Integrating Inter-disciplinary Science Data with Seman

Peter Fox (HAO/ESSL/NCAR)
Krishna Sinha (VT), Rob Raskin (JPL), Deborah McGuinness (RPI)

SESDI - Semantically-Enabled Science Data Integration
Funded by NASA/ACCESS and NASA/ESTO/AIST
Overview

• A little about semantics
• A little about integration
• Use case
• Semantic mis-understanding
• Impact of semantic mediation
• Methodology
• Some details of the integrating concepts
• Now, let’s hook up some data
• Summary and outlook
A little about semantics

• Gives syntax *meaning*
• Basic element is the *triple*: `{subject-predicate-object}`
  - Interferometer is-a optical instrument
  - Optical instrument has focal length
  - An ontology is a representation of this knowledge
• W3C is the primary (but not sole) governing organization for languages, specifications, best practices, etc.
  - RDF - Resource Description Framework
  - OWL 1.0 - Ontology Web Language (OWL 2.0 on the way)
• Encode the knowledge in triples, in a triple-store, software is built to traverse the semantic network, it can be queried or reasoned upon
• Put semantics between/ in your interfaces, i.e. between layers and components in your architecture, i.e. between ‘users’ and ‘information’ to mediate the exchange
A little about integration

- When we integrate, we integrate concepts, terms
- In the past we would ask, guess, research a lot, or give up
- It’s pretty much about meaning
- Semantics can really help find, access, integrate, use, explain, trust…
- What if you…
  - could not only use your data and tools but remote colleague’s data and tools?
  - find and use data you could not before?
  - 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?…
Integrative Use Case

• Determine the statistical signatures of both volcanic and solar forcings on the height of the tropopause
Challenges for Solar Radiation data integration

• Semantic misunderstanding
  – E.g. sunspot number and variations in solar radiation: over 90% of researchers outside the sub-field of solar radiation think: sunspot number is a measure of solar radiation
  – In reality: a sunspot number is a measure of the number of sunspots appearing on the visible solar surface, a sunspot is an indicator of the location of strong solar magnetic fields, strong magnetic fields are collectively known as solar activity, sunspots are observed to produce a localized decrease in the solar radiation output, at some wavelengths, increase at others, etc.

• Interfaces are built by computer scientists with syntax that often works within a discipline but rarely across them
**SESDI Impact:** A Better Way to Access Data

**The Problem**
Scientists only use data from a single instrument because it is difficult to access, process, and understand data from multiple instruments. A typical data query might be:

"Give me the temperature, pressure, and water vapor from the AIRS instrument from Jan 2005 to Jan 2008"

"Search for MLS/Aura Level 2, SO2 Slant Column Density from 2/1/2007"

**A Solution**
Using a simple process, SESDI allows data from various sources to be registered in an ontology so that it can be easily accessed and understood. Scientists can use only the ontology components that relate to their data. An SESDI query might look like:

"Show all areas in California where sulfur dioxide (SO2) levels were above normal between Jan 2000 and Jan 2007"

This query will pull data from all available sources registered in the ontology and allow seamless data fusion. Because the query is measurement related, scientists do not need to understand the details of the instruments and data types.
Semantic Web Methodology and Technology Development Process

- Establish and improve a well-defined methodology vision for Semantic Technology based application development
- Leverage controlled vocabularies, etc.
Volcano-Atmosphere considerations

• Focus on tropopause -> temperature gradients
• Stratospheric and tropospheric aerosols, the tropospheric reservoir
• Quantities/processes: Gas, particles, ejecta, scattering
• Records: Pulses, e.g. in SO2 events
• Related aspects: SO2, H2SO4, O3 chemistry
• Data from: in-situ and remotely sensed observation, proxy, simulation, pseudo-proxy
• Processes: solar, volcanic, GHG, ocean, land-use
• Priors to consider: statistics of variability and extremes
• Main task: detection and attribution
• Solar-Atmosphere considerations are very similar
Components to implement

- An analysis application
- Cross-domain terms, concepts and relations (mediation here)
- Connections to underlying data (registration and mediation)
- Framework to put these together
- Integration connector
Detection and attribution relations…
SWEET 2.0 Ontologies

Math
- Statistics
- Geometry
- Vector
- Units
- Space
- Time

Informatics
- Data
- System

Science
- Biology
  - Plant
  - Animal
- Ecology
- Mechanics
- Thermo
- ElecMag
- Nuclear
- Waves
- Radiative Transfer
- Gephys
- Fluid Dyn
- Physics
- Chemistry
- SocialSci
- Matter
- Element
- Compound

PlanetaryScience
- SolidEarth
- Tectonics
- Volcano
- Seismo
- Bound Layer
- Cloud
- Front
- Precip
- Electric
- Atmosphere
- Hydrosphere
- Ocean
- Cryosphere
- Heliosphere
- Sun
- Surface
- Soil
- Landscape
- Coast

Applications
- Hazards
- Hydrology
- Agriculture
- Infrastructure
- Admin Borders
Data Registration Framework

**Level 1:**
Data Registration at the Discovery Level, e.g. Volcano location and activity

**Level 2:**
Data Registration at the Inventory Level, e.g. list of datasets by types, times, products

**Level 3:**
Data Registration at the Item Detail Level, e.g. access to individual quantities

**Earth Sciences Virtual Database**
A Data Warehouse where Schema heterogeneity problem is Solved; schema based integration

