

# Grids and Semantics: e-Research in Production

Grids: Earth System Grid  
Semantic Data Frameworks/ Virtual Observatories  
Discussion of e-Research

Peter Fox

High Altitude Observatory, NCAR

 OPeNDAP



With thanks to the ESG and VSTO teams  
Funding from DoE/SciDAC, NSF/OCI

# Background - Collaboration

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

From models too...


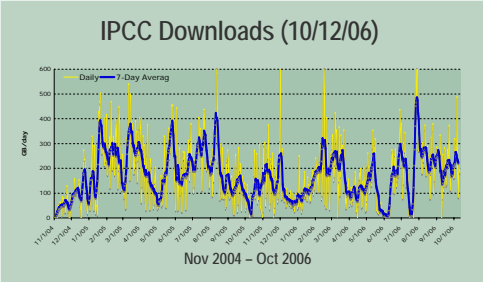
And... there exist(ed) significant levels of semantic heterogeneity, large-scale data, complex data types, legacy systems, inflexible and unsustainable implementation technology...







# ESG facts and figures

| Main ESG Portal   | IPCC AR4 ESG Portal  |
|---|--|
| <p>130 TB of data at four locations</p> <ul style="list-style-type: none"><li>• 840,331 files</li><li>• Includes the past 6 years of joint DOE/NSF climate modeling experiments</li></ul> | <p>35 TB of data at one location</p> <ul style="list-style-type: none"><li>• 77,400 files</li><li>• Generated by a modeling campaign coordinated by the Intergovernmental Panel on Climate Change</li><li>• Model data from 11 countries</li></ul> |
| <p>3,200 registered users</p>   | <p>1,245 registered analysis projects</p>  |
| <p>Downloads to date</p> <ul style="list-style-type: none"><li>• 25 TB</li><li>• 91,000 files</li></ul>   | <p>Downloads to date</p> <ul style="list-style-type: none"><li>• 245 TB</li><li>• 914,400 files</li><li>• 500 GB/day (average)</li></ul>   |
|    |    |
|   | <p>&gt; 300 scientific papers published to date based on analysis of IPCC AR4 data</p>   |

Worldwide ESG user base



# ESG II: experience

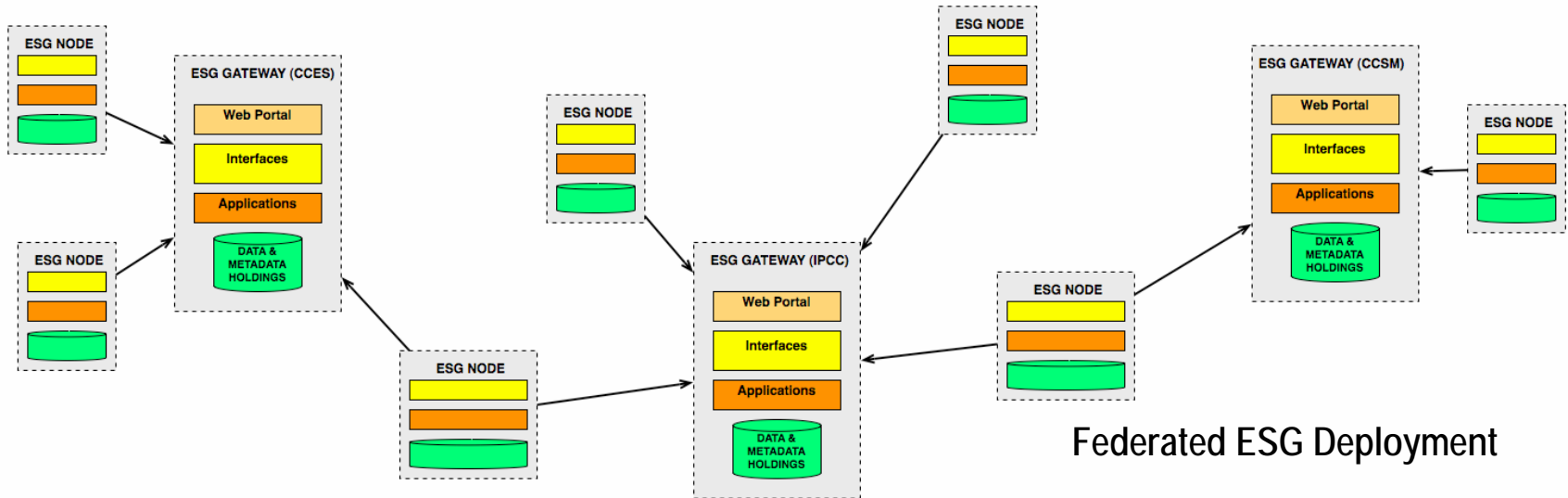
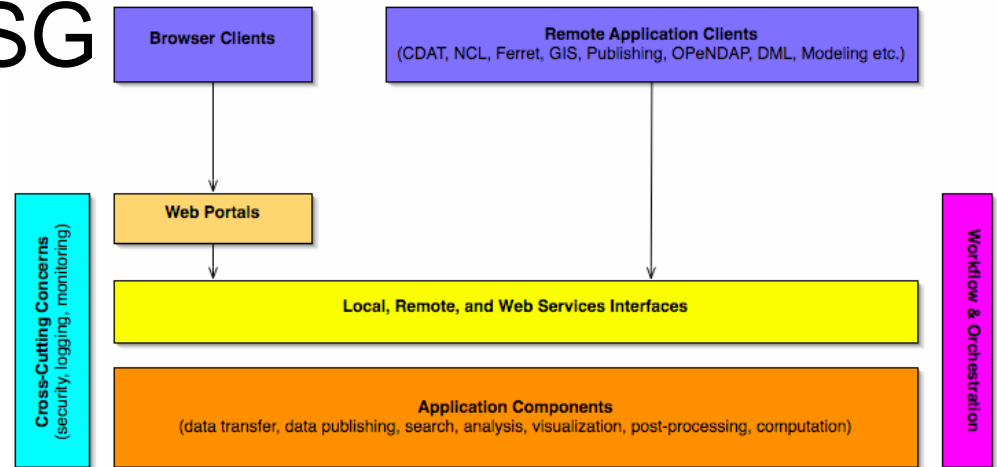
- ESG is a highly collaborative effort - in reality this happened in groups of 2-4 (out of 40)
- Payoffs of this distributed collaborative infrastructure have included:
  - ✓ Distributed data-sharing, RLS works! SRM/HRM work! OPeNDAP-g works!
  - ✓ Simplified data discovery of climate data, the work on metadata paid off! Scalability?
  - ✓ Large-scale climate data processing and analysis via highly integrated portal
  - ✓ Increased collaboration among climate research scientists, people use it!
  - ✓ Aid in climate assessments and estimates of future climate variability and trends, IPCC!
- ↓ Authentication and authorization have been a significant challenge
  - ☒ GSI to CAS (failed)
  - ✓ MyProxy, now moving to Shibboleth, etc.
  - ✓ SAML is working for multi-file batch transfer
- ↓ Transport - GridFTP versus HTTP
  - ☒ Server to server
  - ☒ Very good performance
  - ☒ Clients are not as capable due to 'weight' of globus, revert to HTTP
  - ☒ ESG-OPeNDAP collaboration has had huge benefits to the community
- ↓ Service monitoring
  - ☒ to support the distributed collaborative infrastructure
  - ☒ need lots of monitoring for all services to really make a production environment work
- Many Globus services not used (GRIS, MDS, GIIS, ... )



