Semantically-Enabled Large-Scale Science Data Repositories

Peter Fox*

*HAO/ESSL/NCAR

Deborah McGuinness$#, Luca Cinquini%, Patrick West*, Jose Garcia*, Tony Darnell*, James Benedict$, Don Middleton%, Stan Solomon*

$McGuinness Associates

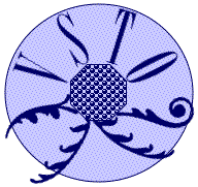
#Knowledge Systems and AI Lab, Stanford Univ.

%SCD/CISL/NCAR

Also Rob Raskin, Krishna Sinha

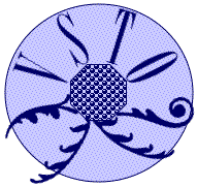
Work funded by NSF and NASA

Outline



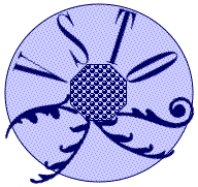
- Terminology and general introduction
- Virtual Observatories and Data Integration
- E.g. Virtual Solar-Terrestrial Observatory
- Semantics and reasoning
- Discussion

Terminology



- 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, domain, community, country
- VxyOs - x and y refer to different disciplines
- NB: VO also refers to Virtual Organization

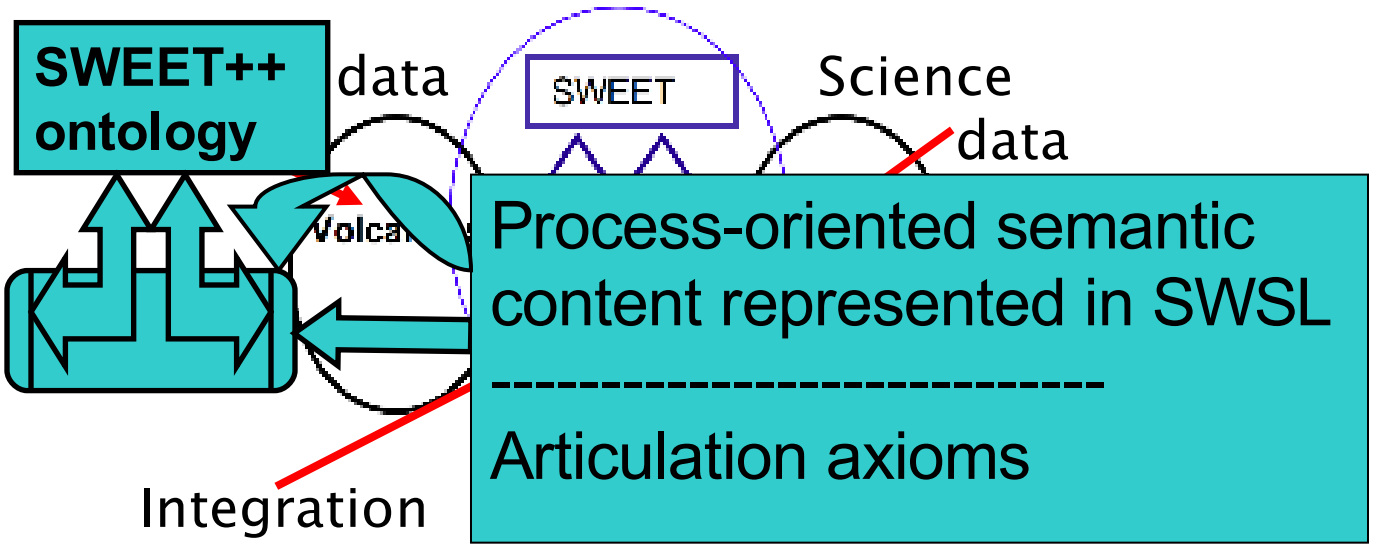
What should a VO do?



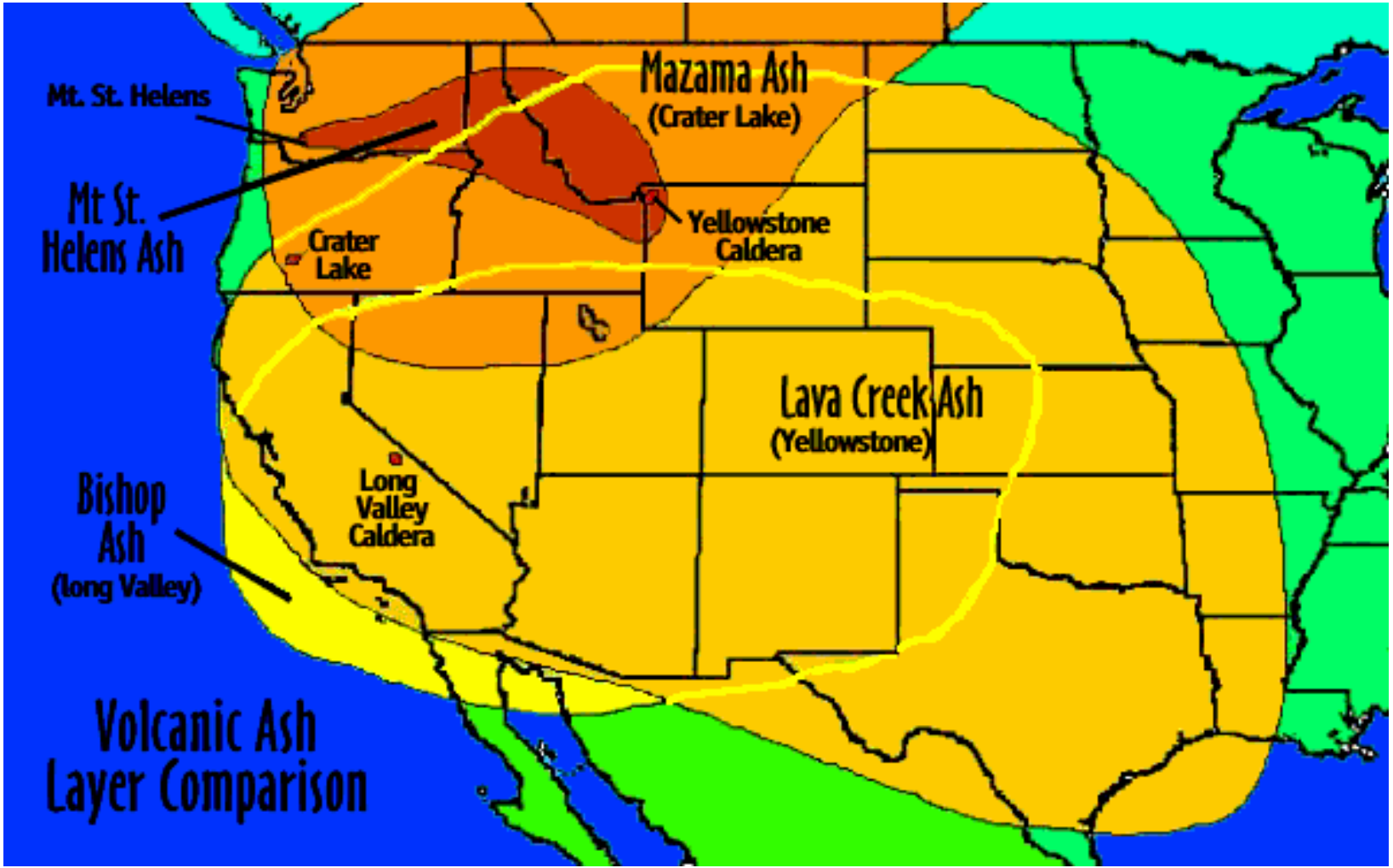
- Make “standard” scientific research much more efficient.
  - Even the principal investigator (PI) teams should want to use them.
  - Must improve on existing services (mission and PI sites, etc.). VOs will not replace these, but will use them in new ways.
  - Access for young researchers, non-experts, other disciplines,
- Enable new, global problems to be solved.
  - Rapidly gain integrated views, e.g. from the solar origin to the terrestrial effects of an event.
  - Find meaningful data related to any particular observation or model.
  - (Ultimately) answer “higher-order” queries such as “Show me the data from cases where a large coronal mass ejection observed by the Solar-Orbiting Heliospheric Observatory was also observed *in situ*.” (science-speak) or “What happens when the Sun disrupts the Earth’s environment” (general public)”

