Life cycle semantics in Earth and space sciences – what's worked (and not) and where are we… a *decade* in…

Ontolog Forum (Earthcube)

Sep. 6, 2012

Peter Fox (RPI) pfox@cs.rpi.edu
Tetherless World Constellation

Background

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 means (instruments, models, analysis) 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. And, it is almost always created in a manner to facilitate its generation *not its* use.

And… there exist(ed) significant levels of semantic heterogeneity, large-scale data, complex data types, legacy systems, inflexible and unsustainable implementation technology…

Origins …

- In 2000-2001 the need for capturing and preserving knowledge in science data became very clear but the barriers were high
- In 2004 we started a virtual observatory project based on semantic technologies
- Use case driven – in solar and solar-terrestrial physics with an emphasis on instrument-based measurements and real data pipelines; we needed implementations
- We knew we also needed integration and provenance (but that came later)
- We aimed to push semantics into our systems to build new ‘prototypes’ but we ‘failed’ ;-)

In 2004

- 2004 – OWL was a W3 recommendation!!
- Protégé 2.x and the Protégé-Java-OWL API
- SWOOP was a viable editor
- Jena and the Jena API were in good shape
- Pellet worked
- SPARQL was still a twinkle in the RDF working group’s eye
- Semantics were still the realm of computer scientists – luckily we had one of the best

Design and Development

- We made a conscious decision only to develop ontologies that were required to answer specific use cases
- We made a conscious effort to use whatever ontologies were available**
- We were pretty sure that rules would be needed
- We ignored query

Use Case example

- Plot the neutral temperature from the Millstone-Hill Fabry Perot, operating in the non-vertical mode during January 2000 as a time series.
- **Plot** the **neutral temperature** *from* the **Millstone-Hill Fabry Perot**, **operating** *in the* **non-vertical mode** *during* **January 2000** *as a* **time series**.
- Objects:
  - Neutral temperature is a (temperature is a) parameter
  - Millstone Hill is a (ground-based observatory is a) observatory
  - Fabry-Perot is a interferometer is a optical instrument is a instrument
  - Non-vertical mode is a instrument operating mode
  - January 2000 is a date-time range
  - Time is a independent variable/ coordinate
  - Time series is a data plot is a data product

Knowledge representation

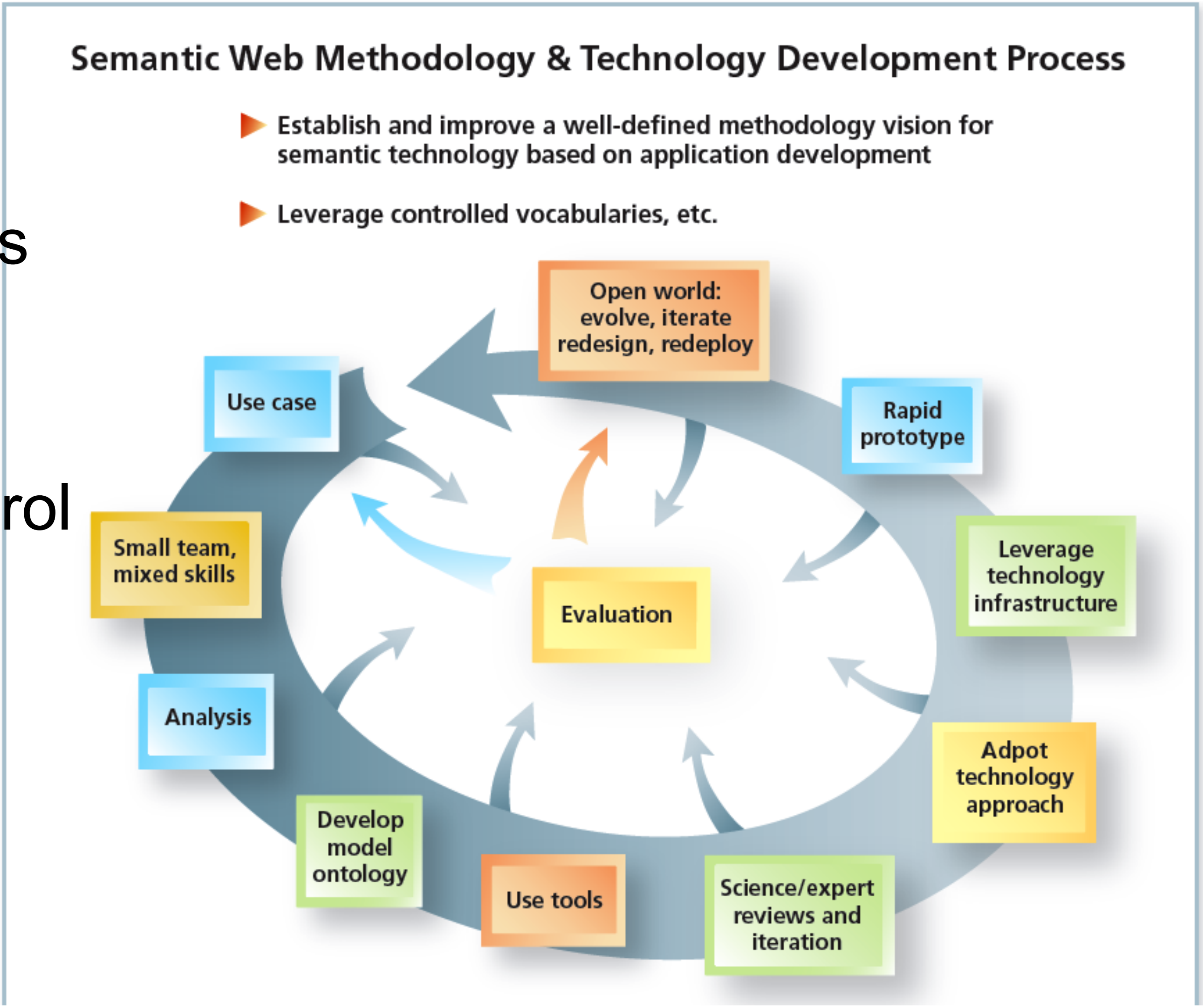
- Statements as triples: {subject-predicate-object}
  - interferometer is-a optical instrument
  - Fabry-Perot is-a interferometer
  - Optical instrument has focal length
  - Optical instrument is-a instrument
  - Instrument has instrument operating mode
  - Instrument has measured parameter
  - Instrument operating mode has measured parameter
  - NeutralTemperature is-a temperature
  - Temperature is-a parameter
- A query*: select all optical instruments which have operating mode *vertical*
- An inference: infer operating modes for a Fabry-Perot Interferometer which measures neutral temperature
- ISWC paper award 2006, IAAI best paper (2007), Fox et al. 2009 in Computers and Geosciences.

