Project Summary

The Tetherless World Constellation at Rensselaer Polytechnic Institute proposes STCI - A Semantic eScience Framework (SESF) to formalize a capability built on a series of prior activities, in virtual observatories, semantic data frameworks, data integration and knowledge provenance funded over the last 4-5 years by NSF and NASA.

The initial activity, the Virtual Solar-Terrestrial Observatory (VSTO) which is now a production semantic data framework, delivers observational and related data in solar physics and aeronomy/upper atmospheric physics to a community of over 600. The VSTO utilizes leading edge knowledge representation, query and reasoning techniques to support knowledge-enhanced search, data access, integration, and manipulation. It encodes term meanings and their inter-relationships in ontologies and uses these ontologies and associated inference engines to semantically enable the data services. The Semantically-Enabled Science Data Integration (SESDI) project implemented data integration capabilities among three sub-disciplines: solar radiation, volcanic outgassing and atmospheric structure using extensions to existing modular ontologies and used the VSTO data framework, while adding smart faceted search and semantic data registration tools. The Semantic Provenance Capture in Data Ingest Systems (SPCDIS) has added explanation provenance capabilities to an observational data ingest pipeline for images of the Sun providing a set of tools to answer diverse end user questions such as “Why does this image look bad?”.

Each of these efforts utilize advanced knowledge representation and reasoning to enable higher levels of semantic capability and interoperability such as explanation, reasoning about rules, and semantic query. In addition, we note that the number of scientific data applications using semantic technologies is growing very rapidly and we are being approached by numerous communities to help them develop new capabilities. We consider both our efforts and those in the community to be sufficiently mature that it is time to work to bring them together into a first generation eScience data framework built on demonstrated semantic methods and technologies as well as a path forward for application to other communities. Thus, the broad goals of this STCI proposal are twofold.

1. Intellectual Merit: To capitalize upon VSTO, SESDI and SPCDIS to design and implement configurable and extensible semantic eScience framework. Configuration will require some research into accommodating different levels of semantic expressivity and user requirements from use cases. Extensibility is best achieved in a modular approach to the semantic encodings (i.e. ontologies) performed in a community setting, i.e. an ontology framework into which specific applications all the way up to communities can extend the semantics for their needs. Over the past few years, semantic technologies have evolved and new tools are appearing. Part of the effort in this project will be to accommodate these advances in the new framework and lay out a sustainable software path for the (certain) technical advances. In addition to a generalization of the current data science interface, we will include an upper-level interface suitable for use by clearingshouses, and/or educational portals, digital libraries, and other disciplines.

2. Broader Impacts: In keeping with our developed semantic web methodology which has proven to applicable across disciplines and end-user levels of expertise, we will fully engage members of the academic and broader communities via a series of workshop which includes existing U.S. national and international focused science programs, semantic technology communities of practice both national and international academic as well as world-wide agency research and implementation efforts. The workshops will span involvement from end-science and non-specialist use, to data system developers, knowledge modelers and ontology developers through to software engineers and tool/application developers. The proposed eScience framework is based on semantics, and software built on and around the semantics. A sustainability path for communities is essential so that use may continue into the future. Sustainability extends beyond software to ontology development and vetting, and communities of practice (scientist, data providers, technical teams). Via extensive community outreach we will be to facilitate the culture change that is required in scientific endeavors to sustain the implementation, evolution and viability of an eScience capability as essential as observing and experimental equipment, supercomputing and high-speed networking infrastructures are now. That culture change requires a demonstration of the value of the outcomes that eScience (data-driven) delivers which is part of our evaluation criteria.

In summary, we propose to take a timely and bold next step to develop and evolve a robust semantic eScience framework in support of broad science applications, leveraging the precursor work and using real specialist and non-specialist data infrastructure needs to drive implementation and research for new capabilities.
1 Results from recent/prior NSF support

a) NSF Award OCI-0431153, Sep. 1, 2004 - Aug. 31, 2009, $1,524,570, b) A Prototype for the Virtual Solar-Terrestrial Observatory (VSTO, PI: Peter Fox, co-I Deborah McGuinness), c) VSTO implemented a distributed, scalable research data framework providing virtual access to specific data, model, tool and material archives containing items from a variety of space- and ground-based instruments and experiments, as well as individual and community modeling and software efforts bridging research and educational use and currently deployed in the fields of solar, solar-terrestrial and space physics. VSTO is built on semantic web technologies an implemented as a web portal, native programming interface and via web service interfaces. VSTO also developed and refined a robust semantic web methodology which is being used in many subsequent projects. d) The VSTO project has produced numerous publications (11 referred, 32 proceedings) and invited presentations at national and international conferences and is also having a broad influence on other virtual observatory projects (the NASA VITMO, for example), educational initiatives, the Electronic Geophysical Year (eGY), the NASA-funded Semantically-Enabled Science Data Integration project, and several NSF-funded environmental observatory projects. Three relevant publications: Fox, P., McGuinness, D.L., Middleton, D., Cinquini, L., Darnell, J.A., Garcia, J., West, P., Benedict, J., Solomon, S. 2006, Semantically-Enabled Large-Scale Science Data Repositories. the 5th International Semantic Web Conference (ISWC06), LNCS, ed. Cruz et al., vol. 4273, pp. 792-805, Springer-Verlag, Berlin. Deborah McGuinness, Peter Fox, Luca Cinquini, Patrick West, Jose Garcia, James L. Benedict and Don Middleton, 2007, The Virtual Solar-Terrestrial Observatory: A Deployed Semantic Web Application Case Study for Scientific Research, in Industrial Applications of Artificial Intelligence 2007, 1730-1737 and AI magazine, 29, #1, 65-76. Peter Fox, Deborah McGuinness, Luca Cinquini, Patrick West, Jose Garcia, and James Benedict, 2009, Ontology-Enabled Semantic Data Frameworks - the Virtual Solar-Terrestrial Observatory Experience, Computers and Geosciences, Geoscience Knowledge Representation for Cyberinfrastructure 35, #4, 724-738. e) Refer to http://tw.rpi.edu/portal/VSTO (project site, including details on publications, presentations, etc.) and http://www.vsto.org (portal). f) n/a

2 Project Description

2.1 The Semantic eScience Framework (SESF)

2.1.1 Why is it needed?

Virtual observatories (VOs), both as a concept, and as implemented technical science infrastructure are changing the ways that researchers and students do science today. In many fields where the concept of an observatory is familiar, they are or are becoming the eScience (Hey and Trefethen, 2005) data framework of
choice. To date, there have been a variety of approaches to developing virtual observatories starting with the original efforts in Astronomy (Szalay 2001) to the substantial growth in geosciences and space sciences today. These approaches are often driven by quite diverse mission and functional requirements as well as underlying data sources. While there are a few definitions of a virtual observatories, one goal seems clear: to find the right balance of data/model holdings coupled with web portals and application client software that enable a researcher to use easily, without effort or interference, as if all the materials were available on his/her local computer. Further this capability must be provided not only to the highly trained researcher - but also to the graduate student, undergraduate, researchers in a related or distant discipline, etc.