Paradigm shift for NASA

- From: Instrument-based
- To: Measurement-based (to find and use data irrespective of which instrument obtained it)
- 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.

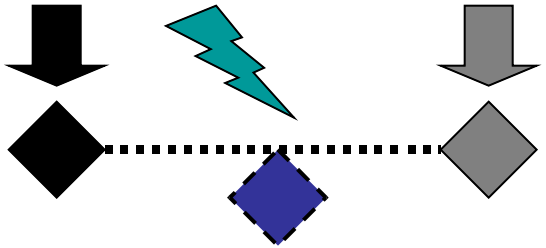


Compilation of distribution of volcanic ash associated with large eruptions. Note the continental scale ash fall associated with Yellowstone eruption ~600,000 years ago. Geologic databases provide the information about the magnitude of the eruption, and its impact on atmospheric chemistry and reflectance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

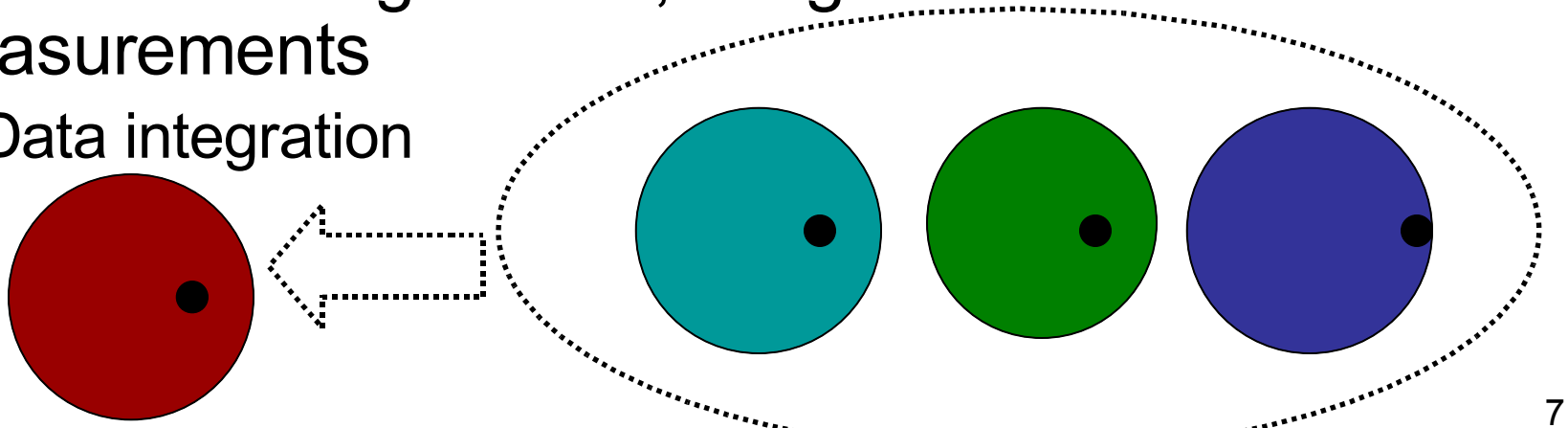


Virtual Observatory schematic

- Conceptual examples:
- In-situ: Virtual measurements
  - Sensors, etc. everywhere
  - Related measurements

