An Earth Science Ontology Dialog

Value Proposition of Ontology and Semantic Technology for the Earth Science Community

How Can Semantics Change Data Practices of the EarthCube Geoscience Community?

Krishna Sinha
Geological Sciences
Virginia Tech
pitlab@vt.edu

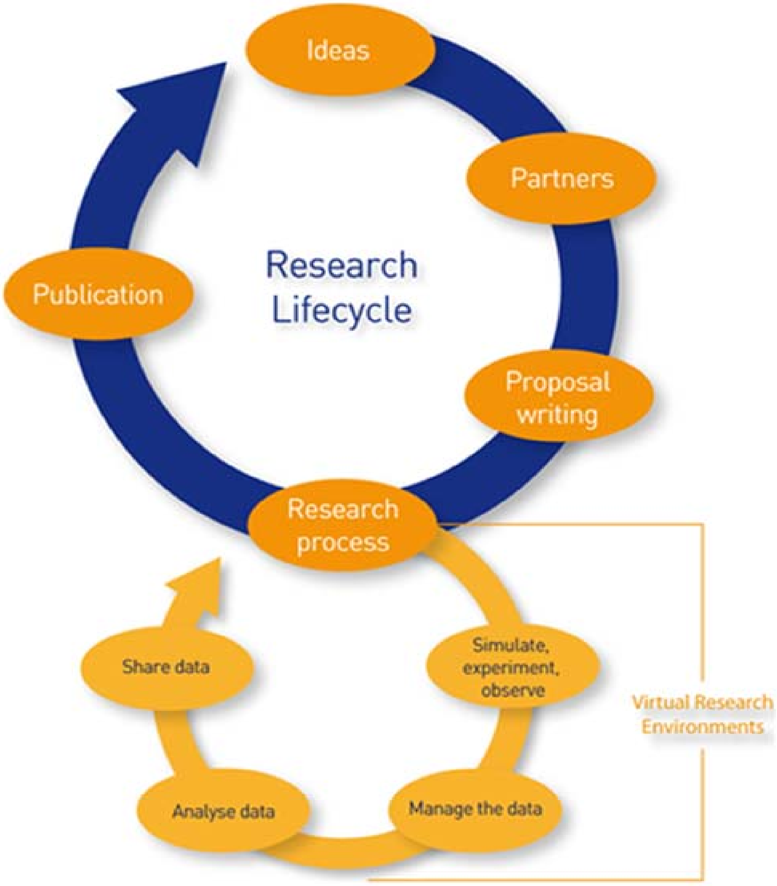
Current Status of Geoscience community data practices

- Data and tool sharing practices face many barriers, and are especially dominant in the world of individual researchers (long tail of science community)
- However, these data are required to understand how natural systems (***earth systems***) change over time through physical, chemical, and biological processes.
- Reasons: lack of time, future publishing opportunities, ownership of data, concerns related to misuse of data, credit for professional advancement, lack of institutional support, and opportunities for commercial applications.

INTRODUCTION

- Data Life Cycle : involvement of individuals
- Complexity of data: earth as a system
- Data Environments : long tail of science
- Sharing data: Credit and motivation
- Motivational Integration scenario: multiple data resources and increased efficiency
- Sharing data at many levels of domain ontology
- Individuals have more than data to share: ontology classes
- Stages in building the semantic infrastructure

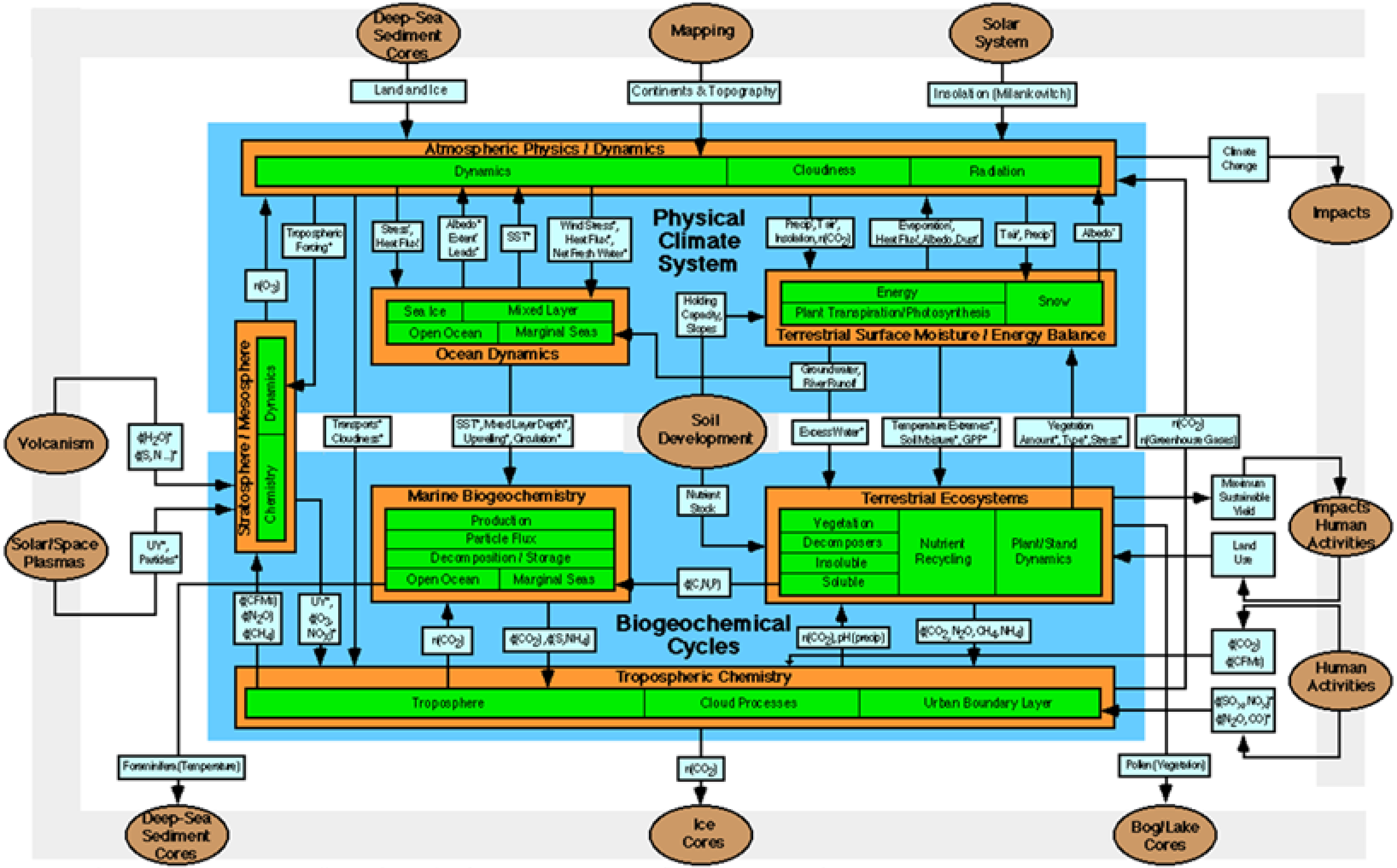
Research and Data Lifecycles



Tenopir C, Allard S, Douglass K, Aydinoglu AU, et al. (2011) Data Sharing by Scientists: Practices and Perceptions. PLoS ONE 6(6): e21101. doi:10.1371/journal.pone.0021101
ne.org/article/info:doi/10.1371/journal.pone.0021101

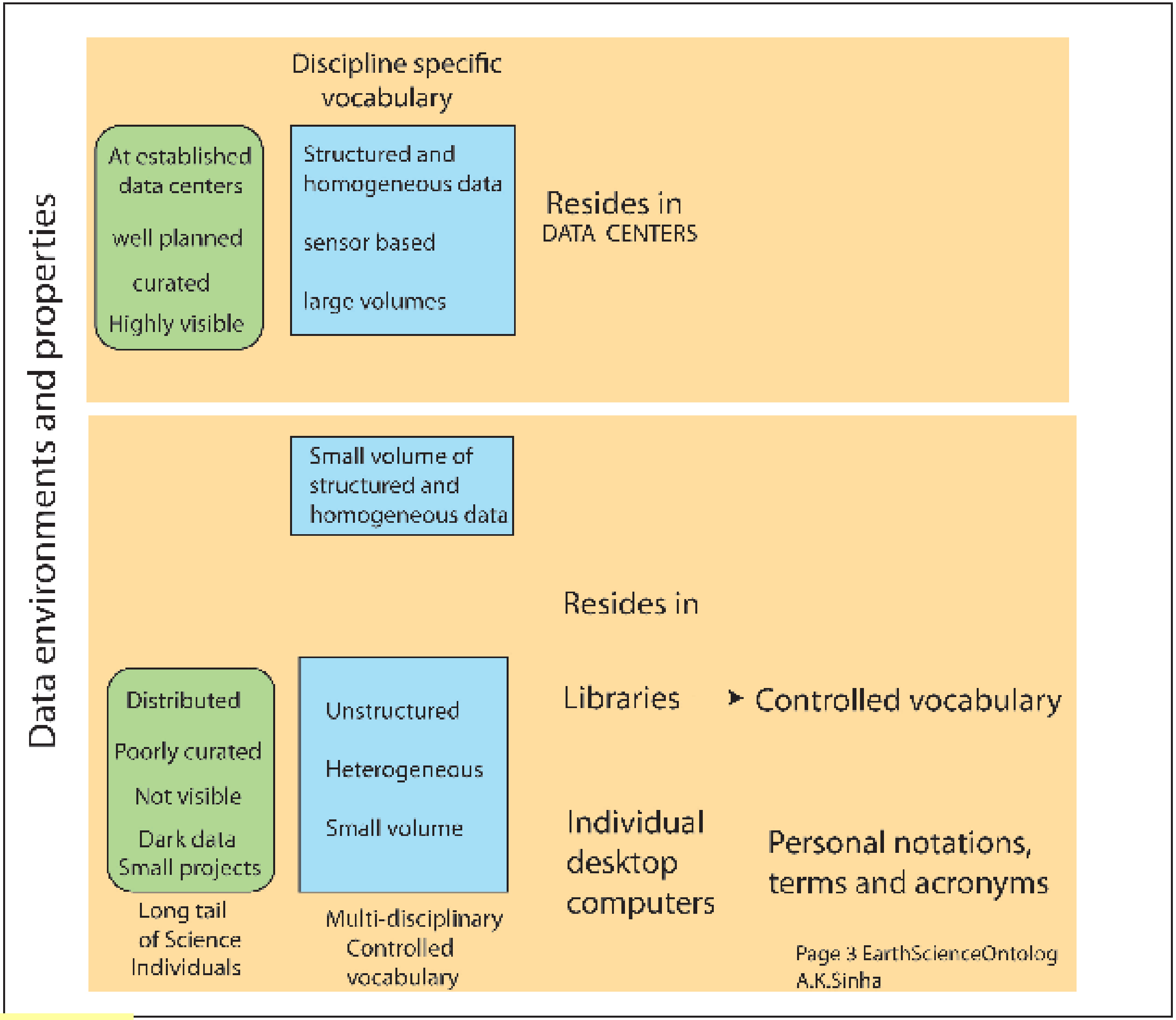
Bretherton Diagram demonstrates the dynamic interaction of systems: the real world complexity of data and sources

CONCEPTUAL MODEL of Earth System process operating on timescales of decades to centuries



' = on timescale of hours to days * = on timescale of months to seasons ϕ = flux n = concentration

Data Environments



Data environments and properties

Discipline specific vocabulary

At established data centers

well planned

curated

Highly visible

Structured and homogeneous data

sensor based

large volumes

Resides in DATA CENTERS

Small volume of structured and homogeneous data

Resides in

Distributed

Poorly curated

Not visible

Dark data

Small projects

Long tail of Science Individuals

Unstructured

Heterogeneous

Small volume

Multi-disciplinary Controlled vocabulary

Libraries ➤ Controlled vocabulary

Individual desktop computers

Personal notations, terms and acronyms

Page 3 EarthScienceOntolog
A.K.Sinha

Discovery, Reuse, and Credit
Mooney and Newton, 2012

- **scientists agree that data sharing is a desirable practice in theory, the fear of receiving no credit and losing funding or publishing opportunities is a serious deterrent to actual practice**
- **fear that sharing data could result in someone else publishing with no reward given to the sharer since there is no system of acknowledgement**,
- **some fairly famous cell lines were generated by obscure people**