# Architecture of the next-generation ESG

- Petascale data archives
- Broader geographical distribution of archives
  - across the United States
  - around the world
- Easier federation of sites
- Increased flexibility and robustness

## Second Generation ESG Architecture



Federated ESG Deployment

# Virtual Observatories

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 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: i.e. *appear to be local and integrated***

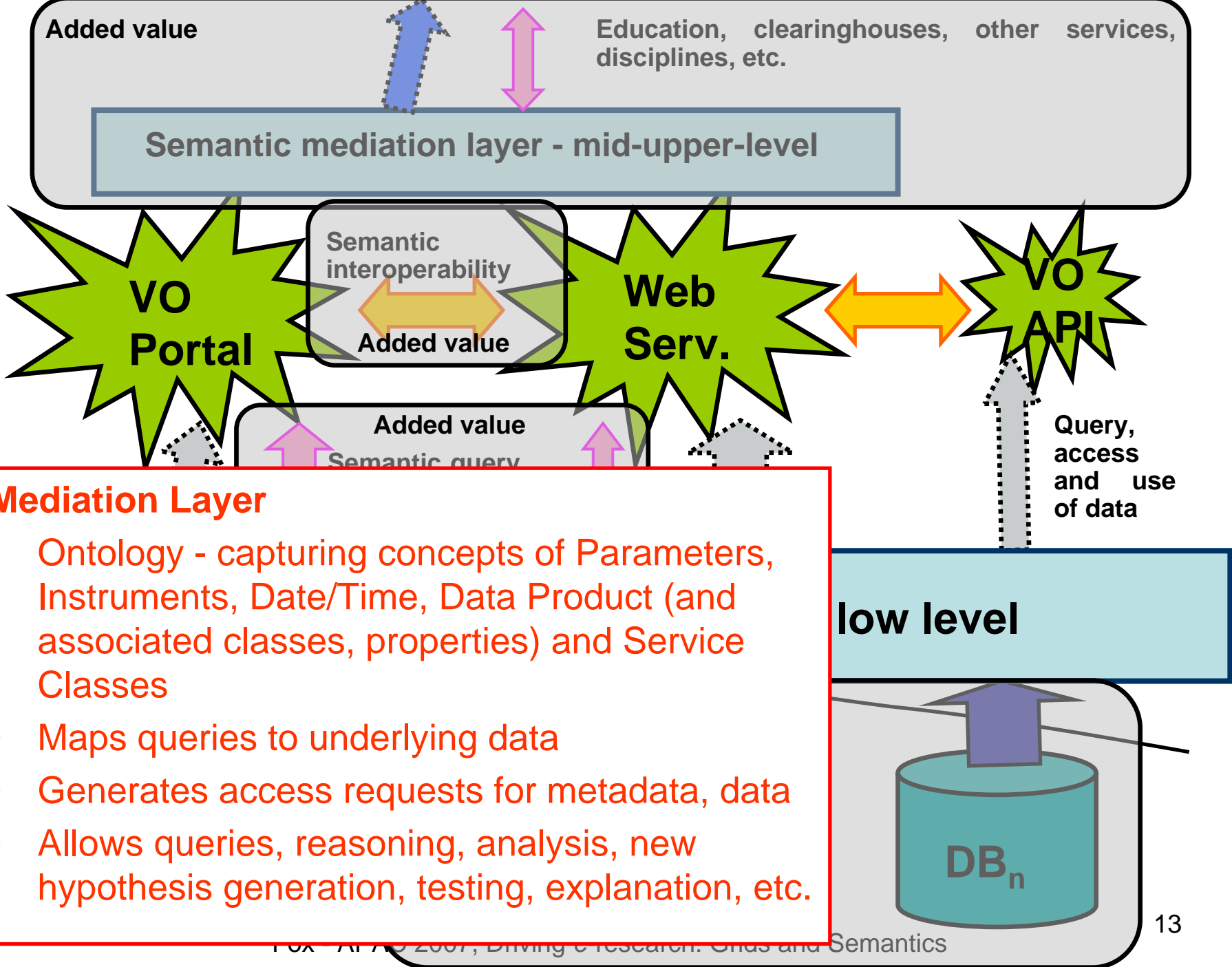
Likely to provide controlled vocabularies that may be used for interoperation in appropriate domains along with database interfaces for access and storage and “smart” tools for evolution and maintenance.

# Science and technical use cases

Find data which represents the state of the neutral atmosphere anywhere above 100km and toward the arctic circle (above 45N) at any time of **high geomagnetic activity**.

- Extract information from the use case - encode knowledge
- Translate this into a complete query for data - inference and integration of data from instruments, indices and models

Provide semantically-enabled, smart data query services via a SOAP web for the Virtual Ionosphere-Thermosphere-Mesosphere Observatory that retrieve data, filtered by constraints on Instrument, Date-Time, and Parameter in any order and with constraints included in any combination.





## Virtual Solar Terrestrial Observatory

[Home](#)
[Data](#)
[Communities](#)
[About Us](#)
[Login](#)
[Start by Instrument](#)
[Start by Dates](#)
[Start by Parameter](#)

### Data

Semantic filtering by domain or instrument hierarchy

#### Data Request Summary

1. Instrument:

2. Start Date:  
Stop Date:

3. Parameters:

#### Input Step 1 of 3: Choose Instrument

Please select an instrument

You may filter the instruments selection by one of the following criteria:

Filter by Physical Domain:  -OR- filter by Instrument Type:

Show Instrument Code



- [?] Instrument:
- OpticalInstrument > Photometer > Chromospheric Helium Imaging Photometer [?]
  - OpticalInstrument > Photometer > MK3-K Coronameter [?]
  - OpticalInstrument > Photometer > MK4-K Coronameter [?]
  - OpticalInstrument > Photometer > H-alpha prominence and solar disk monitor [?]
  - OpticalInstrument > Photometer > MultiChannelPhotometer > Poker Flat 4 Channel Photometer [?]
  - OpticalInstrument > Photometer > MultiChannelPhotometer > Fort Yukon Alaska 4 Channel Photometer [?]
  - OpticalInstrument > Spectrometer > SpectroPhotometer > Davis Antarctica Spectrometer [?]



