Semantically-Enabled (Scientific) Applications

Deborah McGuinness
Acting Director and Senior Research Scientist
Knowledge Systems, AI Laboratory
Stanford University
dlм@ksl.stanford.edu
http://www.ksl.stanford.edu/people/dlm
CEO McGuinness Associates
Themes for semantically enabling science applications

• Provide explanation support for where information came from, how it was manipulated to produce results, and how trustworthy it is.
  – Inference Web *
  – PML (Proof Markup Language) *
  – IW Trust

• Build integrated scientific applications supported by semantic technologies
  – Virtual Solar Terrestrial Observatory (NSF) *
  – Semantically-Enabled Scientific Data Integration (NASA)
  – Semantic Knowledge Integration Framework (NSF)

• Explore privacy and security issues
  – Transparent Accountable Data Mining (NSF). Explain how data was used, whether it was compliant with privacy laws, and where the information came from.
General Explanation Motivation

Provide **interoperable knowledge provenance infrastructure** that supports **explanations** of sources, assumptions, learned information, and answers as an enabler for **trust**.

**Interoperability** – as systems use varied sources and multiple information manipulation engines, they benefit more from encodings that are shareable & interoperable, supporting integration.

**Provenance** – if users (humans and agents) are to use and integrate data from unknown, unreliable, or evolving sources, they need provenance metadata for evaluation.

**Explanation/Justification** – if information has been manipulated (i.e., by sound deduction or by heuristic processes), information manipulation trace information should be available.

**Trust** – if some sources are more trustworthy than others, representations should be available to encode, propagate, combine, and (appropriately) display trust values.
Inference Web Infrastructure
primary collaborators Pinheiro da Silva, Ding, Chang, Fikes, Glass, Zeng

Framework for explaining question answering tasks by
• abstracting, storing, exchanging,
• combining, annotating, filtering, segmenting,
• comparing, and rendering proofs and proof fragments
provided by question answerers.
How PML Works

October 27, 2006

Deborah L. McGuinness
Semantic Web Support for Natural Science Researchers

• Begin with providing what appears to be integrated access
  – Using ontologies and semantic technologies supporting simple workflows
  – Initially done by collaboration between domain scientists and computer scientists

• Expand to include:
  – simple provenance information (using PML-P)
  – explanations of information manipulations (using PML-J)
  – more reasoning, datasets, vocabularies, etc.

• Address distributed community evolution support, including enhanced security, trust, privacy, and collaborative evolution tools
Virtual Observatories

Scientists should be able to access a global, distributed knowledge base of scientific data that:

• appears to be integrated
• appears to be locally available

But… data is obtained by multiple instruments, using various protocols, in differing vocabularies, using (sometimes unstated) assumptions, with inconsistent (or non-existent) meta-data. It may be inconsistent, incomplete, evolving, and distributed
Virtual Observatory Defined

• 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 in Practice

Make data and tools quickly and easily accessible to a wide audience.

Operationally, virtual observatories need to find the right balance of data/model holdings, portals and client software that a researchers can use without effort or interference as if all the materials were available on his/her local computer using the user’s preferred language.

They are likely to provide controlled vocabularies that may be used for interoperation in appropriate domains along with database interfaces for access and storage and “smart” search functions and tools for evolution and maintenance.
Virtual Solar Terrestrial Observatory (VSTO)

- a distributed, scalable education and research environment for searching, integrating, and analyzing observational, experimental, and model databases.
- subject matter covers the fields of solar, solar-terrestrial and space physics
- it provides 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
- 3 year NSF-funded project just beginning the second year
Content: Coupling Energetics and Dynamics of Atmospheric Regions WEB

Community data archive for observations and models of Earth's upper atmosphere and geophysical indices and parameters needed to interpret them. Includes browsing capabilities by periods, instruments, models, ...
Content: Mauna Loa Solar Observatory

Mauna Loa Solar Observatory (MLSO) Website. The MLSO, operated by the High Altitude Observatory in Boulder Colorado, houses several instruments designed to observe the sun at many different wavelengths.