- **What constitutes an adequate data citation**

The Anatomy of a Data Citation: Discovery, Reuse, and Credit

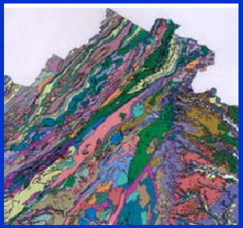
- **Reward structures must be in place to encourage data publication, and citation is the appropriate tool for scholarly acknowledgment. Data citation also allows for the identification, retrieval, replication, and verification of data underlying published studies**
- **Promotion of data citation will foster a scholarly communication system that allows for identification, retrieval,**
- **Repositories publishing data should include appropriate metadata and mandate citations as a condition of reuse.**
- **Normalizing expectations for dataset citation will incentivize data sharing and promote secondary research,**

Incentive through providing easy problem solving environment : use case exemplars resolved through ontologies

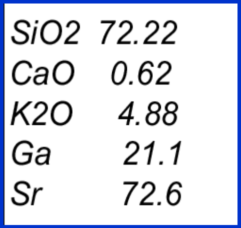
A Geoscientist’s Information Integration Scenario

What is the distribution and U/ Pb zircon ages of A-type plutons in VA?
How about their 3-D geometry using gravity data ?

?
Information Integration

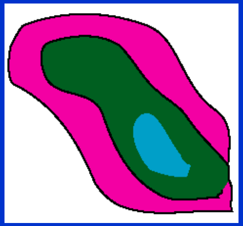


Digital geologic map
Of Virginia
(plutons in Virginia)

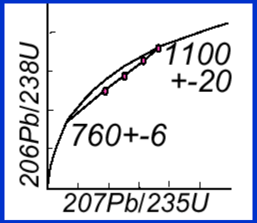


SiO2	72.22
CaO	0.62
K2O	4.88
Ga	21.1
Sr	72.6

Geochemical
Database
(chemical data)



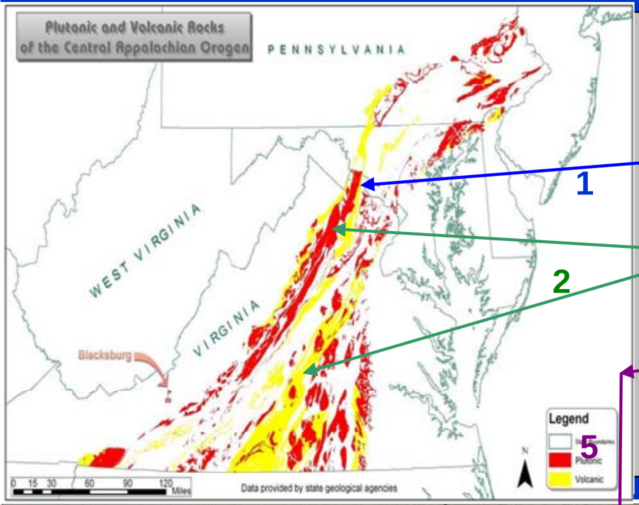
Geophysical Database
(gravity contours)



Geochronologic
database
(Concordia)

Increased efficiency welcomes data sharing

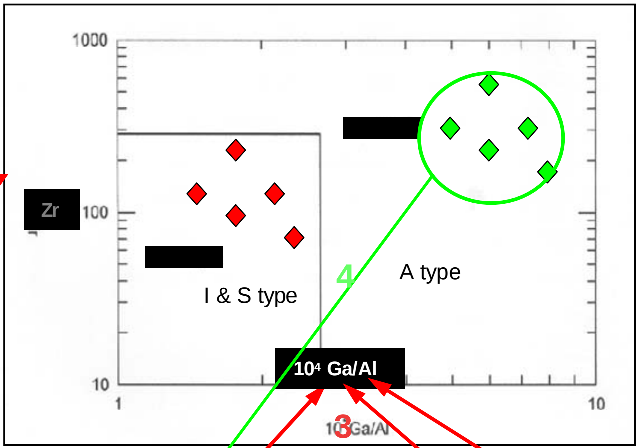
Integration Scenario: Stages for access to data and tools in a workflow environment