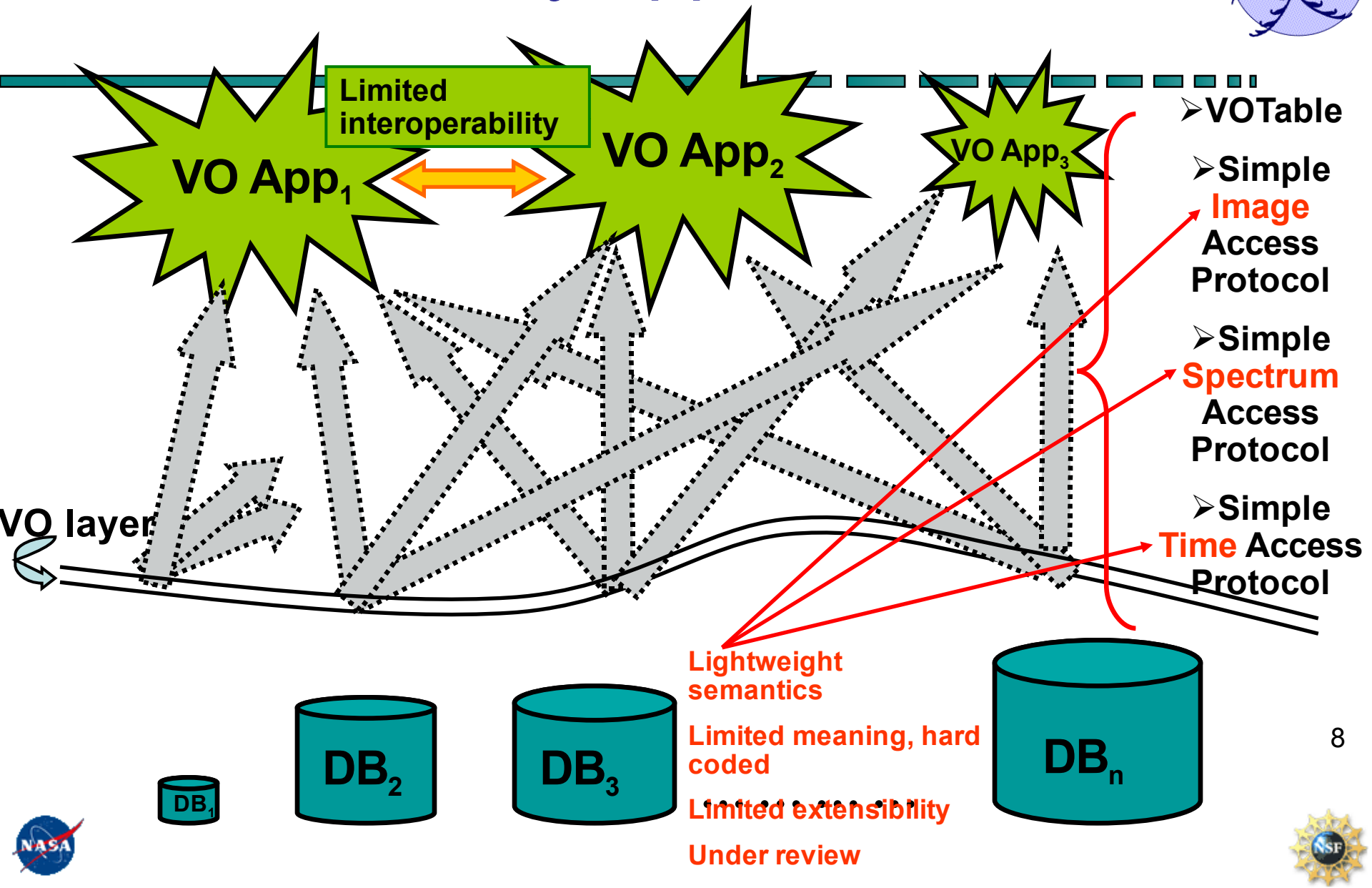


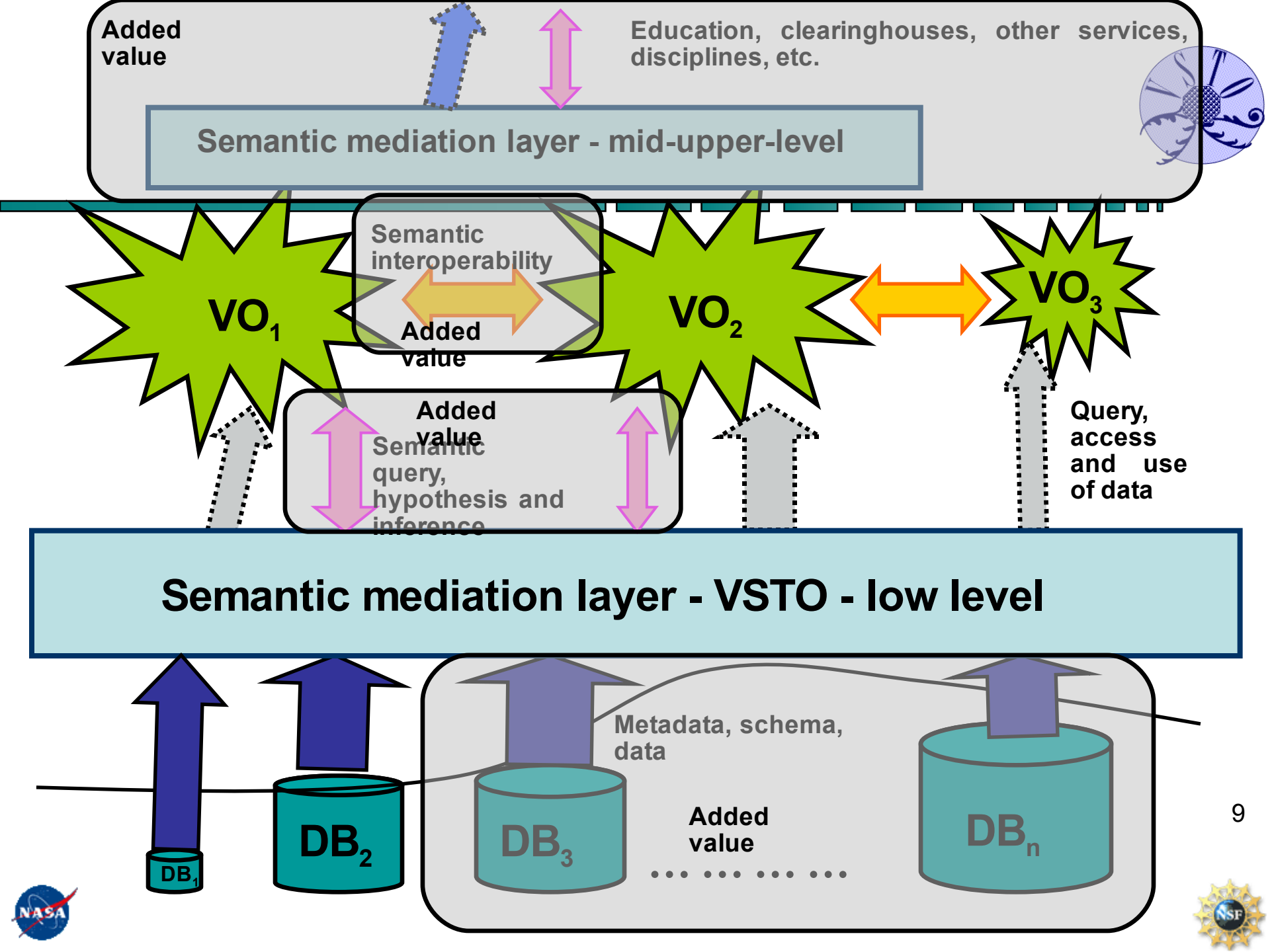
- Remote sensing: Virtual, integrative measurements
  - Data integration



- Both usage patterns lead to additional data management challenges at the source **and** for users; now managing virtual ‘datasets’

The Astronomy approach

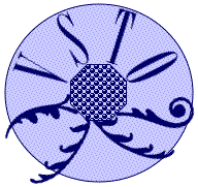



Added value
Education, clearinghouses, other services, disciplines, etc.
Semantic mediation layer - mid-upper-level
VO1
Semantic interoperability
Added value
VO2
VO3
Added value
Semantic query, hypothesis and inference
Query, access and use of data
Semantic mediation layer - VSTO - low level
DB1
DB2
Metadata, schema, data
DB3
Added value
DBn

Semantic Mediation Layer

- Ontology - capturing concepts of Parameters, Instruments, Date/Time, Data Product (and associated classes, properties) and Service Classes
- Maps queries to underlying data
- Generates access requests for metadata, data
- Allows queries, reasoning, analysis, new hypothesis generation, testing, explanation, etc.