Virtual observatories and other distributed data systems offer opportunities to bring large amounts of authentic data, both real-time and historical, to non-traditional audiences, such as policy makers, educators, and the general public. For non-scientists, finding their way through the data in a VO can be challenging and trying to put the data into an educational context can be show-stopping. And yet, the opportunity for virtual observatories to open its doors to these audiences is critical - for both the VOs and for the audiences. For broad audiences, virtual observatories can provide important components to policy debates and decision-making. The data in VOs can provide an important opportunity for students and teachers. Authentic data excites students and teachers use such data to engage students in the excitement of scientific exploration. Students tell us that it means more to them if the data they are using is authentic. And, VOs offer an important pathway into self-exploration for the educated general public. Lastly, the development of access to VOs by a larger audience also allows scientists from other disciplines to access data, increasing the likelihood of emergent, cross-disciplinary discovery. VOs that can successfully bridge the divide between their traditional audiences and new ones stand to benefit from broad exposure to groups that can impact the VOs outlook for long-term sustainability.

2.1.2 What is it?

In developing the Virtual Solar-Terrestrial Observatory (VSTO; originally proposed as a prototype but it became a production system after only ∼ 18 months; see Fox et al. 2006), which targets interdisciplinary data search, access and use and the need for a robust, extensible and smart data framework, we based our approach on the following principles:

- Datasets alone are not sufficient to build a virtual observatory: VSTO needed to integrate tools, models, and data.
- VSTO needed to address the interface problem, effectively and in a scalable fashion.
- VSTO had to address the interdisciplinary metadata and ontology problem - bridging terminology and use of data across disciplines.
- VSTO had to leverage the many forms of data schema that describe the syntax (name of a variable, its type, dimensions, etc. or the procedure name and argument list, etc.), semantics (what the variable physically is, its units, etc.) and pragmatics (e.g. what the procedure does and returns, etc.) of the datasets and tools.
- VSTO was also intended to provide a basis for a framework for building and distributing advanced data assimilation tools.

In developing the following up applications, for data integration and knowledge provenance, among others it became obvious that a clear pattern of development and use was emerging and that we are becoming more adept at re-using substantial parts of what we’d previously developed. This included, methods, ontologies, libraries, tools and even application components. The pattern that emerged was that of a viable semantic eScience framework which we propose here and defined as: a semantically configurable and extensible set of software tools and services that enable smart search and mediated use between a user and the data and information service(s).

As expected we use as a basis and measure of success the definition of what a virtual observatory should do, and its goal(s). Perhaps the most aggressive goal is: To make ‘standard’ scientific research much more efficient (the heart of eScience). An important metric is that the principal investigator (PI) teams should want to (and do) use them. VOs must improve on existing services (all the way from scientific observatories, national data centers, to individual PI sites, etc.):

- VOs will not replace these, but will use them in new ways.
- VOs must enable new, global problems to be solved.
• VOs must enable users to rapidly gain integrated views from a point of origin to the an ultimate effect and find data related to any particular observation.

Ultimately, scientists now want answers to ‘higher-order’ queries such as ‘Show me the data from cases where a large coronal mass ejection observed by the Solar-Orbiting Heliospheric Observatory was also observed in situ.’ (science-speak) or ‘What happens when the Sun disrupts the Earth’s environment’ (general public). Our goal in a semantic eScience framework is to enable this general kind of usage as well, and this would facilitate, for example, a secondary school teacher to develop a lesson plan and also helps the general public find and understand more science data.

2.1.3 How will we do it?

In the VSTO we were led to the need for semantics and pragmatics because we needed our data framework to be smart in a way that we could maintain and evolve it, and more closely reflect the integration of science concepts with the underlying data and other information. We determined that when we integrated, we integrated concepts, terms, relationships and in the past we would ask specialists, guess, research a lot, or give up and end up coding a fraction of the required knowledge. We found that machine processable semantics in the form of ontologies, when developed with an effective methodology, would help (for example a human-in-the-loop) find, access, integrate, use, explain, and trust data.

Figure 1: Our developed Semantic Web Methodology.

VSTO adopted the semantic web as an enabling methodology and technical approach (Benedict et al. 2007) and developed and refine a methodology for its development see Fig. 1. The semantic web is defined as: an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation (http://www.semanticweb.org). Knowledge is encoded in an ontology: An explicit formal specification of how to represent the objects, concepts and other entities that are assumed to exist in some area of interest and the relationships that hold among them. We successfully followed the same method for the Semantically-Enabled Science Data Integration (SESDI; Fox et al. 2007, McGuinness et al. 2008, see Reading List) and the Semantic Provenance Capture in Data Ingest Systems (SPCDIS; Fox et al. 2008, etc.) projects and now teach this method in a graduate course at RPI.

VSTO is operational (e.g. McGuinness et al. 2007, Fox et al. 2009, and see Reading List for further references) and has implemented a modest level of knowledge representation and reasoning. We believe the time has come to take bold and innovative steps to advance the capabilities of the VSTO in response to the needs of science and education data use in and beyond our present communities.
If taken, this will not only have the benefit of increasing the science power and reach of one deployed multi-disciplinary virtual observatory, but possibly even more importantly, it will continue to provide an example of how other VOs can utilize the same infrastructure to help them take the next step with those efforts as well. Finally we note that in cyberinfrastructure project the term VO also refers to virtual organizations (Beyond Being There, 2008). While we will not elaborate on this further we are increasingly finding from very practical experience that virtual observatories and virtual organizations are very similar (VO–VO?).