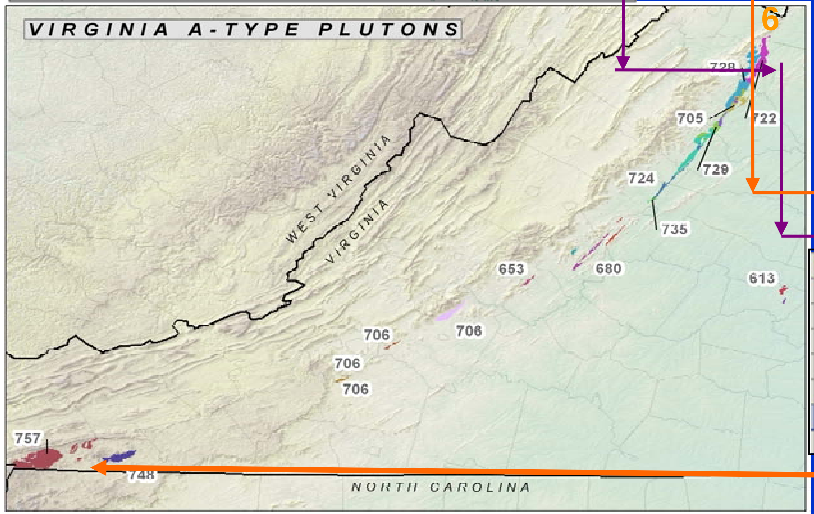

Plutonic and Volcanic Rocks of the Central Appalachian Orogen
PENNSYLVANIA
WEST VIRGINIA
VIRGINIA
Blacksburg
1
2
5
Legend
Plutonic
Volcanic
0 15 30 60 90 120 Miles
Data provided by state geological agencies

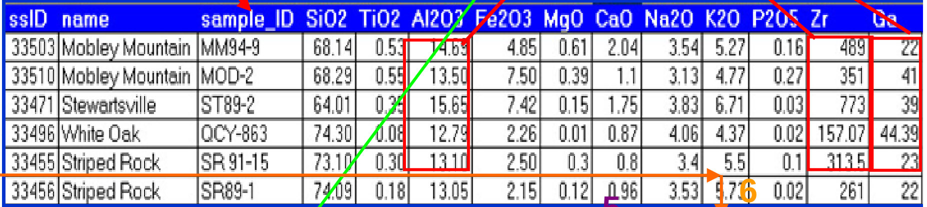
Ontology System

- Location
  - States
    - Virginia
- Classification System
  - Rock Classifiers
    - Igneous
      - Pluton
        - A-type
- Mineral
  - Zircon
- Geologic Time
  - Dating Methods
    - U-Pb Zircon Methods

3

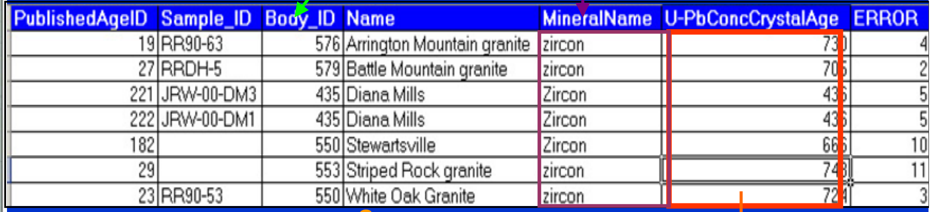






ssID	name	sample_ID	SiO2	TiO2	Al2O3	Fe2O3	MgO	CaO	Na2O	K2O	P2O5	Zr	Ga
33503	Mobley Mountain	MM94-9	68.14	0.53	14.63	4.85	0.61	2.04	3.54	5.27	0.16	489	22
33510	Mobley Mountain	MOD-2	68.29	0.55	13.50	7.50	0.39	1.1	3.13	4.77	0.27	351	41
33471	Stewartsville	ST89-2	64.01	0.35	15.65	7.42	0.15	1.75	3.83	6.71	0.03	773	39
33496	White Oak	QCY-863	74.30	0.08	12.79	2.26	0.01	0.87	4.06	4.37	0.02	157.07	44.39
33455	Striped Rock	SR 91-15	73.10	0.30	13.10	2.50	0.3	0.8	3.4	5.5	0.1	313.5	23
33456	Striped Rock	SR89-1	74.09	0.18	13.05	2.15	0.12	0.96	3.53	5.7[illegible]	0.02	261	22

