Semantic Web for Large scale Integration of Scientific Data

Peter Fox*
James Benedict$
Deborah McGuinness$

*HAO/ESSL/NCAR

$McGuinness Associates

#Knowledge Systems and AI Lab, Stanford Univ.
Semantic Web for Large scale Integration of Scientific Data

Agenda (1 hour)

• Introduction - 10 min (terminology, examples)
• Session Introduction - 5 min
• Remainder: what are interesting aspects of large-scale science data repositories that semantics can enable?
• Wrap-up - 5 min
Semantic Web for Large scale Integration of Scientific Data

Introduction

- What is integration?
  - Virtual Observatories
  - Data Grids
  - Data assimilation

- What does a large-scale, integrated, scientific data repository look like?

- What are the challenging scientific problem, societal problems?

- How are semantics being utilized (or not) right now? Experts are capable of high levels of integration, how can they be helped?

- What can be done with semantics that cannot be done without? Or, what are the examples of mistakes that are made due to misinterpretation?
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.

• VxOs - x is one discipline
Virtual Observatories

• Conceptual examples:
  • In-situ: Virtual measurements
    – Related measurements
  • Remote sensing: Virtual, integrative measurements
    – Data integration
The active Sun, e.g.
Challenges and interoperability

• 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 (10 x #groups + individual spots) of the number of sunspots appearing on the visible solar surface. It is correlated to, but not a measure of, solar radiation.
  – Why does this matter? Low sunspot number does NOT equal low solar radiation…
  – 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

Spectral Filters

Synthesis

Synthetic Spectra

Synthetic Image(s)

Intensity Histograms
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.
Impact: Changing Science

Scientists: 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?
- had support for formulating new, meaningful, hypotheses and questions?
- knew whose results were consistent (or inconsistent) with yours?…

Funders: What if you …
- could identify how one research effort would support other efforts?
- (and your funded investigators) could reuse previous results?
- (and your funded investigators) could really interoperate?

Computer Scientists: What if you…
- could apply your techniques across very large distributed teams of people with related but different apps?
- could compare your techniques with colleagues trying to solve similar problems?
Semantic Web for Large scale Integration of Scientific Data

Session introduction

• We want your description of what integrated results from a large-scale (insert your domain here) distributed data repository looks like.

• What are some extra use-cases - things that you can imagine wanting to do that you cannot do today, or inconvenient to do (i.e. an expert can, non-expert cannot)

• Consider the perspective of virtual observatory - the need for integration within and across discipline areas
  – More than search -> smart search
  – Interpretation
  – Synthesis, leading to …
  – Prediction with causal explanation, e.g. volcano eruption
Exploration
Paradigm shift for NASA?

- From: Instrument-based
- To: Measurement-based
- Requires: ‘bridging the discipline data divide’
- Overall vision: 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.
Semantic connectors

The re-useable component interface. The stub on each end of the connector is based on the Geosciences Network Ontology-Data registration technology and contains articulated axioms derived from the knowledge gained in the unit-level data registration. Includes integrity checks, domain and range, etc.
Wrap-up?

• Watch for geospaceontologies.org - a place to discuss and share ontologies related to data integration in science
• Find a science effort in need of semantics and work with them (ask us)
• Attend a science/informatics meeting
  – Geoinformatics - May 10-12, Reston, VA
  – AGU - May 22-26, Baltimore, MD
  – Ontology workshop - May 26, JHU/APL
  – GSA - October 22-25, Philadelphia, PA
Contacts

• pfox@ucar.edu
• jbeneditc@mcguinnessassociates.com
• dlm@ksl.stanford.edu