## VSTO Guided Workflow: Start by Parameter

### Data Request Summary

1. Parameter: ElectronTemperature [?]

2. Start Date: 2000/03/14

Stop Date: 2000/03/18

3. Instrument: Irkutsk Russia I.S. Radar [?]

### Available Output

The following data products match the current selection:

Data Files: ▶ TAB [?] ▶ FLAT [?] ▶ INFO [?] ▶ DAS [?] ▶ DDS [?] ▶ OPeNDAP [?] ▶ STREAM [?] ▶ IDL [?]

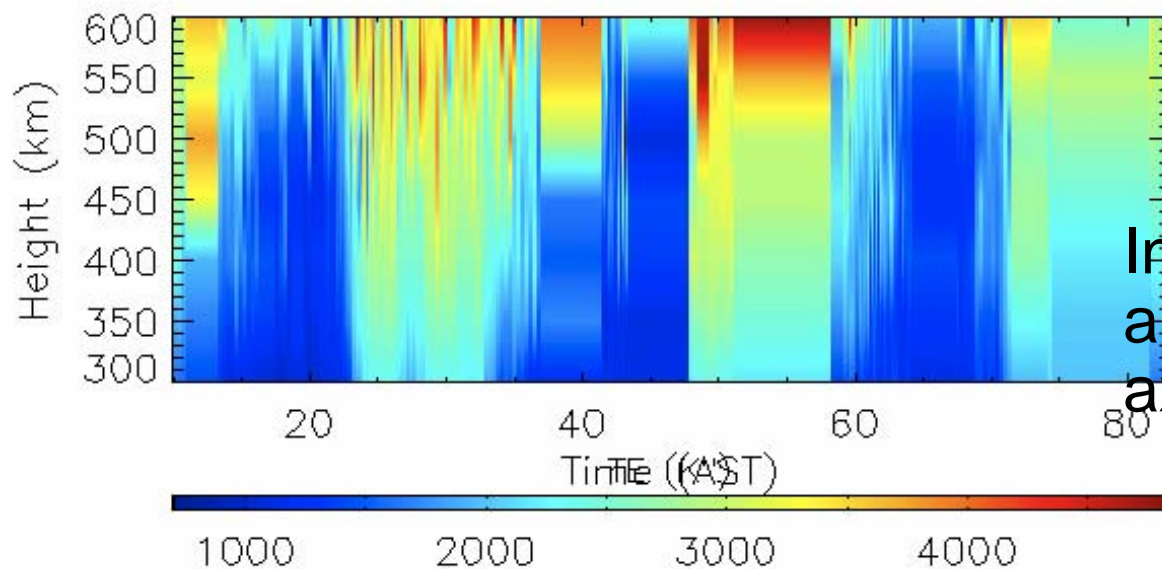
Data Plots: ▶ Height vs Time [?]

### Change Input

Click on the Back button to change your data selection, or Cancel to end the workflow

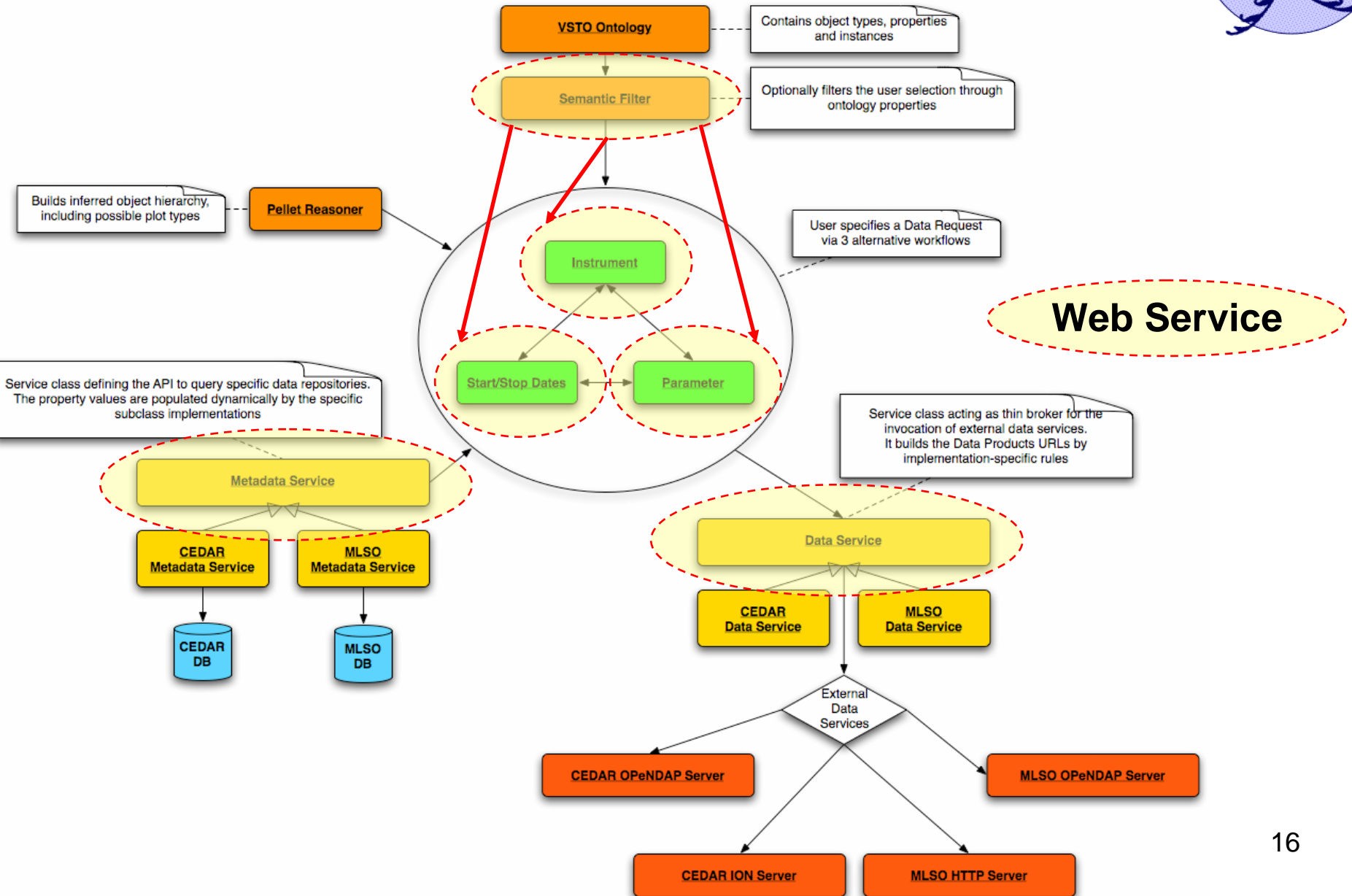
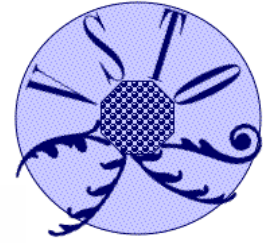
< Back