5 6



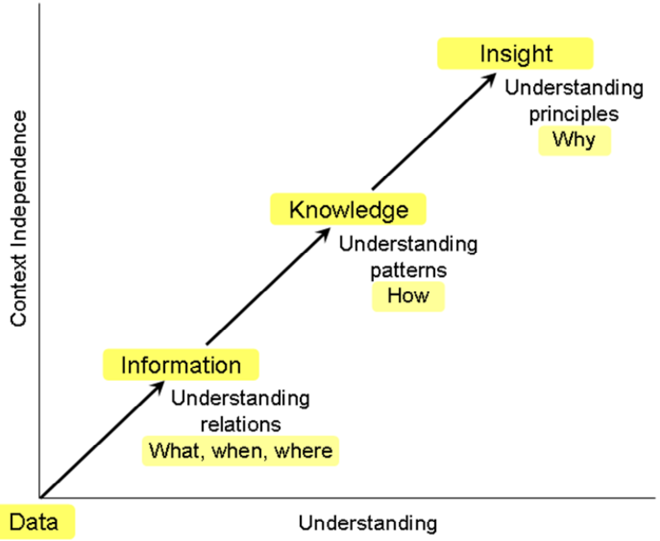
PublishedAgeID	Sample_ID	Body_ID	Name	MineralName	U-PbConcCrystalAge	ERROR
19	RR90-63	576	Arrington Mountain granite	zircon	730	4
27	RRDH-5	579	Battle Mountain granite	zircon	705	2
221	JRW-00-DM3	435	Diana Mills	Zircon	436	5
222	JRW-00-DM1	435	Diana Mills	Zircon	436	5
182		550	Stewartsville	Zircon	666	10
29		553	Striped Rock granite	zircon	748	11
23	RR90-53	550	White Oak Granite	zircon	724	3