Integrative use-cases:

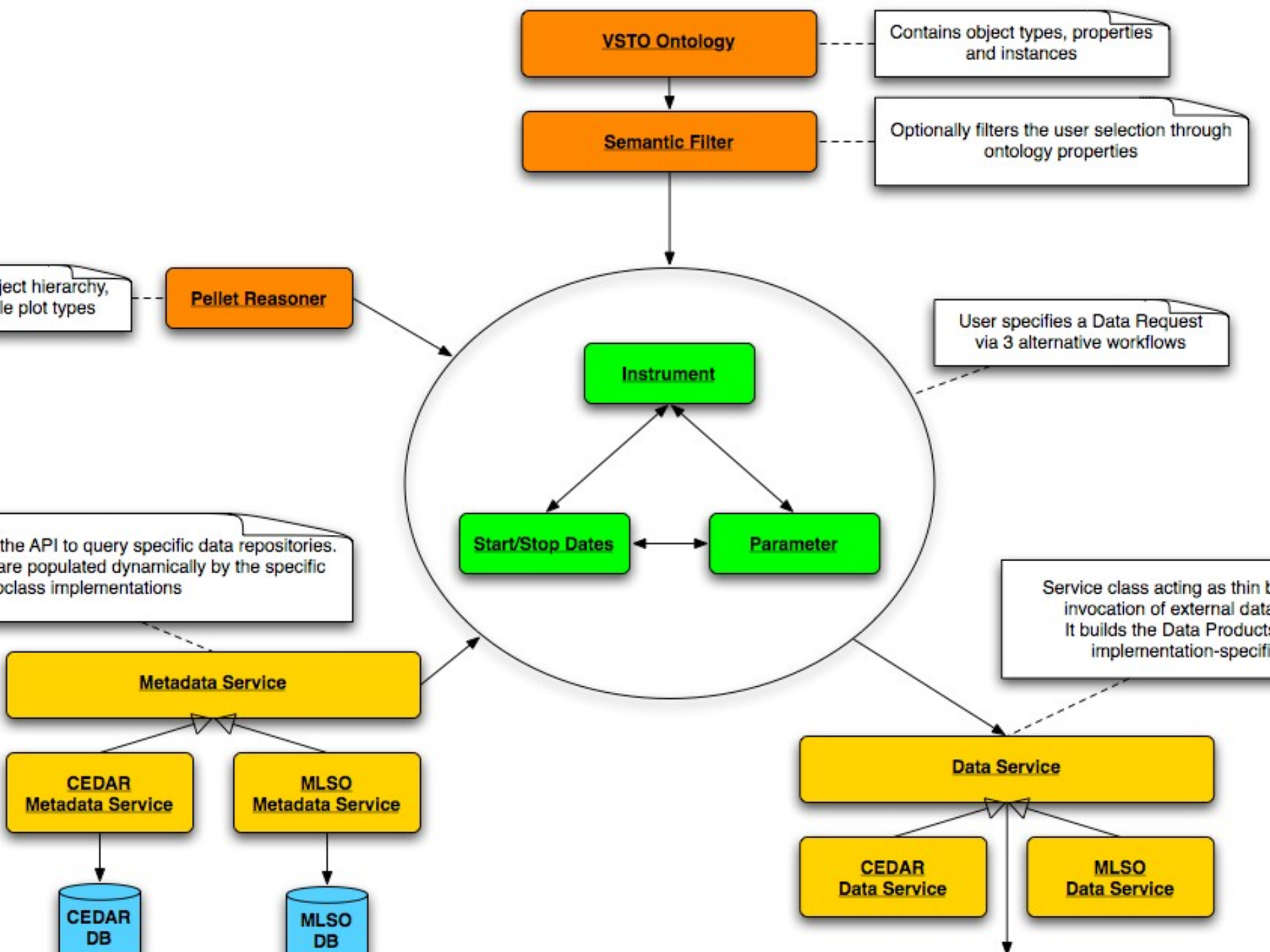


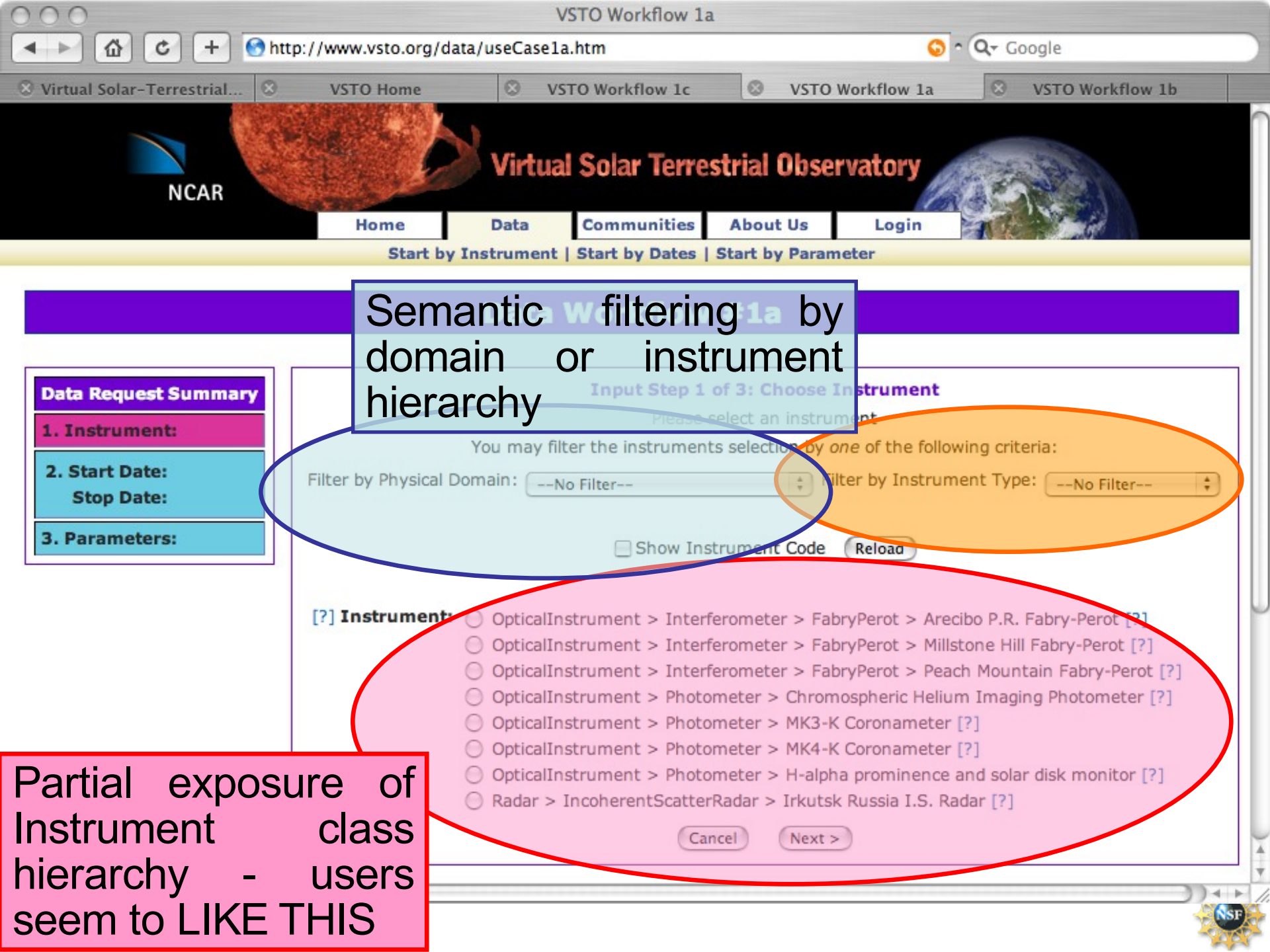
Find data which represents the state of the neutral atmosphere anywhere above 100km and toward the arctic circle (above 45N) at any time of high geomagnetic activity.