Cancel



Inferred plot type  
and return required  
axes data

# VSTO - semantics and ontologies in an operational environment: [vsto.hao.ucar.edu](http://vsto.hao.ucar.edu), [www.vsto.org](http://www.vsto.org)



# Building science cyberinfrastructure -> informatics

- Use case, then requirements
- If collaboration (rather than individual use) is a primary scenario, then write a use case for it!!
- **Then** design, derive architecture and choose technology components
- Build something that works for users from the start
- Get your funding source and community to commit to an evolving architecture
- If you choose a major framework technology, e.g. Globus, OPeNDAP, THREDDS, **peer**-partner with them

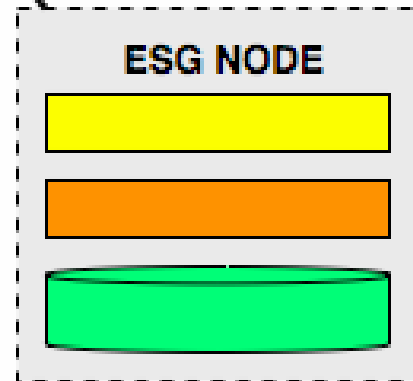
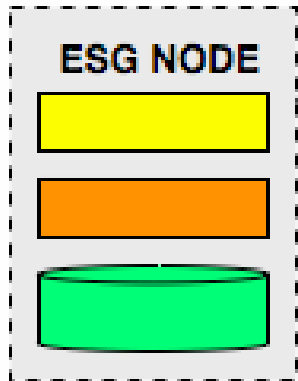
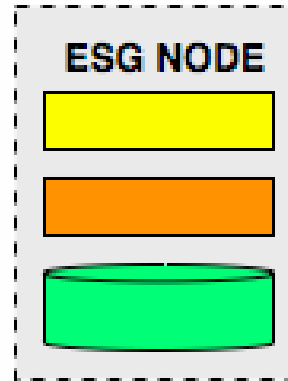
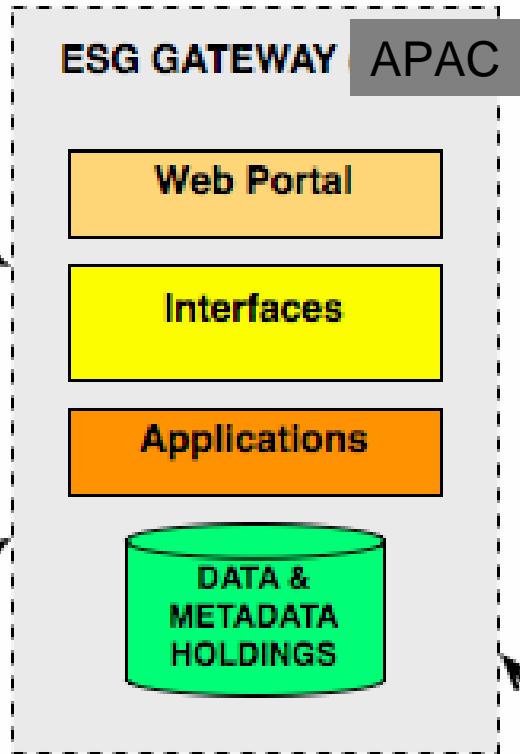
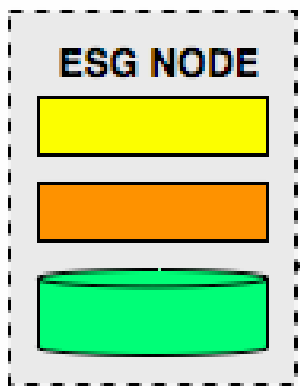
# e-Research opportunities

- Semantics/Ontologies
  - Is an e-research collaboration activity
  - Develop use cases for mediating services to enable, for e.g., script-level access using common terminology
- Virtual data product generation, caching, publishing,
  - Data-mover (LBL)
  - OPeNDAP installation and development
- APAC gateway into ESG (will require)
  - Address the identity exchange requirement
  - Negotiate/federate roles and responsibilities
  - Services to provide a vortal for their communities
- Allow for reciprocal discovery of data and services using semantics (agree on the 3-level approach and the boundaries)
- Participate in the U.S. GeoCollaboratory initiative