- ACOS: Advanced Coronal Observing System. A suite of instruments designed to observe the solar atmosphere at a variety of heights. Includes Chromospheric Helium Imaging Photometer (CHIP, 1083.0nm), H-alpha prominence and solar disk monitor (PICs, 656.2nm), and the MK K-coronameter, which observes the white light K-corona from 1.13-2.79 solar radii.
- ECHO: Experiment for Coordinated Helioseismic Observations. A network of two instruments which observe solar oscillations as seen in the radial velocity of the solar surface.
- PSPT: Precision Solar Photometric Telescope. Observes the solar disk in three bandpasses: 605-810 nm (red), 408-412 nm (blue), and 393 nm (CaliK).

Latest MLSO Images:

- ACOS Mark IV
- ACOS PICs Limb
- ACOS PICs Disc
- ACOS CHIP

K-Corona
720-850 nm
27-Nov-2003 22:56
Movie [merged GIF]
PSPT CaliK

H-Alpha Limb
656.3 nm
26-Nov-2003 21:11
Movie [merged GIF]
PSPT Blue

H-Alpha Disk
656.3 nm
26-Nov-2003 21:04
Movie [merged GIF]
PSPT Red

Helium I
1083 nm
26-Nov-2003 17:39
Movie [merged GIF]
ECHO Sample Velocity Image

Near real-time data from Hawaii from a variety of solar instruments.

Source for space weather, solar variability, and basic solar physics.

Other content used too – CISM – Center for Integrated Space Weather Modeling.
Some Observations about the Virtual Solar-Terrestrial Observatory

- Datasets alone are not sufficient to build a virtual observatory: VSTO integrates tools, models, and data
- VSTO (and all VOs) need to work with interdisciplinary metadata, multiple controlled vocabularies, and multiple interfaces
- VSTO leverages the development of schema that adequately describe the domain information (e.g., the name of a variable, its type, dimensions, procedure names, argument lists, semantics, unit measures, procedure descriptions, dataset descriptions, tool descriptions, ...)
- VSTO provides a basis for a framework for building and distributing advanced data assimilation tools
- Just gone live in two communities: CEDAR & Mauna Loa
- Recent papers at ISWC '06, OWL-ED 06, AGU spring and fall '06, EGU '06, Intl Astronomical Union ‘06
Welcome to the Virtual Solar Terrestrial Observatory

The Virtual Solar Terrestrial Observatory (VSTO) is a unified semantic environment serving data from diverse data archives in the fields of solar, solar-terrestrial, and space physics (SSTSP), currently:

- Upper atmosphere data from the CEDAR (Coupling, Energetics and Dynamics of Atmospheric Regions) archive
- Solar corona data from the MLSO (Mauna Loa Solar Observatory) archive

The VSTO portal uses an underlying ontology (i.e., an organized knowledge base of the SSTSP domain) to present a general interface that allows selection and retrieval of products (ascii and binary data files, images, plots) from heterogeneous external data services.

VSTO Data Access

Acknowledgments

VSTO is a collaboration of the ESSL/HAO (High Altitude Observatory) and CISL/SCD (Scientific Computing Division) divisions at NCAR with McGuinness Associates, funded by the National Science Foundation. This study made use of the CEDAR Database at the National Center for Atmospheric Research which is supported by the National Science Foundation. This study made use of data from the Mauna Loa Solar Observatory operated by the High Altitude Observatory at the National Center for Atmospheric Research which is supported by the National Science Foundation.
Status

• Scratched the surface, but still usable and useful for prototypical use:
• “plot values of parameter1 as recorded by instrument2 subject to constraints in a time period, in a manner that makes sense for the data. E.g., plot neutral temperature as recorded by Millstone Hill Fabry-Perot Interferometer looking in the vertical direction in Jan 2000, in a manner appropriate for the data.”
• Provides integrated access exploiting
  – Domain ontologies (integrated with emerging domain standards (SWEET) and representation language standards (OWL)
  – Semantic web tools (reasoners, editors, sw toolkits)
  – Provides “smarter” queries (in that interfaces help build syntactically and semantically correct queries that require less user input)
• Exploring leveraging current infrastructure
  – Upper level domain ontologies being reused by SESDI (Volcanoes and Climate data)
Impact: Virtual Observatories 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?
- knew whose results were consistent (or inconsistent) with yours?…

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

CS: 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?
Conclusion

Semantic Web languages and tools are evolving and are enabling next generation collaboration and applications including:
- explainable question answering systems
- semantic integration of information
- usable, integrated, explainable virtual observatories

Requires:
- teams including domain and computer scientists
- semantic annotation of data
- usage and analytics-driven ontology design
- (currently) identified data collections

Benefits from:
- use of standards, registry of tools and services
(note – chute’s recommendations are embodied in this effort)

For more info on talk topics:
- Inference Web - iw.stanford.edu (OWL - www.w3.org/TR/owl-features/)
- Virtual Solar Terrestrial Observatory- www.vsto.org
- AGU Session on Earth and Space Science Cyberinfrastructure
- AGU Town Hall on Cyberinfrastructure http://www.agu.org/meetings/fm06/
More Information

http://iw.stanford.edu/2.0/publications.html


Best Requirements paper (from interviewing Intelligence Analysts)
Cowell, McGuinness, Varley, &Thurman. **Knowledge-Worker Requirements for Next Generation Query Answering and Explanation Systems.** Workshop on Intelligent User Interfaces for Intelligence Analysis, (IUI 2006), Sydney, Australia


Extras
Research Theme: Making KR&R More Usable

Ontology Level
- Languages (CLASSIC, DAML-ONT, DAML+OIL, OWL, IKL, …)
- Environments (FindUR, Chimaera, OntoBuilder/Server, Sandpiper …)
- Standards efforts (NAPLPS, …, W3C’s WebOnt, W3C’s Semantic Web Best Practices, EU/US Joint Committee, OMG ODM, …)

Rules - SWRL (previously CLASSIC Rules, …)
Logic - Description Logics, FOL, …

Proof - PML, Inference Web Services and Infrastructure

Trust - IWTrust, Collaborative Information
Repository Trust, NSF TAMI with W3C/MIT

Applications
- VSTO, SESDI, SKIF, SSOA, BISTI, …
- Domain ontologies & environments
- Tools academic & industry(Sandpiper, …)
- Startups: BlueChip, Katalytik, …
Annotations: (Add)
owl:versionInfo (en): version 2.0, Authors: Deborah L. McGuinness, Li Ding, Paulo Pinheiro da Silva, Alyssa Glass and Cynthia Chang (Delete)
rdfs:label (en): PML2 Justification ontology (Delete)
rdfs:comment (en): The justification part of PML2 ontology. It is a fundamental component of PML2 ontology. (Delete)

Imports: (Add)
http://inferenceweb.stanford.edu/2006/05/aml-provenance.owl (Delete)

Total Number of Classes: 37 (Defined: 12, Imported: 25)
Total Number of Datatype Properties: 32 (Defined: 9, Imported: 23)
Total Number of Object Properties: 36 (Defined: 19, Imported: 17)
Total Number of Annotation Properties: 3 (Defined: 3, Imported: 0)
Total Number of Individuals: 1 (Defined: 0, Imported: 1)
A Sample PML encoding

http://inferenceweb.stanford.edu/2006/02/example1-iw-wiki.owl

```
  <iw:hasConclusion>
    <iw:isConsequentOf>
      <iw:InferenceStep>
        <iw:hasRule rdf:resource="http://inferenceweb.stanford.edu/registry/DPR/Told.owl#Told"/>
        <iw:hasInferenceEngine rdf:resource="http://inferenceweb.stanford.edu/registry/IE/CitationTrust.owl#CitationTrust"/>
        <iw:hasSourceUsage>
          <iw:SourceUsage>
            <iw:hasSource>
            </iw:hasSource>
          </iw:SourceUsage>
        </iw:hasSourceUsage>
      </iw:InferenceStep>
    </iw:isConsequentOf>
  </iw:hasConclusion>
</iw:NodeSet>

<iw:AggregatedTrustRelation>
  <iw:hasTrustValue rdf:datatype="http://www.w3.org/2001/XMLSchema#float">0.1766</iw:hasTrustValue>
</iw:AggregatedTrustRelation>

<iw:AggregatedTrustRelation>
  <iw:hasTrustedParty rdf:resource="http://inferenceweb.stanford.edu/wp/registry/PER/Alexandrov.owl#Alexandrov"/>
  <iw:hasTrustValue rdf:datatype="http://www.w3.org/2001/XMLSchema#float">0.1766</iw:hasTrustValue>
</iw:AggregatedTrustRelation>
```

author trust
Requirements gathered from…

DARPA Agent Markup Language (DAML)
Enable the next generation of the web

DARPA Personal Assistant that Learns (PAL)
Enable computer systems that can reason, learn, be told what to do, explain what they are doing, reflect on their experience, & respond robustly to surprise

DARPA Rapid Knowledge Formation (RKF)
Allow distributed teams of subject matter experts to quickly and easily build, maintain, and use knowledge bases without need for specialized training

DTO Novel Intelligence for Massive Data (NIMD)
Avoid strategic surprise by helping analysts be more effective (focus attention on critical information and help analyze/prune/refine/explain/reuse/…)

DTO IKRIS – Interoperable knowledge representation for intelligence apps

NSF & NASA Scientific Data Integration (NSF Virtual Observatories (VSTO), NSF GEON, NASA SESDI, NASA SKIF, …)

NSF Cybertrust Transparent Accountable Data Mining (TAMI)
Govt Classified applications that must defend their conclusions

October 27, 2006  Deborah L. McGuinness
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.
NASA Application

- One trend in science: moving 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.

Volcano  

SWEET  

Climate

SESDI
Intelligence Tool Explanation
(similar to other applications including sensor-fed scientific applications)
Follow-up: Metadata

Current Question
- List phoneCalls during Year_2003.

Selected Answer
- A phoneCall from Ramazi to Galab occurred on April_22_2003_1422 EST.
- Ramazi had access to phone703-65 April_22_2003_1422 EST Dir
- A phoneCall from phone703-65 April_22_2003_1422 EST Dir
- Galab had access to phone804

Told assertions supporting selected answer

Assertion Sources
1. Case Study 4
   - URL: file:/ksl/u5/millar/wrk/jtp/cs4.kif
   - Description in English: Domain knowledge for JMIC's Case Study 4 for use during Devil's Advocate component demonstration.

Follow up action: (Please select a follow up question)
Follow-up: Assumptions

Current Question

- List phoneCalls during Year_2003.

Selected Answer

- A phoneCall from Ramazi to Galab occurred on April_22_2003_1422_EST.

Assumptions supporting current selected answer

- Typically there may be phoneCall from ?person1 to ?person2 during ?when.
- Typically ?person1 who has access to phone804-774-8920 during Year_2003 may use it.
- Typically ?person2 who has access to phone703-659-2317 during Year_2003 may use it.