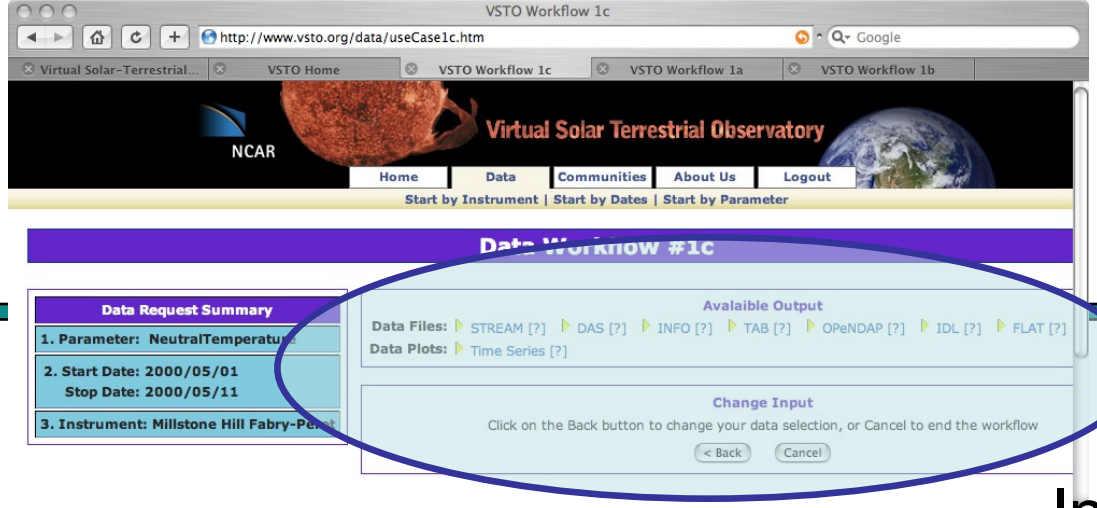
Translate this into a complete query for data. Was all the needed information recorded?

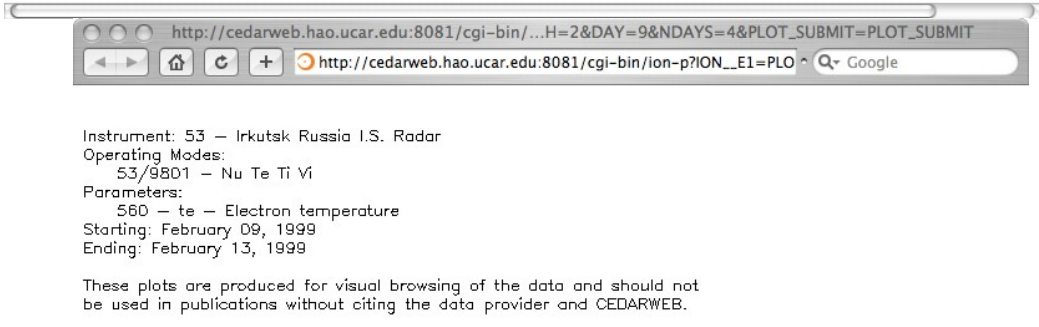
Information needs to be inferred (and integrated) from the use-case

What is returned: Data from instruments, indices and models.

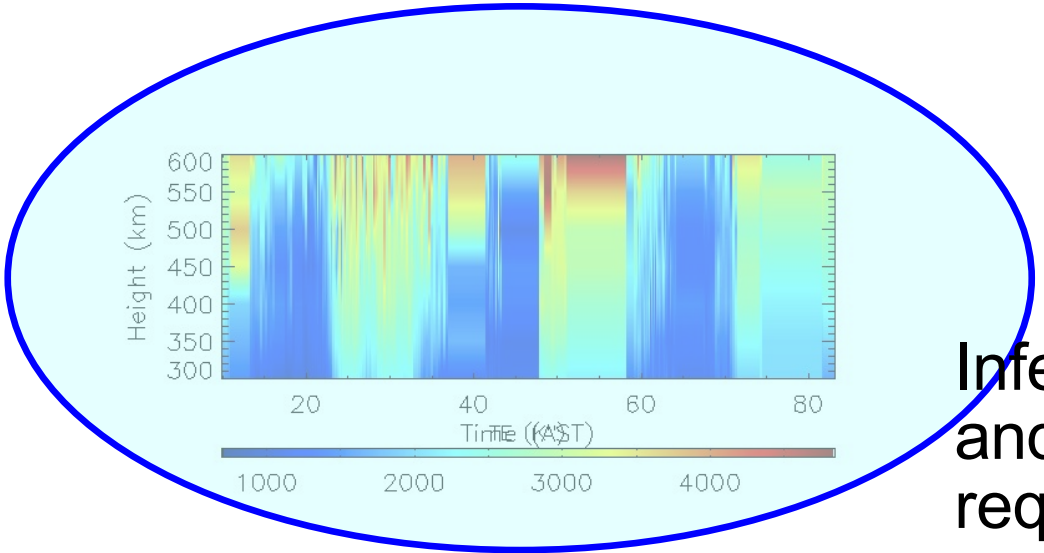

VSTO Ontology
Contains object types, properties and instances
Semantic Filter
Optionally filters the user selection through ontology properties
ject hierarchy, le plot types
Pellet Reasoner
User specifies a Data Request via 3 alternative workflows
Instrument
Start/Stop Dates
Parameter
the API to query specific data repositories. are populated dynamically by the specific oclass implementations
Service class acting as thin b invocation of external data It builds the Data Products implementation-specifi
Metadata Service
Data Service
CEDAR Metadata Service
MLSO Metadata Service
CEDAR Data Service
MLSO Data Service
CEDAR DB
MLSO DB