The specific tasks proposed are approximately divided into the following areas:

- Current software framework evaluation
- Cross-domain and multi-expertise use cases
- Software framework re-design and re-factorization
- Advanced semantics for SESF
- Services, tools, portlets, and programming interfaces
- Migration of VSTO, SESDI and SPCDIS to the new framework, development of non-specialist 'VO' toolkit
- Outreach workshops for developers to also migrate
- Education for the current and next generation of users and developers of such a framework
- Regular formal evaluation studies of the semantic web methodologies and technical approach utilized

We will demonstrate these developments in the current science application areas of VSTO, SESDI and SPCDIS as well as selected others. We will publish extensive papers and reports on all of our results in national and international settings and integrate these into our semantic eScience course website (see tw.rpi.edu).

2.1.4 What is the benefit?

**Intellectual Merit: Getting Authentic and Useful Data to a Spectrum of End Users**

What if you:

- could not only use your data and tools but remote colleague’s data and tools?
- understood their assumptions, constraints, etc. and could evaluate applicability?
- knew whose research currently (or in the future) would benefit from your results?
- knew whose results were consistent (or inconsistent) with yours?

![Figure 2: The spectrum of strategic uses of data.](image-url)

The proposed SESF will facilitate new and better community science since the semantics that enable it will have built into it, a mid-level ontology, i.e. an integration framework allowing data and information sources, including from models, to be brought into consideration for an investigation without the user having to know what to ask for or how to find related data. As more and diverse datasets are added, together with
the ability to find them with advance reasoning, many opportunities are available at a fraction of the cost
and time of development within existing service offerings. Increasingly, advancing science and non-specialist
use involves data integration, data assimilation, enabling science workflows and service chaining, explaining
sources and origins of data, and much more.

Beyond research communities, it is clear that data has lots of uses. A NASA internal report on science
education entitled: “Why EPO” (Education and Public Outreach; NASA 2005), clearly highlights the
spectrum of strategic uses of data. At the apex of Figure 2 are the more strategic uses, descending to the
base where less strategic uses are indicated. All of these uses are important; including those of researchers
and undergraduate/graduate students for the purpose of attaining and producing new knowledge.

What is rare in today’s modern technology infrastructure is a scalable and sustainable information and
information product framework (also infrastructure) that can span this spectrum of use. Information prod-
ucts (for example a weather map) that can be adapted all the way from educational use to decision support
with high quantitative integrity and appropriate fitness for purpose and provenance are now a responsibility
of both mission agencies in the U.S. (NOAA, EPA, USGS, NASA, ...) as well as research enterprises, such
as modeling, research observation networks (for example the new NSF Ocean Observing Initiative (OOI,
especially their cyberinfrastructure effort - http://www.oceanobservatories.org/) among numerous others.

Across many Earth system science disciplines the importance of multi-channel and interrelated data
streams from ground-based and remotely sensed observatories is growing at a rate that is unprecedented
and without an end in sight (in a similar way to the night time astronomy community; Szalay 2007). Time series
data, CCD imagery, in multiple wavelengths with rapid time sampling, models (analysis and simulation) are
being produced routinely.

In terrestrial middle and upper atmospheric research, the circumstances are almost the same with a rich
interplay between instruments, models, and interpretation. Space weather, which intersects these fields, re-
quires that a researcher be able to ingest solar data into a magnetospheric model, for example. Across these
fields, space physics including space weather and related interdisciplinary studies emphasize the real time or
near-real time requirement to get to the data and analyze and combine it in a way that enhances our under-
standing of the Sun, the upper atmosphere of the Earth (including but not limited to solar influences) and
many aspects of the Sun-Earth environment. NASA, via their Virtual Observatories for Heliophysical Data
(VOHD) and the European Union’s recently funded FP7 program: HELIO - the Heliophysical Observatory
(www.helio-vo.eu), many of which highlight the need for rich semantics are further evidence of the readiness
for a coherent framework. These successes are spreading into all areas of geoscience, biology, ecology and
environmental engineering.

Virtual observatories have gone a long way toward getting these diverse data to researchers, yet the
digital divide remains. Educators are grappling with getting authentic forms of data into different levels of
the educational spectrum. There are notable successes, such as SkyView (http://skyview.gsfc.nasa.gov/) for
astronomy and the IRIS consortium (http://www.iris.edu) for seismology.

The audiences for data contain several challenges for reaching the non-specialist: which include the non-
specialist scientists, educators, and general public/policy makers. The importance of connecting science and
education in the classroom cannot be over-emphasized or under-estimated as a challenge to our modern
informatics infrastructure as is the importance of making any infrastructure scalable and sustainable, i.e. an
eScience capability.

2.1.5 Pushing the Broader Impact

The challenges faces by VOs in their quest to serve a broader audiences base comes in several flavors. First,
in order to develop an effective non-specialist use case, a set of specific learning goals or outcomes need to
be defined. These outcomes are defined based on a clear a priori understanding of what the user is likely to
desire. In the case of educators, this prospect is made less daunting by the science education standards in
place in most classrooms. In the US, the National Science Education Standards and 2061 Benchmarks guide
the state standards that are in place in most classrooms, giving a clear set of educational goals at each grade
level that can be used as the starting point for determining what kinds of information teachers and students
are most likely to come to the VO for. In the case of the VSTO, the choice of an educational use case that
began with a question about the aurora (see Fig.3) was an obvious one, as it met several national standards
for middle school regarding the Sun-Earth connection and would be an obvious place for many teachers to
start. What works for teachers in terms of finding accessible science information relevant to particular topics
of popular or political interest (such as global warming or northern lights) is likely to work for policy makers, managers, etc., as well as scientists from different fields, moving us towards a model of truly interdisciplinary knowledge discovery.

Once again, the reasons why we were led to semantic methodologies and technologies for VSTO also apply to the digital divide and even to a greater extent. Educators at all levels aim to teach concepts often by example - concrete example. Meaning is important. Real data (and models) are important. Ontologies and semantic data frameworks are now recognized as a very viable answer.

We’ve made progress by developing semantically enabled computer (software) infrastructure which goes far beyond the traditional syntactic implementations (which limit usability unless there is a human-in-the-loop). Increasingly we see situations where there are “humans on the loop” i.e. people who do not have enough detailed knowledge to be tightly integrated into the loop (with all of the decisions involve) but want to have visibility into the process. Instead, effective and efficient processing and use of data streams will now possible in SESF. VSTO and SESDI have developed ontologies from the use cases and embedded data models and schemas that describe the syntax, semantics and pragmatics of the datasets and tools that use them as well as used ontologies from existing sources, e.g. SWEET (Raskin and Pan 2005).

However, it is not enough. Figure 3 shows a schematic of a use case we are currently exploring for a US 9th grade teacher preparing curriculum materials and a lesson plan on the area of Sun-Earth studies. The schematic demonstrates the divide between the research virtual observatory and the educational user/non-domain literate/expert who often fails to cross the divide.

| Teacher selects the link: Search for "Aurora" in Terrestrial Phenomena |
|-----------------------------|--------------------------|
| Learns that Aurora are part of a set of terrestrial manifestations of 'space weather' and gets a set of websites (educational and research), glossary terms, and hints on related phenomena and a connection to solar phenomena and time periods when aurora are most visible. |

Teacher receives four groupings of search results:

<table>
<thead>
<tr>
<th>Educational materials:</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://www.meted.ucar.edu/topics_spaceen.php">link</a> and <a href="http://www.meted.ucar.edu/faq/aurora/">link</a></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Research, data and tools: via three virtual observatories, knows to search for brightness, or green/red line emission</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Did you know?:</strong> Aurora is a phenomena of the upper terrestrial atmosphere (ionosphere) also known as Northern Lights</td>
</tr>
<tr>
<td><strong>Did you mean?:</strong> Aurora Borealis or Aurora Australis, etc.</td>
</tr>
</tbody>
</table>

Teacher accesses internet goes to An Educational Virtual Observatory and enters a search for "Aurora".

<table>
<thead>
<tr>
<th>VO interface needed</th>
</tr>
</thead>
</table>

Teacher selects: models of the aurora during high solar activity

VO returns a link to an animation of a solar coronal mass ejection impacting the Earth's Magnetosphere, ... demonstrating the electron precipitation that results in the northern auroral oval in the ionosphere.

An link is provided to download the movie and search for data of brightness on a latitude/longitude grid.

| Teacher selects: the parameter Auroral Brightness over a date range corresponding to a period of high solar activity. |

Typical VO interface

Teacher gets a link to some data and downloads it.

| Teacher loads the data into a spreadsheet program and plots a color contour of the aurora. The teacher then starts developing the lesson plan for the students. |

Figure 3: Schematic of educational use case workflow which highlights the education-science divide created by existing VO terminology.

To bridge a divide such as this (which we see as representative of the broader class of non-specialist use) we posit that three elements are required:

1. Developing the mid-level semantics to achieve the closing of the knowledge gap between education and science, this requires educational use cases
2. Extension of the existing knowledge encoding, particularly aimed at supporting additional reasoning.
   This task will require the identification of additional terms, concepts and relations, representation of them, and perhaps the building of tools to handle the extended reasoning and information manipulation.
3. A formal evaluation component to assess and document which activities enhance end-user access to and use of data to advance science and education and ongoing community input (see 2.3.7 for additional details).
We elaborate on the merits of some of these elements in section 2.3 of the proposal.

2.2 Outcomes

2.2.1 Intellectual Merit: Strategic Approach to Bridging Science and Education

In the present mode of development for the VSTO, SESDI and SPCDIS, we utilize use case methodology to transcend the design and functional spectrum from end user to data use. We form small teams of about 4-8 by engaging a small number of users; domain experts, Knowledge representation experts, computer science and cyberinfrastructure developers and engineers, and a facilitator. Why should the data being made available for scientists be separated from the data being made available to non-specialist in general? Our strategy is to span the science–education divide with the semantic mediation (coupled with reasoning) of ontologies.

The process we utilize for development of the semantic data framework and its implementation will also be used to address the educational data access requirements. The associations encoded and also inferred in the ontology do bridge the two areas in the same way that we accommodate the interdisciplinary needs within VSTO and SESDI; viz. upper atmosphere and solar physics. Within the simple educational use case we have prepared as an example, we have developed an initial concept map of how the terms and associations can be made (not shown here).

For the non-specialist use cases we follow the same procedure as for the science use cases; i.e. we develop both a high-level and detailed description of the use case, e.g. one of the first we implemented for VSTO was: Plot the observed/measured Neutral Temperature (Parameter) looking in the vertical direction for the Millstone Hill Fabry-Perot Interferometer (Instrument) from January 2000 (in a way that makes sense for the data). A later use case was: Find data which represents the state of the neutral atmosphere anywhere above 100km and toward the arctic circle (above 45N) at any time of high geomagnetic activity. The full development requires us to translate this into a complete query for data and in the cases were not all the needed information was recorded, we use inference to fill in the gaps. We analyze the vocabulary and concepts, identify class and subclasses, properties and associations usually in a graphical manner using tools like CMAP (http://cmap.ihmc.us) and then convert the visual ontology into a formal ontology in OWL using Protege.

In such use cases, we do not overspecify the provided information since we take advantage of gaps in knowledge that need to be inferred (and integrated) from the use-case, and thus the underlying associations. As we move to more general applications, such as education and non-expert use, the requirement for inference (deduction) increases and thus we are led to an advanced level of semantic representation.

2.2.2 Broader Impacts: on eScience

Virtual Observatories (VOs) are now becoming common. For example the Virtual Astronomical Observatory (VAO; is about to go into an operational phase) has been very successful in delivering data to astronomers world-wide (Szalay 2001).

The VSTO was designed as a research and educational capability utilized in conjunction with existing space weather activities and capabilities. Too often new and valuable research data or models are developed which take considerable time to integrate into operational, especially forecast, systems. In the same way that a specific and semantically meaningful set of SESF interfaces enables researchers, students and educators to access diverse science data resource it will also rapidly and cost-effectively make these available to other community and government agencies to possibly greater enhance their capabilities.