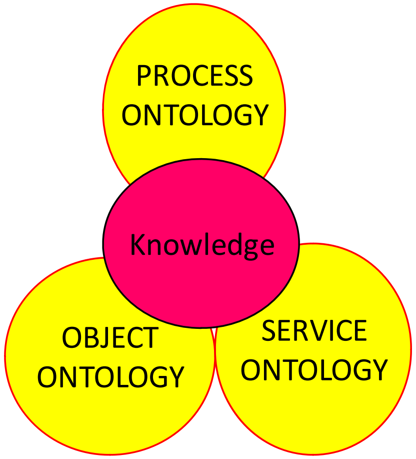
6

Krishna Sinha, Geological Sciences, Virginia Tech, pitlab@vt.edu

Providing the ontology framework

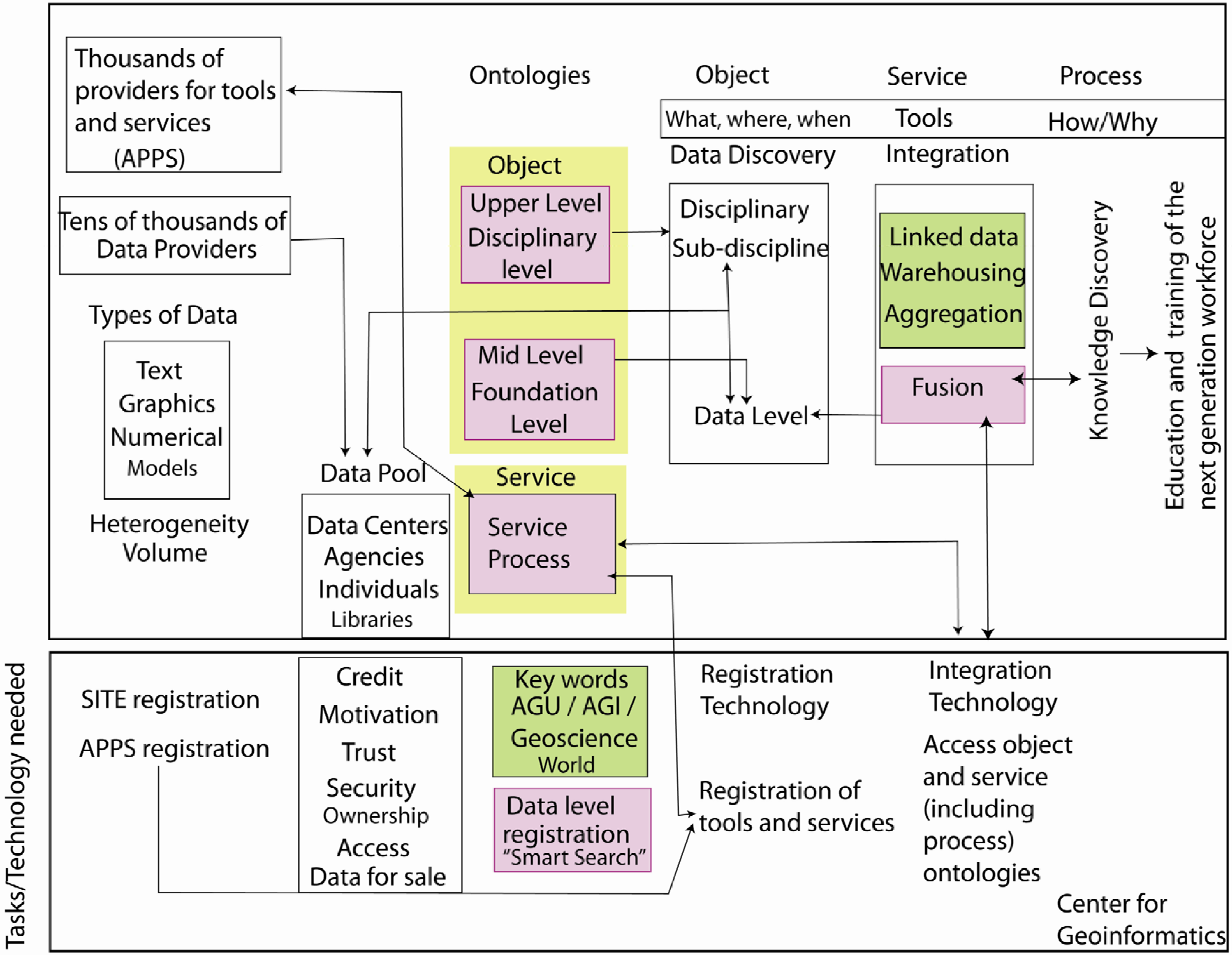
Sharing Geoscience Knowledge in a semantic world: Requires three classes of ontology frameworks





- **Objects** represent our understanding of the state of the system when the data were acquired, while **processes** capture the physical and chemical forcings on objects that may lead to changes in state and condition over time. **Service** provides tools (e.g., simulation models and analysis algorithms) to assess multiple hypotheses, including inference or prediction.
- These three classes of ontologies within the semantic layer of the scientific cyberinfrastructure are thus required to enable automated discovery, analysis, utilization, and understanding of data through both induction and deduction

Semantic infrastructure development stages: data to knowledge pathway



Thousands of providers for tools and services (APPS)

Tens of thousands of Data Providers

Types of Data

Text
Graphics
Numerical
Models

Heterogeneity
Volume

Data Pool

Data Centers
Agencies
Individuals
Libraries

Ontologies

Object

Upper Level Disciplinary level

Mid Level Foundation Level

Service

Service Process

Object | Service | Process

What, where, when | Tools | How/Why

Data Discovery

Disciplinary
Sub-discipline
Data Level

Integration

Linked data
Warehousing
Aggregation

Fusion

Knowledge Discovery

Education and training of the next generation workforce

Tasks/Technology needed

SITE registration

APPS registration

Credit
Motivation
Trust
Security
Ownership
Access
Data for sale

Key words AGU / AGI / Geoscience World

Data level registration "Smart Search"

Registration Technology

Registration of tools and services

Integration Technology

Access object and service (including process) ontologies

Center for Geoinformatics

Legacy data Services

Standards Services

Open Access Services

Ontology Services

Registration services

Integration Services

Preservation Services

Suggestions

- This community to endorse data citation
- Use real world use cases that long tail scientists relate to, and recognized increase in efficiency through semantics is likely to promote data sharing (aspects of cost benefit)