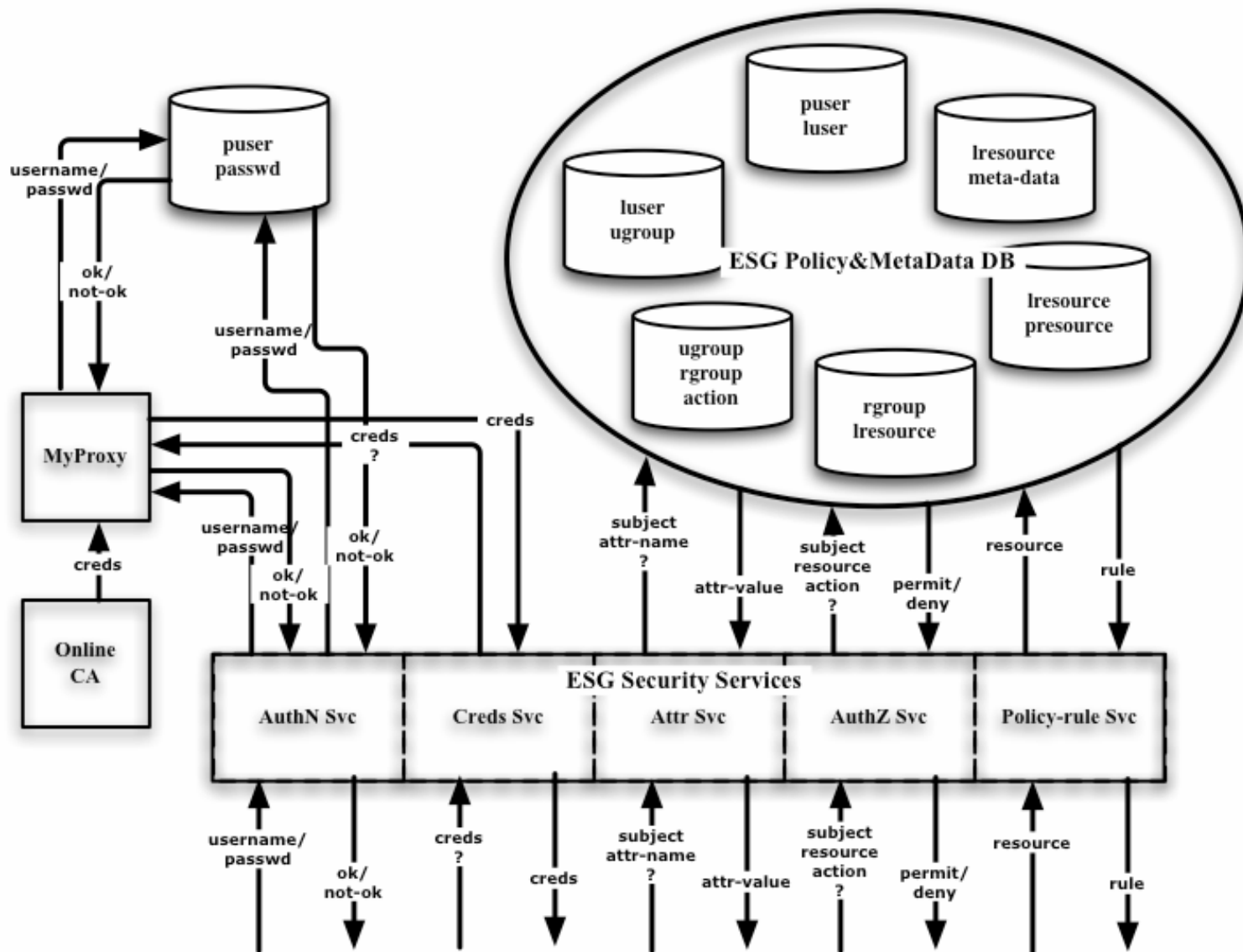


# e-Research opportunities

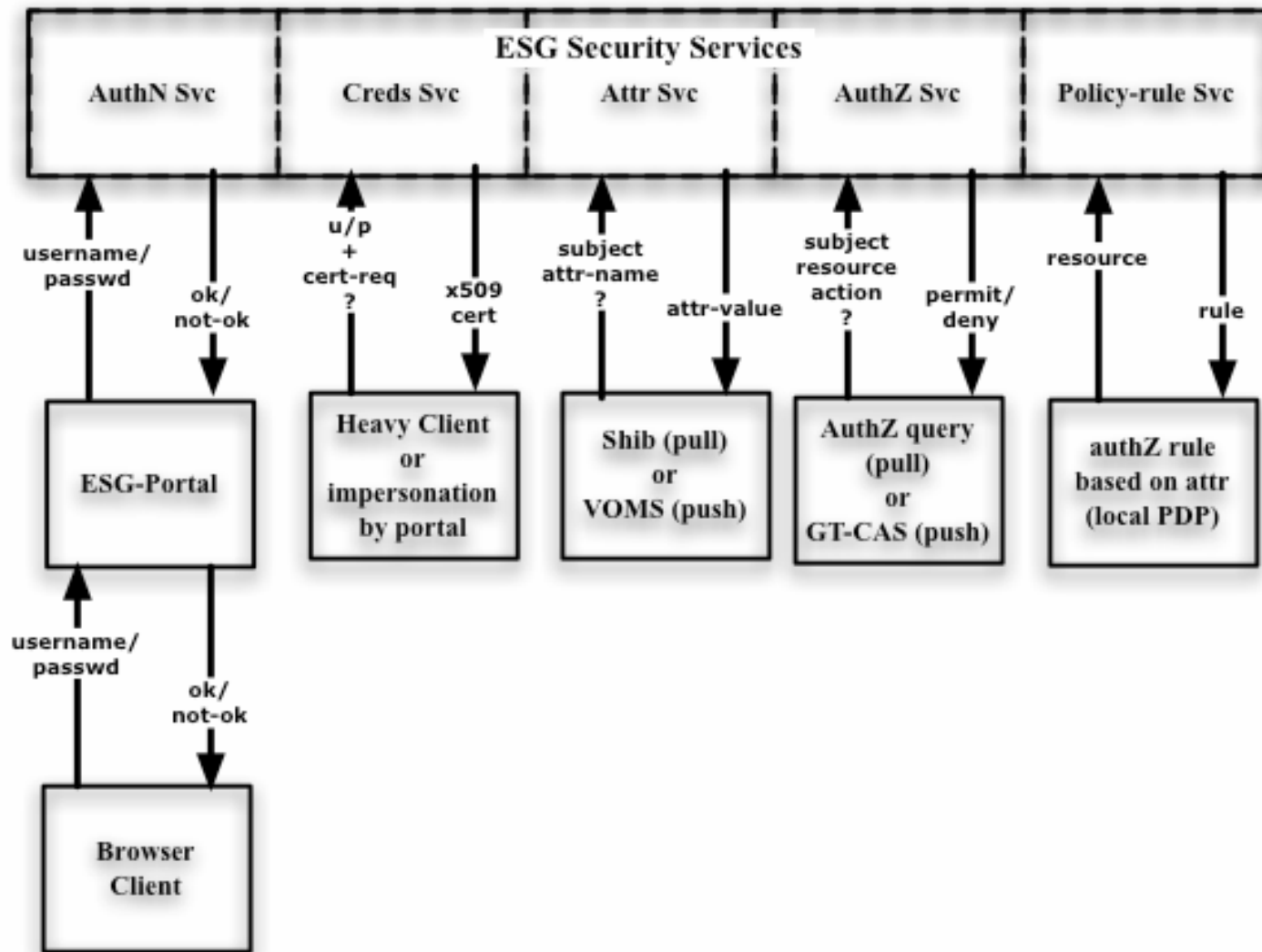
- Semantics/Ontologies
  - Is an e-research collaboration activity
  - Develop use cases for mediating services to enable, for e.g., script-level access using common terminology
- Virtual data product generation, caching, publishing,
  - Data-mover (LBL)
  - **OPeNDAP** installation and development - plumbing the net
- APAC gateway into ESG (will require)
  - Address the identity exchange requirement
  - Negotiate/federate roles and responsibilities
  - Services to provide a vortal for their communities
- Allow for reciprocal discovery of data and services using semantics (agree on the 3-level approach and the boundaries)
- Participate in the U.S. GeoCollaboratory initiative



