Semantically-enabled (large-scale) Scientific Data Integration (SESDI)

Peter Fox*
Deborah McGuinness$#
Robert Raskin%
Krishna Sinha@

*HAO/ESSL/NCAR
$McGuinness Associates
#Knowledge Systems and AI Lab, Stanford Univ.
% JPL/NASA
@ Virginia Tech

+ others from NCAR and the VSTO project

Work funded by NSF/CISE/SCI and NASA/ACCESS and NASA/ESTO.
Outline

- Virtual Observatories as a context
- Some examples of data integration within and across disciplines; the emerging need

- Challenges and integration
- NASA needs
- The enabling technology – guess what?
- What’s ahead
Recent definition

➤ Workshop: A Virtual Observatory (VO) is a suite of software applications on a set of computers that allows users to uniformly find, access, and use resources (data, software, document, and image products and services using these) from a collection of distributed product repositories and service providers. A VO is a service that unites services and/or multiple repositories (from NASA VO workshop, October 2004)

➤ VxOs – x is one discipline, you may have one and not call it a VO

➤ Increasingly, the need is for VxyO – i.e. interdisciplinary (or VxyzO, etc.) and emphasizes integration (data, processes, etc.)
Virtual Observatories

- Conceptual examples:
  - In-situ: Virtual measurements
    - Related measurements

- Remote sensing: Virtual, integrative measurements
  - Data integration
The active Sun
Center for Integrated Space–Wx Modeling

Goal: To create a physics-based numerical simulation model that describes the space environment from the Sun to the Earth.

**THE USES OF SPACE WEATHER MODELING**

A scientific tool for increased understanding of the complex space environment.

A specification and forecast tool for space weather prediction.

An educational tool for teaching about the space environment.

Integrates/assimilates data.
Challenges for 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, etc.*
  » How to ‘explain’ this to a computer?

» Interfaces are built by computer scientists with syntax that often works within a discipline but rarely across them
Synthesizing the solar spectrum

RT model
  └ Empirical atmos model
      └ Atomic/ Molec. data

Spectral Database(s)
  └ Feature distribution on the solar disk: MASKS, parametric

Synthesis
  └ Synthetic Spectra
  └ Synthetic Image(s)
      └ Intensity Histograms
Paradigm shift for NASA?

- From: Instrument based
- To: Measurement based
- Requires: ‘bridging the discipline data divide’
- Overall vision for SESDI: To integrate information technology in support of advancing measurement-based processing systems for NASA by integrating existing diverse science discipline and mission-specific data sources.

Semantically-Enabled Science Data Integration (SESDI) and The Virtual Solar–Terrestrial Observatory (VSTO)
The SESDI re-useable component interfaces. The stub on each end of the connector is based on the GEON Ontology–Data registration technology and contains articulation axioms derived from the knowledge gained in the unit-level data registration. Includes integrity checks, domain and range, etc.

SWEET – Semantic Web for Earth Environmental Terminology
Compilation of distribution of volcanic ash associated with large eruptions. Note the continental scale ash fall associated with Yellowstone eruption ~600,000 years ago. Geologic databases provide the information about the magnitude of the eruption, and its impact on atmospheric chemistry and reflectance associated with particulate matter requires integration of concepts that bridge terrestrial and atmospheric ontologies.
VSTO ONTOLOGY

Semantically-Enabled Science Data Integration (SESDI) and The Virtual Solar-Terrestrial Observatory (VSTO)
What’s ahead?

- Filling out the volcano and atmospheric ontologies
- Encoding them and building a callable framework
- Filling out the instrument ontology
- Mapping the parameters to SWEET
- Identifying the databases and register them with the ontologies
- Re-configure an existing analysis application to utilize the search/retrieve framework for integration

- Repeat for solar irradiance and climate
- Generalized the semantic connector