Follow up action: (Please select a follow up question)
Explaining Extracted Entities (Techies)

Sentences in English

Sentences in annotated English

Sentences in logical format, i.e., KIF
Trustworthiness of Extracted Entities

A trustworthy conclusion from IBM STAG KDD-model Annotator

A highly trustworthy conclusion from IBM EAnnotator

The combined conclusion is highly trustworthy
Estimated trustworthiness of the IBM extraction and integration components

IBM Cross-Annnotator Coreference Resolver 0.82
IBM Cross-Document Coreference Resolver 0.63
IBM EAnnotator 0.91
IBM GlossOnt 0.33
IBM J Resporator 0.31
IBM KANI holdsDuring Relation Detector 0.20
IBM Knowledge Integrator 0.88
IBM Knowledge Structures Group's Relation Detector 0.94
IBM Statistical Text Analytics Group's ACE-model Annotator 0.80
IBM Statistical Text Analytics Group's KDD-model Annotator 0.73
IBM TAF/Talent plus a collection of miscellaneous TFST grammars 0.78
IBM Talent time annotator 0.83
Trustworthiness of Ramazi report

Ramazi

Report Date: 1 April, 2003. Abdul Ramazi is the owner of the Select Gourmet Foods shop in Springfield Mall, Springfield, VA. (Phone number 703-659-2317). First Union National Bank lists Select Gourmet Foods as holding account number 1070173749003. Six checks totaling $35,000 were written from this account in the past four months and are recorded in the bank's records. There is an account at the Pyramid Bank of Cairo, Egypt and the Dubai Bank of the United Arab Emirates. Both of these banks have just been listed as possible conduits in money laundering schemes.


Report Date 27 April, 2003. From a laptop computer captured in Afghanistan, a computer named Sahim Albakri, who fought with the Taliban in 1990-1992, traveled by an alias Bagwant Dhaliwal. On this same computer was found the name Muhammed bin Harazi, who served with the Taliban from 1987-1993. Records on this computer reveal that Muhammed bin Harazi entered the USA in March, 1993 and uses the alias Abdul Ramazi.
Trustworthiness of revised Ramazi report

This text fragment changed from “Neutral” to “Trustworthy” after it was revised by an analyst (the phone number was corrected).

Ramazi

Report Date: 1 April, 2003. Abdul Ramazi, owner of the Select Gourmet Foods shop in Springfield Mall, Springfield, Virginia (phone number 703-659-2317). First Union National Bank lists Select Gourmet Foods as holding account number 1070173749003. Six checks totaling $2,000 have been deposited in this account in the past four months and described as having been drawn on accounts at the Pyramid Bank of Cairo, Egypt and the Central Bank of Dubai, United Arab Emirates. Both of these banks have just been listed as possible conduits in money laundering schemes.


Report Date 27 April, 2003. From a laptop computer captured in Afghanistan it was learned that a Pakistani named Sahim Albakri, who fought with the Taliban in 1990-1992, travels using an Indian passport in the name Bagwant Dhaliwal. On this same computer was found the name Muhammed bin Harazi, who served with the Taliban from 1987-1993. Records on this computer reveal that Muhammed bin Harazi entered the USA in March, 1993 and uses the alias Abdul Ramazi.
Trust Representation, Calculation, & Propagation

- Begin with simple representation of trust
- Present trust coloring view
- Calculation & propagation research options
Knowledge Provenance Elicitation

Provenance information may be essential for users to trust answers.

Data provenance (aka data lineage) is defined and studied in the database literature.

[Buneman et al., ICDT 2001]
[Cui and Widom, VLDB 2001]

Knowledge provenance extends data provenance by adding data derivation provenance information

IWTrust: Trust in Action

Trust can be inferred from a Web of Trust.

IWTrust provides infrastructure for building webs of trust.

The infrastructure includes a trust component responsible for computing trust values for answers. IWTrust is described in [Zaihrayeu, Pinheiro da Silva & McGuinness, iTrust 2005]
Course Type: NON-SPICY-RED-MEAT

"Pairs well with dry red varieties. Medium-bodied wines match especially well."

The local knowledge base particularly recommends the following:

- MOUNTADAM PINOT NOIR
- FORMAN CABERNET SAUVIGNON
- SAUCELITO CANYON ZINFANDEL
- GARY FARRELL MERLOT
- MARIETTA OLD VINES RED
- PAGE MILL WINERY CABERNET SAUVIGNON
- CHIANTI CLASSICO
- MARIETTA PETITE SYRAH
- WHITEHALL LANE CABERNET FRANC
- MARIETTA CABERNET SAUVIGNON
- MARIETTA ZINFANDEL
- KATHY KENNEDY LATERAL

The recommended wines can be found below, along with some comparable selections:

Web Inventory Search

Alternatively, the following varieties include many suitable matches:

- PINOT NOIR
- MERLOT
KSL Wine Agent
Semantic Web Integration Example

Uses emerging web standards to enable smart web applications

Given a meal description
• Deborah’s Specialty

Describe matching wines
• White, Dry, Full bodied…

Retrieve some specific options from web
• Forman Chardonnay from DLM’s cellar, ThreeSteps from wine.com, ….

Info:  http://www.ksl.stanford.edu/people/dlm/webont/wineAgent/
KSL Wine Agent
Semantic Web Integration Technology

- **OWL:** for representing a domain ontology of foods, wines, their properties, and relationships between them
- **JTP theorem prover:** for deriving appropriate pairings
- **Chimaera:** ontology diagnostics and ontology merging
- **DQL/OWL QL:** for querying a knowledge base
- **Inference Web:** for explaining and validating answers (descriptions or instances)
- **Web Services:** for interfacing with vendors
- **Connections to online web agents/information services**
- **Utilities for conducting and caching the above transactions**
Course Type: PASTA-WITH-NON-SPICY-RED-SAUCE

"Pairs well with dry red varieties. Medium-bodied wines featuring moderate flavors match especially well."

The local knowledge base particularly recommends the following:

- MARIETTA PETITE SYRAH
- SAUCELITO CANYON ZINFANDEL
- MARIETTA ZINFANDEL
- PAGE MILL WINERY CABERNET SAUVIGNON
- GARY FARRELL MERLOT
- MARIETTA CABERNET SAUVIGNON
- CHIANTI CLASSICO
- MOUNTADAM PINOT NOIR
- MARIETTA OLD VINES RED
- WHITEHALL LANE CABERNET FRANC

The recommended wines can be found below, along with some comparable selections:

Web Inventory Search

Alternatively, the following varieties include many suitable matches:

- PINOT-NOIR
- MERLOT
- RED-BORDEAUX
NodeSet: Executing(GS)

SupportsTopLevelGoal(x) & IntentionPreconditionMet(x) & TerminationConditionNotMet(x) => Executing(x)

NodeSet: IntentionPrecondition(GS)

TopLevelGoal(y) & Supports(x,y) => SupportsTopLevelGoal(x)

NodeSet: TopLevelGoal(BL)

ParentOf (x,y) & Supports(y,z) => Supports (x,z)

NodeSet: SupportsTopLevelGoal(GS)

NodeSet: Supports(GS, BL)

NodeSet: ParentOf(GS, GA)

NodeSet: ParentOf(GA, BL)

NodeSet: ParentOf(GA, BL)

NodeSet: Supports(BL, BL)

Supports (x,x)

GS: GetSignature
BL: BuyLaptop
GA: GetApproval
Advantages to ICEE Approach

- **Unified framework** for explaining task execution and deductive reasoning.
- Architecture for reuse among many task execution systems.
- **Introspective predicates** and software wrapper that extract explanation-relevant information from task reasoner.
- Reusable action schema for representing task reasoning.
- A version of InferenceWeb for generating formal justifications.
Task Explanation - sample prototype for Command Post of the Future (Stanford, SRI)

- Explanations of end-to-end task processing
- Initial dialog (limited follow-up capability)
- PML representation for complex tasks
- Design in process for explaining learned task modifications

Initial explanation, with links indicating follow-up queries (why haven’t you completed xxx) and alternate strategies.