# ESG Security Services



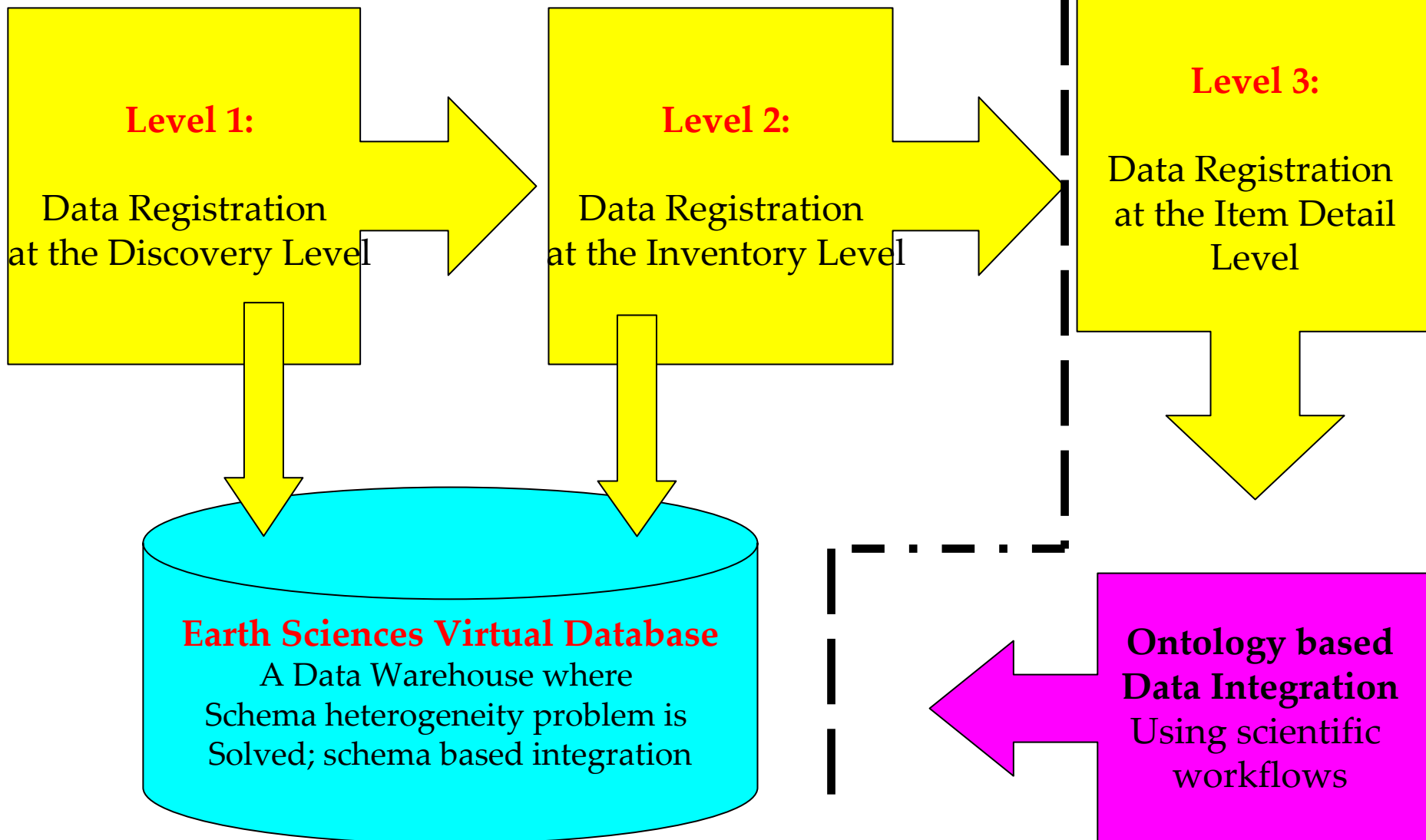
# ESG Security Services - Client Interaction



# Data Registration Framework

## Data Discovery

## Data Integration

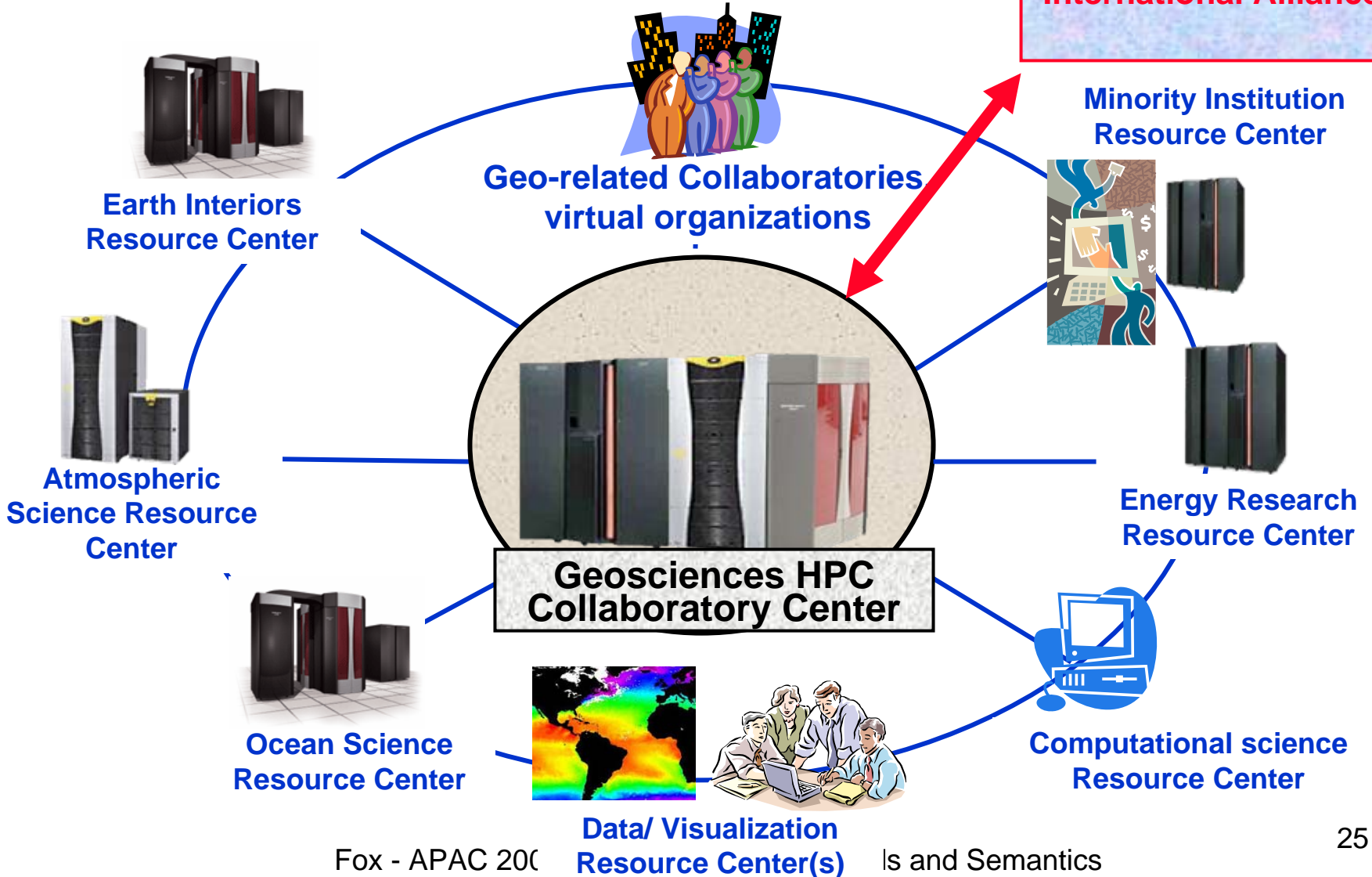


# A Geosciences Collaboratory

Serving the NSF Geosciences Research Community...

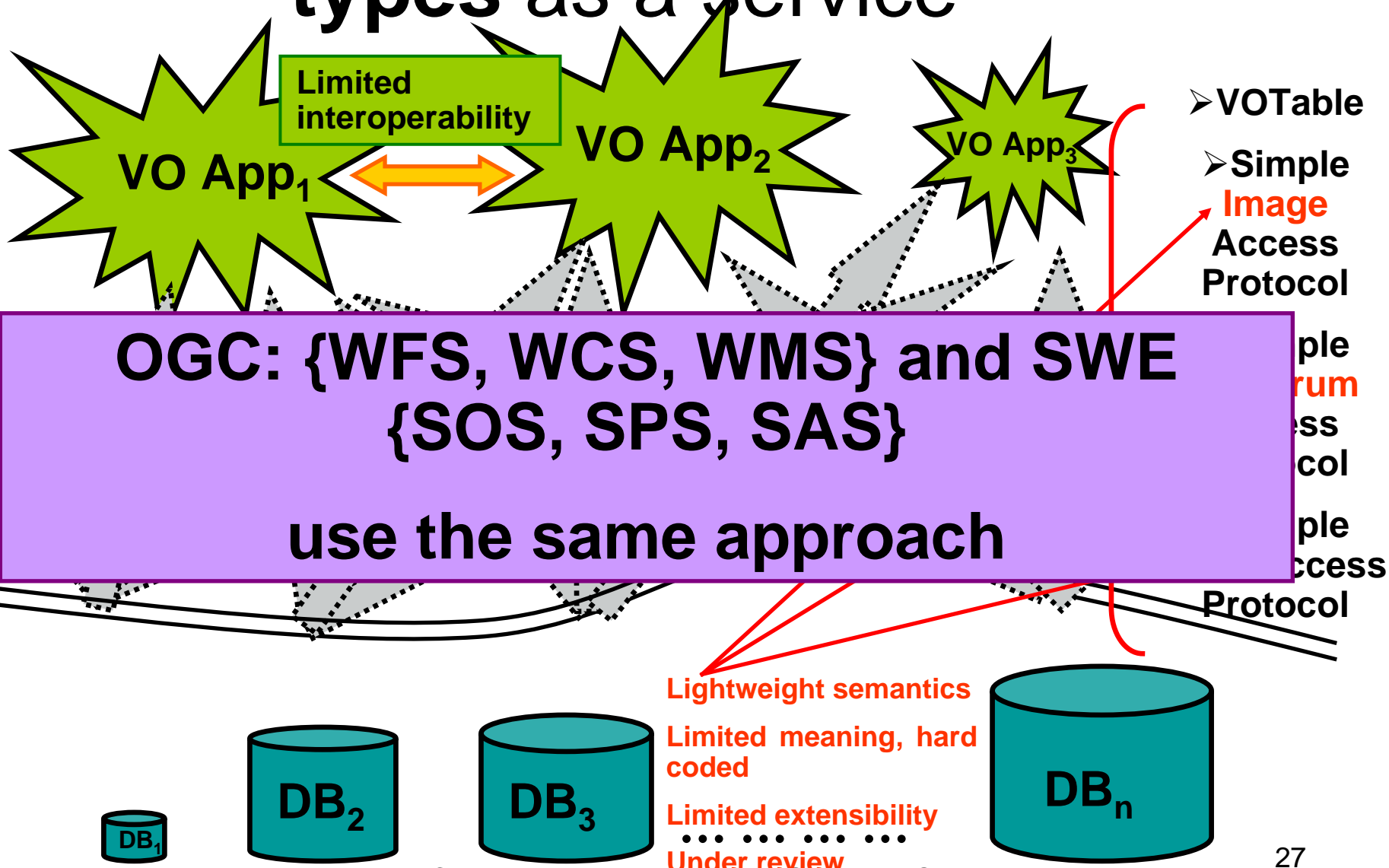


Other National Labs and Supercomputer Centers, Grids and International Alliances



# Back shed

# The Astronomy approach; **data-types as a service**



# Semantic Web Services

VSTO Query Instrument Web Service

NCAR

Virtual Solar Terrestrial Observatory

Home Data Communities About Us Login

Guided Workflows: Start by Instrument | Start by Dates | Start by Parameter Web Services: Query Instrument | Query Parameter | Query Data

**VSTO Web Services**

## Query Instrument Web Service

**Description:** Web Service used to query the VSTO ontology to retrieve all the Instrument instances matching one or more optional constraints.

**Input:** String parameterClass (optional, must be valid Parameter class name from VSTO ontology)  
String startDate (optional, formatted as yyyy-mm-dd)  
int nDays (required if startDate is used, must be  $1 < nDays < 31$ )  
String domain (optional, must be 'CEDAR' or 'MLSO')  
String instrumentClass (optional, must be valid instrument class name from VSTO ontology)

**Output:** XML/OWL document containing the Instrument instances matching the query. The XML is serialized as a String.

**Exception:** Thrown if invalid input is used in the query

**Endpoint:** <http://www.vsto.org:8080/services/VSTOQueryService>

**WSDL:** <http://www.vsto.org:8080/services/VSTOQueryService?wsdl>



**Example:** Find all Instruments that measure Neutral Temperature

Input: parameterClass='NeutralTemperature', startDate=null, ndays=0, domain=null, instrumentClass=null

**Example:** Find all Instruments of type Interferometer that measured data in August 1999

Input: parameterClass=null, startDate='1999-08-01', ndays=31, domain=null, instrumentClass='Interferometer'

## Query Input

Use the following interface to perform a live test of the VSTO Query Instrument Web Service:

**Parameter Type:**  Optional: return only instruments that measured this type of parameter

Select from list:

**Start Date:**  (yyyy-mm-dd) **Number of Days:**  Optional: return only instruments that measured data within this time interval

