SQL queries with inferences over domain ontologies we sought to identify which areas of the scientific research process are adequately addressed by semantic search and which parts may still be lacking. For this process, we utilized the SPASE

Please cite this article as: Narock, T., Fox, P., From science to e-Science to Semantic e-Science: A Heliophysics case study. Computers & Geosciences (2011), doi:10.1016/j.cageo.2011.11.018

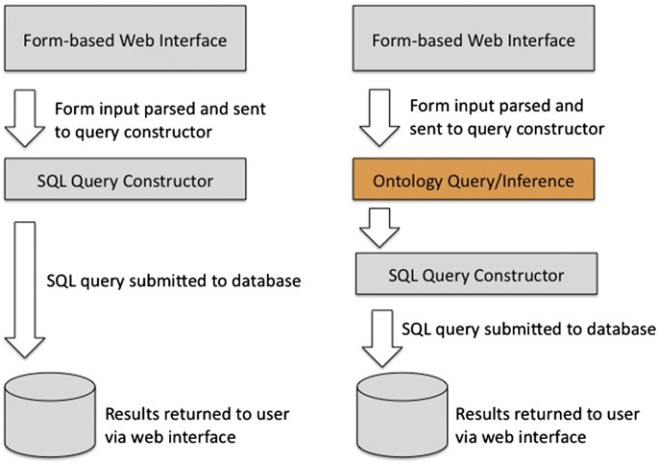

Fig. 1. Graphical depiction of system complexity within the e-Science and Semantic e-Science scenarios.

ontology[12] with its VHO[13] and VMO[14] instantiations. These ontologies are OWL encodings (Narock et al., in press) of the XML-based schema developed by the Space Physics Archive Search and Extract (SPASE) consortium (Harvey et al., 2008). Specifically, these ontologies have Description Logic (Baader et al., 2003) expressivity *ALCOIN(D)*, 35 classes, 20 object properties, 45 datatype properties, and on the order of 20,000 instances.

The aforementioned SPASE consortium, which consists of heliophysics researchers, software developers and data providers, was founded to aid in the integration of heliophysics data. The intent of SPASE is to create a metadata schema that would lead to standardized descriptions of NASA Heliophysics data resources. Within this case study the On-To-Knowledge methodology (Staab et al., 2001) was used (Narock et al., in press) to continuously convert releases of the SPASE metadata model into OWL. It should be noted that knowledge acquisition and knowledge representation have historically been challenging and time consuming efforts. However, On-To-Knowledge (Staab et al., 2001) and other emerging methodologies (e.g. Benedict et al., 2007; Fox and McGuinness, submitted for publication) are beginning to lead to quick deployment, rapid community buy-in, and reduced effort on the end-users part.

We utilized the final versions of the VHO and VMO search systems that now utilize formal ontologies (Narock et al., in press). Fig. 1 illustrates the how the VHO and VMO search systems were augmented by the addition of semantics.

Evident from Fig. 1 is that minimal changes were made to the system architecture. Specifically, the system now consults the aforementioned ontologies prior to constructing SQL queries. Domain logic is captured within the ontologies and can be easily manipulated and inferenced with using Semantic Web tools. This is opposite of the initial system design in which this information had to be hard-coded into the query algorithms. Previous research has shown (Narock et al., in press) that such a scenario reduces complexity, query execution time, and system maintenance. In addition, we can pose our complete query to the system, have it answer the portions it is capable of answering, and infer to which other system to send the remaining portions of the query.

In what has become a catch phrase of the Semantic Web, Hendler (2007) mused, "A little Semantics Goes a Long Way". Such is certainly the case in semantic eScience. The addition of basic Semantic Web technologies – ontologies, reasoner, and software toolkit (Jena[15]) – reduces the data discovery process in our applications from ~2 h to ~15 min. Such a reduction occurs because we no longer need to manually visit 2 sites and more importantly we no longer need to artificially break our query into sub-domain specific questions. In the Semantic e-Science implementation incoming queries are converted from a standard XML representation (Narock and King, 2008) into SPARQL (Prud'hommeaux and Seaborne, 2008), a semantic query language approved as a W3C[16] Recommendation. It is at this stage that the benefits of Semantic e-Science become apparent. Using our domain ontology, which spans all of Heliophysics, we can infer to which sub-domain each component of the query refers. Further inferences can then be made as to which information system within the NASA data environment addresses this sub-domain. At present, this capability is limited to the VHO and VMO as they are the only systems utilizing semantics. Nevertheless, this limited capability illustrates an advantage over e-Science—namely the artificial decomposition of queries.

4. Current Semantic e-Science limitations and infrastructural challenges

While basic Semantic Web technologies reduced the search time by orders of magnitude they are not sufficient to offer a "complete" answer. That is, the Semantic e-Science approach did not retrieve all of the results found within the ***science*** approach. While the Semantic e-Science approach offered previously unknown events to ponder, the semantic search only returned one of the two original Slavin events.

More importantly the results lacked provenance information regarding certain aspects of the query.

While the semantic search correctly identified data sets of interest, calls to services that search the underlying numerical values failed to turn up one of the two original events. This was due to the resolution of data that the service searched over and was not a fault of any semantic inferencing. However, this highlights the importance of coupling semantic search to related research in web services and Semantic Web services. In order to be completely successful these technologies need to be components in frameworks rather than independent streams of research. This observation is consistent with previous research (Lim et al., 2010), which noted that e-Science lacks "a single backbone that can support the entire spectrum of requirements that are essential to undertake cross-disciplinary scientific collaborations." Additionally, Abbott (2009) and (Goodman and Wong, 2009) have commented that current and future scientific challenges can only be addressed with a focus on systems thinking and generalized frameworks, respectively. Such groundwork has been laid (Fox et al., 2009b) and is being advanced by many groups in the Earth and Space Sciences communities.

4.1. Provenance

Also missing from the semantic search results is provenance information. One definition of provenance is "the description of the origins of a piece of data and the process by which it arrived in a database" (Buneman et al., 2001, p. 316). Metadata does exist for all data sets returned by the systems, and this information is readily available to the user. However, subtleties in the execution

[12] http://vho.nasa.gov/ontology/spase.owl.
[13] http://vho.nasa.gov/ontology/vho.owl.
[14] http://vho.nasa.gov/ontology/vmo.owl.
[15] http://jena.sourceforge.net/.
[16] http://w3c.org/.

Please cite this article as: Narock, T., Fox, P., From science to e-Science to Semantic e-Science: A Heliophysics case study. Computers & Geosciences (2011), doi:10.1016/j.cageo.2011.11.018

of queries can extrapolate to the point that results are ambiguous or even unusable. For example, our query required a spatial restriction that was expressed by a keyword—"Near-Earth Heliosphere". Within the domain semantics this term has specific meaning and one can utilize various services to determine if a spacecraft is inside or outside of this spatial region. However, how the data in the service was generated is absent from the result, as is almost always true for current web-based search engines; only limited provenance on the search result itself is often available. The lack of formal identification that provenance metadata is indeed related to, and must be returned with, a search result is a serious omission in the search for relevant scientific data. Absence of this information further leads to questions of accuracy as the computational models (e.g. Peredo et al., 1995; Merka and Szabo, 2004) that predict the spatial region have inherent biases, assumptions, and errors known to the scientists who are most likely to have initiated the search. This particular case is easily remedied by supplying more information to the results, however, as queries become more complex these subtle pieces of information increase and the user's complete understanding of the results diminishes.

Provenance is an actively researched topic (Pinheiro da Silva et al., 2004) that is now finding its place on the Semantic Web (McGuinness et al., 2007b), and specifically within semantic eScience—known as knowledge provenance (Fox et al., 2008a, 2008b; Zednik et al., 2009; Zednik et al., in press), and we see this as a key component to providing users with a "complete" answer to their queries. Complete in the sense that essential explanations, verifications and justifications are given to satisfy a scientist's questions about any returned result.

In the area of knowledge provenance, the encoding of provenance interlinguas such as the Proof Markup Language (PML; Pinheiro da Silva et al., 2004; McGuinness et al., 2007b) and the Open Provenance Model (Moreau et al., 2011) as ontologies (Michaelis and McGuinness, 2010) brings new expressive power to the important but previously disconnected provenance metadata. However, languages for encoding are really only a foundation for provenance. What really address a scientist's curiosity are application tools that work with the encoded provenance. At present these tools include provenance specific search browse and visualization (McGuinness and Pinheiro da Silva, 2004a) as well as integrated applications such as provenance-aware smart faceted search (Fox et al., in preparation). These tools work primarily with PML but similar tools are geared to work with OPM (e.g. the KARMA system; Cao et al., 2009). Many of these tools however are still clumsy for users to use and much work to improve them remains. Of particular significance for convergence in this area of web-based provenance is the formation and activity of a W3C incubator group for provenance.[17]

This progress not-withstanding, 'complete' answers to user queries require a multi-domain knowledge base that is an intersection and super-set of domain knowledge, provenance, and often other data product related information (Zednik et al., 2009; Fox et al., in preparation).

4.2. Ontology mapping and ontology integration

The VHO and VMO share a common base ontology (Narock et al., in press) and as a result queries passed amongst the two systems are easily interchangeable and executable. However, we found this not to be the case within the broader community. Our case study query could have benefitted from other data sources, specifically ones available via the Virtual Solar–Terrestrial Observatory (VSTO) (McGuinness et al., 2007a, Fox et al., 2009b). Specifically, Slavin (2002) utilized an auxiliary data source in order to confirm initial assumptions. Such an effort could have been replicated via VSTO data; yet, that system operates with its own independent ontology and associated web services (Fox et al., 2007). At present there is no formal mapping between VSTO and VHO/VMO ontologies. This is not a criticism of either system. Rather, it is an empirical confirmation of the broad tendency (Cummings et al., 2008) to create one-off systems without taking time or effort for harmonization and as a result the lack of interoperability (in this case, semantic interoperability) can hinder system-level science.

Semantic e-Science needs to have a concerted effort to standardize, align, and map ontologies. Such needs have recently been recognized with several National Science Foundation projects devoted to semantic interoperability, such as the SONET (Madin et al., 2007) project. Also the Semantic eScience Framework (SESF; Fox et al., 2009b; McGuinness et al., 2010) project mentioned earlier has among its goals to enable configurable semantic data frameworks using modular approaches to ontologies and the bridging of disciplines as well as application level integration, also using ontologies to encode meaning and relations for such applications (Rozell et al., 2010). Further, SESF advances the key notion of semantically aware application-level tools as an essential part of the semantic ecosystem, integrating not only data and information sources but provenance as well.

5. Summary and conclusions

Admittedly, one case study can only begin to address the questions surrounding the emergence of Semantic e-Science. Yet, our case study has shed light on emerging issues and provides empirical evidence in support of some of the founding tenets (Fox and Hendler, 2009a) of Semantic e-Science. We now return to consider our initial research questions.

- How efficiently and effectively does semantic search allow us to facilitate ingrained data discovery techniques?
- Does semantic search arrive at the same (or better) conclusions as researchers using manual data discovery means?

Semantic search and basic semantic technologies (i.e. RDF/OWL, reasoners, and toolkits) are providing great benefit to the scientific research process. As evidenced by this case study these basic technologies allow one to replicated previous studies in fractions of the time. While this is but one of many possible metrics (McGuinness et al., 2007c) it is one that busy scientists often care about most. Moreover, these technologies do not require the artificial decomposition of complex questions. Reasoning and inferencing capabilities are providing great benefit to both the user and the system maintainer. This study, as well as previous work (Narock et al., in press), has shown evidence of easier system maintenance and easier query execution when transitioning from e-Science to Semantic e-Science although further quantitative studies are required to substantiate this claim.

While semantic search is a vital component of Semantic e-Science it is not the only component. Interspersed with ontologies are non-semantic technologies such as Web Services and relational databases. How and when these technologies interact is just beginning to be understood. For example, provenance information is often thought of as the lineage of the data and something to present at the end of execution. Yet the benefits of using provenance to enhance the search process can be seen in previous research (Fox et al., in preparation) as well as in evidence

[17] http://www.w3.org/2005/Incubator/prov/charter.

Please cite this article as: Narock, T., Fox, P., From science to e-Science to Semantic e-Science: A Heliophysics case study. Computers & Geosciences (2011), doi:10.1016/j.cageo.2011.11.018

presented earlier in this paper. Similar scenarios should be sought in future research.

Semantic search returned mixed results within our case study. While it pointed us toward previously unconsidered data it also failed to find one of the events from the ***science*** approach. Again, this was not the result of the search processes itself, but rather the ongoing challenges of connecting semantic search to other cyberinfrastructure capabilities.

- Is semantic search by itself sufficient for data discovery in scientific research?

This case study indicates that semantic search as implemented in our case study is insufficient for "complete" data discovery. This conclusion is reached for a number of reasons. First, the massive quantities of scientific data (Gray et al., 2005) that apply to Heliophysics, make it unlikely that all, possibly not even most, data values will reside as ontology instances; let alone as Linked Data (Bizer et al., 2009). Thus, semantic search is only as good as the cyberinfrastructure services to which it connects. This was evidenced in our study where the ontology returned the correct data set, however, data services were unable to retrieve all relevant time periods. Second, provenance plays an important role in semantic search. Without the complete Semantic Web stack (Berners-Lee, 2006, i.e. Provenance, Rules, and Trust) the capabilities of semantic search appear to be artificially limited.

- What affect does semantic search have on the research process?

Our case study has shown quantitative benefits in query execution time. Similar results were found in a previous study (Szabo et al., 2009) with researchers conducting studies in a fraction of the time previously required. We believe that this will slowly, but steadily, accelerate the pace of heliophysics research. Time, money, and resources saved on one research problem can be devoted to other problems. The cumulative effect is acceleration in the creation of new knowledge.

- What general statements of applicability can be made to other science areas?

The benefits of using provenance to enhance the search process have been evidenced within this study as well as in other domains (Fox et al., in preparation). This tenet extends the traditional data lineage view of provenance and appears to have general applicability. Moreover, Semantic e-Science is more than just semantic search. The integration of search with web services, relational databases, and other cyberinfrastructure, has applicability as a generalized research area within Semantic e-Science. Massive data volumes (Gray et al., 2005) prohibit, or at least deem it very unlikely, that all information will reside as ontology instances or Linked Data. Thus, specific and targeted research into component and framework-based systems is vitally needed.

The future of Semantic e-Science is very promising. However, there are numerous challenges that still need to be overcome. Semantic search is presently one of the main drivers, however, it cannot exist in isolation. We need to move beyond closed systems to broader semantic access. Data access, integration, provenance, and processing services all need semantic representation (Fox and Hendler, 2009a). This case study has shown the benefits, and limitations, of semantically replicating data discovery in one case study in the physical sciences. It has attempted to provide a much needed extension to the current technological discussion. Namely, we have investigated the socio-technical implications of deploying semantic technologies in real-world applications. As we have shown, one can closely replicate manual discovery methods, yet a "complete" answer will require the entire Semantic Web stack and the full emergence of Semantic e-Science.

References

Abbott, M.R., 2009. A new path for science? In the fourth paradigm: data-intensive scientific discovery. Microsoft Research, 111–116.

Baader, F., Calvanese, D., McGuinness, D.L., Nardi, D., Patel-Schneider, P.F., 2003. The Description Logic Handbook: Theory, Implementation, Applications. Cambridge University Press, Cambridge, UK.

Benedict, J.L., McGuinness, D.L., Fox, P., 2007. A Semantic Web-based methodology for building conceptual models of scientific information. American Geophysical Union (Fall Meeting (AGU2007) (Eos Transaction AGU 88(52), Fall Meeting Supplement, Abstract IN53A-0950)).

Berners-Lee, T., Hendler, J., Lassila, O., 2001. "The Semantic Web". Scientific American (May 17).

Berners-Lee, T., 2006. Artificial Intelligence and the Semantic Web, AAAI 2006, 〈http://www.w3.org/2006/Talks/0718-aaai-tbl/Overview.html〉#(14).

Bizer, C., Heath, T., Berners-Lee, T., 2009. Linked data—the story so far. International Journal on Semantic Web Information Systems 5 (3), 1–22. doi:10.4018/jswis.2009081901.

Brin, S., Page, L., 1998. The anatomy of a large-scale hypertextual Web search engine Source. Computer Networks and ISDN Systems 30 (1–7) (April).

Buneman, P., Khanna, S., Tan, W.-C., 2001. Why and Where: A Characterization of Data Provenance, Database Theory, Lecture Notes in Computer Science Book Series, vol. 1973. Springer, Berlin/Heidelberg.

Cao, B., Plale, B., Girish S., Robertson, E., Simmhan, Y., 2009. Provenance information model of karma, version 3. In: Proceedings of the IEEE 2009 Third International Workshop on Scientific Workflows (SWF'09).

Cummings, J., Finholt, T., Foster, I., Kesselman, C., Lawrence, K., 2008. Beyond Being There: A Blueprint for Advancing the Design, Development, and Evaluation of Virtual Observatories, Final Report From Workshops on Building Effective Virtual Organizations, National Science Foundation, May, 2008. Available at: 〈http://www.ci.uchicago.edu/events/VirtOrg2008〉.

Dalton, R., 2007. Geophysicists combine forces. Nature 447 (7148), 1037.

Finin, T., Sachs, J., 2004. Will the Semantic Web change science? Science Next Wave. American Association for the Advancement of Science (September 15, 2004).

Fox, P., Zednik, S., West, P. Semantic provenance in support of explaining, justifying and verifying science data on the web. Journal of Web Semantics, in preparation.

Fox, P., Hendler, J., 2009a. Semantic e Science: encoding meaning in next-generation digitally enhanced science, in the fourth paradigm: data-intensive scientific discovery. Microsoft Research, 147–153.

Fox, P., McGuinness, D.L., Cinquini, L., West, P., Garcia, J., Benedict, J., Middleton, D., 2009b. Ontology-supported scientific data frameworks: the virtual solar-terrestrial observatory experience. Computers and Geosciences 35 (4), 724–738.

Fox, P., McGuinness, D.L., Pinheiro da Silva, P., Zednik, S., Garcia, J., Ding L., Del Rio, N., Chang, C., 2008a. Semantic provenance for image data processing. In: Proceedings of Geoinformaitcs 2008 Data to Knowledge, June.

Fox, P., McGuinness, D.L., Pinheiro da Silva, P., 2008b. Knowledge provenance in virtual observatories: application to image data pipelines. In: Proceedings of the International Semantic Web Conference.

Fox, P., Cinquini, L., McGuinness, D., West, P., Garcia, J., Benedict, J.L., Zednik, S., 2007. Semantic Web Services for interdisciplinary scientific data query and retrieval. In: Proceedings of the AAAI Workshop on Semantic eScience. doi:10.11.144.7441.

Fox, P. McGuinness, D.L., 2011b. An Open-world iterative methodology for the development of semantically-enabled applications. Computers and Geosciences, submitted for publication.

Gray, J., Liu, D., Nieto-Santisteban, M., Szalay, A., DeWitt, D., Heber, G., 2005. Scientific Data Management in the Coming Decade, ACM SIGMOD Record, vol. 34, Issue 4 December 2005, (pp. 34–41).

Goodman, A.A., Wong, C.G., 2009. Bringing the night sky closer: discoveries in the data deluge, in the fourth paradigm: data-intensive scientific discovery. Microsoft Research, 39–44.

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

Hendler, J., 2007. The dark side of the Semantic Web. IEEE Intelligent Systems (January/Febuary).

Hevner, A.R., March, S.T., Park, J., Ram, S., 2004. Design science in information systems research. MIS Quarterly 28 (1), 75–105 (March).

Hey, T., Trefethen, A.E., 2005. Cyberinfrastructure for e-Science. Science 308 (5723), 817–821 (May).

Hotz, R., 2009. A data deluge swamps science historians. The Wall Street Journal (August 28, 2009) available at: 〈http://online.wsj.com/article/SB125139942345664387.html〉.

King, T., Narock, T., Walker, R., Merka, J., Joy, S., 2008. A brave new (virtual) world: distributed searches, relevance scoring and facets. Earth Science Informatics 1 (1). doi:10.1007/s12145-008-0002-7 (April).

Please cite this article as: Narock, T., Fox, P., From science to e-Science to Semantic e-Science: A Heliophysics case study. Computers & Geosciences (2011), doi:10.1016/j.cageo.2011.11.018

Lee, A., 2000. Systems thinking, design science, and paradigms: heeding three lessons from the past to resolve three dilemmas in the present to direct a trajectory for future research in the information systems field, keynote address. In: Proceedings of the Eleventh International Conference on Information Management, Taiwan, May. (available online at: 〈http://www.people.vcu.edu/~aslee/ICIM-keynote-2000〉.

Lim, H.B., Iqbal, M., Yao, Y., Wang, W., 2010. A smart e-Science cyberinfrastructure for cross-disciplinary scientific collaborations. Annals of Information Systems 11, 2010. doi:10.1007/978-1-4419-5908-9 (Chapter 3).

Madin, J., Bowers, S., Schildhauer, M., Drivov, S., Pennington, D., Villa, F., 2007. An ontology for describing and synthesizing ecological observation data. Ecology Information 2 (3), 279–296.

McGuinness, D.L., Fox, P.A., West, P., Rozell, E., Zednik, S., Chang, C., 2010. Progress toward a Semantic eScience framework; building on advanced cyberinfrastructure. EOS Transactions (IN22A-02).

McGuinness, D.L., Fox, P., Brodaric, B., Kendall, E., 2009. The emerging field of semantic scientific knowledge integration. IEEE Intelligent Systems 24 (1), 25–26.

McGuinness, D.L., Fox, P., Cinquini, L., West, P., Benedict, J., Garcia, J., 2007a. Current and Future Uses of OWL for Scientific Data Frameworks: Successes and Limitations, In OWLED.

McGuinness, D.L., Ding. L., Pinheiro da Silva, P., Chang, C., 2007b. PML 2: a modular explanation interlingua. In: Proceedings of the 2007 Workshop on Explanation-aware Computing (ExaCt-2007). Vancouver, Canada, July 22–23.

McGuinness, D.L., P. Fox, L. Cinquini, P. West, J. Garcia, J.L. Benedict, D. Middleton, 2007c. The virtual solar-terrestrial observatory: a deployed Semantic Web application case study for scientific research. In the proceedings of the 19th Conference on Innovative Applications of Artificial Intelligence (IAAI). Vancouver, BC, Canada, July 2007, pp. 1730–1737 and AI magazine, 29, #1, pp. 65–76.

McGuinness, D.L., Pinheiro da Silva, P., 2004a. Explaining answers from the Semantic Web: the inference web approach. Journal of Web Semantics 1, 397–413.

McGuinness, D.L., van Harmelen, F., 2004b. OWL Web Ontology Language Overview, W3C Recommendation, 2004, available at: 〈http://www.w3.org/TR/owl-features〉/.

Merka, J., Narock, T., Szabo, A., 2008a. Navigating through SPASE to heliospheric and magnetospheric data. Earth Science Informatics 1 (1). doi:10.1007/s12145-008-0004-5 (April).

Merka, J., Szabo, A., Narock, T., Walker, R., King, T., Slavin, J., Imber, S., Karimabadi, H., Faden, J., 2008b. Using the virtual heliospheric and magnetospheric observatories for geospace studies. Eos Transactions AGU 89 (53) (Fall Meeting Supplement, Absract SA51B-05).

Merka, J., Szabo, A., 2004. Bow shock's geometry at the magnetospheric flanks. Journal of Geophysical Research 109 (A12224). doi:10.1029/2004JA010567.

Michaelis, J., McGuinness, D.L., 2010. Towards provenance and collective knowledge for web applications. In: Proceedings of the Third International Provenance and Annotation Workshop (IPAW 2010), pp. 265–273.

Moreau, L., Clifford, B., Freire, J., Futrelle, J., Gil, Y., Groth, P., Kwasnikowska, N., Miles, S., Missier, P., Myers, J., Plale, B., Simmhan, Y., Stephan, E., denBuss-che, J.V., 2011. The open provenance model core specification (v1.1). Future Generation Computing Systems 27 (6), 743–756.

Narock, T., King, T., 2008. Developing a SPASE query language. Earth Science Informatics 1 (1). doi:10.1007/s12145-008-0007-2 (April, 2008).

Narock, T., Yoon, V., Merka, J., Szabo, A., The Semantic Web in federated information systems: a space physics case study. Journal of Information Technology Theory and Application, in press.

Neuman, A., 2005. A life science semantic web: are we there yet? Science STKE 2005 (283), pe22. doi:10.1126/stke.2832005pe22.

Peredo, M., Slavin, J.A., Mazur, E., Curtis, S.A., 1995. Three-dimensional position and shape of the bow shock and their variation with Alfvénic, sonic and magnetosonic Mach numbers and interplanetary magnetic field orientation. Journal of Geophysical Research 100, 7907–7916.

Pinheiro da Silva, P., McGuinness, D.L., Fikes, R.E., 2004. A proof markup language for Semantic Web services. In: Bell, David, Bussler, Christoph, Yang, Jian (Eds.), The Semantic Web and Web services: Information Systems. 31(4–5), 381–395.

Prud'hommeaux , E., Seaborne, A., 2008. 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.

Rozell, E., Fox, P., Maffei, A., Belliau, S., 2010. S2S an application integration framework for distributed oceanographic repositories. Eos Transaction AGU (Fall Meeting Supplement, Abstract IN23A-1349).

Slavin, J.A., Fairfield, D.H., Lepping, R.P., Hesse, M., Ieda, A., Tanskanen, E., Østgaard, N., Mukai, T., Nagai, T., Singer, H.J., Sutcliffe, P.R., 2002. Simultaneous observations of earthward flow bursts and plasmoid ejection during magnetospheric substorms. Journal of Geophysical Research 107 (A7). doi:10.1029/2000JA003501.

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

Szabo, A., Merka, J., Narock, T.W., 2009. Multispacecraft observations of solar wind gradients near L1: VHO case studies. American Geophysical Union (Fall Meeting 2009, abstract #SH51B-1271).

Szalay, A., Gray, J., 2001. The World-Wide Telescope. Science 293, 2037–2038 (14 September 2001).

Zednik, S., Fox, P., McGuinness, Deborah L., Pinheiro da Silva, P. Chang, C., 2009. Semantic provenance for science data products: application to image data processing. In: Juliana Freire, Paolo Missier, Satya Sanket Sahoo (Eds.), Proceedings of the First International Workshop on the Role of Semantic Web in Provenance Management (SWPM 2009), Washington DC, USA, October 25, 2009. CEUR Workshop Proceedings 526, CEUR-WS.org.

Zednik, S., Fox, P., McGuinness, D.L. System transparency, or how i learned to worry about meaning and love provenance. In: Proceedings of the 3rd International Provenance and Annotation Workshop (IPAW2010), Troy, NY, USA (June 15–16, 2010), LNCS, in press.

Please cite this article as: Narock, T., Fox, P., From science to e-Science to Semantic e-Science: A Heliophysics case study. Computers & Geosciences (2011), doi:10.1016/j.cageo.2011.11.018