Semantic Web Benefits

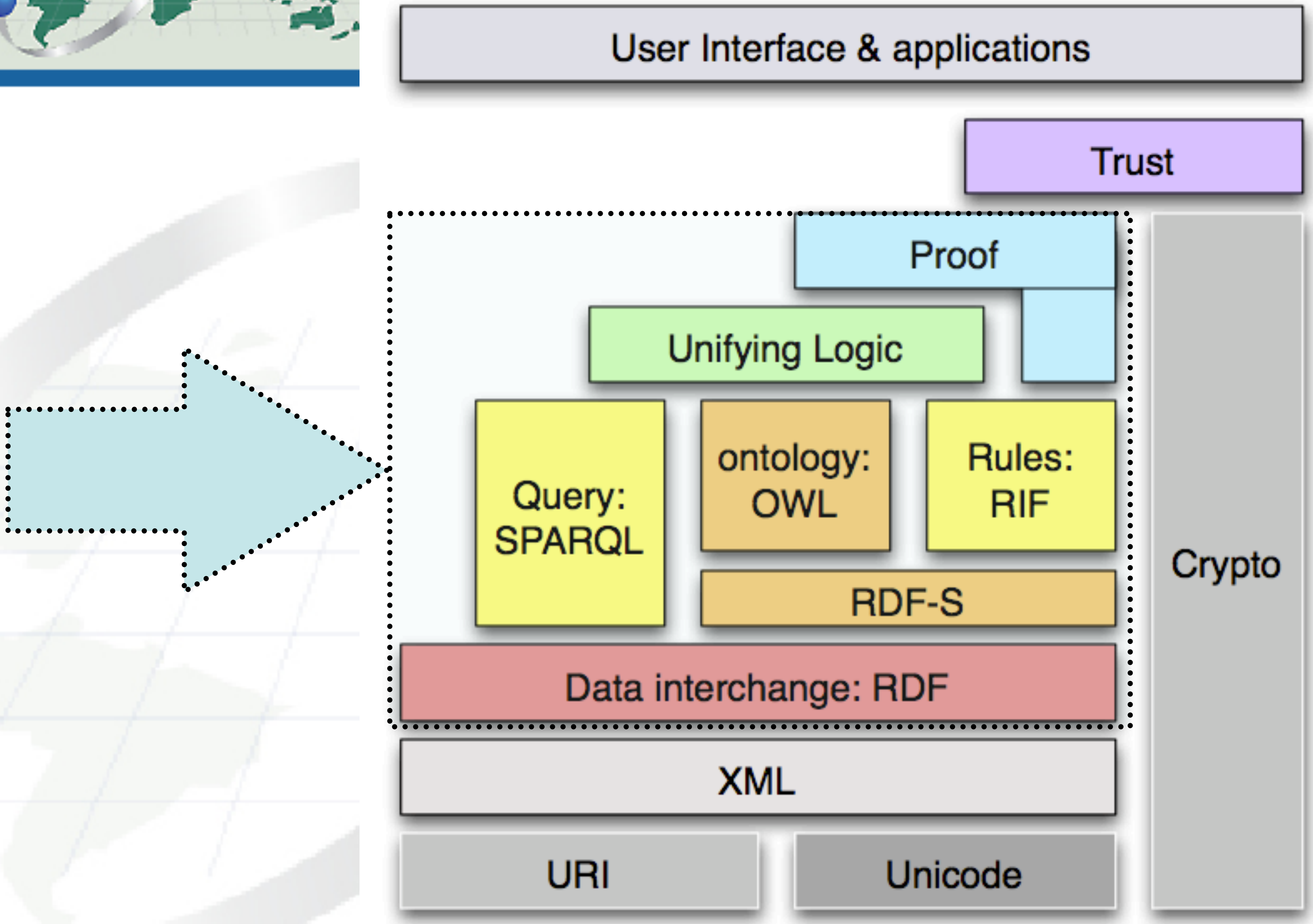
- Unified/ abstracted query workflow: Parameters, Instruments, Date-Time
- Decreased input requirements for query: in one case reducing the number of selections from **eight** to **three**
- Generates only syntactically correct queries: which was not always insurable in previous implementations without semantics
- Semantic query support: by using background ontologies and a reasoner, our application has the opportunity to only expose coherent query (portal and services)
- Semantic integration: in the past users had to remember (and maintain codes) to account for numerous different ways to combine and plot the data whereas now semantic mediation provides the level of sensible data integration required, and exposed as smart web services
  - understanding of coordinate systems, relationships, data synthesis, transformations.
  - returns independent variables and related parameters
- A broader range of potential users (PhD scientists, students, professional research associates and those from outside the fields)

Semantics - Modern informatics enables a new scale-free** framework approach

- Use cases
- Stakeholders
- Distributed authority
- Access control
- Ontologies
- Maintaining Identity



Semantic Web Layers



http://www.w3.org/2003/Talks/1023-iswc-tbl/slide26-0.html, http://flickr.com/photos/pshab/291147522/

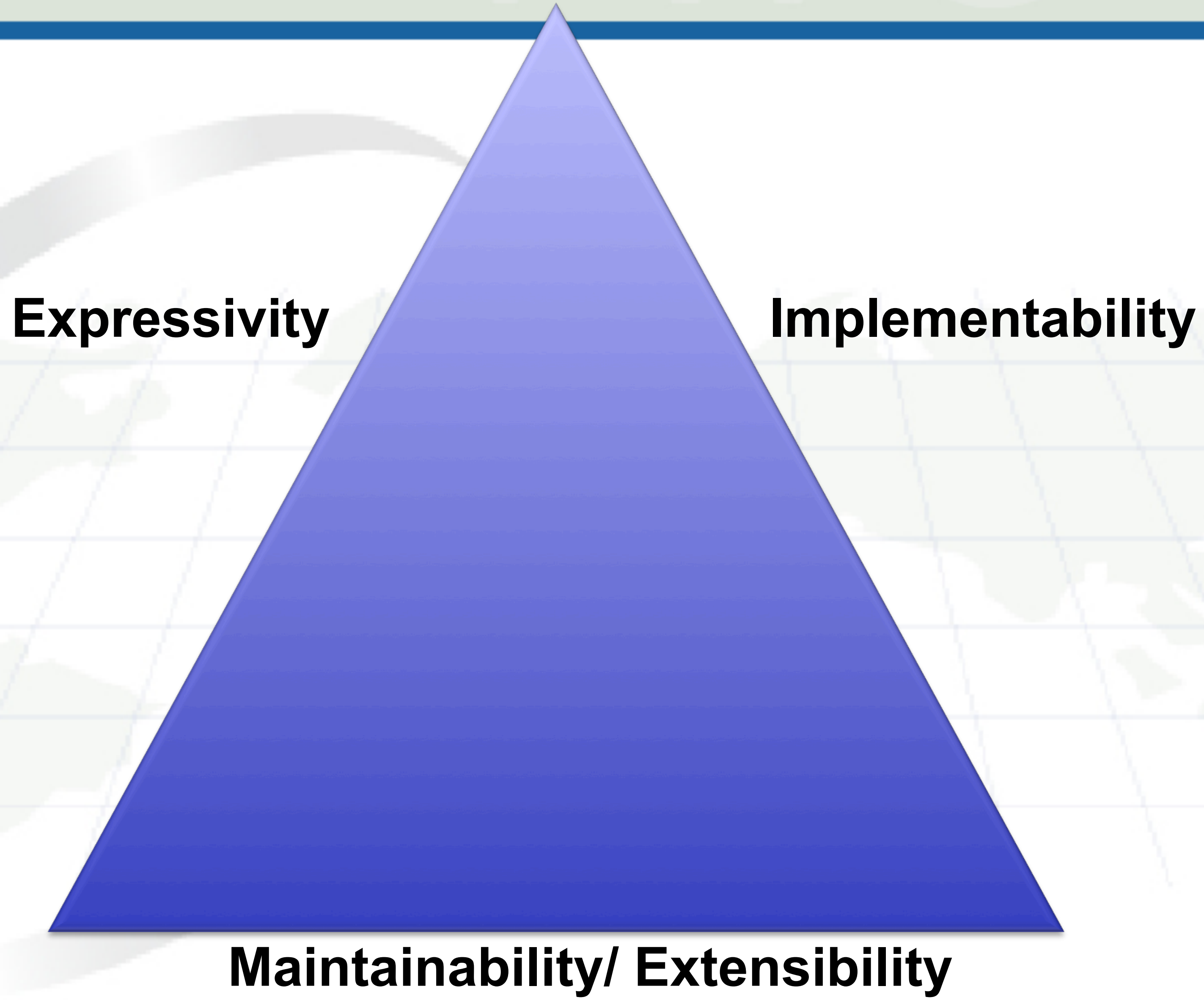
Developing ontologies (c. 2005)

- Use cases and small team (7-8; 2-3 domain/ data experts, 2 knowledge experts, 1 software engineer, 1 facilitator, 1 scribe)
- Identify classes and minimal properties (leverage controlled vocab.)
  - Start with narrower terms, generalize when needed or possible
  - Adopt a suitable conceptual decomposition (e.g. SWEET)
  - Import modules when concepts are orthogonal
  - Add service classes and properties where needed
- Review, vet, publish
- Only code them (in RDF or OWL) when needed (CMAP, …)
- Ontologies: small and modular

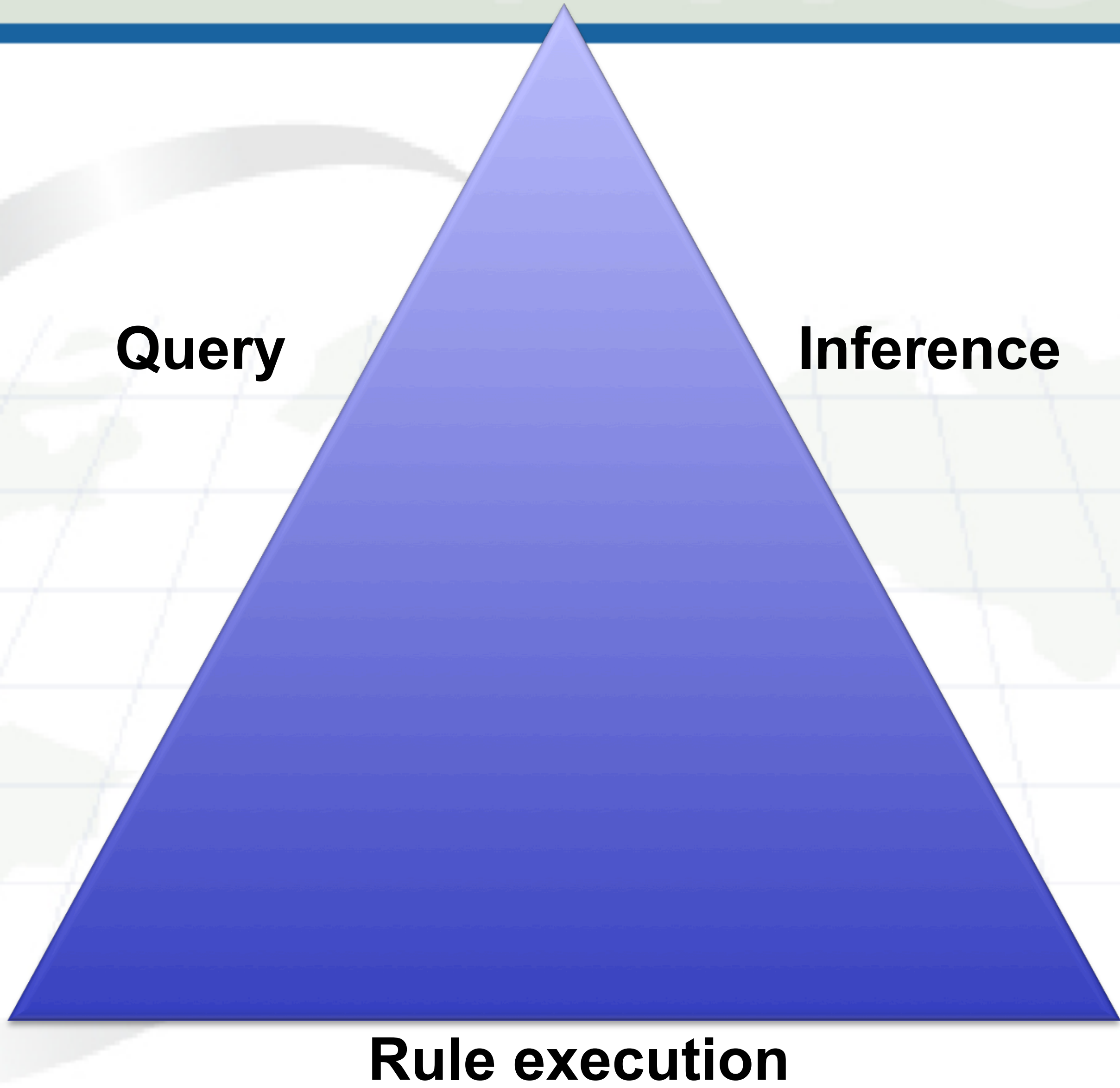
Semantics between 2004 and 2009

- Ontologies were needed for data integration and provenance and mediation for data mining
- Protégé 3.x and then 4.0 came out
- SWOOP development was interrupted
- Cmap added OWL predicate support*
- SPARQL became a recommendation
- Triple stores exploded in use and capability
- Linked Open Data started to take off
- Pellet 2.0 came out
- We invaded OWLED 2006, 2007, 2009, (2010)

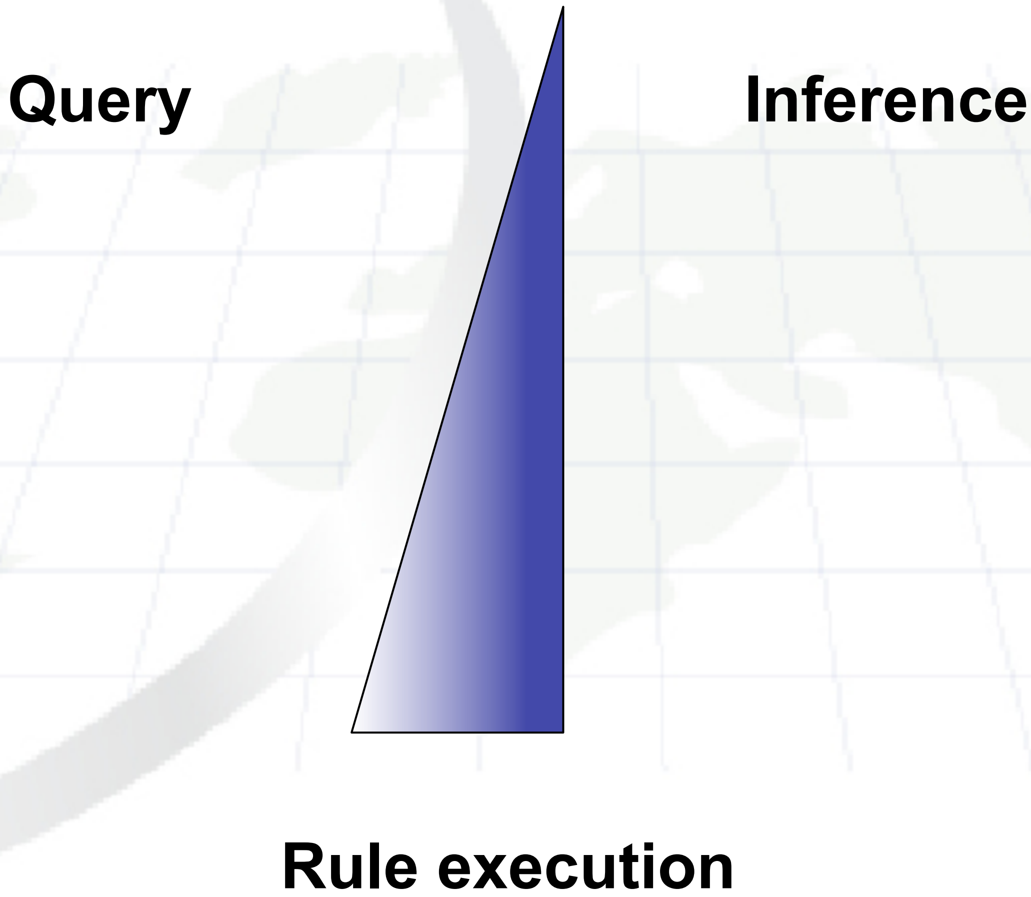
Working with knowledge



Working with knowledge



Or it may be this …



Expressivity/ Implementation

Declarative

Procedural

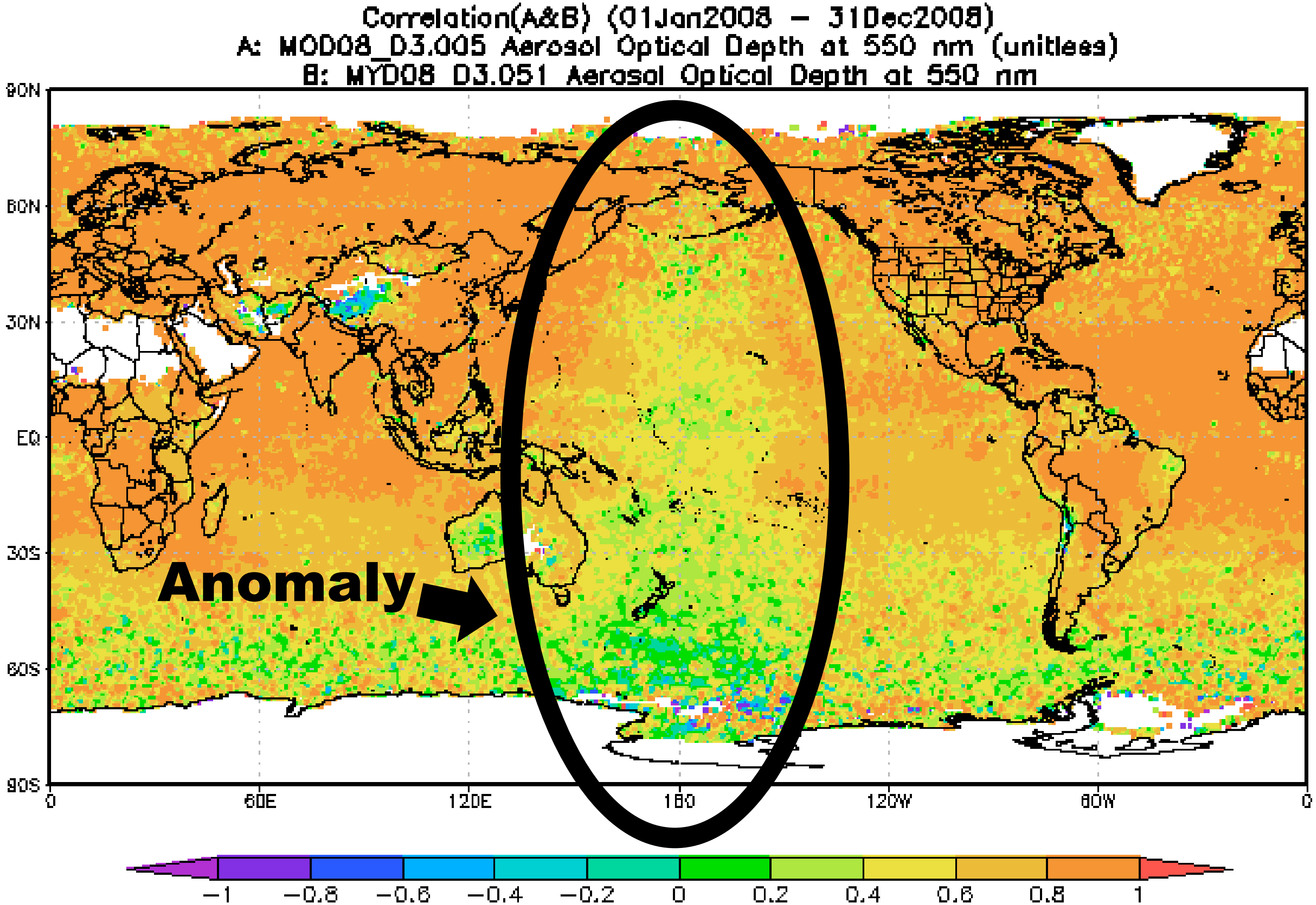
Linked open data
URI/http/RDF *

Ontology encoded

Semantics between 2009 and 2012

- Semantic data framework (SeSF)
- Substantial knowledge provenance work
- Data quality, uncertainty and bias representations and applications (oh, these are in production at NASA)
- Multi-sensor advisor:

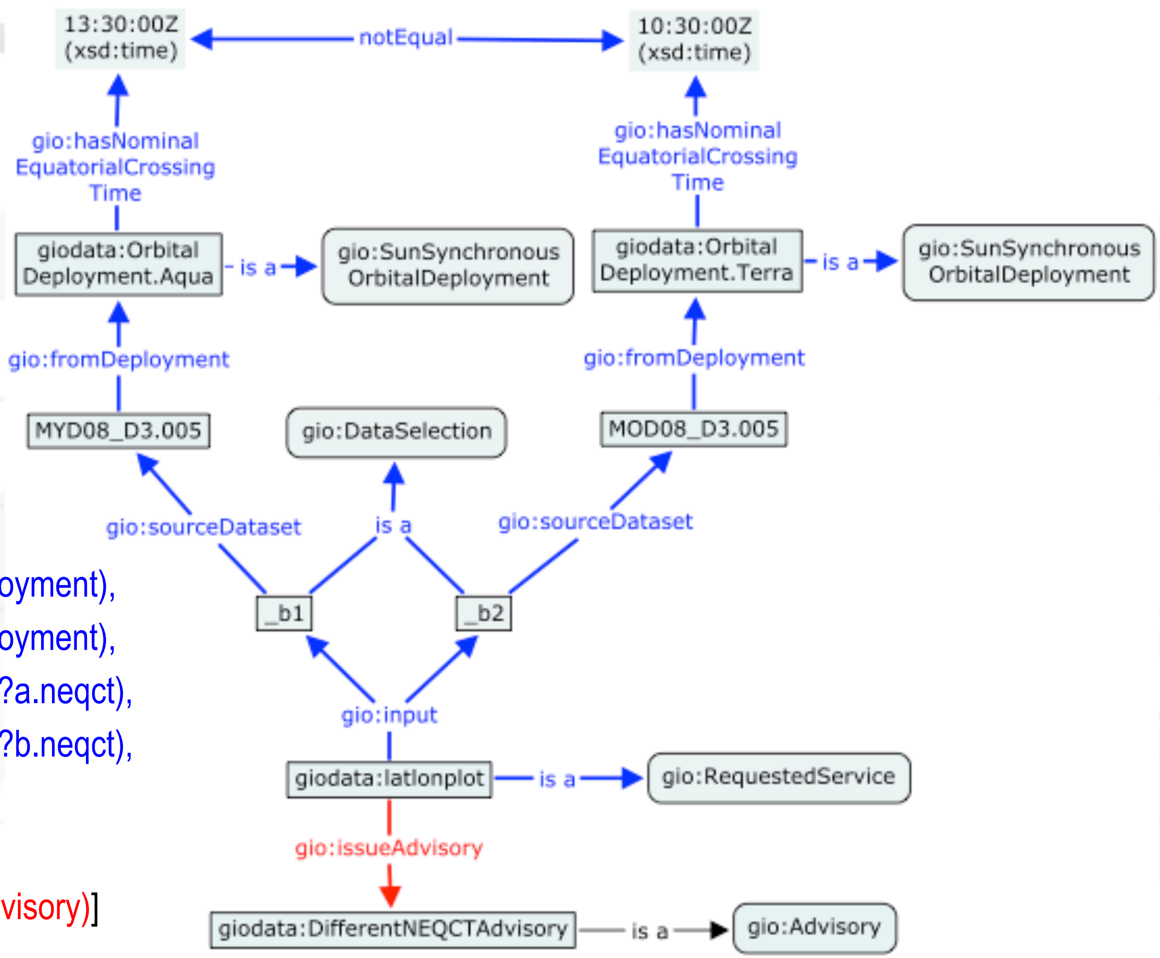
Anomaly Example: South Pacific Anomaly



MODIS Level 3 dataday definition leads to artifact in correlation

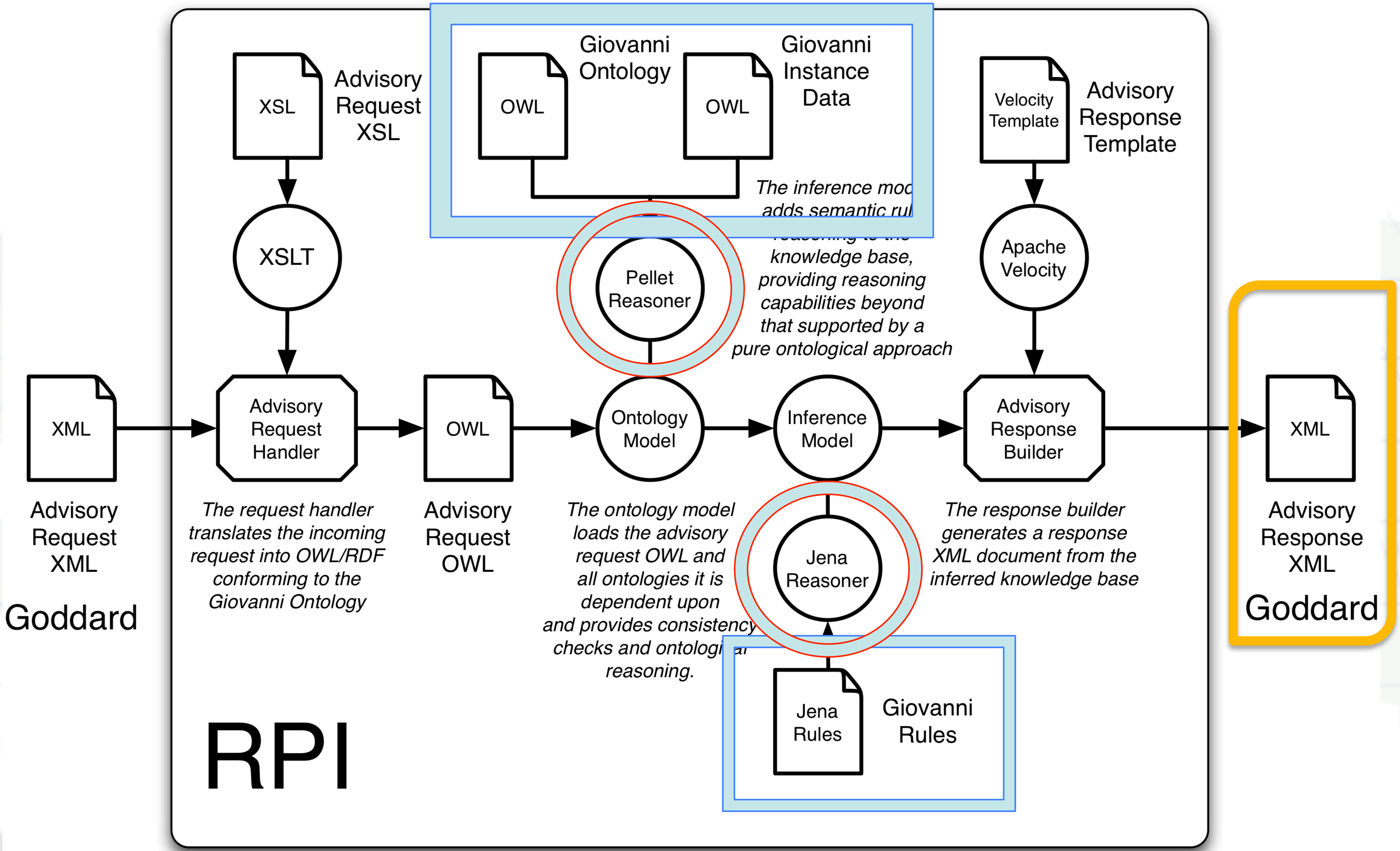
RuleSet Development

```
[DiffNEQCT:
(?s rdf:type gio:RequestedService),
(?s gio:input ?a),
(?a rdf:type gio:DataSelection),
(?s gio:input ?b),
(?b rdf:type gio:DataSelection),
(?a gio:sourceDataset ?a.ds),
(?b gio:sourceDataset ?b.ds),
(?a.ds gio:fromDeployment ?a.dply),
(?b.ds gio:fromDeployment ?b.dply),
(?a.dply rdf:type gio:SunSynchronousOrbitalDeployment),
(?b.dply rdf:type gio:SunSynchronousOrbitalDeployment),
(?a.dply gio:hasNominalEquatorialCrossingTime ?a.neqct),
(?b.dply gio:hasNominalEquatorialCrossingTime ?b.neqct),
notEqual(?a.neqct, ?b.neqct)
->
(?s gio:issueAdvisory giodata:DifferentNEQCTAdvisory)]
```



Multi-sensor Data Synergy Advisor (NASA), Leptoukh, Lynnes, Zednik, et al.

Semantic Advisor Architecture

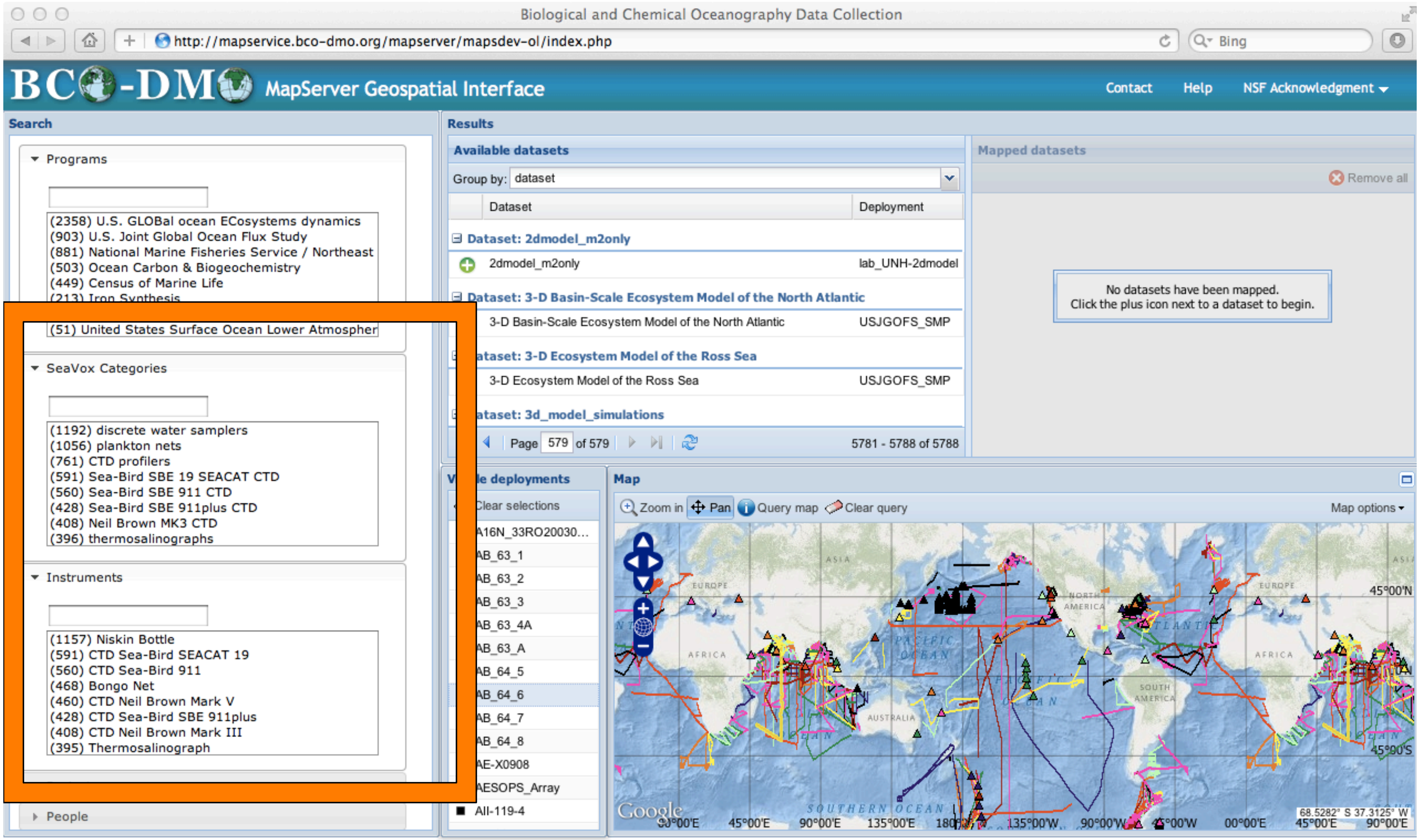


Multi-sensor Data Synergy Advisor (NASA), Leptoukh, Lynnes, Zednik, et al.

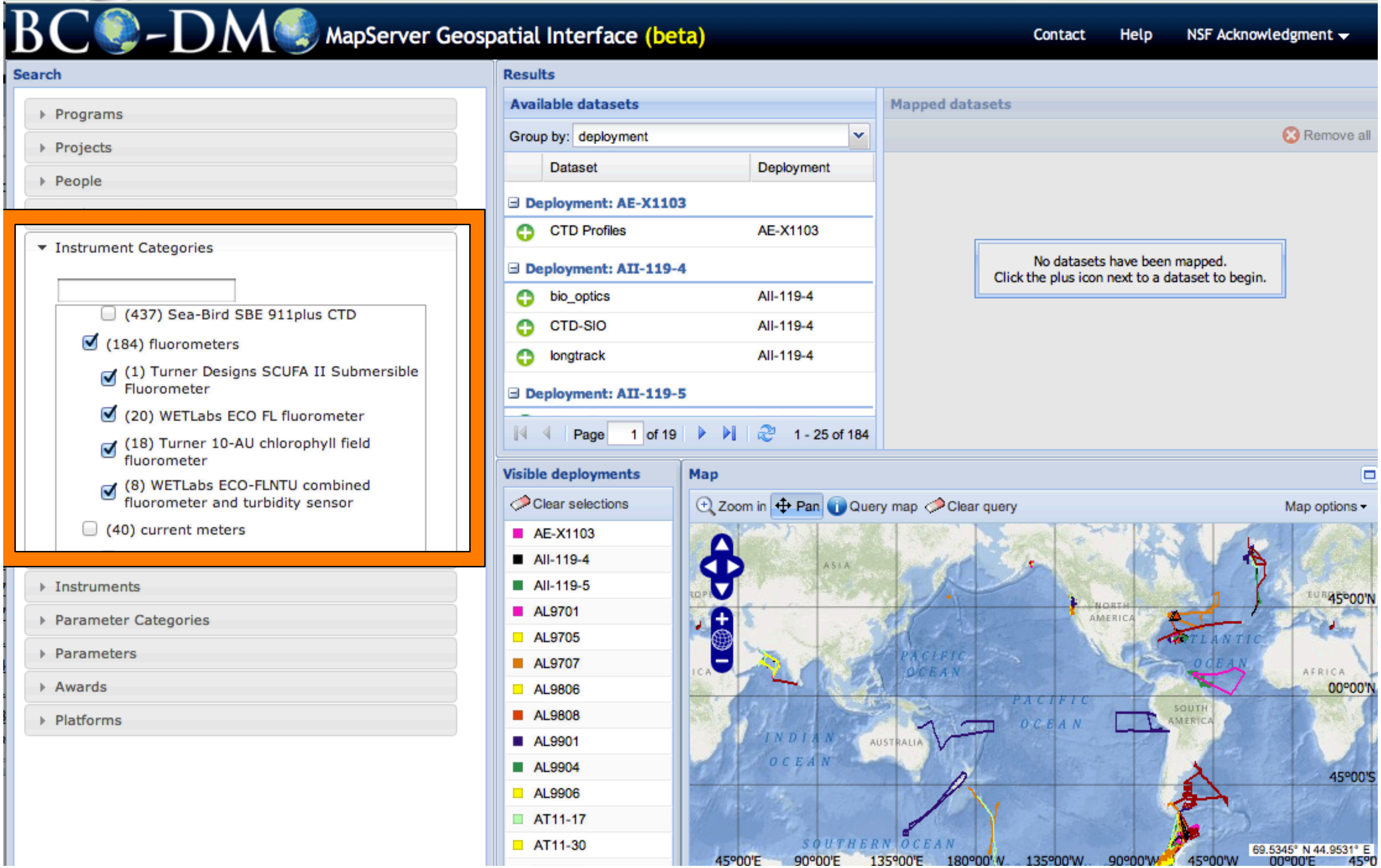
Semantics between 2009 and 2012

- Semantic data framework (SeSF)
- Substantial knowledge provenance work
- Data quality, uncertainty and bias representations and applications (oh, these are in production at NASA)
- Multi-sensor advisor
- Applications:
  – Sea Ice, Carbon Observatory, Integrated Ecosystem Assessments, globalchange.gov, ocean.data.gov, energy.data.gov ….