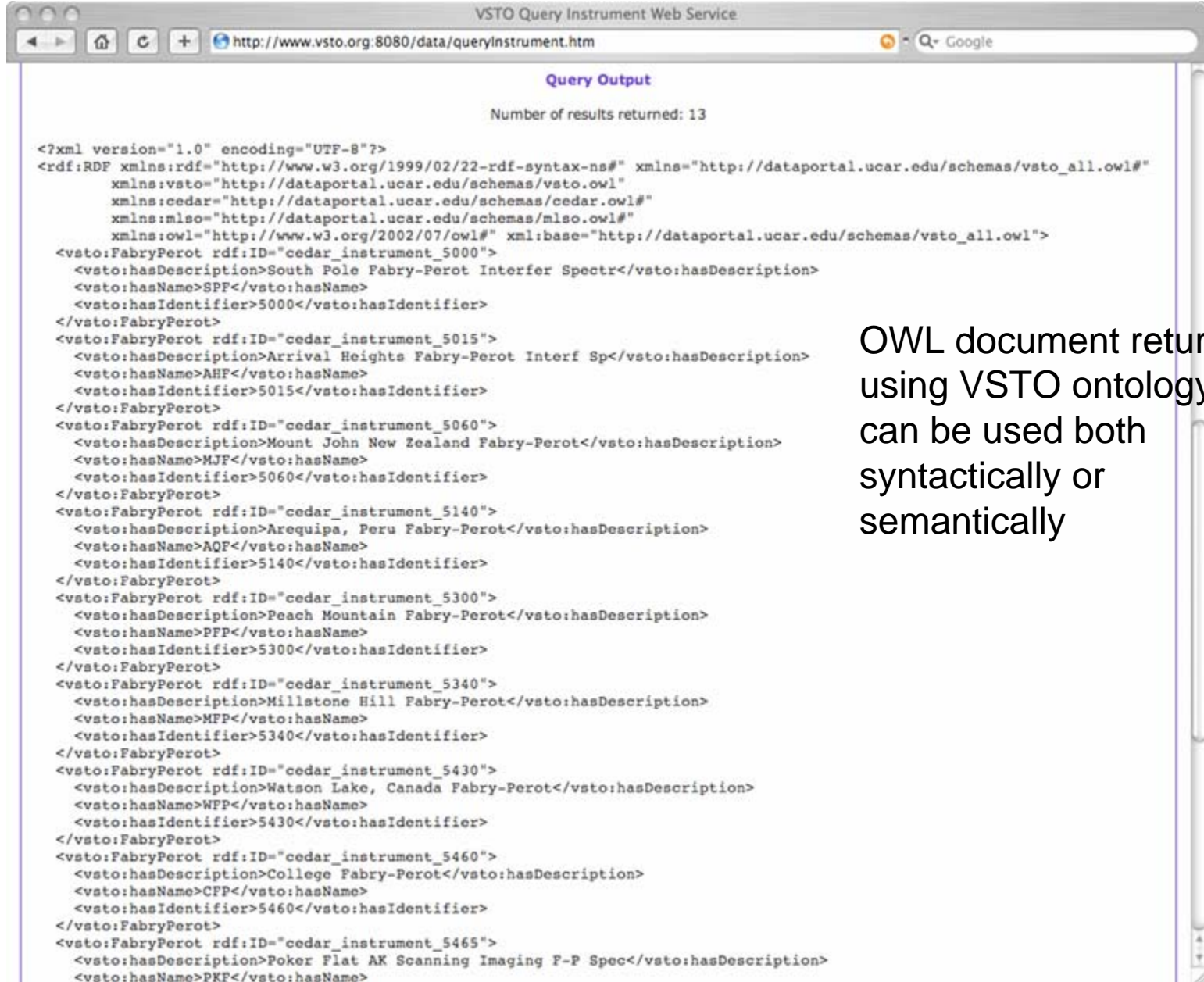
**Domain:**  Optional: return only instruments in this domain

**Instrument Type:**  Optional: return only instruments of this kind

Select from list:

Submit

# Semantic Web Services



Query Output

Number of results returned: 13

```
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#" xmlns="http://dataportal.ucar.edu/schemas/vsto_all.owl#"
  xmlns:vsto="http://dataportal.ucar.edu/schemas/vsto.owl"
  xmlns:cedar="http://dataportal.ucar.edu/schemas/cedar.owl#"
  xmlns:mlso="http://dataportal.ucar.edu/schemas/mlso.owl#"
  xmlns:owl="http://www.w3.org/2002/07/owl#" xml:base="http://dataportal.ucar.edu/schemas/vsto_all.owl">
  <vsto:FabryPerot rdf:ID="cedar_instrument_5000">
    <vsto:hasDescription>South Pole Fabry-Perot Interfer Spectr</vsto:hasDescription>
    <vsto:hasName>SPP</vsto:hasName>
    <vsto:hasIdentifier>5000</vsto:hasIdentifier>
  </vsto:FabryPerot>
  <vsto:FabryPerot rdf:ID="cedar_instrument_5015">
    <vsto:hasDescription>Arrival Heights Fabry-Perot Interf Sp</vsto:hasDescription>
    <vsto:hasName>AHF</vsto:hasName>
    <vsto:hasIdentifier>5015</vsto:hasIdentifier>
  </vsto:FabryPerot>
  <vsto:FabryPerot rdf:ID="cedar_instrument_5060">
    <vsto:hasDescription>Mount John New Zealand Fabry-Perot</vsto:hasDescription>
    <vsto:hasName>MJF</vsto:hasName>
    <vsto:hasIdentifier>5060</vsto:hasIdentifier>
  </vsto:FabryPerot>
  <vsto:FabryPerot rdf:ID="cedar_instrument_5140">
    <vsto:hasDescription>Arequipa, Peru Fabry-Perot</vsto:hasDescription>
    <vsto:hasName>AQF</vsto:hasName>
    <vsto:hasIdentifier>5140</vsto:hasIdentifier>
  </vsto:FabryPerot>
  <vsto:FabryPerot rdf:ID="cedar_instrument_5300">
    <vsto:hasDescription>Peach Mountain Fabry-Perot</vsto:hasDescription>
    <vsto:hasName>PPF</vsto:hasName>
    <vsto:hasIdentifier>5300</vsto:hasIdentifier>
  </vsto:FabryPerot>
  <vsto:FabryPerot rdf:ID="cedar_instrument_5340">
    <vsto:hasDescription>Millstone Hill Fabry-Perot</vsto:hasDescription>
    <vsto:hasName>MFP</vsto:hasName>
    <vsto:hasIdentifier>5340</vsto:hasIdentifier>
  </vsto:FabryPerot>
  <vsto:FabryPerot rdf:ID="cedar_instrument_5430">
    <vsto:hasDescription>Watson Lake, Canada Fabry-Perot</vsto:hasDescription>
    <vsto:hasName>WFP</vsto:hasName>
    <vsto:hasIdentifier>5430</vsto:hasIdentifier>
  </vsto:FabryPerot>
  <vsto:FabryPerot rdf:ID="cedar_instrument_5460">
    <vsto:hasDescription>College Fabry-Perot</vsto:hasDescription>
    <vsto:hasName>CFP</vsto:hasName>
    <vsto:hasIdentifier>5460</vsto:hasIdentifier>
  </vsto:FabryPerot>
  <vsto:FabryPerot rdf:ID="cedar_instrument_5465">
    <vsto:hasDescription>Poker Flat AK Scanning Imaging F-P Spec</vsto:hasDescription>
    <vsto:hasName>PKF</vsto:hasName>
```

OWL document returned using VSTO ontology - can be used both syntactically or semantically