To fully realize the significant potential of the broader impacts of this project, we will undertake specific outreach efforts into the digital libraries, library science, and information retrieval communities. We plan to engage science communities

Between all the investigators for this proposed project we have scientific society affiliations (e.g. AGU, GSA, EGU) and significant program associations that will provide the platforms for dissemination, e.g. under the path being forged by the Electronic Geophysical year (eGY), the Earth Science Information Partnership (ESIP) federation, the group on earth observations (GEO) and many NASA programs in advance technologies for collaboration (ACCESS), discipline virtual observatories (VxO), sensor web, the Global Change Master Directory (GCMD) and the Distributed Active Archive Centers (DAAC) as well as NSF-sponsored programs such as the Geosciences Network (GEON) and Linked Environments for Atmospheric Disturbance (LEAD) and emerging environmental observatory and eco-informatics projects. Additionally we will plan outreach
into the semantic web, web science, and AI communities leveraging our significant program associations with Association for the Advancement of Artificial Intelligence (AAAI), Web Science Research Initiative (WSRI), SemTech as well as the more research oriented Semantic Technologies research venues of ISWC and ESWC.

2.3 Current and Required Work

2.3.1 Data Integration and more

The PIs have engaged in scientific data integration in numerous projects, but none more directly than in the SESDI project - Semantically-Enabled Scientific Data Integration (Fox, et al, 2006, McGuinness et al, 2008). The broad goal of that effort is to enable the next generation of interdisciplinary scientific research by supporting semantically-informed data integration. The specific science project aims to integrate heterogeneous volcanic and atmospheric chemical compound data in support of assessing the atmospheric effects of a volcanic eruption. In SESDI, we have leveraged existing infrastructure from our interdisciplinary virtual observatory work along with ontologies from VSTO and from GEON to inform the integration. This project takes that project forward along a few directions. First, we need to include additional support for process modeling is included - initially focused on forcings and their interrelationships. Second, we need to design and develop registration facilities that (1) are aimed at a broader class of users; (2) have support for registering process-related information.; and (3) provide guidance for generation and use of modular ontologies for integration and reuse.

2.3.2 Provenance and explanation

The goal of our semantic provenance capture for data ingest systems is to design and implement an extensible provenance solution that is deployed at the time scientific data is collected and provided to data systems. We have designed and implemented a system that works in the domain of solar coronal physics. Our claim is both that the design and implementation are useful for the particular scientific image data services we designed for, but further that the design provides an operational specification for other scientific data applications. This project takes the next logical step beyond our work on multi-disciplinary virtual observatories by providing a much more complete system that integrates provenance with the science data. The resulting system provides users with the ability to obtain data such as solar images along with surrounding meta information supporting questions such as what were the weather and observing conditions for this data image and why does it look bad? The system leverages and expands on our explanation infrastructure - Inference Web - which aims to provide actionable systems by providing support for explanations along with system answers. SPCDIS (ISWC Ref and Geoinformatics Ref) uses the PML (REF) proof interlingua to encode and share metadata. Our current proposal will migrate our initial research and implementation into the eScience Framework and generalize the infrastructure to provide provenance web services that are more reusable by science applications.

2.3.3 Ontologies

Many best practice semantic technology-based efforts include domain-specific ontologies that capture term definitions and relationships in form that is easy for computer programs to manipulate. There are also a number of existing community generated and maintained quality ontologies that are ripe for reuse. The PI team has a history of (a) reusing best practice ontologies where possible (b) building science ontologies that simultaneously designed to solve use cases AND to be reused. Our experiences have strengthened our belief that ontologies should be designed in a modular form from the start and should also be designed with reuse in mind from the beginning. In this effort, we will design metrics for evaluating ontologies for their modularity and potential for reuse and generate validation tools for evaluating ontologies with respect to these metrics. We will also document features to look for when looking for best practice science ontologies. We will draw on our long history of designing and implementing ontology tools and evolution environments (e.g., cite knowl enhanced search, chimaera, vertical net’s ontology environment, inference web, - McGuinness, 1998, McGuinness et al, 2000, Das, Wu, and McGuinness, 2001, McGuinness and Pinheiro da Silva, 2004) as well as our multi-year effort to reuse a best practice domain ontology (SWEET) and the resulting modularization our effort generated.
2.3.4 Advanced Semantics

The goal of our work on knowledge representation support for scientific data to be used by scientists and educators is to enable systems that are transparent and trustworthy. Our work on multi-disciplinary virtual observatories aims to use machine understandable encodings of term meanings to power data integration. However, if these systems are to be broadly used, they need to be able to explain what they are doing, have done, how they obtained their input data, and the conditions under which it was collected. The application of ontologies and knowledge bases have largely unexplored possibilities in the realm of complex, and interdisciplinary data models and interfaces, tool and algorithm classification and use. In this project, we aim to identify patterns of desired reasoning support and design ontologies that capture the appropriate formal specifications of term meanings that can be used by reasoners to infer conclusions and then later be used by explanation modules to explain the system conclusions. We also aim to encode the semantics in languages, such as the OWL 2 profiles (subsets of OWL chosen to facilitate particular kinds of usage) for which efficient reasoners exist and also for which learning algorithms may be used to learn descriptions.

The knowledge that exists in people’s minds is one of the more difficult elements to capture and represent within cyberinfrastructure. This problem is compounded in interdisciplinary settings in a similar way that it is harder to find scientists who can communicate across discipline boundaries. Knowledge can be built by having excellent tools that enable rapid learning and discovery. Ontologies are organized collections of human knowledge. Large-scale realizations are becoming an essential component of many applications including standard search (e.g. Google), e-commerce (e.g. Amazon), configuration (e.g. Dell), and intelligent government agent programs such as DARPA’s Personal Assistant that Learns (PAL) program and National Cyber Range (NCR) programs along with multiple intelligence analyst toolkit programs, including those initiating from programs like DTO’s Novel Intelligence for Massive Data (NIMD) program.

More recently, ontologies are also being used in learning systems, e.g., in DARPA’s Integrated Learning Infrastructure program with the integration teams focusing efforts around learning while exploiting ontologies. While common in eCommerce, various expert and recommender systems, and even in Geographic Information Systems (GIS), there has been relatively little use or application to the challenging area of solar, solar-terrestrial and space physics. The VSTO set of ontologies is designed to meet the representational and reasoning needs of a virtual observatory that provides access to observational data (and model data and educational materials) to a broad, interdisciplinary audience.

A key technology development will be an operational system that both serves to accomplish enhanced VSTO services, but also serves as an operational prototype of how ontologies with rule encodings can be used to provide scientific data integration and access. Currently a small set of science applications are using simple ontologies to support relatively simple access, but we envision this system to go noticeably beyond taxonomies with a reasonable set of properties and a small set of reasoning paradigms.

We will use the W3 Rule Interchange Format (RIF) currently in last call before becoming a recommendation and the OWL profile aimed at ontologies for use in rule applications (OWL-RL) also currently in last call status. This will enable us to use best practice techniques for representation along with the emerging reasoning tools being developed to support those representation paradigms. The resulting system will provide a template for ontology and rule integration in deployed scientific applications.

2.3.5 Knowledge representation and reasoning

Our initial design will use ontologies to capture the basic science terms and their inter-relationships. We will use the evolving best in class service languages, OWL-S, SWSL, and WSMO if required, to encode information about the interoperation services. These services can “advertise” what they are capable of doing, what they require as input, and what they will deliver. The explanation capabilities will also be able to explain how these services succeed as well as why they may fail to achieve a goal (McGuinness et al. 2005).

As part of the SESDI project we extensively used and evolved the SWEET ontology (http://sweet.jpl.nasa.gov/2.0) and made significant strides toward understanding the balances between modularity, expressiveness and reasoning needs. We are acutely aware of the need and challenges in such tasks, in particular merging, modularized and importation/assembly of lower and mid-level ontologies. As part of our effort to evolve the VSTO ontology to create the SESF ontology we will need to evaluate the need existing approaches to mid and upper-level ontologies (Brodaric and Probst 2009).
Broader Impacts: Technology Capability

A key technology development will be enhancements to the set of linked, interdisciplinary-enabled, ontologies built for VSTO. The starting point for these ontologies is a family of ontologies that have already seen reuse well beyond their initial science targets. This process requires tools that support broad ranges of users in (1) merging of ontological terms from varied sources, (2) diagnosis of coverage, correctness, and style of ontologies, and (3) maintaining ontologies over time. The knowledge engineering and ontology tool expertise of Deborah McGuinness, will be utilized to help domain experts create ontologies for longevity and reuse and evolve and integrate existing ontologies.

An over-arching goal is to design a methodology for building and maintaining user-centric semantic web artifacts (such as ontologies and web services built for reuse. We will document the methodology in writing and in tools that can be used to evaluate the objective assessments of the present implemented technologies that work. This effort would add to the present anecdotal and some qualitative reports on modern data frameworks, user interfaces, science progress achieved, etc. that is currently captured within papers, presentations and in software. We thus expect that the results of this project will have a wide degree of applicability to similar community and government programs.

Evaluation of Software, Methods and Ontologies

Based on our accumulated experience and several informal evaluations (McGuinness et al 2007) we are certain that evaluation studies hold the key to a robust approach to evolving capabilities of a combined knowledge and software system that forms SESF. To provide a more formal basis, we draw on the work of Twidale, Randall and Bentley (1994) who define evaluation studies consisting of several components:

1. An assessment of the overall effectiveness of a piece of software, ideally yielding a numeric measure by which informed cost-benefit analysis of purchasing decisions can be made.
2. An assessment of the degree to which the software fulfills its specification in terms of functionality, speed, size or whatever measures were pre-specified.
3. An assessment of whether the software fulfills the purpose for which it was intended.
4. An assessment of whether the ideas embodied in the software have been proved to be superior to an alternative, where that alternative is frequently the traditional solution to the problem addressed.
5. An assessment of whether the money allocated to a research project has been productively used, yielding useful generalisable results.
6. An assessment of whether the software proves acceptable to the intended end-users.
7. An assessment of whether end-users continue to use it in their normal work.
8. An assessment of where the software fails to perform as desired or as is now seen to be desirable.
9. An assessment of the relative importance of the inadequacies of the software.

The multitude of techniques that can be used (according to a number of orthogonal dimensions) adds to the complexity of the evaluation task. These dimensions are:

- summative <-> formative
- quantitative <-> qualitative
- controlled experiments <-> ethnographic observations
- formal and rigorous <-> informal and opportunistic

In this work, we proposed to construct a set of evaluation that are summative, mostly quantitative, and formal but not in controlled situations (virtual observatories and data frameworks in the ‘wild’ or in real and often unplanned use).

Outcomes will be measured using a combination of data gathering processes, including surveys, interviews, focus groups, document analysis and observations that will yield both qualitative and quantitative results. Evaluation questions are identified in the evaluation section of the proposal and will be used to determine the degree to which the SESF has enhanced search, access, and use of data for scientific and educational needs and effectively utilized and implemented a template for user-centric utilization of the semantic web methodology.

We plan to issue a small sub-contract call for an external evaluator with the criteria of having evaluation experience and a strong background in Science, Technology, Engineering, and Mathematics programs. The evaluation design for this project involves both qualitative and quantitative methods, with dual goals.
of yielding evidence of the impact of the SESF and educational case use program and formative evaluation information to highlight successes and areas for improvement and to document variables associated with the greatest impact. The chosen external evaluator will be request to use a combination of data gathering processes, including interviews and focus groups with faculty and staff at participating institutions. Surveys of SESF users will be used to gather baseline and follow-up data about the nature and extent of their access to and use of data, perceptions of impact, and progress toward outcomes reflected in the evaluation questions. Documents, including case studies, will be collected and reviewed to gather additional information about the educational uses, collaborative activities, and related activities. Finally, observations of SESF and educational activities will provide evidence of the user-centric utilization of the semantic web methodology and technologies that work. The evaluator will provide annual reports with recommendations to support continuous program improvement. A final summative report will document progress toward program goals over the course of the project highlighting effective strategies, and will provide recommendations for sustaining, improving, and replicating practices. They will collaborate with the SESF Investigators, project staff to ensure data, findings, and recommendations serve to enhance the overall quality of the project and potential sustainability. The following evaluation questions may be collected with Interviews/ focus groups, surveys, document analysis and observation methods - To what extent do(es) the SESF:

1. Activities enhance end-user access to and use of data to advance science and education needs?
2. Activities enable higher levels of semantic capability and interoperability such as explanation reasoning on rules, and semantic query?
3. Contribute to the development and support of community resources, virtual observatories and data systems and provision of results from diverse observing systems using semantically-enabled technologies?
4. Template contribute to the reports on modern data frameworks, user interfaces, and science progress achieved?
5. Incorporate user experiences in the redesign and development cycles of the SESF?
6. How do SESF activities affect IHE faculty and staff from participating institutions (e.g., changes to virtual observatories and data sources, results from diverse observing systems using semantically-enabled technologies and institutional collaboration activities?)