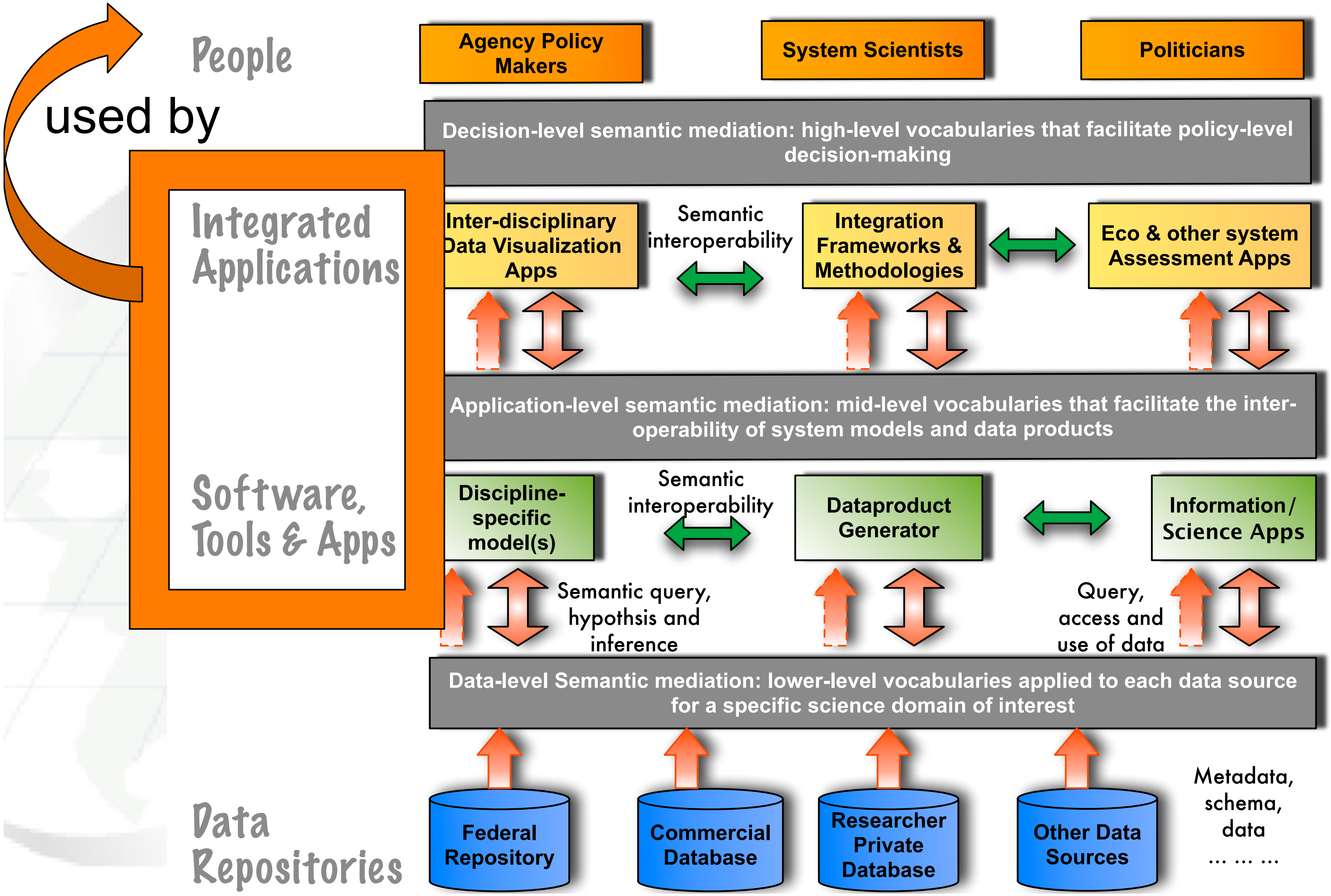
Respect and Mediation … how



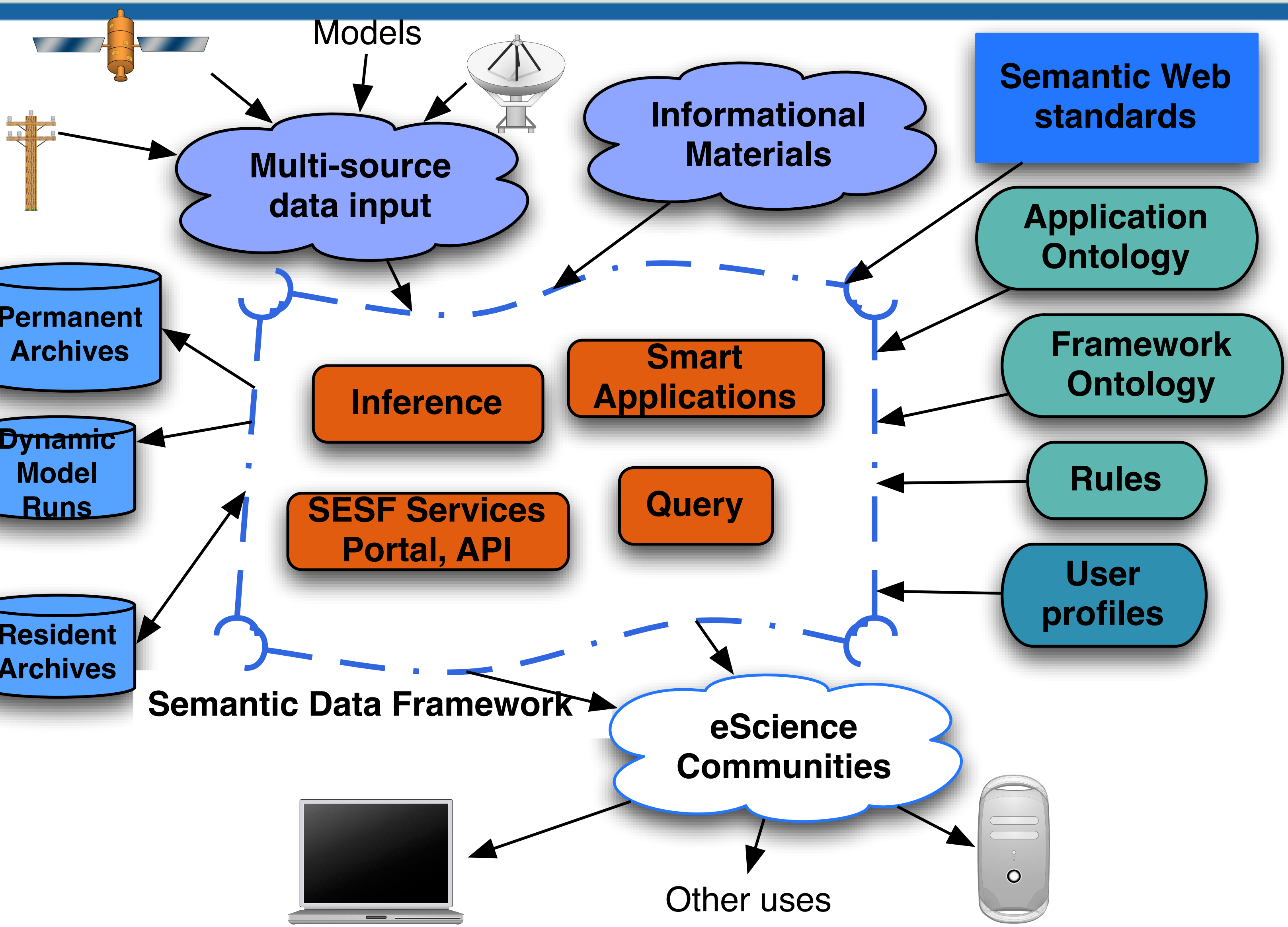
Discovering new data



Core and Framework Semantics - Multi-tiered interoperability



High-level architecture



Summary

- In 2004 we set out to build a prototype and ended up with a production semantic data framework
  - Languages and tools served us well
- Even with modest expressivity we challenged the tools of the time and made many compromises
- All along the way, we evaluated our ontology developments and implementations to gauge the benefits of semantics
- Maintainability, esp. modularization is driving new expressivity needs
- Xinformatics and a repeatable methodology is the key (information models) - we continue to need to bridge computer science and application communities (“It’s the language stupid”, i.e. semantics)

Contact

- pfox@cs.rpi.edu
- http://tw.rpi.edu
- @taswegian

- See also wiki.esipfed.org (Semantic Web Cluster)