VSTO Workflow 1a
http://www.vsto.org/data/useCase1a.htm
Google
Virtual Solar-Terrestrial...
VSTO Home
VSTO Workflow 1c
VSTO Workflow 1a
VSTO Workflow 1b
NCAR
Virtual Solar Terrestrial Observatory
Home
Data
Communities
About Us
Login
Start by Instrument | Start by Dates | Start by Parameter
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
You may filter the instruments selection by one of the following criteria:
Filter by Physical Domain: --No Filter--
Filter by Instrument Type: --No Filter--
Show Instrument Code
Reload
[?] Instrument:
OpticalInstrument > Interferometer > FabryPerot > Arecibo P.R. Fabry-Perot [?]
OpticalInstrument > Interferometer > FabryPerot > Millstone Hill Fabry-Perot [?]
OpticalInstrument > Interferometer > FabryPerot > Peach Mountain Fabry-Perot [?]
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 [?]
Radar > IncoherentScatterRadar > Irkutsk Russia I.S. Radar [?]
Cancel
Next >
Partial exposure of Instrument class hierarchy - users seem to LIKE THIS





Inferred plot type and return formats for data products

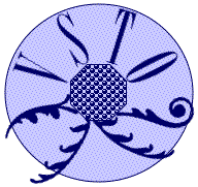


Inferred plot type and return required axes data

VSTO Notable progress

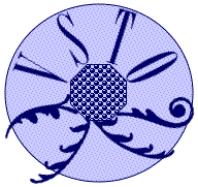
- Conceptual model and architecture developed by combined team; KR experts, domain experts, and software engineers
- Semantic framework developed and built with a small, cohesive, carefully chosen team in a relatively short time (deployments in 1st year)
- **Production** portal released, includes security, etc. with community migration (and so far endorsement)
- VSTO ontology version 0.4, (vsto.owl)
- Web Services encapsulation of semantic interfaces being documented
- More Solar Terrestrial use-cases to drive the completion of the ontologies - filling out the instrument ontology
- Evaluating ontologies for broader use (volcanoes, climate, …)

What has KR done for us?



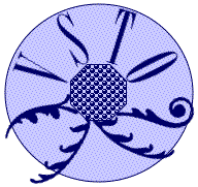
- In addition to valued added noted previously - some of which is transparent
- Reduced the need for 8 steps to query to 3 and reduced choices at each stage
- Allowed scientists to get data from instruments they never knew of before
- Allowed augmentation and validation of data
- Useful and related data provided without having to be an expert to ask for it
- Integration and use (e.g. plotting) based on inference
- Ask and answer questions not possible before

Knowledge engineering, etc.



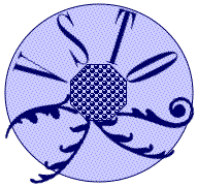
- Many controlled vocabularies as starting points.
- Strive for compatibility with “best practice” controlled vocabularies, taxonomies, and ontologies.
- Designed our own ontologies as dictated by use-case needs constantly with the goal of reusability and extensibility. (Provided VSTO modules back to at least one ontology suite with a much broader scope.)
- Early design HIGHLY collaborative in design and implementation – critical and continued contributions from domain scientists and knowledge representation.
- Result is fairly extensible by entire team.

Issues for Virtual Observatories



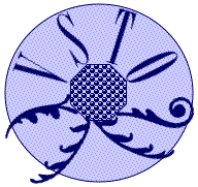
- Scaling to large numbers of data providers
- Crossing disciplines
- Security, access to resources, policies
- Branding and attribution (where did this data come from and who gets the credit, is it the correct version, is this an authoritative source?)
- Provenance/derivation (propagating key information as it passes through a variety of services, copies of processing algorithms, …)
- Data quality, preservation, stewardship, rescue
- Interoperability at a variety of levels (~3)

Final remarks

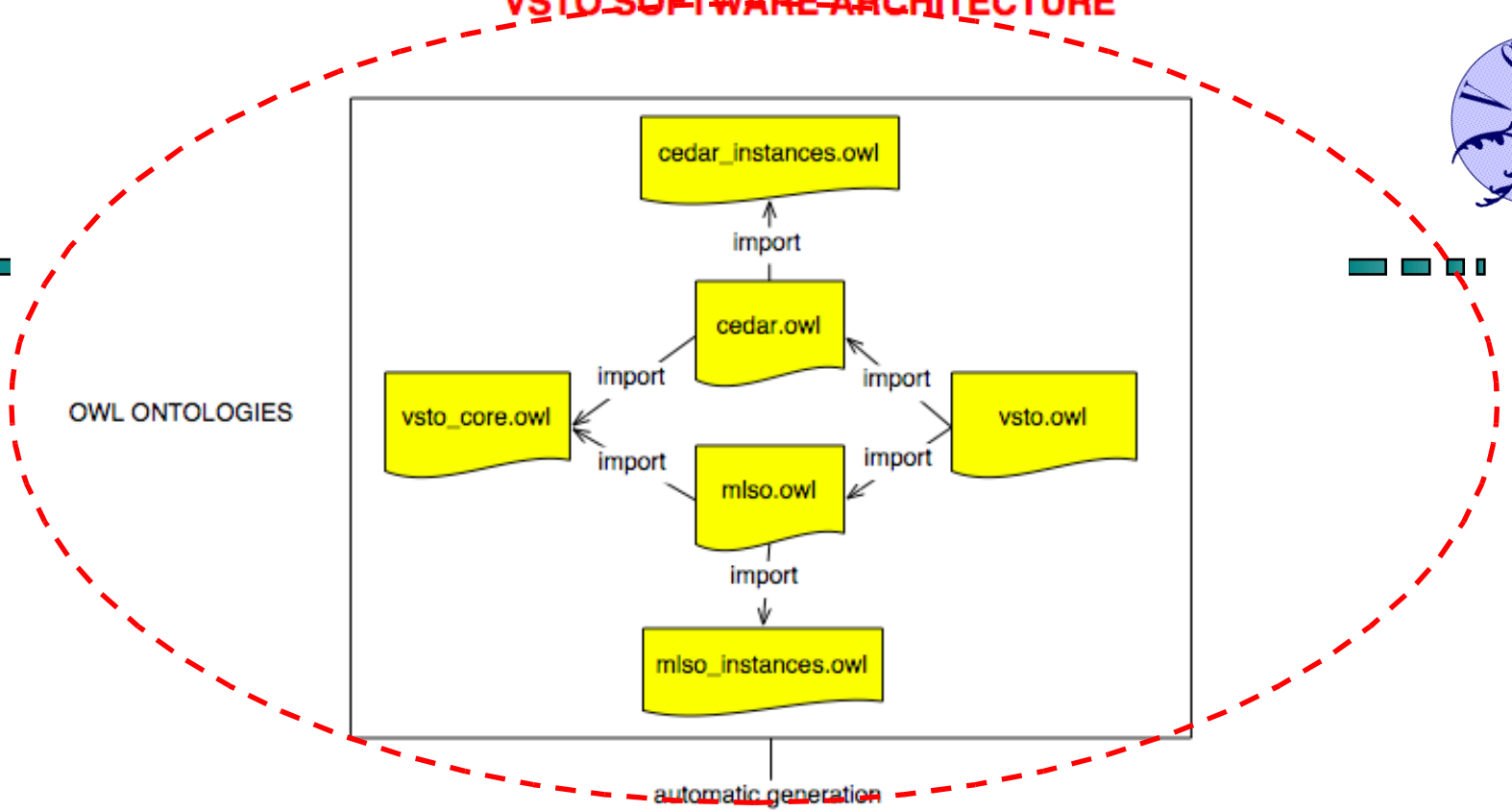


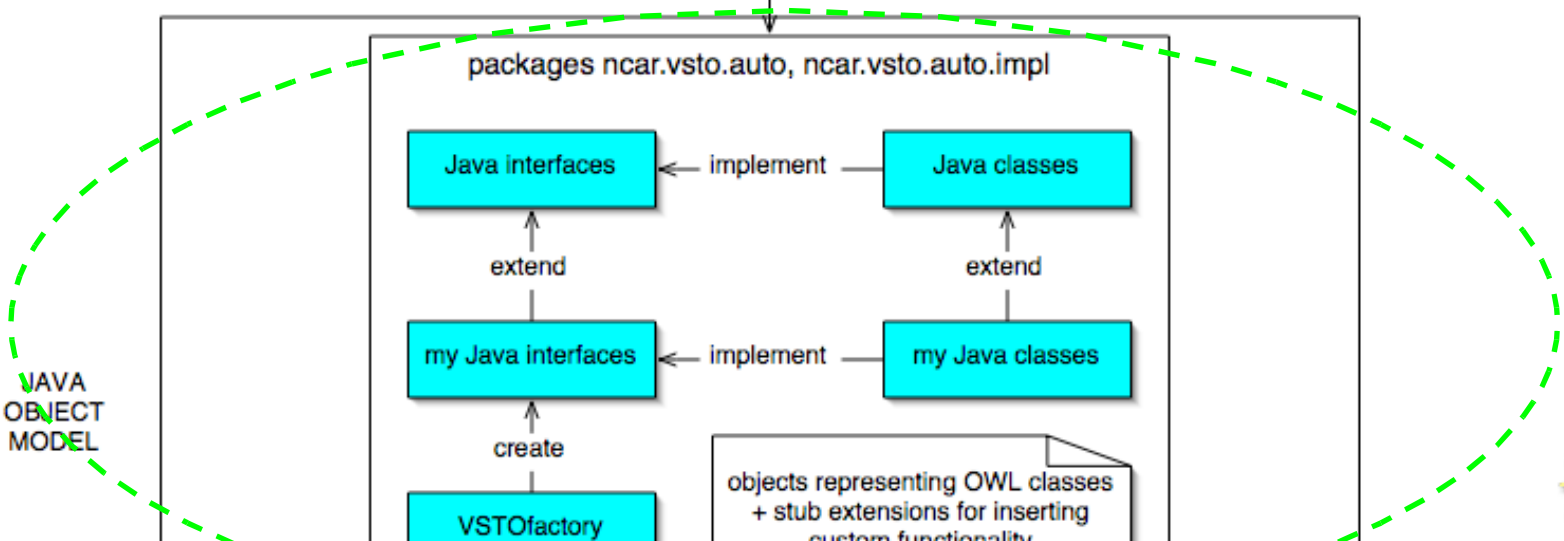
- Many geoscience VOs are in production
- Informatics efforts in Geosciences are exploding
  - GeoInformatics Town Hall at Fall AGU meeting Dec. 11 2006 in San Francisco, many cyberinfrastructure sessions
  - VO conference - April 2007 in Denver, CO
  - *e*-monograph to document state of VOs
  - NEW Journal of Earth Science Informatics
  - Special issue of Computers and Geosciences: “Knowledge Representation in Earth and Space Science Cyberinfrastructure”
- Ongoing activities for VOs through 2008 under the auspices of the Electronic Geophysical Year (*e*GY; www.egy.org)
- Contact pfox@ucar.edu, dlm@cs.stanford.edu

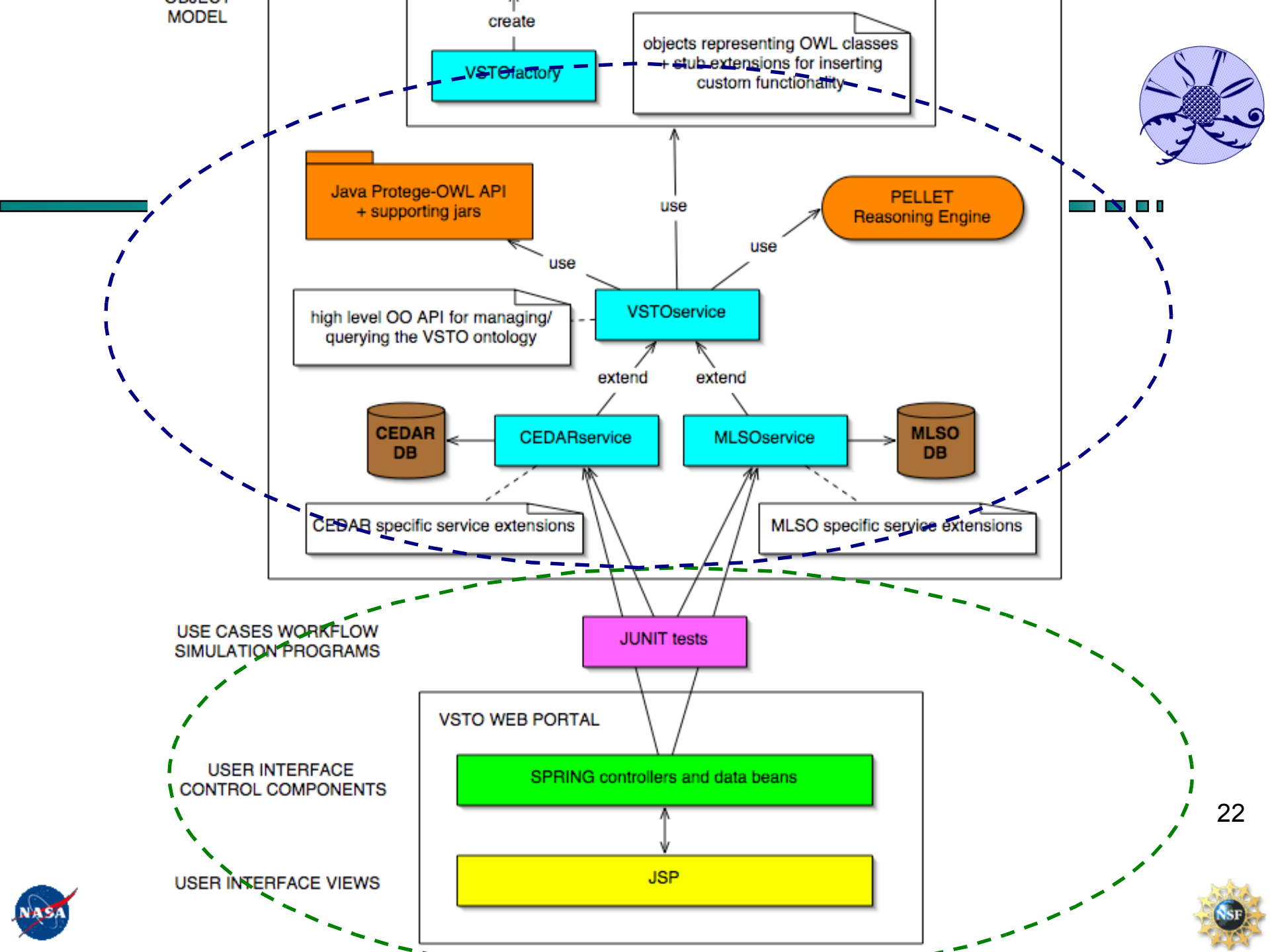
Garage

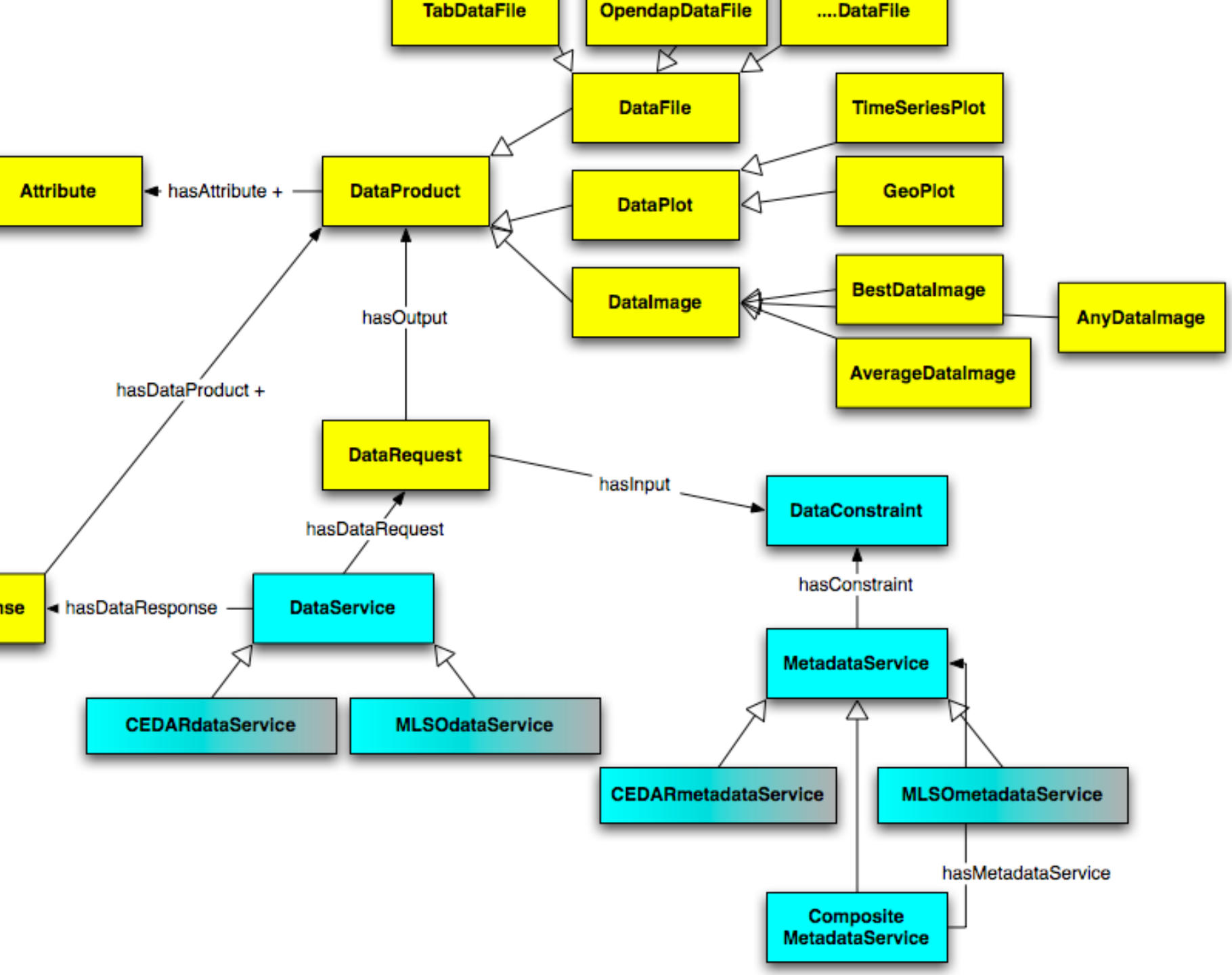


VSTO SOFTWARE ARCHITECTURE

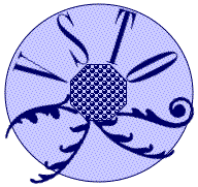





MODEL
create
VSTOfactory
objects representing OWL classes
+ stub extensions for inserting
custom functionality
Java Protege-OWL API
+ supporting jars
use
use
use
PELLET
Reasoning Engine
high level OO API for managing/
querying the VSTO ontology
VSTOservice
extend
extend
CEDAR
DB
CEDARservice
MLSOservice
MLSO
DB
CEDAR specific service extensions
MLSO specific service extensions
USE CASES WORKFLOW
SIMULATION PROGRAMS
JUNIT tests
VSTO WEB PORTAL
USER INTERFACE
CONTROL COMPONENTS
SPRING controllers and data beans
USER INTERFACE VIEWS
JSP


TabDataFile
OpendapDataFile
....DataFile
DataFile
TimeSeriesPlot
Attribute
hasAttribute +
DataProduct
DataPlot
GeoPlot
DataImage
BestDataImage
AnyDataImage
AverageDataImage
hasOutput
hasDataProduct +
DataRequest
hasInput
DataConstraint
hasDataRequest
hasConstraint
hasDataResponse
DataService
MetadataService
CEDARdataService
MLSOdataService
CEDARmetadataService
MLSOmetadataService
hasMetadataService
Composite
MetadataService

Why we were led to semantics



- When we integrate, we integrate concepts, terms
- In the past we would ask, guess, research a lot, or give up
- It's pretty much about **meaning**
- Semantics can really help find, access, **integrate, use, explain, trust…**
- 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?…