HAO
Exploring the Sun and its effects on the Earth's atmosphere and physical environment.
HIGH ALTITUDE OBSERVATORY

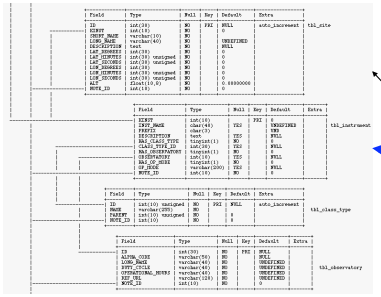
Infusing semantic web into operational data systems: real application experience

Patrick West[1] (pwest@ucar.edu), Peter Fox[2] (pfox@cs.rpi.edu), Deborah McGuinness[2,3] (zednik@ucar.edu), Stephan Zednik[1] (zednik@ucar.edu), ([1]HAO/NCAR P.O. Box 3000, Boulder, CO, 80307 United States), ([2]Rensselaer Polytechnic Institute 110 8th St., Troy, NY, 12180 United States), ([3]McGuinness Associates 4 Shaker Bay Rd., Latham, NY, 12110 United States)

Abstract

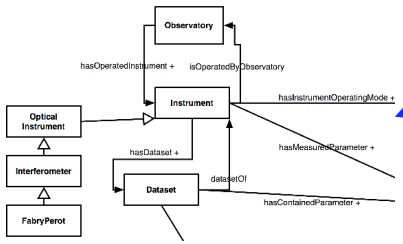
As part of our semantic data framework activities across disciplines from solid-earth, lower, middle and upper terrestrial atmosphere, and solar atmosphere, to integrative subjects such as climate response and space weather, we have collected a set of experiences: technical, collaboration and social that relate to how easy, or hard the infusion process has been. We cover both the semantic web and knowledge infusion as well as underlying service infusion such as catalogs and OPeNDAP data servers.

Catalogs/Ontology Integration



CEDAR Catalog Entity Relationship Diagram. You can see here how many of the tables and relationships in the CEDAR catalog relate directly to classes and properties in the VSTO ontology (partially represented below). We expanded on this to add more relationships, more properties, more semantics into these entities. The work in CEDAR has helped to define the ontology for VSTO, and the work on the ontology for VSTO has helped to define a better Catalog for CEDAR.

VSTO Ontology



Above is a portion of the VSTO ontology, which can be accessed at http://dataportal.ucar.edu/schemas/vsto_all.owl.

Gloassary:
VSTO - Virtual Solar Terrestrial Observatory
OPeNDAP - Open-source Project for Network Data Access Protocol
CEDAR - Coupled Energetics and Dynamics of Atmospheric Research
MLSO - Mauna Loa Solar Observatory

Knowledge Infusion

Inter-discipline support: VSTO science domains are solar physics, space physics, and solar-terrestrial physics. There are many data collections, spanning disciplines, and growing in volume and complexity. Major communities include those interested in solar images from the Mauna Loa Solar Observatory (MLSO), and the NSF-funded Coupled Energetics and Dynamics of Atmospheric Regions (CEDAR). These collections provided a good focus for virtual observatory work since the datasets are of significant scientific value to a set of researchers and capture many, if not all, of the challenges inherent in complex, diverse scientific data.

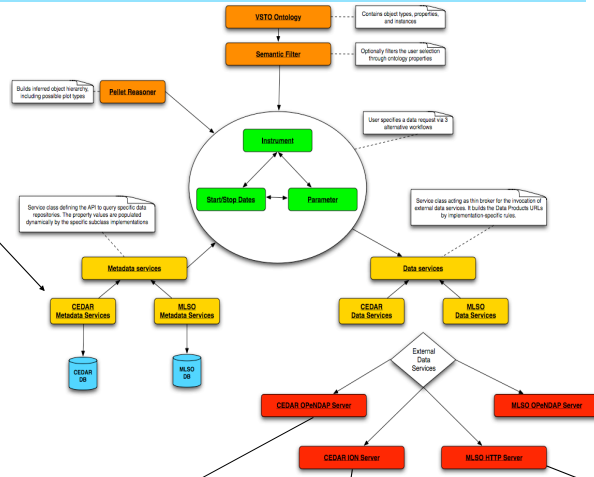
Use Cases: Use cases are used to indicate a specific capability that drives both what knowledge is to be represented and used and also what software and interfaces are built for the user and to the underlying data. A use case appears to be a brief statement but in practice is accompanied by detailed descriptions including functional and non-functional requirements, success and failure scenarios, etc.

Ontologies: Ontologies created in OWL to define concepts, relations, terms etc. in order to utilize their precise formal definitions for semantic search and interoperability. Use case sentences examined to identify initial concepts and relations between them. Hierarchies became apparent, and important properties, as well as restrictions on the values for certain concepts within a given context.

Integration

One significant benefit of semantics has been in facilitating and enabling integration of data and information sources. This integration occurs in several ways: finding and using independent variable data from sources which may not be in the data source (e.g. sometimes time coordinate values are in the header records for the CEDAR database), finding and fusing data measured by different instruments (e.g. temperature and density for different parts of the middle and upper atmosphere from optical instruments and radars) possibly from very different disciplines and sub-disciplines (e.g. chemical constituents of the atmosphere measured by remote sensing instruments and in-situ communities). One of the major benefits of ontologies is to formalize the relations between physical quantities, instruments that measure them, and important ancillary metadata.

VSTO Workflow



Experience

Technical: Free, or open-source software tools, packages and development environments are used with the intent of documenting an end-to-end methodology that would be reproducible and usable by others without the need for significant investments in time and resources. Web-browser access was required, via a portal, web services as well as native application programming interface (API) access. A web-architecture using Java is used as well as utilizing interface-level access to existing services that provide access to data, graphical representations, extant catalogs, etc. OWL-DL is used for ontology development (as opposed to OWL-Full) in order to leverage efficient reasoning tools for OWL-DL. Protégé's automatic Java and factory classes generation capabilities are used. The Pellet reasoning engine is used to support multiple workflow scenarios.

Collaboration: To enable explicit semantic interoperability, extensive engagement of the end user; the domain scientist, associate scientists, students and professional assistants was a key element. In order to provide a scientific infrastructure that is usable and extensible, VSTO required contributions concerning semantic integration, and knowledge representation while requiring depth in each of the science areas. In developing and analyzing the use cases, a small team made up of domain literate experts, data and instrument providers, knowledge representation and engineering experts, computer science/software engineers and a facilitator are utilized.

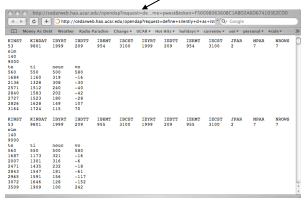
Social: An implication of adding semantics is the strong need and role for domain literate members of the team to develop and then vet the knowledge encoding. This turns out to be a social exercise as much as it is a technical and methodical activity since the team comprises people from multiple disciplines. As such they are used to different modes of working, vocabularies and incentive and reward systems. The interaction of domain scientists, computer scientists, software engineers, data managers and often social scientists. Personalities are important to consider and the leader of this activity must be aware and accommodate many dimensions of participation if the infusion activity is to be successful. One result of this is that it is very difficult to conduct the infusion simply as a remote or virtual activity unless it is substantially founded on face-to-face meetings when social forces come into play. Social considerations also affect how project are sustained and we are just beginning to understand these consequences and will report on them in the future.

Infusion

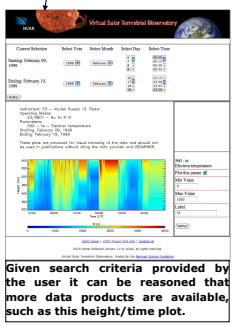
Easy:
• Identification of domain experts was key: two to four carefully chosen experts were sufficient (and preferable) in the use case development and knowledge engineering. At later stages a larger group can be effective, especially in getting community support and buy-in for the utilization of semantic technologies.

Hard:
• Face-to-face meetings of team members to participate in infusion activities.
• Mapping of certain components of data catalogs from the specific domains to classes and instances within the ontology, such as CEDAR's lack of specification of independent variables in datasets.
• Gaining access to data holdings and providing external data access for virtual observatories data services

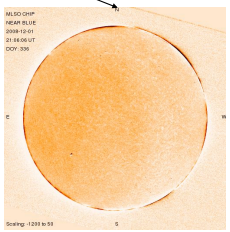
Data Integration



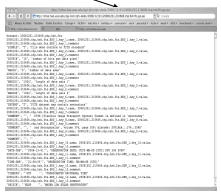
OPeNDAP Back-End Server Apache Module accessed via URL built by the CEDAR data service class.



Given search criteria provided by the user it can be reasoned that more data products are available, such as this height/time plot.



MLSO CHiP velocity jpg image after processing by scientists. This is a service provided by MLSO via a URL.



MLSO has an OPeNDAP Server 3 CGI programming available to access its data. This is specified in the MLSO data service class.