What factors impede or facilitate progress toward SESF goals is best answered in interviews and observation. The question: What progress has been made toward sustaining and ‘scaling up ’ SESF activities may be answers also interactively and via document analysis

2.3.8 Workshops

Based on our experience with use case development and refinement, community engagement, both specialistic and non-specialist to provide ontology vetting, We have found a workshop format (ranging from 6 up to 25 participants, depending on the desired outcomes and scope) is a very effective mechanism to make rapid progress. We include a funding request in the budget for this effort to facilitate workshops each year of the project. The workshops can be part of a larger meeting, stand-alone or partly virtual (via remote telecommunication). We have found (for example, in our data integration work) that domain experts in particular are extremely willing to participate in these workshops. Together Fox and McGuinness have run more than 30 such workshops over the last 5 years. Workshops will also be held for developers and those broadly engaged in tool/ application research.

2.3.9 Intellectual Merit: Education and Workforce Development and Broader Impacts: on Educational Communities

Two clear impacts are:

- Impact of the non-specialist use case on bringing data to new communities
- Impact of training a cadre of undergraduates with the SESF model of advanced semantics

Virtual Observatories and eScience frameworks are recognizing the need to reach beyond the scientific audience that they traditionally serve to policy makers, educators, and to a broader scientific community with the goal of anticipated emergent opportunities. The development of a successful model for a non-specialist
use case can offer a variety of VOs the possibility of reaching that broader audience using a set of pathways that are demonstrated to be effective.

We will be employing a series of undergraduates from computer science, and many areas of chemistry, biology and physics, and education at RPI to develop and test the non-specialist use case model. In the process, we hope to train a cadre of students that can effectively work across boundaries in support of virtual observatories.

Provided we meet our project goals for the education and workforce development, we will propose at a later time to conduct a longitudinal study to follow up with students in ~ 2-3 years time that will build on the yearly evaluations and the longer term consequences.

2.3.10 SESF Sustainability: Evolution and Viability

The existing VSTO framework (Fox et al. 2006, Fox et al. 2009) will evolve as a part of this project. A broad schematic of SESF is presented in Fig. 4. Since VSTO currently is a production environment with four public (two portal and two web services) and one internal (Java API) interface and SESDI and SPDCDIS capabilities are being used by scientists daily, it is essential that this service will not be disrupted by the augmentations we have proposed in this project.

As a result of adding to the mid-level and application ontology, we will add to the internal API and smart application tool suite. A user of the portal will begin to see evidence of more annotation to the web pages displayed in response to queries.

2.3.11 Configurator

To enable configuration and extensions for specific applications we will provide discipline and level specific data framework built from a configuration procedure by partitioning the ontology, generating the class interfaces, etc. that can be assembled with the existing user interface front ends (portal and web services). For example, we may utilize the semantic filters of FieldofStudy=“SolarPhysics” and Instrument=“SpectroPolarimeter” (or better yet, Parameter=“MagneticField”) to query the ontology, coupled with reasoning to extract knowledge representations of the measurement of magnetic fields for the Sun and that would include all the terminology, subclasses, properties and instances that are relevant.
2.3.12 Semantic Technologies

Over the past few years, semantic technologies have evolved, e.g. OWL 2, and new tools (Jiao et al. 2009, etc.) are appearing. Part of the effort in this project will be to accommodate these advances in the new framework and lay out a sustainable software path for the (certain) technical advances. To ensure the framework is usable, based on our accumulated experience with the three projects, users and various levels of evaluation studies, we will package user tools, and applications based on the framework. Examples of this may be a semantic faceted search (e.g. similar to mspace or Jspace) with OWL encoding and reasoning, and an OWL-RL (new Rule Interchance Format subset encoded in OWL) reasoner.

In addition to a generalization of the current data science interface, we will include an upper-level interface suitable for use by clearinghouses, and/or educational portals, digital libraries, and other disciplines. These interfaces will be manifest both as semantically-enabled web service remote portlets (WSRP), i.e. web interface components ready to integrate into a variety of end-user applications and modalities (hand-held, classroom, ...), and as a robust high-performance production web service provision accompanied by a published suite of ontologies.

2.3.13 Broad design

Sustaining and evolving: As noted earlier we will utilize to every extent possible existing work in our prototyping and re-design and re-build only when it is necessary to meet a functional requirement. We do expect to re-use many elements of all of the existing project efforts (VSTO, SESDI, and SPCDIS) but we will research and augment and instrument them with the required semantic configuration and annotation requirements to serve many disciplines.

‘User’/‘Science’ focus: The investigator team has substantial experience with using use case driven requirements development which is how we plan to translate user requirements into an effective system. We plan to focus on the diverse science data, models, tools, etc. community of collaborators in gathering the user requirements. We have several test-bed data collections (production and developing) to design and build the first instance of the SESF.

Rapid prototyping: We will continue with a rapid prototyping approach which we successfully devised and adopted for VSTO, SESDI and SPCDIS. All the existing semantic web languages, tools and services are available for immediate use as are the auxiliary data and catalog services, triple store, etc.

Software Engineering: RPI professional software engineers and the community will develop software for SESF in addition to altering existing software. Software that may require partial development will include the calibration processing programs, and tools for populating catalogs and creating and querying the triple store leveraging emerging efforts such as D2RQ and D2RServer.

Ontology work: We will develop a set of descriptive ontologies leading to an extension of the current VSTO specification of semantic concepts and properties for data, data types, data quality, data processing, data access and transport infrastructure including support of new data formats and the necessary security software to identify and track users and usage where required.

2.4 Proposed work

To evolve VSTO+SESDI+SPCDIS to SESF in year 1 we will (objective=major task):

- hire post-doctoral fellow, engage graduate student, familiarize them with the project
- (objective) team meeting
- (objective) develop and refine first set of non-specialist use case in science area
- (objective) plan and hold workshop to involve computer scientists, scientists, educators and non-
specialist; present use cases, ontology and design
- assess readiness of mid-level ontologies
- assess availability and capability of educational standard ontologies
- SESF design and software engineering plan
- (objective) refactor VSTO, SESDI ontologies and redeploy frameworks and user interfaces for each
  application
- incude SESF progress into RPI Semantic eScience graduate class
- develop user access and interface requirements for non-specialist
- develop/augment mid-level ontology (leverage SWEET 2.0 and VSTO, SESDI, others)
- perform tool and application support design and prototype
present and publish progress of project work at national/international meetings and in publications

In year 2:

• (objective) deploy first prototype framework of SESF using VSTO including new user access interface for non-specialist use case (public release)
• develop and refine second set of non-specialist use cases
• assess tool and application support
• (objective) perform SESF (including software engineering) full design
• (objective) deploy second prototype framework of SESF using SESDI application
• RPI Semantic eScience graduate class uses and evaluates developed SESF, solicit new use cases from students
• (objective) plan and hold training workshop to involve developers of tools and applications to present programming interfaces, tools, configuration options and design
• (objective) plan and hold workshop to involve non-specialist and computer scientists; present use cases, ontology and design
• (objective) review design, and re-implement and review
• review tool and application support design and prototype
• (objective) review SESF design, assess and augment SESF ontology configuration
• (objective) publish SESF ontologies
• augment mid-level ontology framework in community settings (e.g. ESIP)
• (objective) deploy second prototype framework of SESF
• (objective) capability demonstration at regional or national science meeting
• present and publish progress of project work at national/international meetings and in publications

In year 3:

• (objective) second public release of SESF
• select second science discipline/community for engagement
• review upper-level ontologies for use
• RPI Semantic eScience graduate class uses and evaluates re-designed SESF
• release SESF interface specification and ontology after mid-year
• (objective) full review and evolution of updated SESF ontology
• publish updated SESF ontology, augment mid-level ontology
• carry out and publish informal evaluation study on SESF implementation
• (objective) capability demonstration at regional or national science meeting and publish progress
• final report

2.5 The personnel to address the challenges

Professor Fox will serve as overall PI, chief architect and technical and natural science lead on the project providing local expertise in domains, lead the use case and science ontology modeling efforts, and connect the implementations to the science data services. Professor McGuinness will serve as co-PI, share technical coordination responsibilities for the overall project and provide leadership in computer science, knowledge acquisition, ontology modeling, ontology environments, and knowledge provenance. Professor Hendler (senior personnel) will serve as advisor and un-funded senior personnel and provide guidance on technical leadership in scalable web systems, application tools and services. The investigators are supported by a post-doctoral research scientist, 2 PhD students, and 1.5 FTE software engineer. The post-doctoral fellow will have regular interactions with the students in building the ontologies and implementing the services as well as in collaborations with the scientific and developer communities. For post doctoral mentoring we will provide the training and opportunity to develop skills in research project management, writing research papers for publication, and in international virtual observatory research development. The fellow will also participate in the graduate semantic eScience class each Fall. The investigator team exploits highly complementary expertise in ontologies and ontology environments (McGuinness), domain science (Fox) and scalable web science (Hendler). Each of the investigators has a history of working on application and development of fundamental applications within and between their fields and the post-doctoral fellow will be part of that scientific research environment. Hendler, McGuinness and Fox have recently joined Rensselaer Polytechnic Institute and formed the innovative Tetherless World constellation. The constellation focuses on next generation web technologies and their use, particularly for eScience. Further, each investigator has significant and extensive domestic and foreign collaborator connections and broad community connections.
things, McGuinness was a leading researcher in the current W3C recommended ontology language and in using building and using ontologies in large sponsored research efforts. Prior to coming to RPI, Dr. Fox was Chief Computational Scientist at the NCAR High Altitude Observatory and is a science and informatics leader in Virtual Observatories in professional societies (AGU, EGU) and international unions/associations (IUGG and COSPAR) as well as the new International Council for Science’s Strategic Coordinating Committee on Information and Data (SCCID). Dr. McGuinness was Acting Director of the Knowledge Systems Laboratory at Stanford where she led numerous ontology-based multi-million dollar government sponsored research efforts. Hendler initiated the DARPA program that led to OWL, and McGuinness and Fox have a five year history of joint government funded interdisciplinary semantic research for science and McGuinness and Hendler have a ten year joint history in funded knowledge representation and reasoning infrastructure research and individually have 20+ year histories in KR&R and scalable systems respectively. Fox has a 23+ year history in cutting edge scientific informatics research in addition to his domain science research and is PI for four major informatics/ cyberinfrastructure projects bringing semantic web to science data holdings. Additionally the investigators have a long history of sponsored research work and enjoy doing both fundamental research and providing implemented reusable results that can be used by a broader community.

We also draw on two experienced professional software engineers, Patrick West and Stephan Zednik who are actively working on VSTO, SESDI and SPCDIS project (as well as data intensive project such as OPeNDAP and the Earth System Grid). To provide a higher level of academic connection we will hire one new Post-doctoral fellow for effort intended to bridge science data and model expertise.

Candidate external evaluators for this project include: RMC Research Corporation of Denver, CO has 35+ years of evaluation experience and a strong background in Science, Technology, Engineering, and Mathematics programs and the experts at the School of Information at the University of Michigan. We will determine the final evaluator using an RFP mechanism.

We base the estimates of all personnel efforts on direct experience with previous or current projects such as OPeNDAP, ESG, VSTO and CEDAR. We also will draw on a set of existing collaborations with technology projects and where previously indicated build new associations and collaborations with peer projects in the community.
Selected Reading and References


CEDAR data services, http://cedarweb.hao.ucar.edu/


JESS; http://www.jessrules.com/

Jena; http://jena.sourceforge.net/

Knowledge Interchange Format (KIF), http://www-ksl.stanford.edu/knowledge-sharing/kif/

LEAD, Linked Environments for Atmospheric Disturbance, http://www.leadproject.org/


MLSO, Mauna Loa Solar Observatory, http://mslo.hao.ucar.edu/


Open Knowledge Base Connectivity (OKBC), http://www.ksl.stanford.edu/software/OKBC/

OWL (Web Ontology Language) http://www.w3.org/TR/owl-ref/


Rules Interchange Format (RIF), http://www.w3.org/2005/rules/

SESDI, Semantically-Enabled Science Data Integration, http://sesdi.hao.ucar.edu/


Jiao Tao, Li Ding, Deborah L. McGuinness, Instance Data Evaluation for Semantic Web-Based Knowledge Management Systems, 42th Hawaii International Conference on System Sciences (HICSS-42), January