**Ontology based Data Integration**
• The data is organized under the following categories:

- AtmosphericScienceDiscipline
- AtmosphericPhenomenon
- AtmosphericDynamics
- AtmosphericThermodynamics
- AtmosphericPhase
- AtmosphericChemistry
- AtmosphericContaminant
- AtmosphericComposition
- AtmosphericForcing

- AtmosphericForcing is a type of Aerosol
- Aerosol is a type of Watervapor
SEDRE: Semantically Enabled Data Registration Engine

• **SEDRE**: a system that enables scientists to semantically register data sets for optimal querying and semantic integration

  • **SEDRE** enables mapping of heterogeneous data to concepts in domain ontologies

A. K. Sinha, A. Rezgui, Virginia Tech
NASA ESTC 2008 Fox Semantic Data Integration
Semantic Registration in SEDRE: An Overview

- SEDRE is a desktop application
- Users download and install SEDRE
- SEDRE accesses domain ontologies
- Users map data attributes (e.g., SO$_2$) to concepts in ontologies without ‘knowing it’
Example 1: Registration of Volcanic Data

Location Codes:
• U - Above the 180° turn at Holei Pali (upper Chain of Craters Road)
• L - Below Holei Pali (lower Chain of Craters Road)
• UL - Individual traverses were made both above and below the 180° turn at Holei Pali
• H - Highway 11

SO₂ Emission from Kilauea east rift zone - vehicle-based (Source: HVO)

Abreviations: t/d=metric tonne (1000 kg)/day,
SD=standard deviation, WS=wind speed, WD=wind direction east of true north, N=number of traverses
Loading Volcanic Data into SEDRE
Registering Volcanic Data (1)
Registering Volcanic Data (2)

- No explicit lat/long data
- Volcano identified by name
- Volcano ontology framework will link name to location
Example 2: Registration of Atmospheric Data

Satellite data for $SO_2$ emissions

<table>
<thead>
<tr>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
<th>K</th>
<th>L</th>
<th>M</th>
<th>N</th>
<th>O</th>
<th>P</th>
<th>Q</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>-40.487</td>
<td>-40.412</td>
<td>52.913</td>
<td>52.847</td>
<td>52.06</td>
<td>51.994</td>
<td>52.443</td>
<td>46.464</td>
<td>33.584</td>
<td>48.756</td>
<td>0.293</td>
</tr>
<tr>
<td>3</td>
<td>-40.371</td>
<td>-40.303</td>
<td>52.06</td>
<td>51.994</td>
<td>51.244</td>
<td>51.178</td>
<td>51.61</td>
<td>46.838</td>
<td>29.258</td>
<td>48.365</td>
<td>-0.443</td>
</tr>
<tr>
<td>4</td>
<td>-40.261</td>
<td>-40.194</td>
<td>51.244</td>
<td>51.178</td>
<td>50.488</td>
<td>50.422</td>
<td>50.826</td>
<td>47.195</td>
<td>24.866</td>
<td>48.004</td>
<td>0.875</td>
</tr>
<tr>
<td>5</td>
<td>-40.153</td>
<td>-40.089</td>
<td>50.488</td>
<td>50.422</td>
<td>49.778</td>
<td>49.711</td>
<td>50.095</td>
<td>47.533</td>
<td>20.499</td>
<td>47.661</td>
<td>1.021</td>
</tr>
<tr>
<td>6</td>
<td>-40.049</td>
<td>-39.986</td>
<td>49.778</td>
<td>49.711</td>
<td>49.107</td>
<td>49.039</td>
<td>49.405</td>
<td>47.856</td>
<td>16.159</td>
<td>47.338</td>
<td>-1.422</td>
</tr>
<tr>
<td>9</td>
<td>-39.738</td>
<td>-39.681</td>
<td>47.833</td>
<td>47.763</td>
<td>47.215</td>
<td>47.143</td>
<td>47.488</td>
<td>47.876</td>
<td>3.249</td>
<td>46.117</td>
<td>1.735</td>
</tr>
<tr>
<td>10</td>
<td>-39.631</td>
<td>-39.575</td>
<td>47.215</td>
<td>47.143</td>
<td>46.599</td>
<td>46.526</td>
<td>46.871</td>
<td>46.334</td>
<td>5.067</td>
<td>132.086</td>
<td>0.558</td>
</tr>
<tr>
<td>11</td>
<td>-39.521</td>
<td>-39.467</td>
<td>46.599</td>
<td>46.526</td>
<td>45.983</td>
<td>45.907</td>
<td>46.255</td>
<td>49.366</td>
<td>5.345</td>
<td>133.58</td>
<td>-0.208</td>
</tr>
</tbody>
</table>

Abbreviation: SCD: Slant Column Density (in Dobson Unit (DU))
Loading Atmospheric Data into SEDRE
Registering Atmospheric Data (1)
Registering Atmospheric Data (2)
Semantic framework indicating how volcano and atmospheric parameters and databases can immediately be plugged in to the semantic data framework to enable data integration.
Summary and outlook

• Semantic data frameworks/technologies are changing the landscape of providing data to scientists (in a good way)
• Tools for data registration are soon to be available
• Applications to perform data integration mediated by semantics are available
• Initial results - applied to two volcanoes - led to correlation of SO2 concentration from volcano and in the atmosphere and relation to H2SO4
• Solar radiation ontologies and data sources are in progress
SEDRE+DIA: Overview

DIA: Web-based System for Data Discovery, Integration and Analysis

(Developed at Virginia Tech through NSF funding)
General applicability

• To apply to another use case: