Integrating Inter-disciplinary Science Data with Semantic Mediation

Peter Fox (HAO/ESSL/NCAR)

Krishna Sinha (VT), Rob Raskin (JPL), Deborah McGuinness (RPI)

SESDI - Semantically-Enabled Science Data Integration

Funded by NASA/ACCESS and NASA/ESTO/AIST

Overview

- A little about semantics
- A little about integration
- Use case
- Semantic mis-understanding
- Impact of semantic mediation
- Methodology
- Some details of the integrating concepts
- Now, let’s hook up some data
- Summary and outlook

A little about semantics

- Gives syntax *meaning*
- Basic element is the *triple*: {subject-predicate-object}

  Interferometer is-a optical instrument

  Optical instrument has focal length

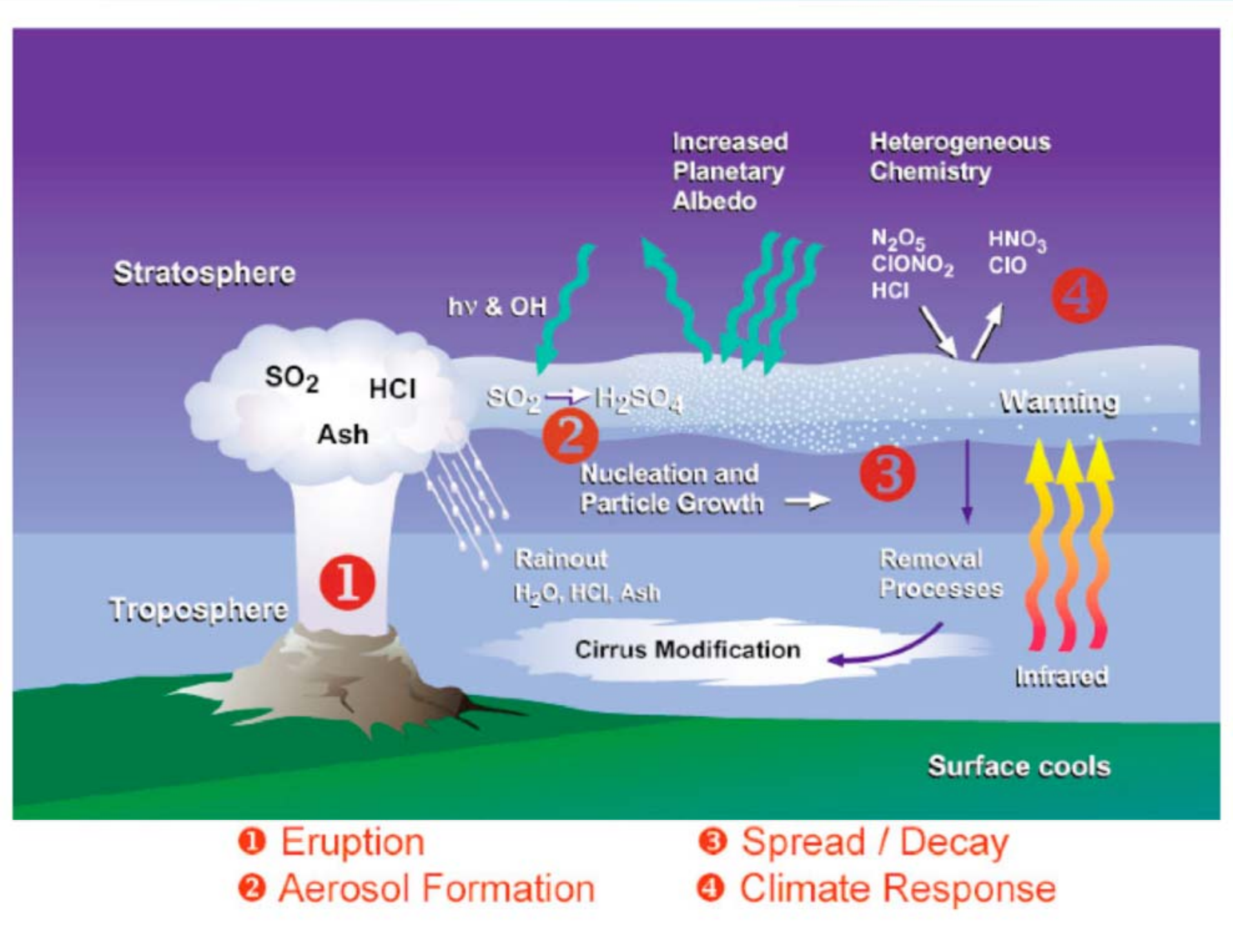
  An ontology is a representation of this knowledge
- W3C is the primary (but not sole) governing organization for languages, specifications, best practices, etc.
  - RDF - Resource Description Framework
  - OWL 1.0 - Ontology Web Language (OWL 2.0 on the way)
- Encode the knowledge in triples, in a triple-store, software is built to traverse the semantic network, it can be queried or reasoned upon
- Put semantics between/ in your interfaces, i.e. between layers and components in your architecture, i.e. between ‘users’ and ‘information’ to mediate the exchange

A little about integration

- 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?
  - find and use data you could not before?
  - 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?…

Integrative Use Case

- Determine the statistical signatures of both volcanic and solar forcings on the height of the tropopause



Challenges for Solar Radiation data integration

- 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 of the number of sunspots appearing on the visible solar surface, a sunspot *is an* indicator of the location of strong solar magnetic fields, strong magnetic fields are collectively known as solar activity, sunspots are observed to produce a localized *decrease* in the solar radiation output, at some wavelengths, *increase* at others, etc.
- Interfaces are built by computer scientists with syntax that often works within a discipline but rarely across them

SESDI Impact: A Better Way to Access Data

The Problem

Scientists only use data from a single instrument because it is difficult to access, process, and understand data from multiple instruments.

A typical data query might be:

> **“Give me the temperature, pressure, and water vapor from the AIRS instrument from Jan 2005 to Jan 2008”**
>
> **“Search for MLS/Aura Level 2, SO2 Slant Column Density from 2/1/2007”**

A Solution

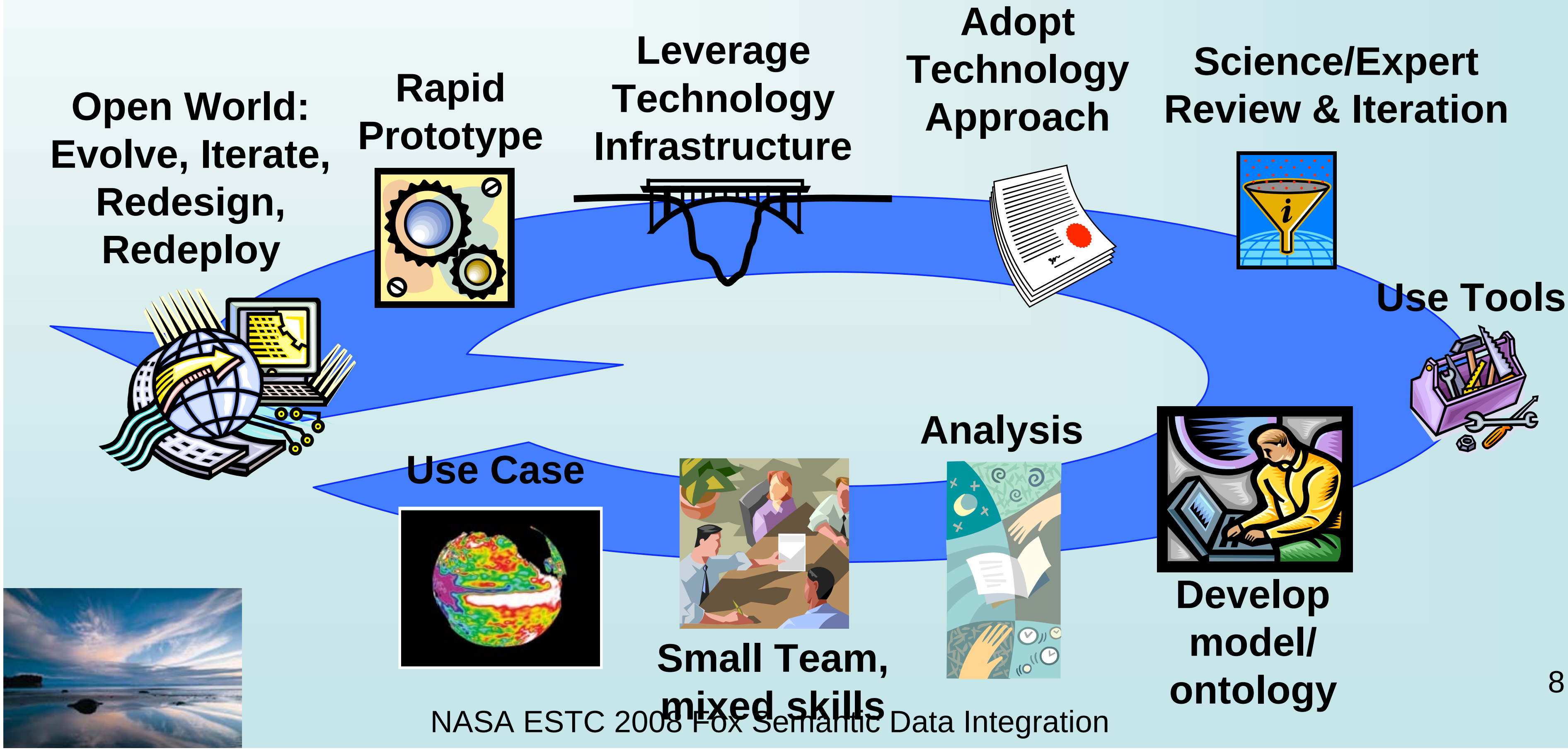
Using a simple process, SESDI allows data from various sources to be registered in an ontology so that it can be easily accessed and understood. Scientists can use only the ontology components that relate to their data. An SESDI query might look like:

> **“Show all areas in California where sulfur dioxide (SO2) levels were above normal between Jan 2000 and Jan 2007”**

This query will pull data from all available sources registered in the ontology and allow seamless data fusion. Because the query is measurement related, scientists do not need to understand the details of the instruments and data types.

Semantic Web Methodology and Technology Development Process

- Establish and improve a well-defined methodology vision for Semantic Technology based application development
- Leverage controlled vocabularies, etc.



Open World: Evolve, Iterate, Redesign, Redeploy

Rapid Prototype

Leverage Technology Infrastructure

Adopt Technology Approach

Science/Expert Review & Iteration

Use Tools

Develop model/ ontology

Analysis

Small Team, mixed skills

Use Case

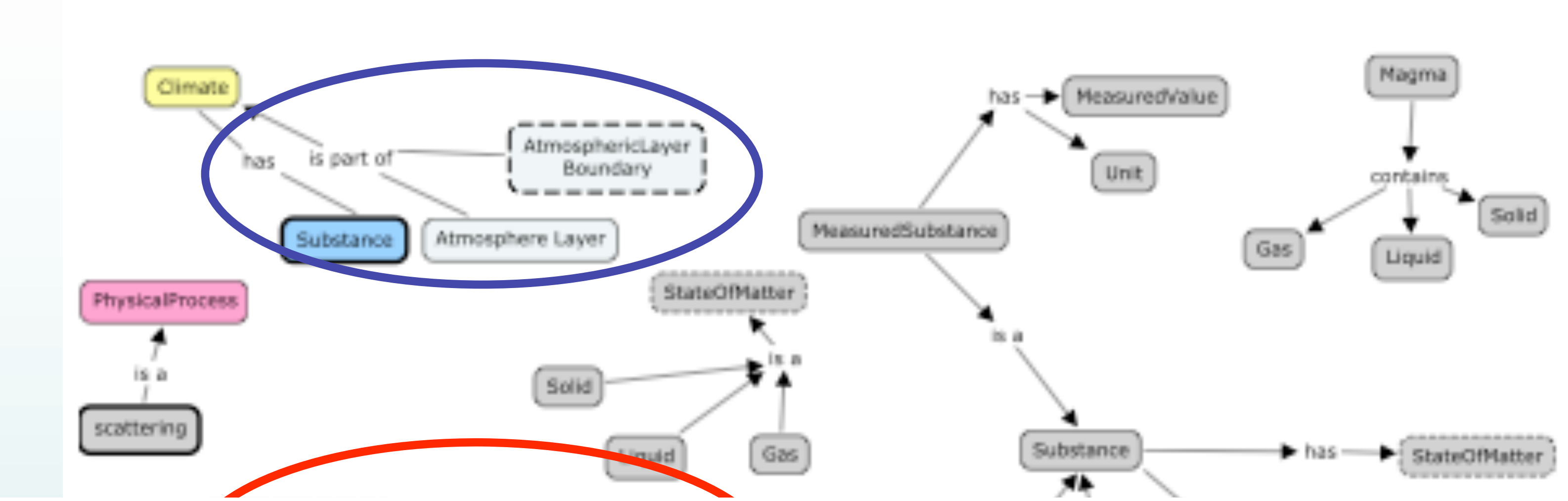
Volcano-Atmosphere considerations

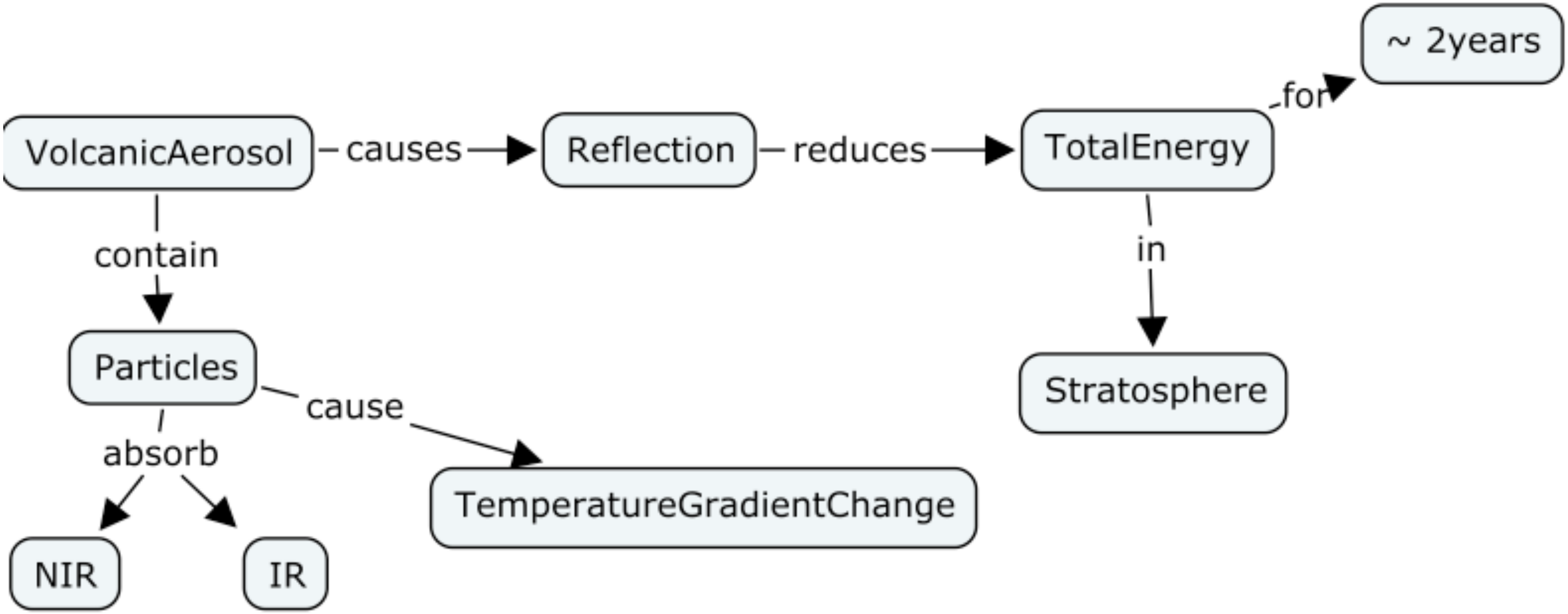
- Focus on tropopause -> temperature gradients
- Stratospheric and tropospheric aerosols, the tropospheric reservoir
- Quantities/processes: Gas, particles, ejecta, scattering
- Records: Pulses, e.g. in SO2 events
- Related aspects: SO2, H2SO4, O3 chemistry
- Data from: in-situ and remotely sensed observation, proxy, simulation, pseudo-proxy
- Processes: solar, volcanic, GHG, ocean, land-use
- Priors to consider: statistics of variability and extremes
- Main task: **detection and attribution**
- Solar-Atmosphere considerations are very similar

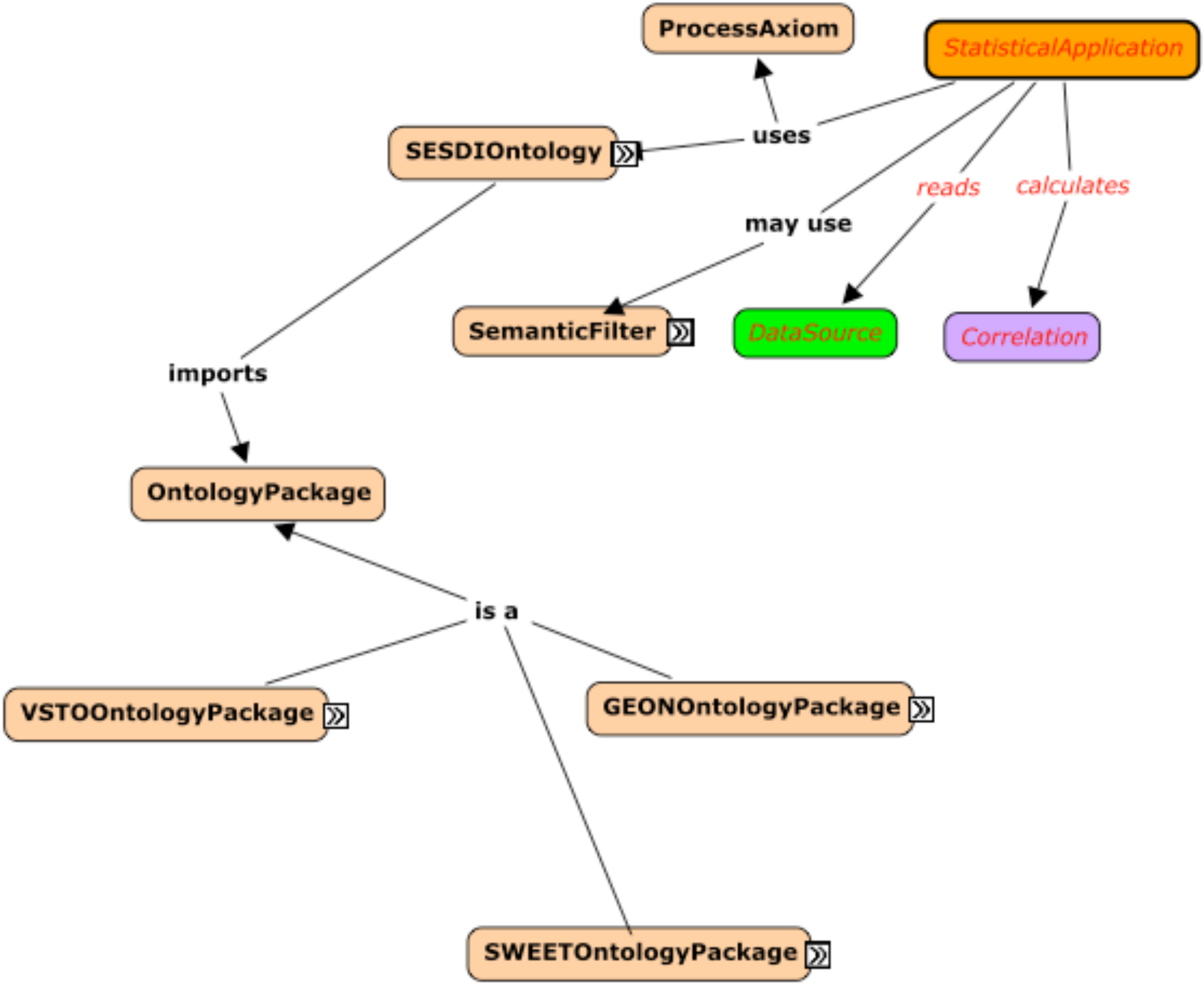
Components to implement

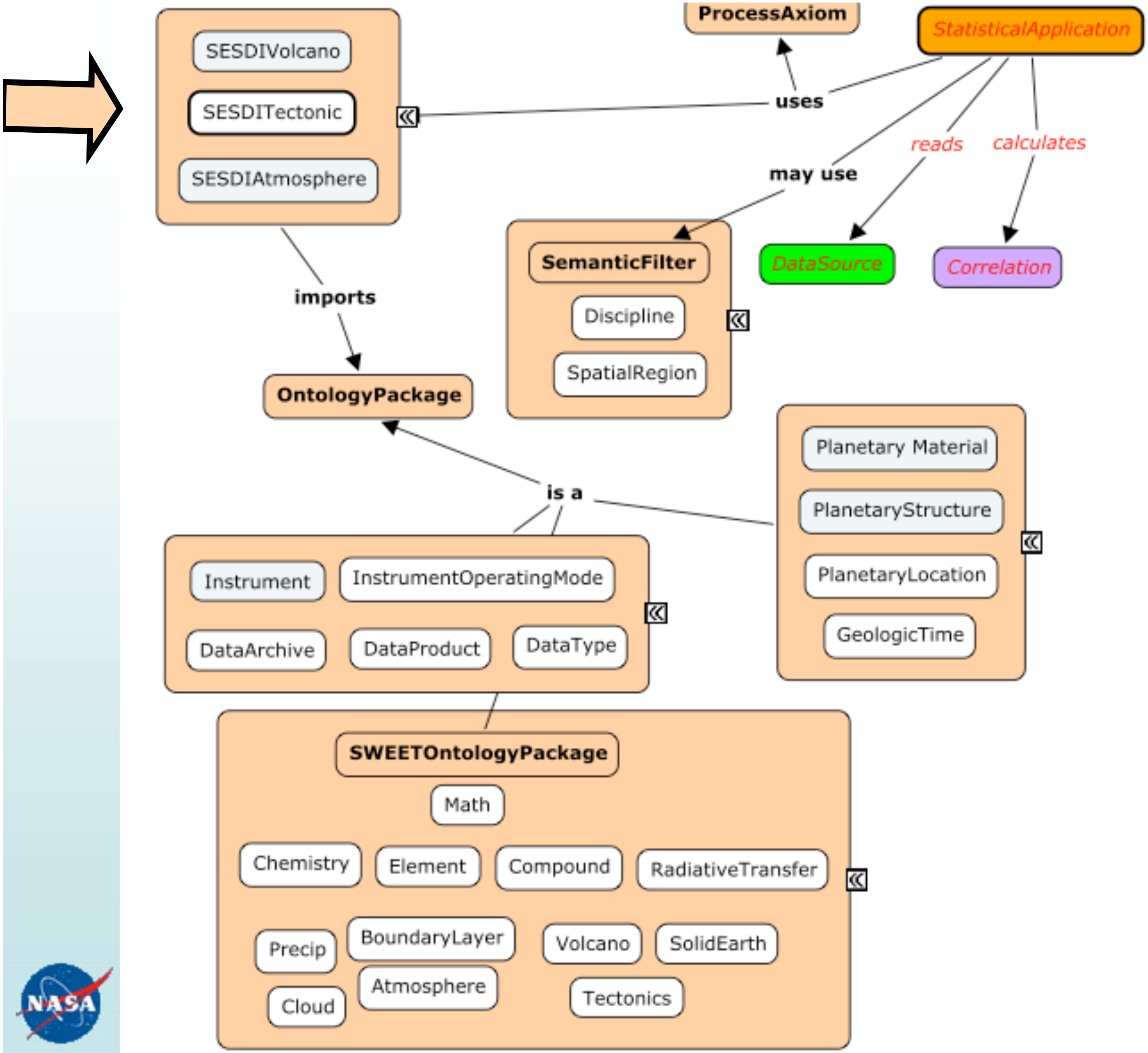
- An analysis application
- Cross-domain terms, concepts and relations (mediation here)
- Connections to underlying data (registration and mediation)
- Framework to put these together
- Integration connector

Detection and attribution relations…






ProcessAxiom
StatisticalApplication
SESDIOntology
uses
reads
calculates
may use
SemanticFilter
DataSource
Correlation
imports
OntologyPackage
is a
VSTOOntologyPackage
GEONOntologyPackage
SWEETOntologyPackage



SESDIVolcano

SESDITectonic

SESDIAtmosphere

ProcessAxiom

StatisticalApplication

uses

may use

reads

calculates

imports

SemanticFilter

Discipline

SpatialRegion

DataSource

Correlation

OntologyPackage

is a

Planetary Material

PlanetaryStructure

PlanetaryLocation

GeologicTime

Instrument

InstrumentOperatingMode

DataArchive

DataProduct

DataType

SWEETOntologyPackage

Math

Chemistry

Element

Compound

RadiativeTransfer

Precip

BoundaryLayer

Volcano

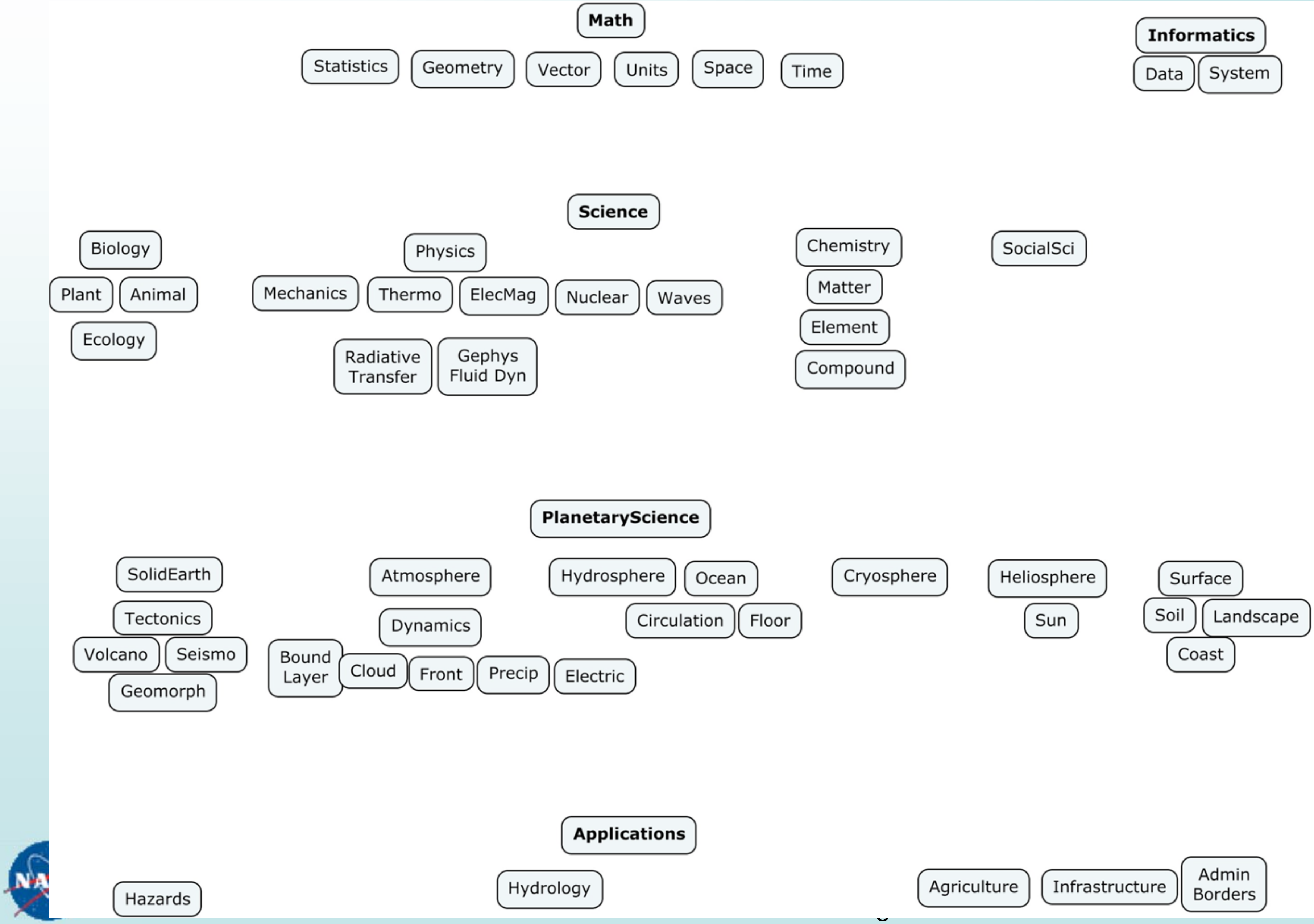
SolidEarth

Cloud

Atmosphere

Tectonics

SWEET 2.0 Ontologies



Math

Statistics Geometry Vector Units Space Time

Informatics

Data System

Science

Biology
Plant Animal
Ecology

Physics
Mechanics Thermo ElecMag Nuclear Waves
Radiative Transfer Gephys Fluid Dyn

Chemistry
Matter
Element
Compound

SocialSci

PlanetaryScience

SolidEarth
Tectonics
Volcano Seismo
Geomorph

Atmosphere
Dynamics
Bound Layer Cloud Front Precip Electric

Hydrosphere Ocean
Circulation Floor

Cryosphere

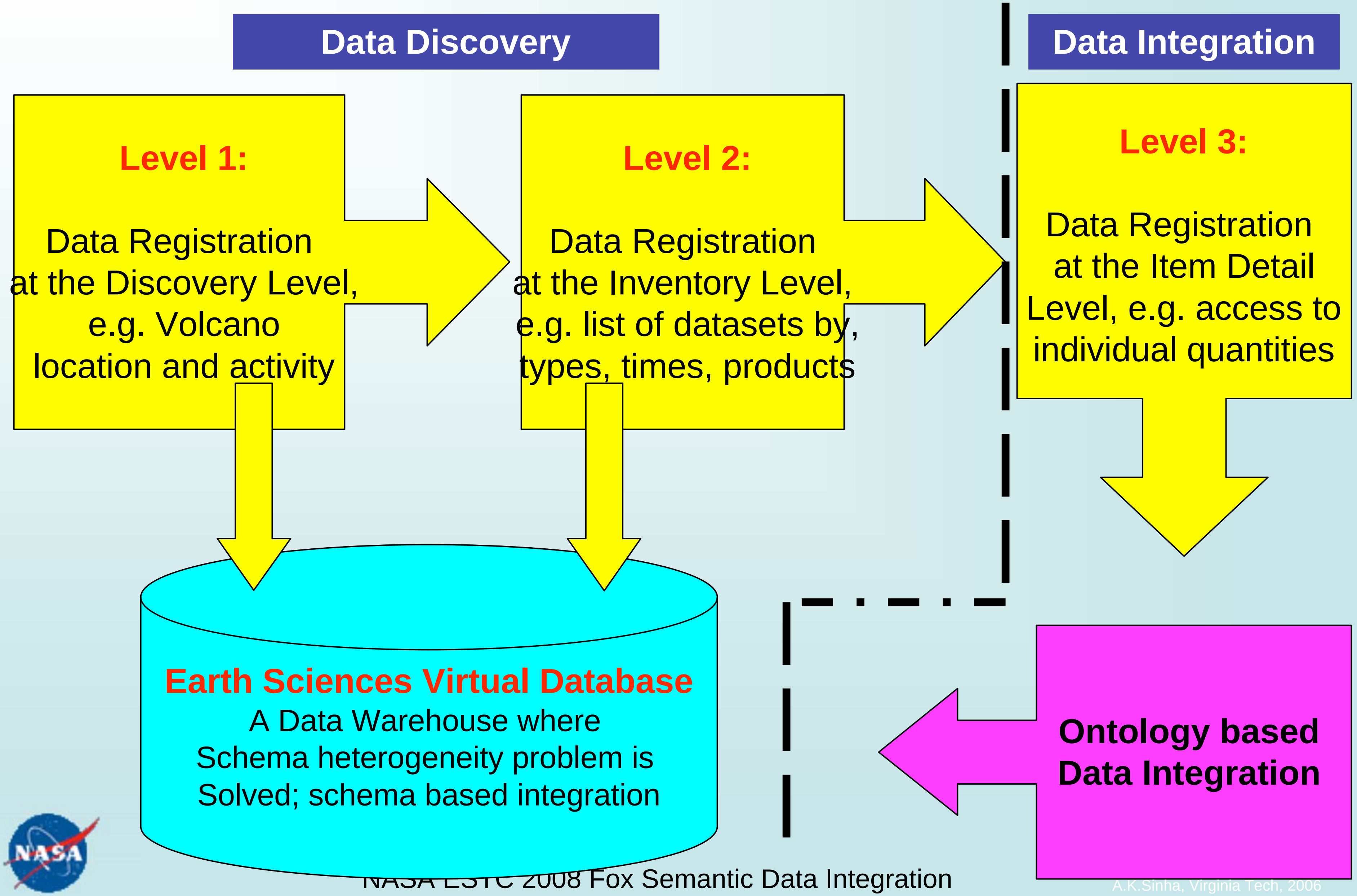
Heliosphere
Sun

Surface
Soil Landscape
Coast

Applications

Hazards Hydrology Agriculture Infrastructure Admin Borders

Data Registration Framework



Data Discovery

Data Integration

Level 1:

Data Registration at the Discovery Level, e.g. Volcano location and activity

Level 2:

Data Registration at the Inventory Level, e.g. list of datasets by, types, times, products

Level 3:

Data Registration at the Item Detail Level, e.g. access to individual quantities

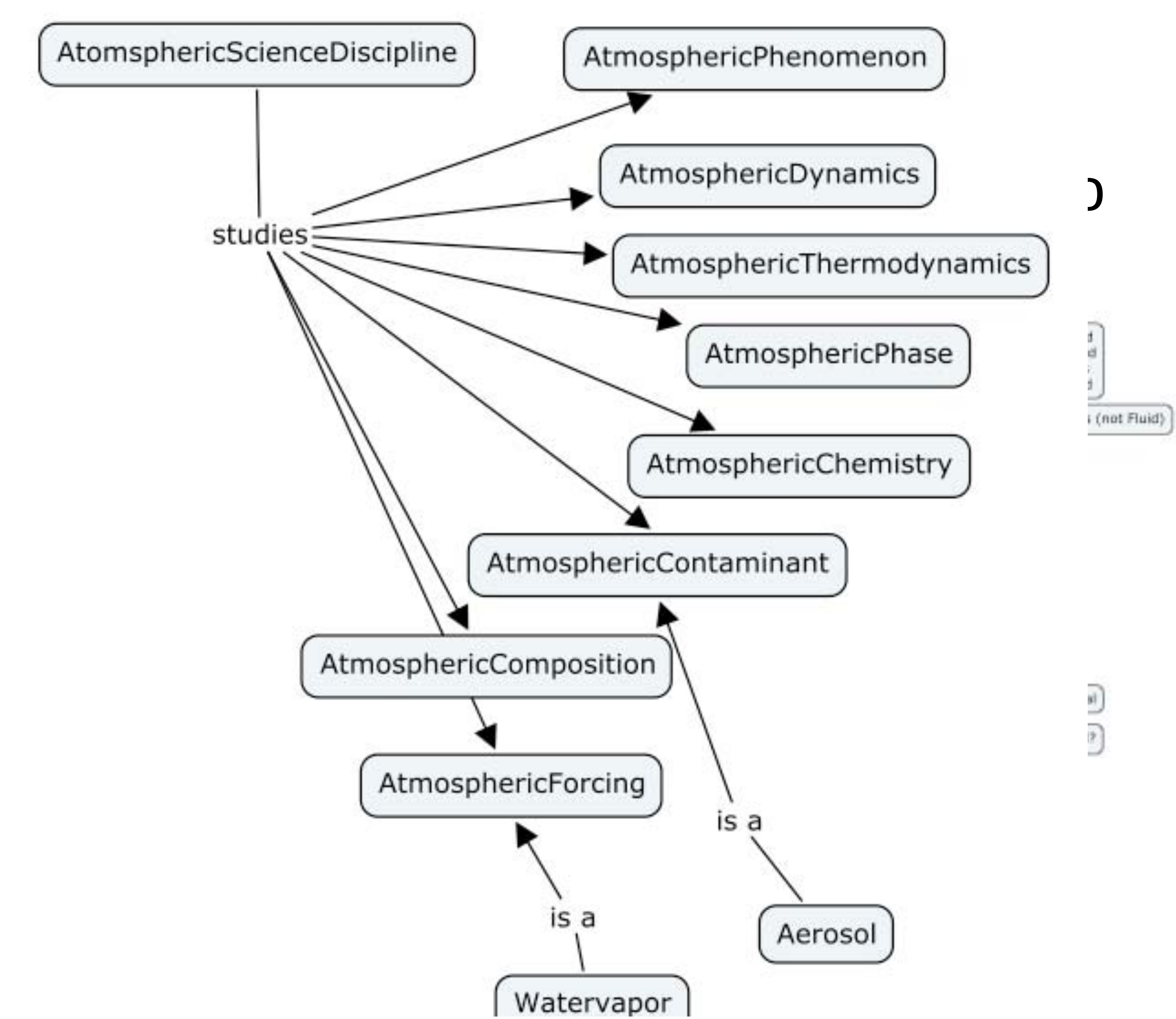
Earth Sciences Virtual Database

A Data Warehouse where Schema heterogeneity problem is Solved; schema based integration

Ontology based Data Integration

A.K.Sinha, Virginia Tech, 2006

- Th


AtomsphericScienceDiscipline
studies
AtmosphericPhenomenon
AtmosphericDynamics
AtmosphericThermodynamics
AtmosphericPhase
AtmosphericChemistry
AtmosphericContaminant
AtmosphericComposition
AtmosphericForcing
is a
is a
Aerosol
Watervapor
(not Fluid)

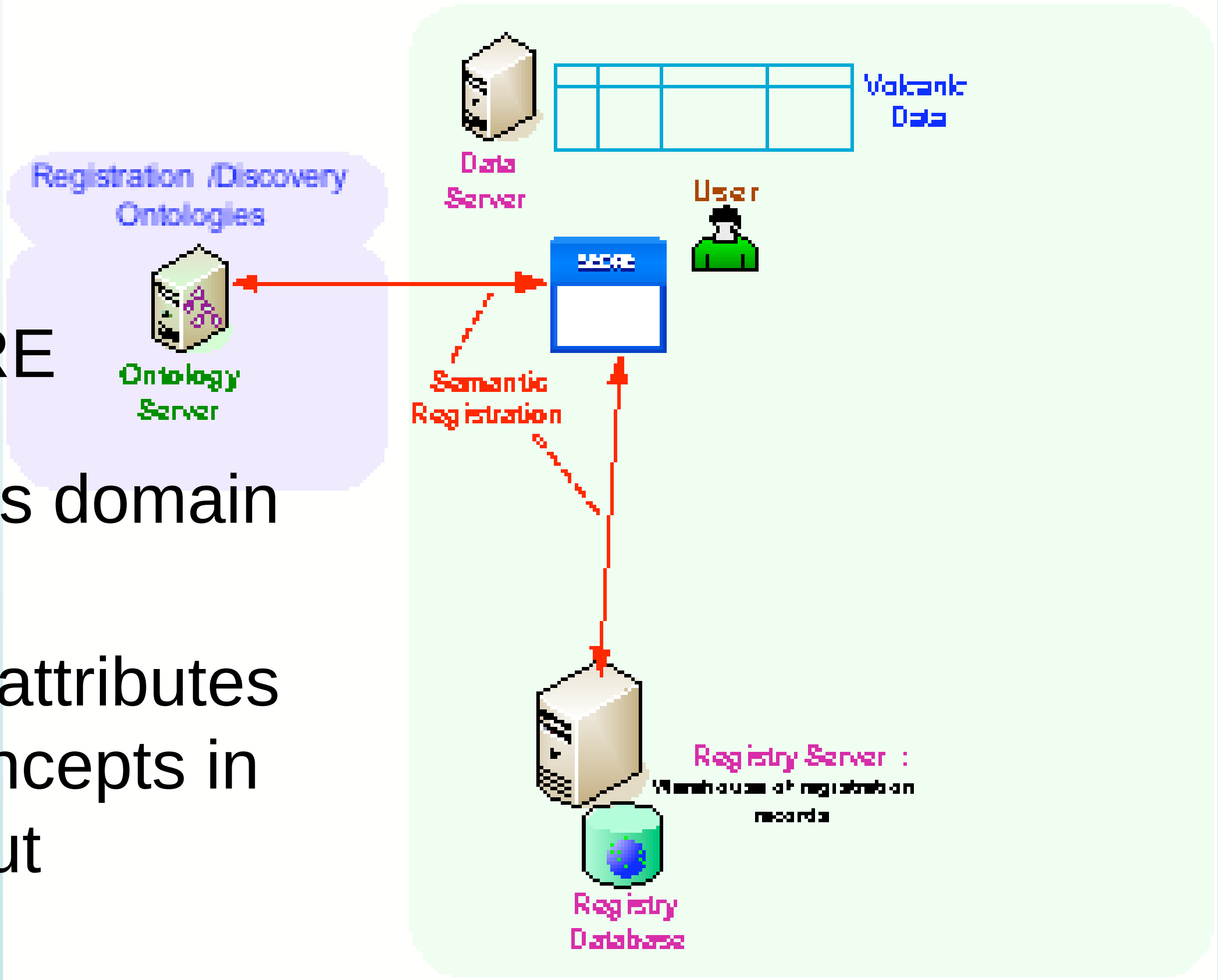
SEDRE: Semantically Enabled Data Registration Engine

- SEDRE: a system that enables scientists to semantically register data sets for optimal querying and semantic integration
- SEDRE enables mapping of heterogeneous data to concepts in domain ontologies

A. K. Sinha, A. Rezgui, Virginia Tech

Semantic Registration in SEDRE: An Overview

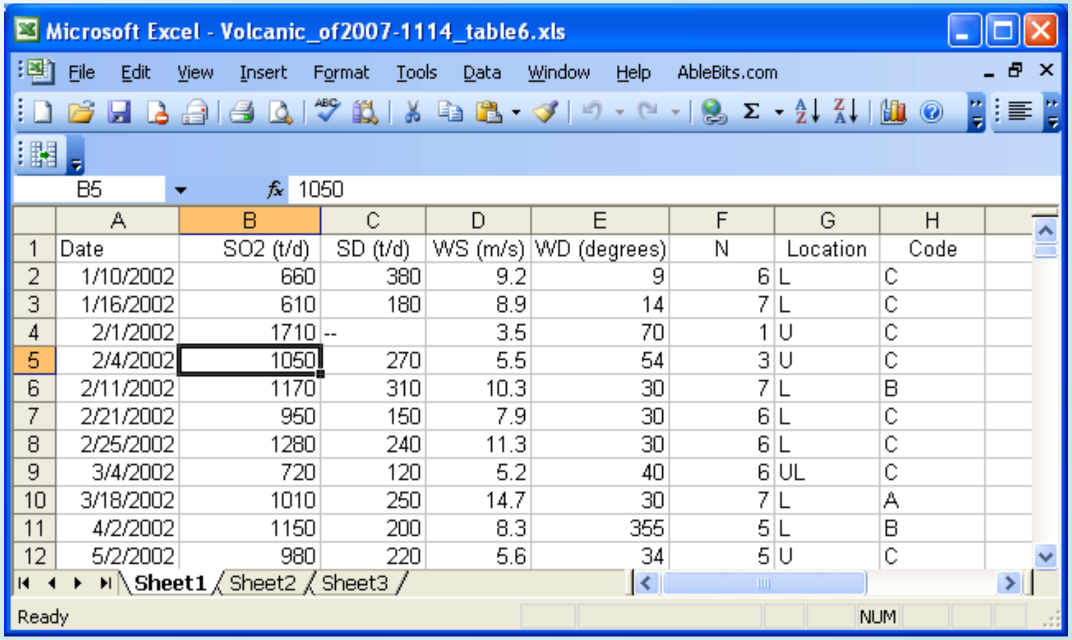
- SEDRE is a desktop application
- Users download and install SEDRE
- SEDRE accesses domain ontologies
- Users map data attributes (e.g., SO_2) to concepts in ontologies without ‘knowing it’



Example 1: Registration of Volcanic Data

Location Codes:

- U - Above the 180° turn at Holei Pali (upper Chain of Craters Road)
- L - Below Holei Pali (lower Chain of Craters Road)
- UL - Individual traverses were made both above and below the 180° turn at Holei Pali
- H - Highway 11

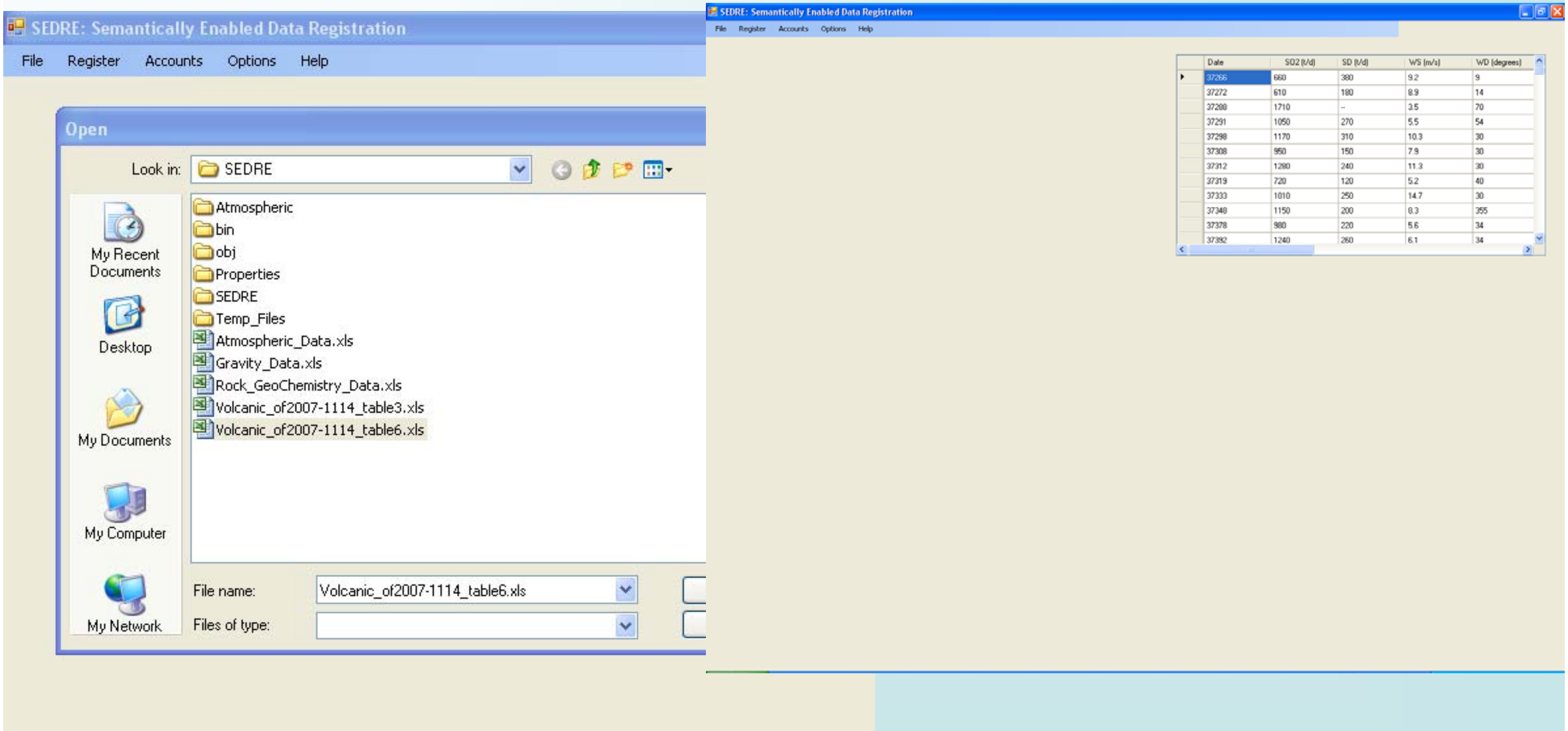

Microsoft Excel - Volcanic_of2007-1114_table6.xls
File Edit View Insert Format Tools Data Window Help AbleBits.com
B5
1050
Sheet1
Sheet2
Sheet3
Ready
NUM

	A	B	C	D	E	F	G	H
1	Date	SO2 (t/d)	SD (t/d)	WS (m/s)	WD (degrees)	N	Location	Code
2	1/10/2002	660	380	9.2	9	6	L	C
3	1/16/2002	610	180	8.9	14	7	L	C
4	2/1/2002	1710	--	3.5	70	1	U	C
5	2/4/2002	1050	270	5.5	54	3	U	C
6	2/11/2002	1170	310	10.3	30	7	L	B
7	2/21/2002	950	150	7.9	30	6	L	C
8	2/25/2002	1280	240	11.3	30	6	L	C
9	3/4/2002	720	120	5.2	40	6	UL	C
10	3/18/2002	1010	250	14.7	30	7	L	A
11	4/2/2002	1150	200	8.3	355	5	L	B
12	5/2/2002	980	220	5.6	34	5	U	C

SO_2 Emission from Kilauea east rift zone - vehicle-based (Source: HVO)

Abreviations: t/d=metric tonne (1000 kg)/day, SD=standard deviation, WS=wind speed, WD=wind direction east of true north, N=number of traverses

Loading Volcanic Data into SEDRE



Look in: SEDRE

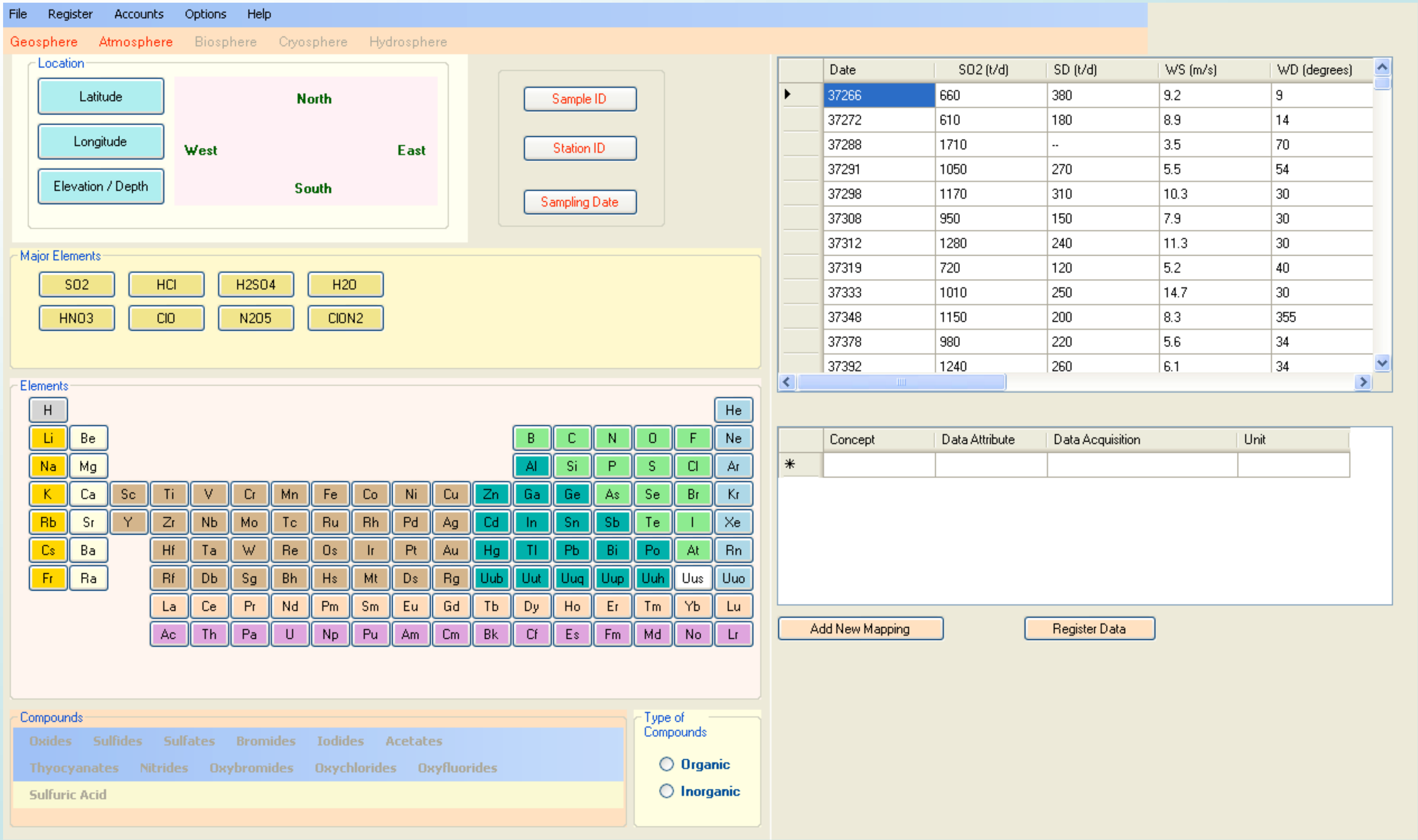
- Atmospheric
- bin
- obj
- Properties
- SEDRE
- Temp_Files
- Atmospheric_Data.xls
- Gravity_Data.xls
- Rock_GeoChemistry_Data.xls
- Volcanic_of2007-1114_table3.xls
- Volcanic_of2007-1114_table6.xls

File name: Volcanic_of2007-1114_table6.xls

Files of type:

Date	SO2 (t/d)	SD (t/d)	WS (m/s)	WD (degrees)
37266	660	380	9.2	9
37272	610	180	8.9	14
37288	1710	--	3.5	70
37291	1050	270	5.5	54
37298	1170	310	10.3	30
37308	950	150	7.9	30
37312	1280	240	11.3	30
37319	720	120	5.2	40
37333	1010	250	14.7	30
37348	1150	200	8.3	355
37378	980	220	5.6	34
37392	1240	260	6.1	34

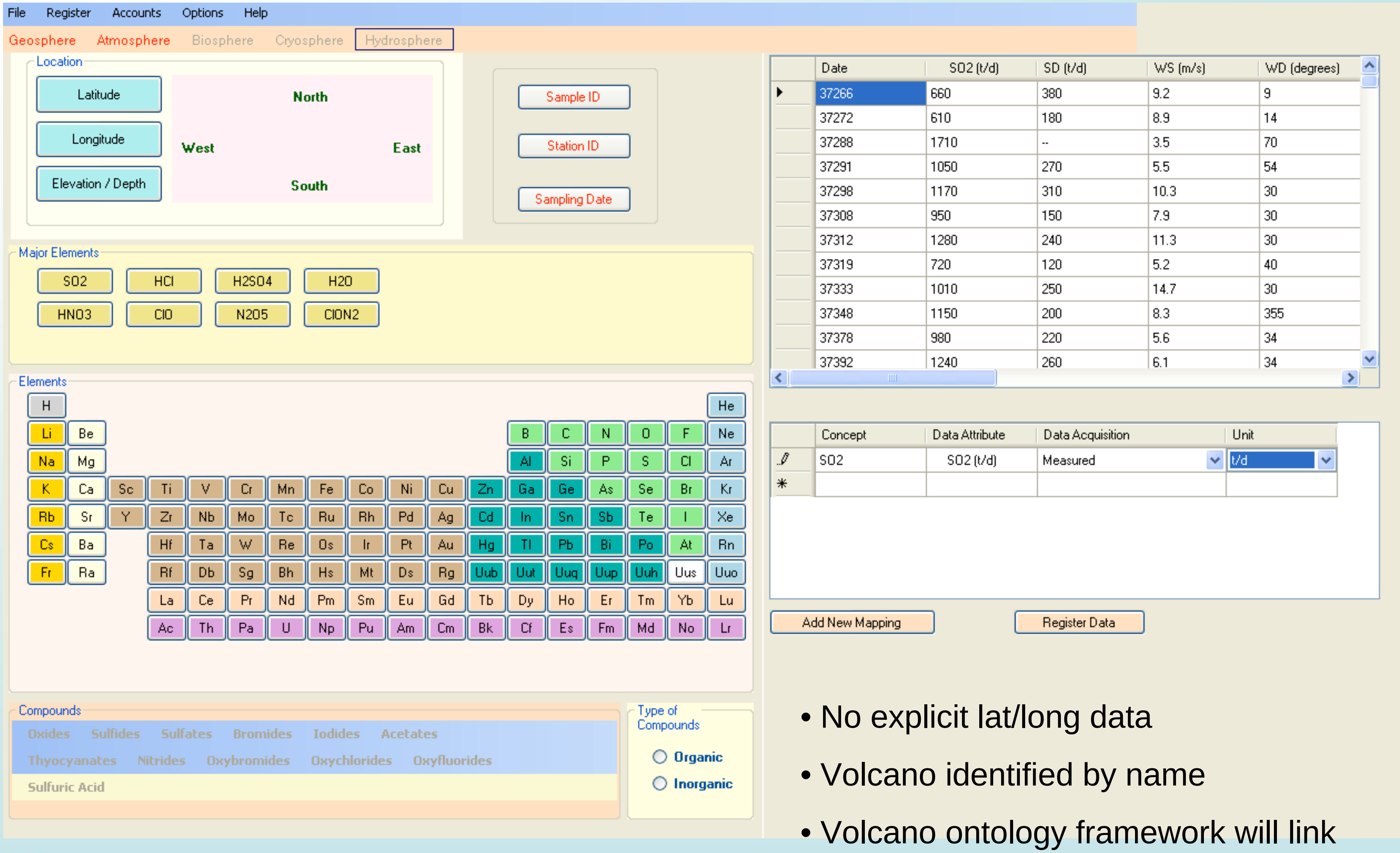
Registering Volcanic Data (1)



Date	SO2 (t/d)	SD (t/d)	WS (m/s)	WD (degrees)
37266	660	380	9.2	9
37272	610	180	8.9	14
37288	1710	--	3.5	70
37291	1050	270	5.5	54
37298	1170	310	10.3	30
37308	950	150	7.9	30
37312	1280	240	11.3	30
37319	720	120	5.2	40
37333	1010	250	14.7	30
37348	1150	200	8.3	355
37378	980	220	5.6	34
37392	1240	260	6.1	34

Concept	Data Attribute	Data Acquisition	Unit

Registering Volcanic Data (2)

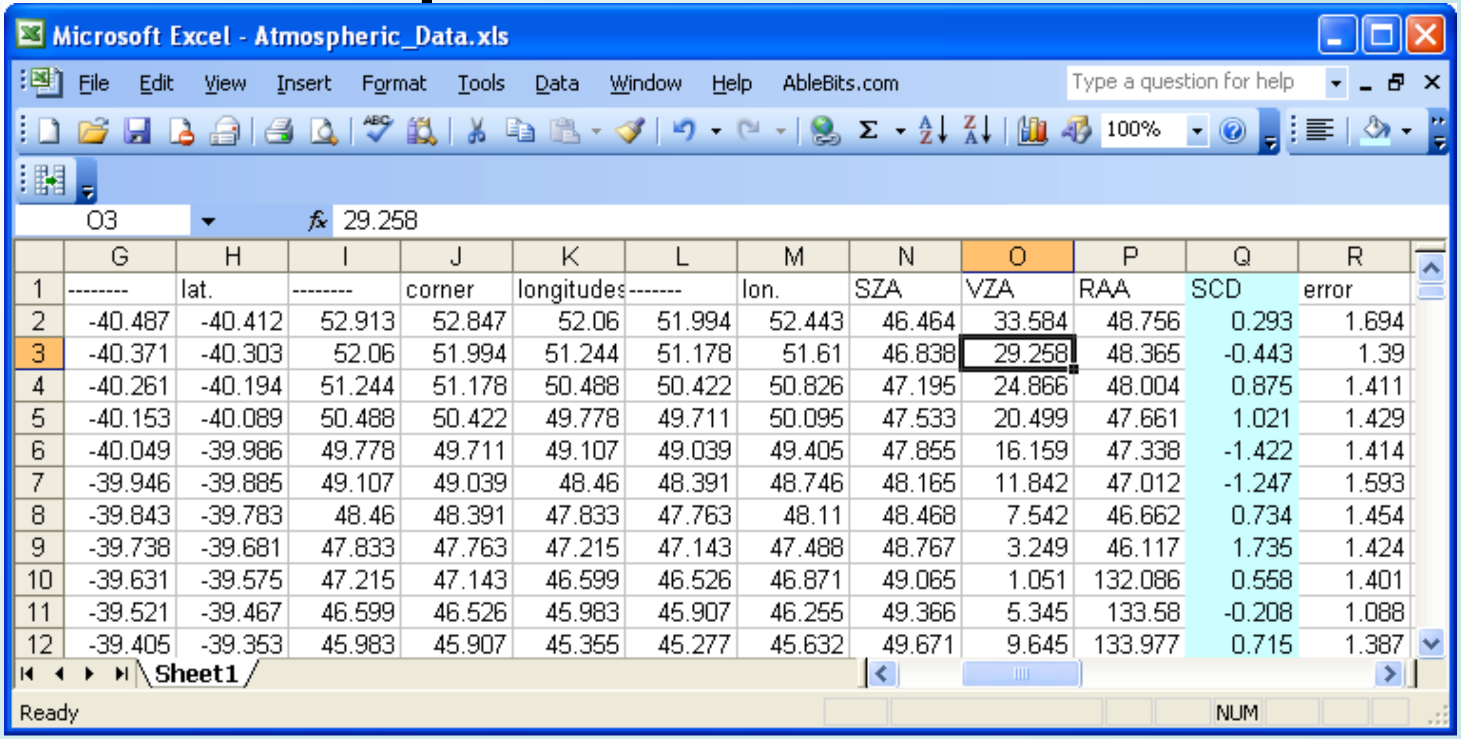


Date	SO2 (t/d)	SD (t/d)	WS (m/s)	WD (degrees)
37266	660	380	9.2	9
37272	610	180	8.9	14
37288	1710	--	3.5	70
37291	1050	270	5.5	54
37298	1170	310	10.3	30
37308	950	150	7.9	30
37312	1280	240	11.3	30
37319	720	120	5.2	40
37333	1010	250	14.7	30
37348	1150	200	8.3	355
37378	980	220	5.6	34
37392	1240	260	6.1	34

Concept	Data Attribute	Data Acquisition	Unit
SO2	SO2 (t/d)	Measured	t/d

- No explicit lat/long data
- Volcano identified by name
- Volcano ontology framework will link name to location

Example 2: Registration of Atmospheric Data



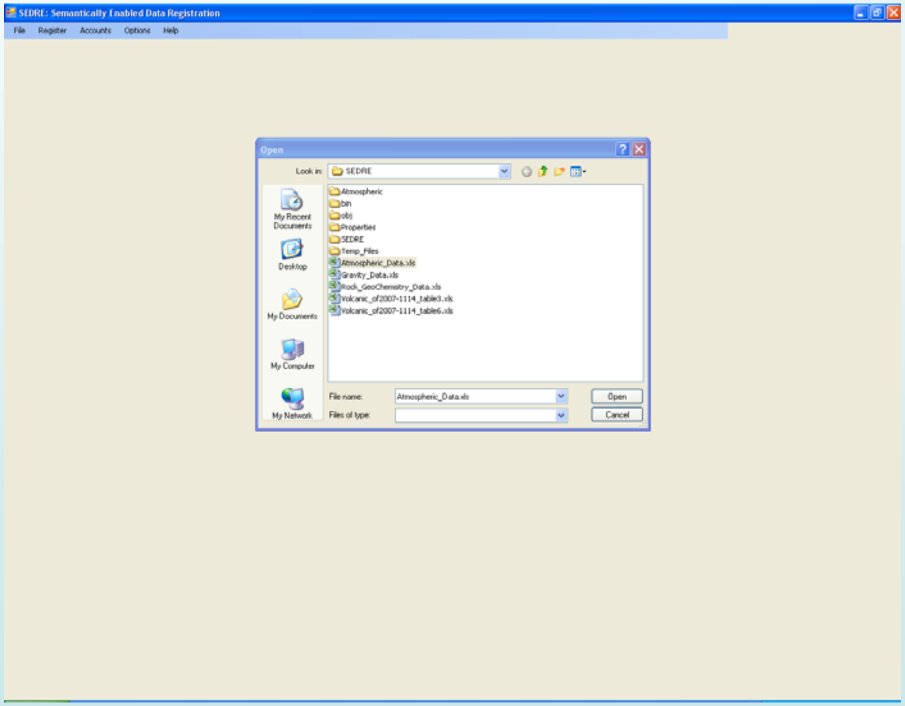
O3 fx 29.258

	G	H	I	J	K	L	M	N	O	P	Q	R
1	--------	lat.	--------	corner	longitudes	-------	lon.	SZA	VZA	RAA	SCD	error
2	-40.487	-40.412	52.913	52.847	52.06	51.994	52.443	46.464	33.584	48.756	0.293	1.694
3	-40.371	-40.303	52.06	51.994	51.244	51.178	51.61	46.838	29.258	48.365	-0.443	1.39
4	-40.261	-40.194	51.244	51.178	50.488	50.422	50.826	47.195	24.866	48.004	0.875	1.411
5	-40.153	-40.089	50.488	50.422	49.778	49.711	50.095	47.533	20.499	47.661	1.021	1.429
6	-40.049	-39.986	49.778	49.711	49.107	49.039	49.405	47.855	16.159	47.338	-1.422	1.414
7	-39.946	-39.885	49.107	49.039	48.46	48.391	48.746	48.165	11.842	47.012	-1.247	1.593
8	-39.843	-39.783	48.46	48.391	47.833	47.763	48.11	48.468	7.542	46.662	0.734	1.454
9	-39.738	-39.681	47.833	47.763	47.215	47.143	47.488	48.767	3.249	46.117	1.735	1.424
10	-39.631	-39.575	47.215	47.143	46.599	46.526	46.871	49.065	1.051	132.086	0.558	1.401
11	-39.521	-39.467	46.599	46.526	45.983	45.907	46.255	49.366	5.345	133.58	-0.208	1.088
12	-39.405	-39.353	45.983	45.907	45.355	45.277	45.632	49.671	9.645	133.977	0.715	1.387

Satellite data for SO_2 emissions

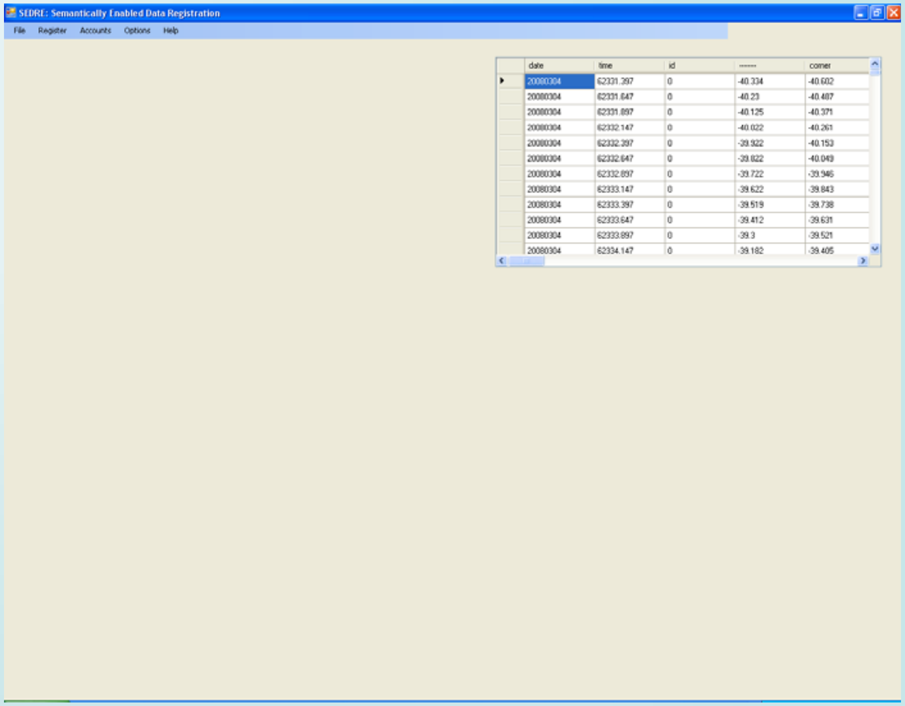
Abbreviation: SCD: Slant Column Density (in Dobson Unit (DU))

Loading Atmospheric Data into SEDRE



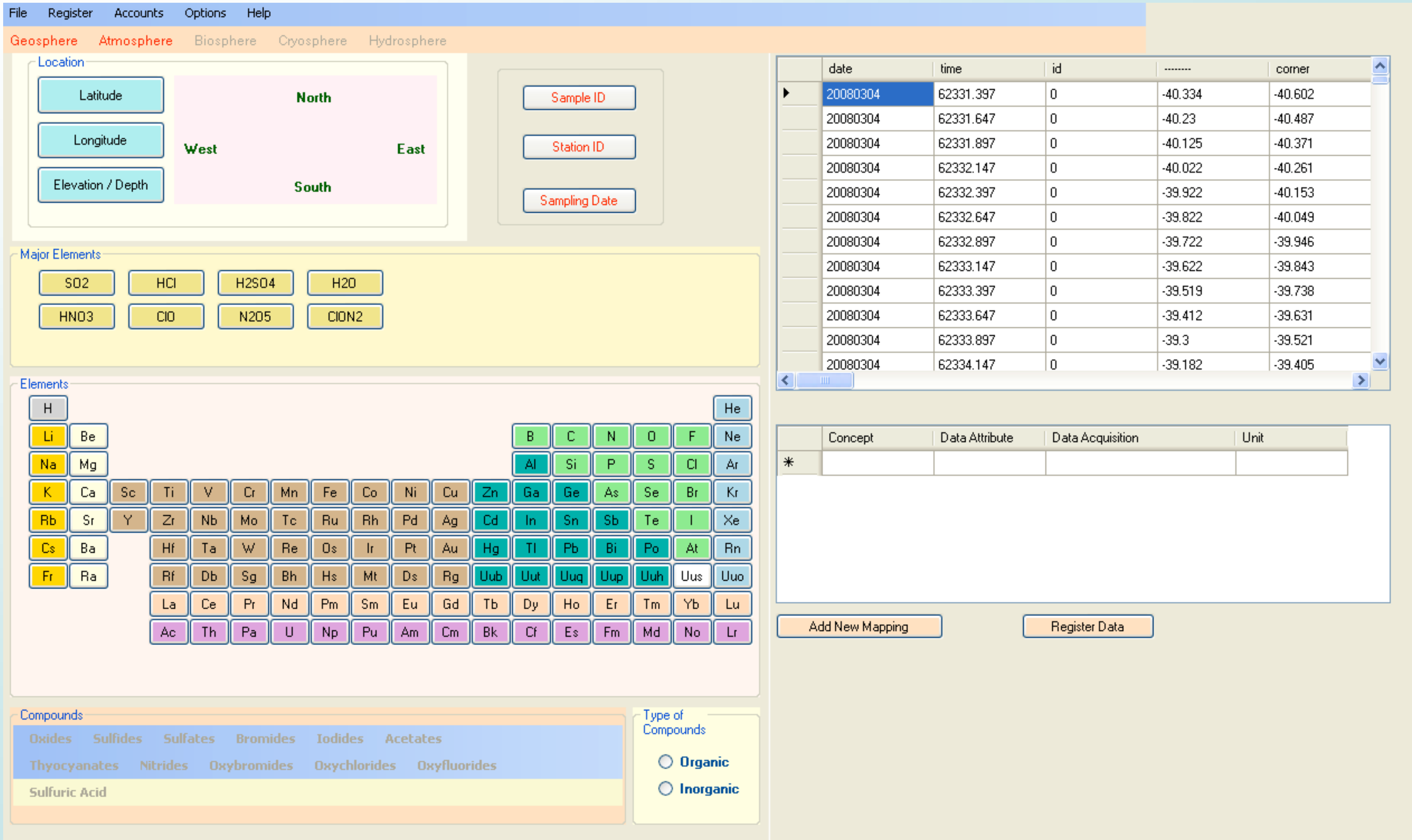
File name: Atmospheric_Data.xls

Files of type:



date	time	id	-------	corner
20080304	62331.397	0	-40.334	-40.602
20080304	62331.647	0	-40.23	-40.487
20080304	62331.897	0	-40.125	-40.371
20080304	62332.147	0	-40.022	-40.261
20080304	62332.397	0	-39.922	-40.153
20080304	62332.647	0	-39.822	-40.049
20080304	62332.897	0	-39.722	-39.946
20080304	62333.147	0	-39.622	-39.843
20080304	62333.397	0	-39.519	-39.738
20080304	62333.647	0	-39.412	-39.631
20080304	62333.897	0	-39.3	-39.521
20080304	62334.147	0	-39.182	-39.405

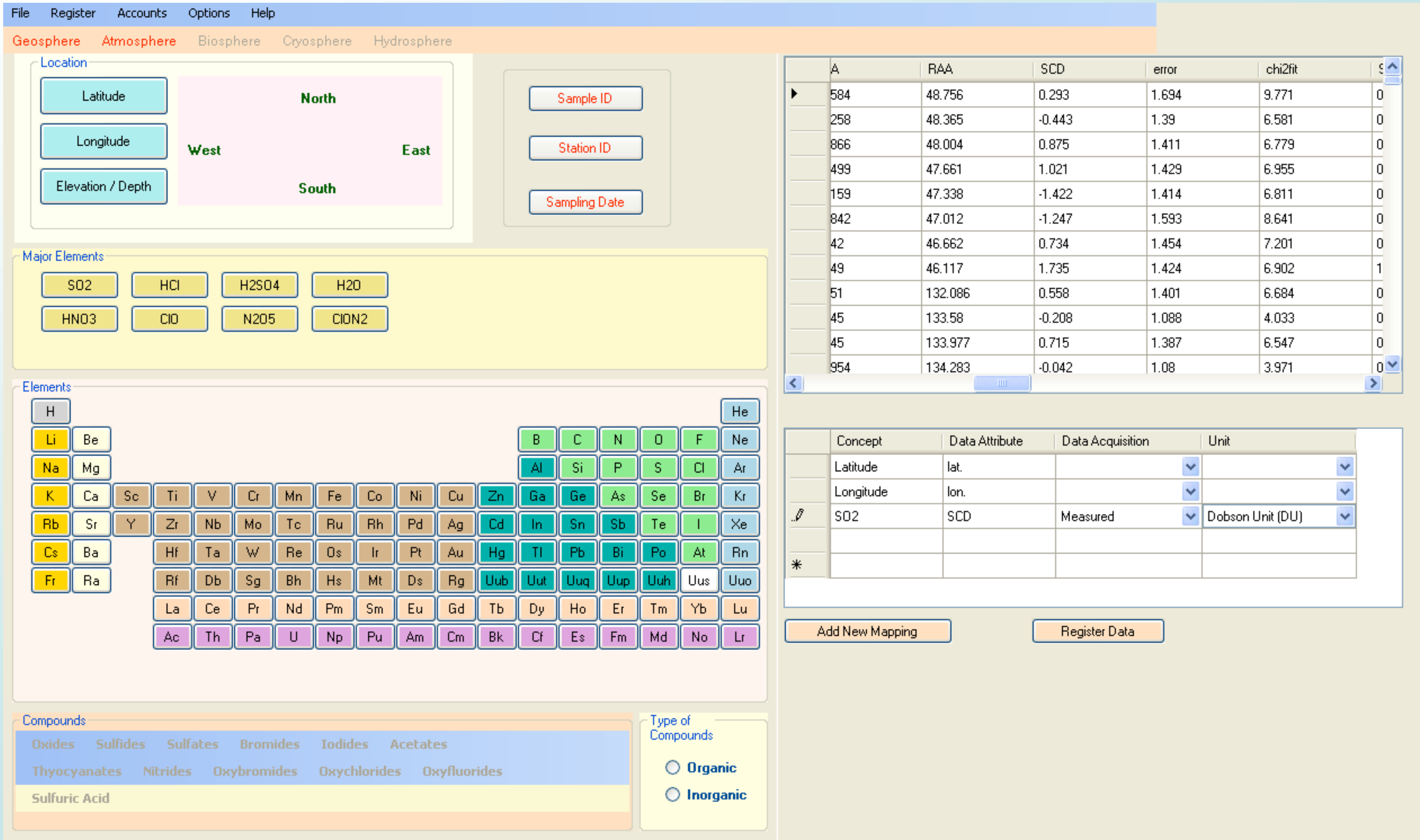
Registering Atmospheric Data (1)



	date	time	id	--------	corner
▶	20080304	62331.397	0	-40.334	-40.602
	20080304	62331.647	0	-40.23	-40.487
	20080304	62331.897	0	-40.125	-40.371
	20080304	62332.147	0	-40.022	-40.261
	20080304	62332.397	0	-39.922	-40.153
	20080304	62332.647	0	-39.822	-40.049
	20080304	62332.897	0	-39.722	-39.946
	20080304	62333.147	0	-39.622	-39.843
	20080304	62333.397	0	-39.519	-39.738
	20080304	62333.647	0	-39.412	-39.631
	20080304	62333.897	0	-39.3	-39.521
	20080304	62334.147	0	-39.182	-39.405

	Concept	Data Attribute	Data Acquisition	Unit
*				

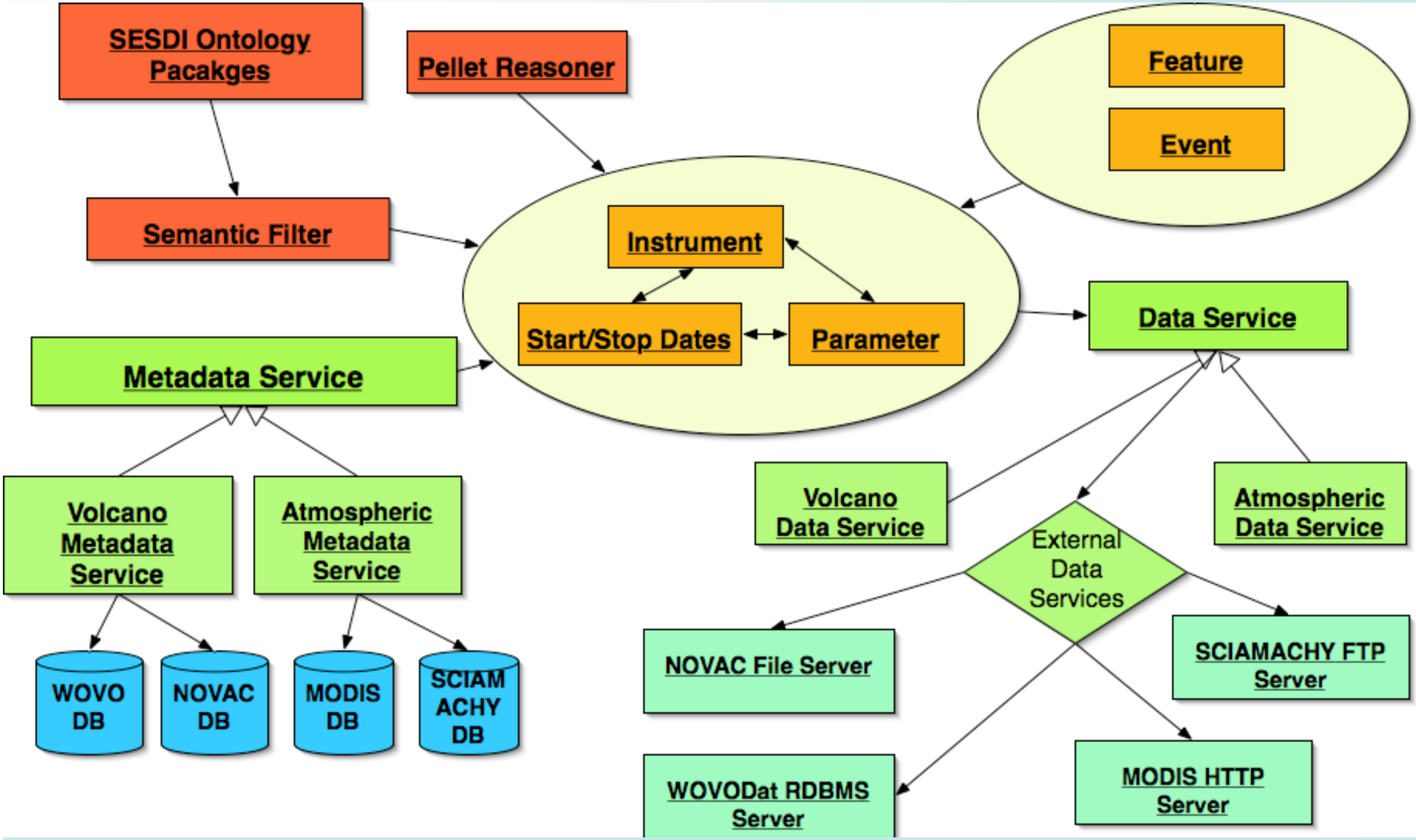
Registering Atmospheric Data (2)



A	RAA	SCD	error	chi2fit	S
584	48.756	0.293	1.694	9.771	0
258	48.365	-0.443	1.39	6.581	0
866	48.004	0.875	1.411	6.779	0
499	47.661	1.021	1.429	6.955	0
159	47.338	-1.422	1.414	6.811	0
842	47.012	-1.247	1.593	8.641	0
42	46.662	0.734	1.454	7.201	0
49	46.117	1.735	1.424	6.902	1
51	132.086	0.558	1.401	6.684	0
45	133.58	-0.208	1.088	4.033	0
45	133.977	0.715	1.387	6.547	0
954	134.283	-0.042	1.08	3.971	0

Concept	Data Attribute	Data Acquisition	Unit
Latitude	lat.		
Longitude	lon.		
SO2	SCD	Measured	Dobson Unit (DU)

Semantic framework indicating how volcano and atmospheric parameters and databases can immediately be plugged in to the semantic data framework to enable data integration.



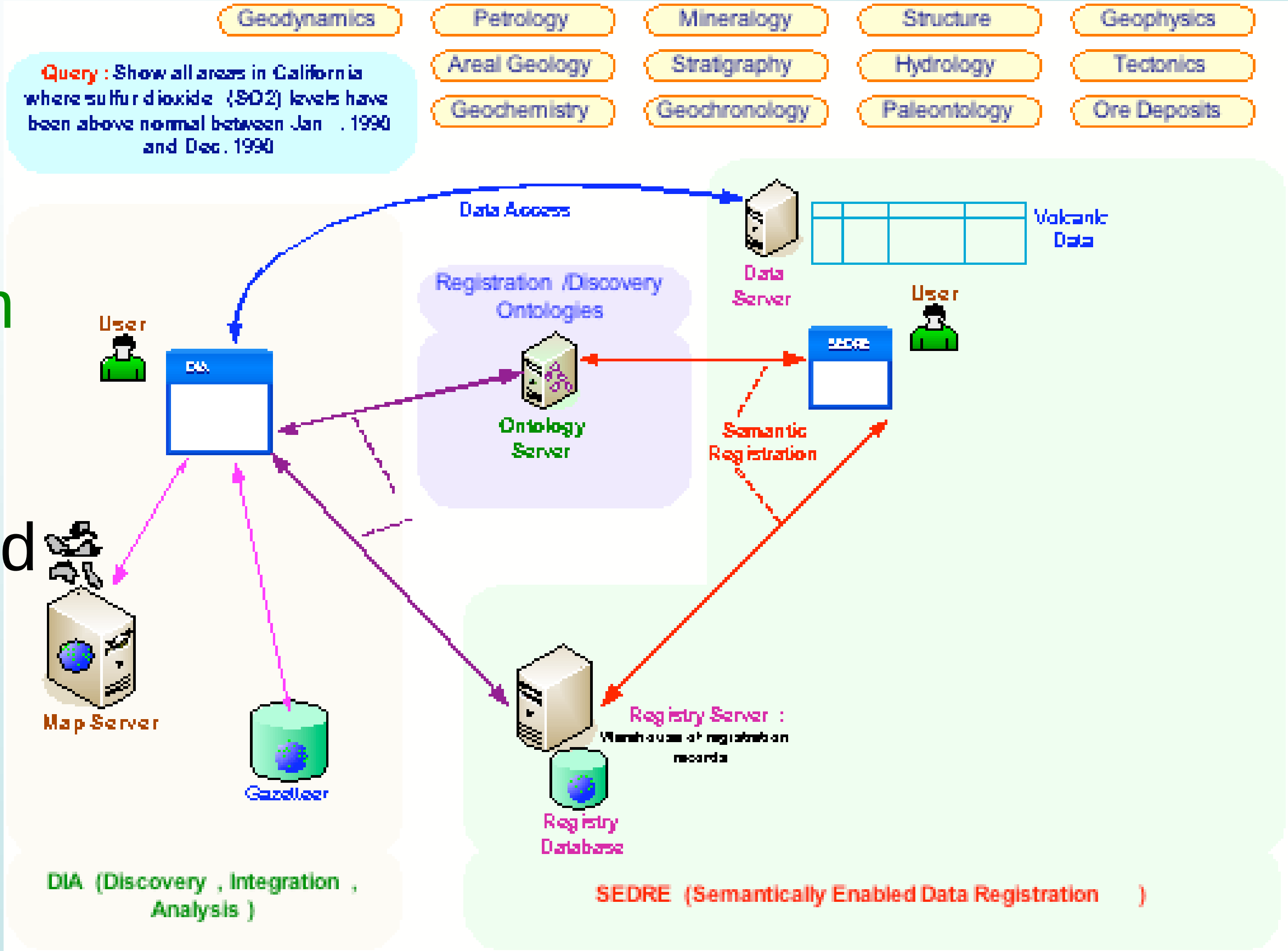
Summary and outlook

- Semantic data frameworks/ technologies are changing the landscape of providing data to scientists (in a good way)
- Tools for data registration are soon to be available
- Applications to perform data integration mediated by semantics are available
- Initial results - applied to two volcanoes - led to correlation of SO2 concentration from volcano and in the atmosphere and relation to H2SO4
- Solar radiation ontologies and data sources are in progress

SEDRE+DIA: Overview

DIA: Web-based System for Data Discovery, Integration and Analysis

(Developed at Virginia Tech through NSF funding)



Query : Show all areas in California where sulfur dioxide (SO2) levels have been above normal between Jan. 1990 and Dec. 1990

General applicability

- To apply to another use case: