Managing Scientific Data

From Data Integration to Scientific Workflow

Annies Ductan

Discussion Outline

- **Introduction to Scientific Data Management**
- **Integration Challenges**
- **Integration Examples**
  - Data-centric: *Geologic-Map Data Integration*
  - Process-centric: *Mineral Classification Workflow*
- **Data Integration**
  - Mediator Approach
  - Semantic Mediation
  - Semantic Data Registration
- **Scientific Workflows**
  - Scientific Workflows In Kepler
  - Gravity Modeling Workflow
  - Semantic Workflow Extensions

Introduction

Why is there a need for managing scientific data:

There are a number of integration challenges that need to be overcome:

- Enabling more data-driven “e-science”
- Providing scientist with the right tools so that they spend less time on labor-intensive, error-prone manual data management.

This paper is an overview of the different kinds of data integration and interoperability challenges in scientific data management

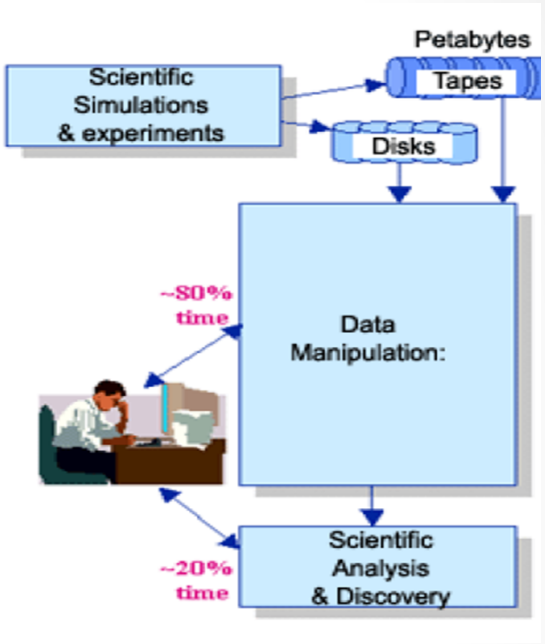
Introduction

Importance of Interoperability:

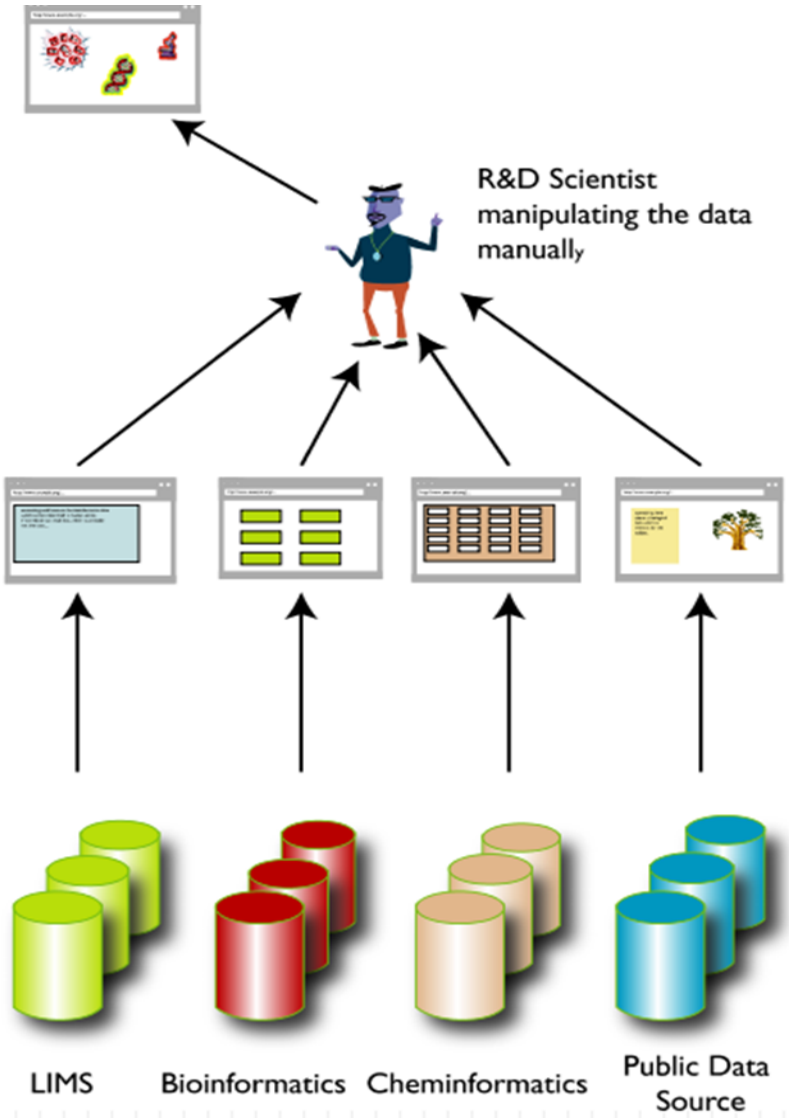
- Reduces operational cost and complexity:
  - Customers will continue to have mixed environments
  - Enabling systems to work together reduces the cost of building and supporting a heterogeneous infrastructure
- Enables “best-of-breed” deployments:
  - Customers may have business requirements that can only be delivered with specific applications or platforms
- Leverages existing investments:
  - Customers have a large and diverse range of systems installed in their environments. Moving to new platforms needs to be gradual

Introduction

- Current integration strategies:
  - Scientist often repeat tasks and data management steps:
    - Select appropriate analysis methods to address a question
    - Search relevant data and create new data
    - Determine whether existing data can be used for the desired analyses
    - Pre-processing and integrating data as needed



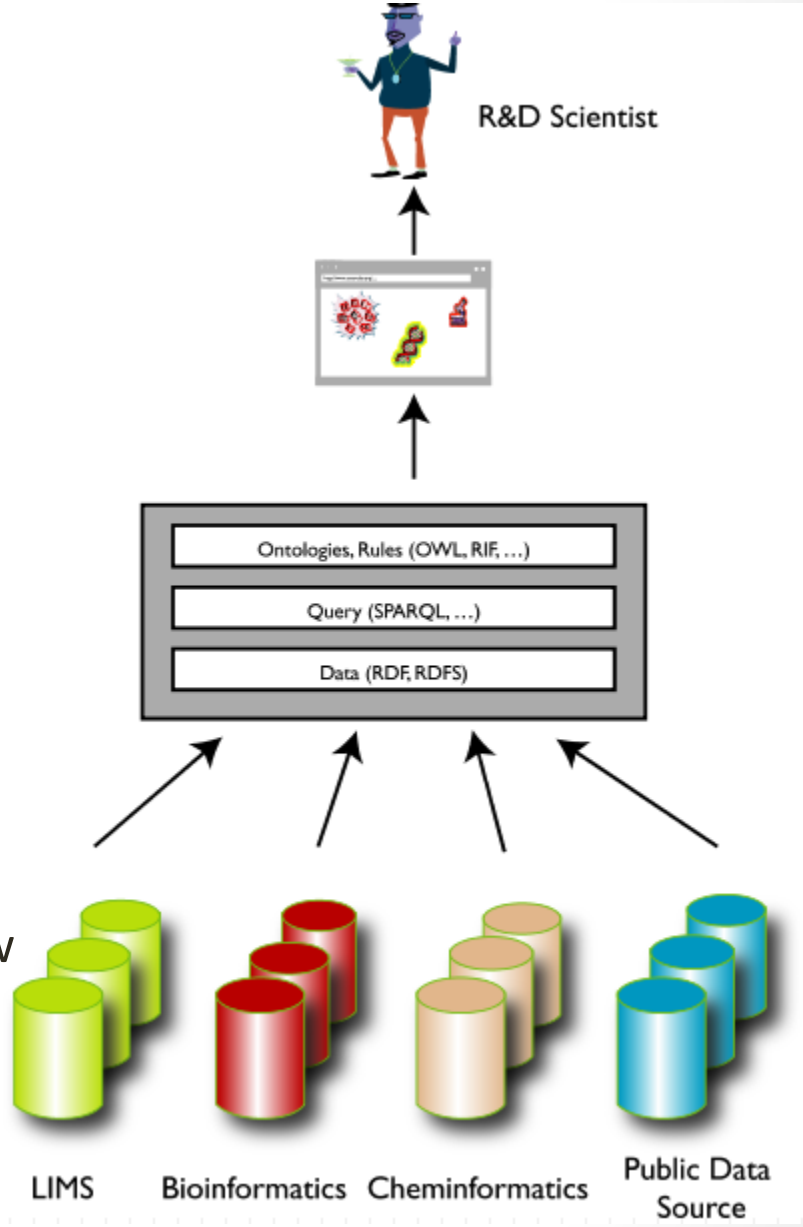
Integration Challenges



- Scientist need to access data and information provided via:
  - Community databases
  - Analytical tools developed by community members
- Easy way to make use of the increasing number of databases, analytical tools, and computational services
- Creating techniques to leverage these resources for data integration

Integration Challenges

- The goal:
  - Provide a framework that can be utilized by scientist and engineers for creating data transformation steps between semantically compatible, but structurally incompatible analytical steps
- Benefits:
  - System aware of connections between workflow components
  - System aware of connections between data sources and workflow components



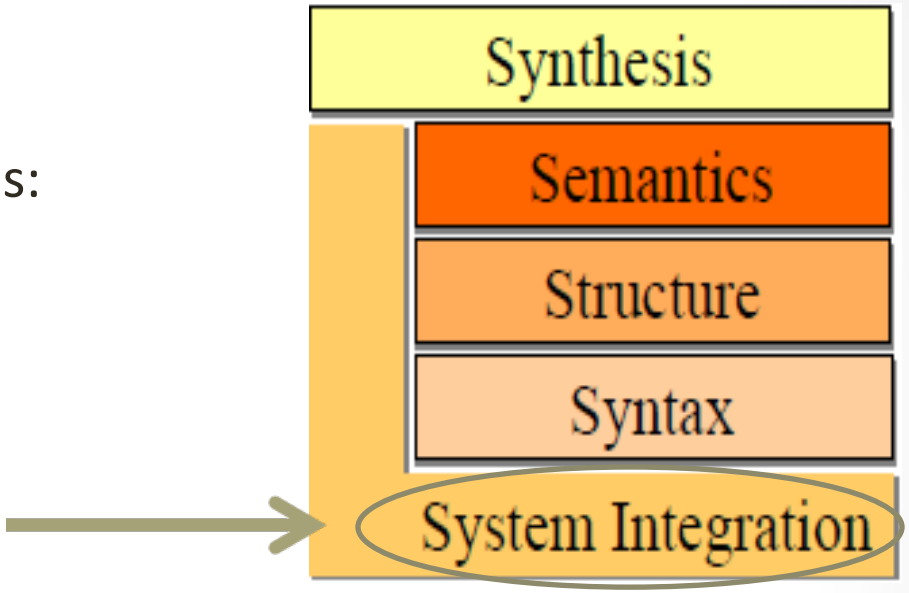
Integration Challenges

- Issues arise due to heterogeneities that occur across systems, data sources and services that make up the scientific data management infrastructure
- Traditional Heterogeneity:
  - Syntax
  - Structure
  - Semantics

- Scientific data analysis and information-integration involve two additional levels:
  - System integration
  - Synthesis

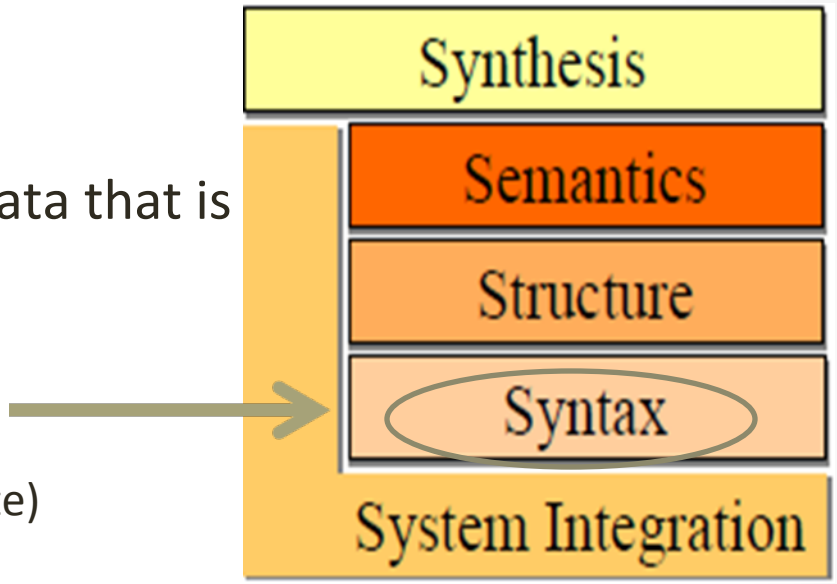
Integration Challenges

- Refers to low level issues such as:
  - **Network protocol**
    - HTTP
    - FTP
  - **Platforms**
    - Operating system
  - **Remote execution Methods**
    - Web services
  - **Authorization & Authentication mechanisms**
    - Kerbros (network authentication protocol)



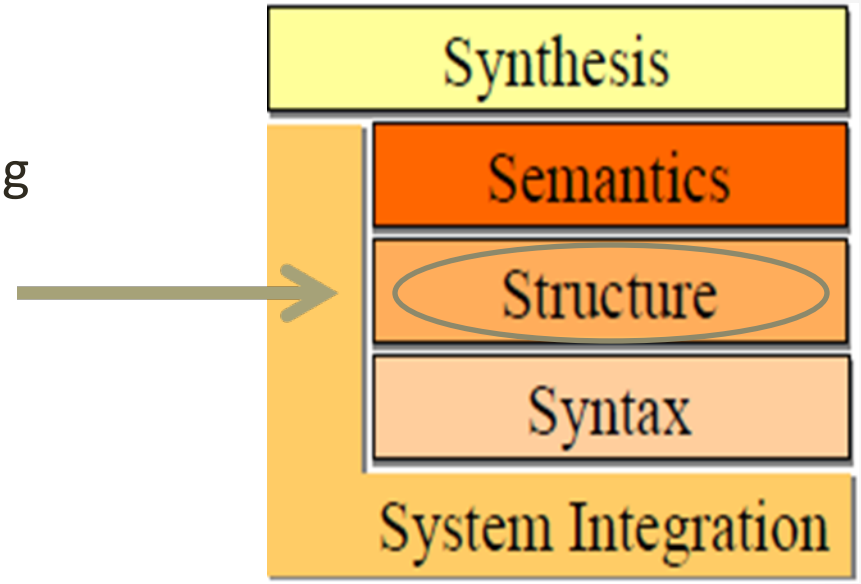
Integration Challenges

- Data not stored in a database or data that is exchanged between applications:
  - Raster or vector
  - File formats:
    - *netCDF* (Weather forecast & climate)
    - *HDF* (Scientific data format)
    - *ESRI* Shapfiles (Environmental System Research Institute)



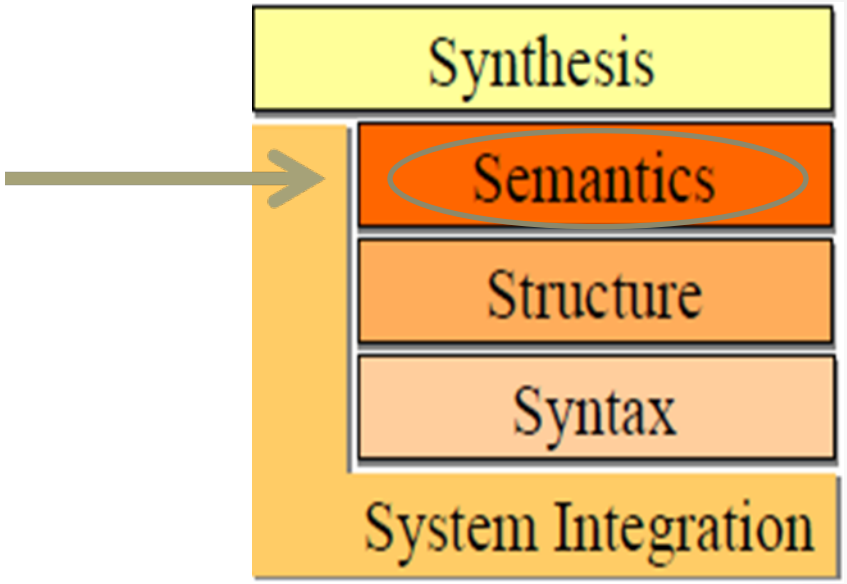
Integration Challenges

- Similar data represented using different schemas
- Structured entities requiring management:
  - database schemas & queries

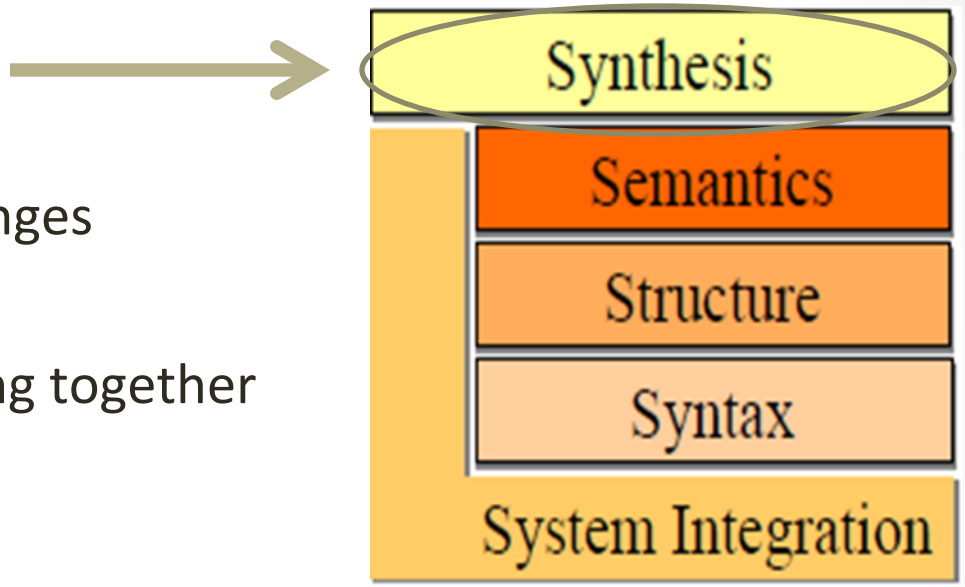


Integration Challenges

- Storing scientific data in a database system to provide solutions to a number of technical challenges



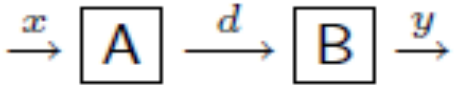
Integration Challenges



- Encompasses all the challenges
- Refers to the issue of putting together databases:
  - Semantic expressions
  - Queries & Transformations
  - Other computational services

Integration Challenges

- **Synthesis example:**
  - If scientist want to put together two process steps in a simple analysis pipeline:

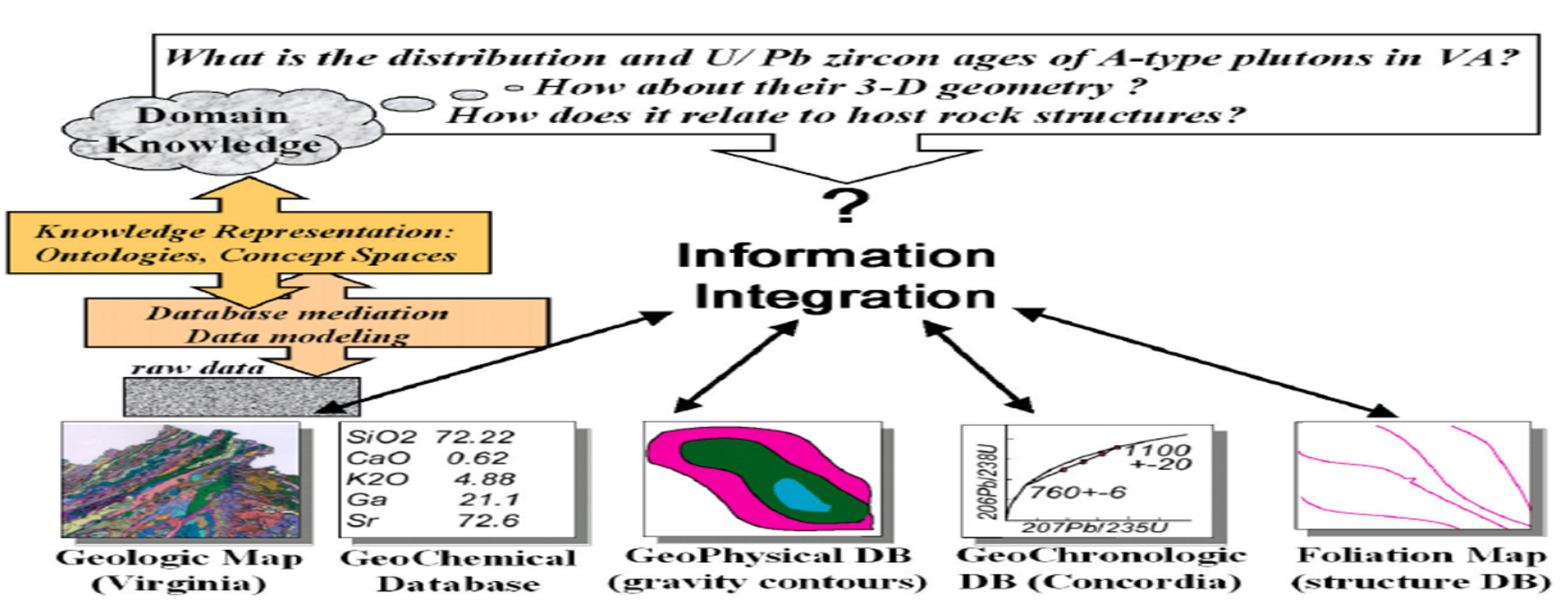


- **Syntactic and structural heterogeneities:**
  - Does output **d** of **A** directly fit format **B** or is data transformation necessary

- **System and semantic:**
  - What mechanisms should be used to invoke processes and how should data be shipped from **A** to **B**
  - Is it meaningful and valid to connect **A** and **B** in this way

Integration Challenges

Multi-world integration



- Distribution of a certain rock type
- Specific regions
- 3D geometry of the regions and relation to the host rock structures
- Through databases and analytical tools, scientist can gather valuable information to answer scientific questions

Integration Examples

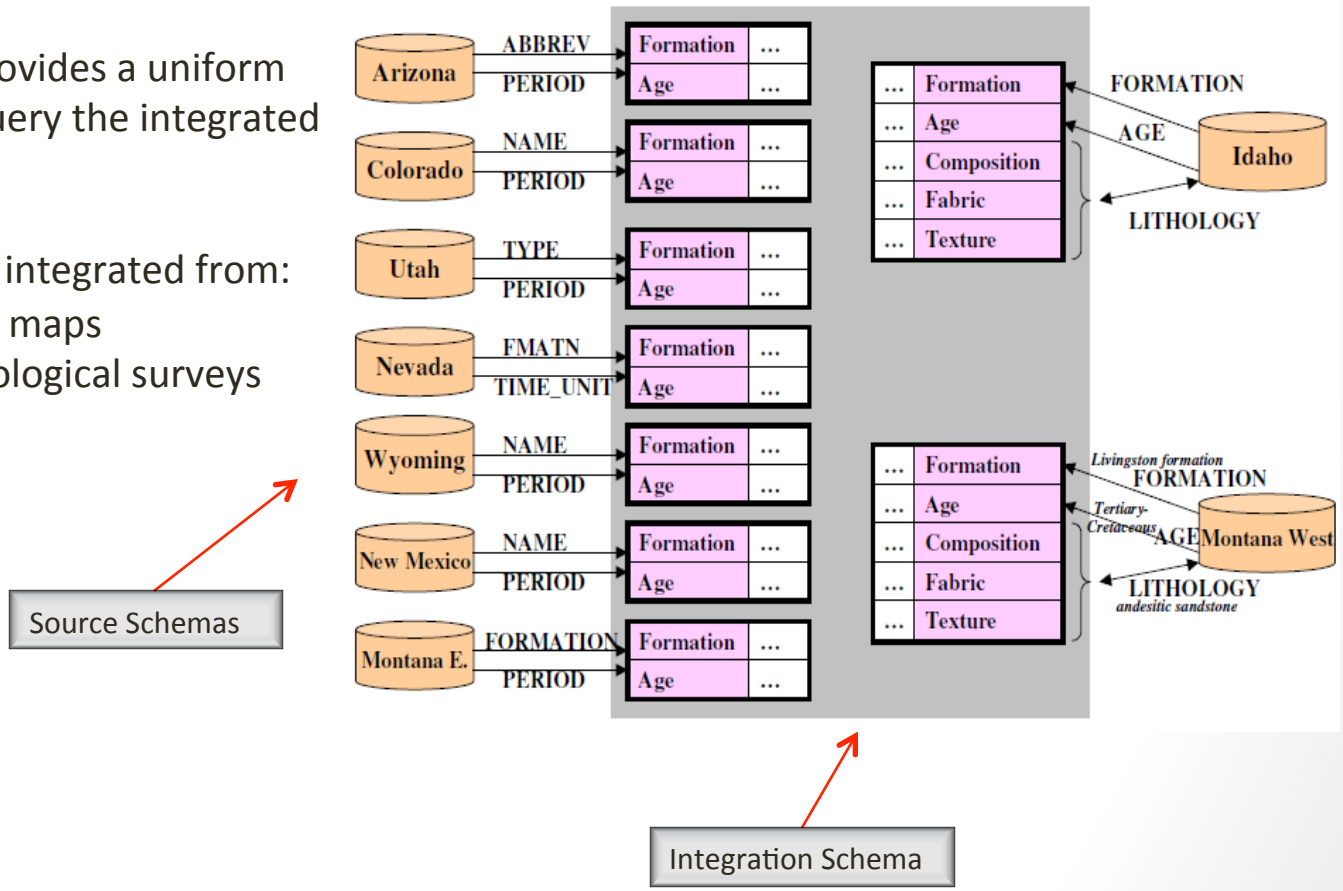
- **Ontology:**
  - **Intended for modeling knowledge about objects, their attributes, and their relationships to other objects**
  - A set of representational primitives used to model a domain of knowledge
    - Primitives are:
      - *Classes* (or sets)
      - *Attributes* (or properties)
      - *Relationships* (or relations among class members)

Integration Examples

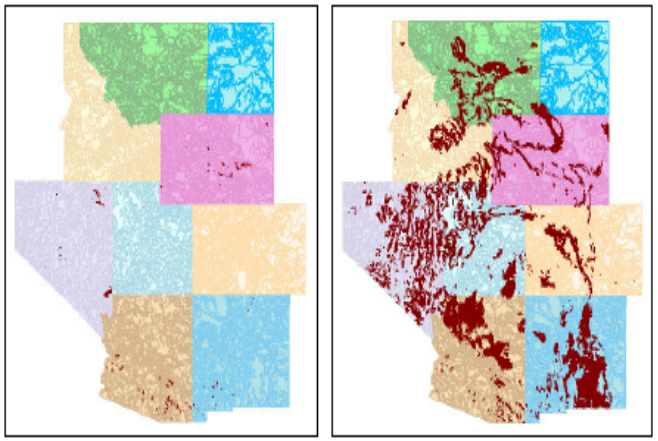
- **Data-centric:** Geologic-Map Data Integration:

Using an ontology to answer a conceptual-level query

- The system provides a uniform interface to query the integrated information.
- Information is integrated from:
  - Geologic maps
  - State geological surveys



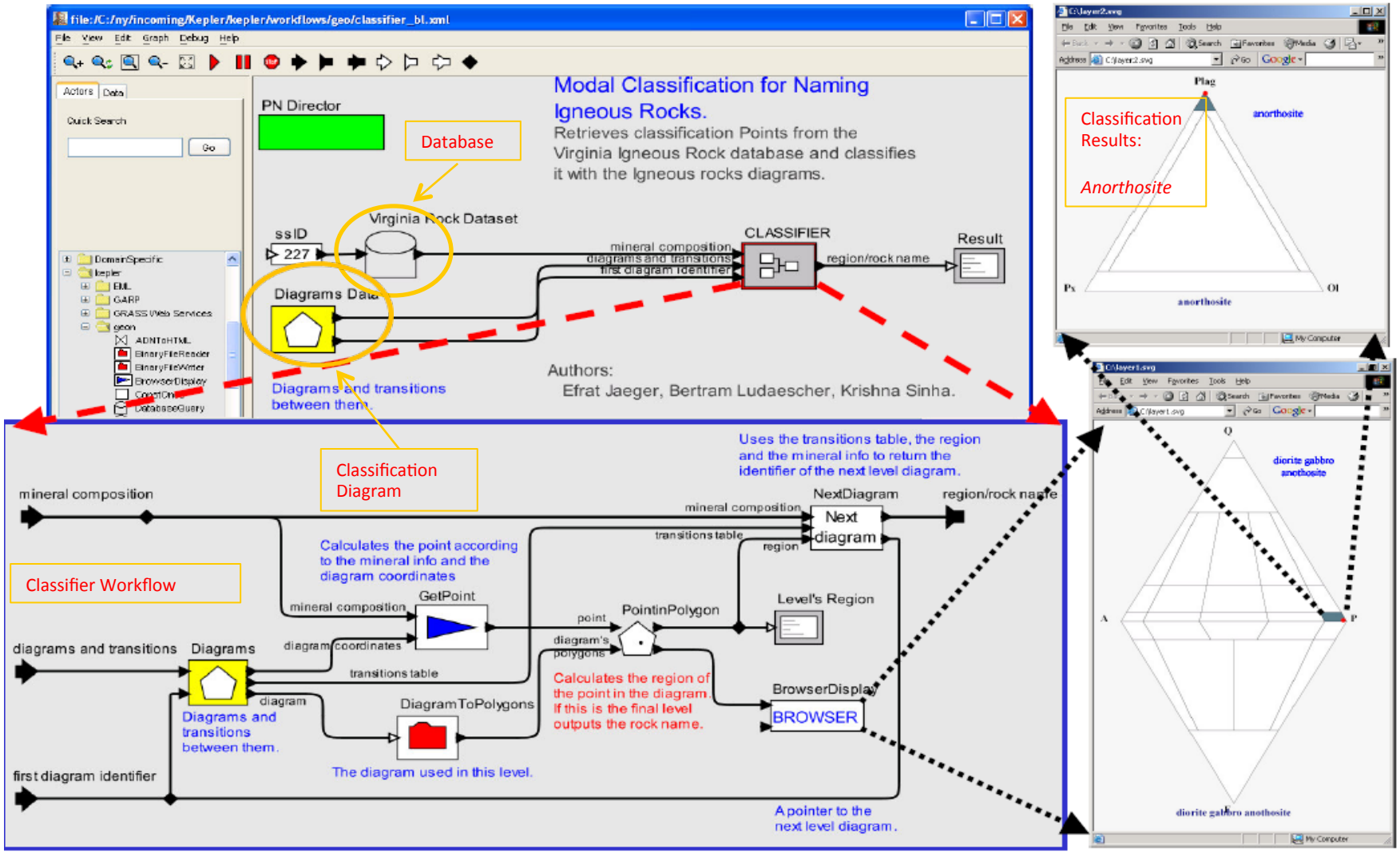
Integration Examples



- **Data-centric:** Geologic-Map Data Integration:
  - Results of a query for regions with Paleozoic age:
  - **Without ontology:**
    - Few results return because missing domain knowledge of other geologic ages:
      - *Paleozoic*
        - *Cambrium and Devon*
  - **With ontology:**
    - A more complete set of regions returned
    - The system re-writes the user query to look for *Paleozoic* and all its sub-ages
  - Preliminaries:
    - Requires *semantic data registration*
    - Data objects(Polygons) must be associated with concepts from previously registered ontologies

Integration Examples

- **Process-centric:** Mineral classification & Interactive display



Scientific workflow system for automating manual data-analysis procedure, or reengineering an existing data-analysis tool in more generic and extensible environment.

Integration Examples

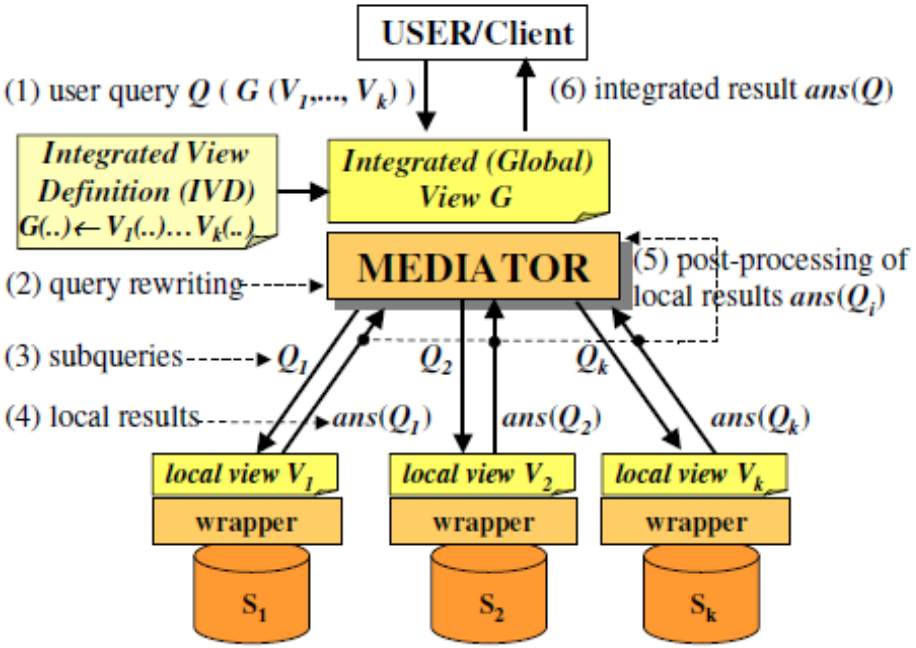
Kepler Workflow System:

- Develops generic solutions to process and application-integration challenges of scientific workflows
- An extension of the PTOLEMY II System
  - An open-source software framework supporting experimentation with actor-oriented design.
    - Actors are software components that execute concurrently and communicate through messages sent via interconnected ports.
    - A model is a hierarchy of interconnections of actors
  - In Ptolemy II semantics is determined by a software component in the model called a director which implements a model of computation

Data Integration

Traditional Mediator Approach:

1. User or application programmer defines a query (***Q***) against ***integrated view G***
2. The ***mediator*** takes ***Q*** and the ***integrated view definition G(…)←…Vi*** and rewrites them into a query plan with a number subqueries Q_1……Q_k For the different sources
3. Subqueries Q_i are sent to the local source for evaluation
4. The local ans(Q_i) are sent back to the mediator
5. Further processing occurs
6. The integrated result ans(Q) is returned to the user



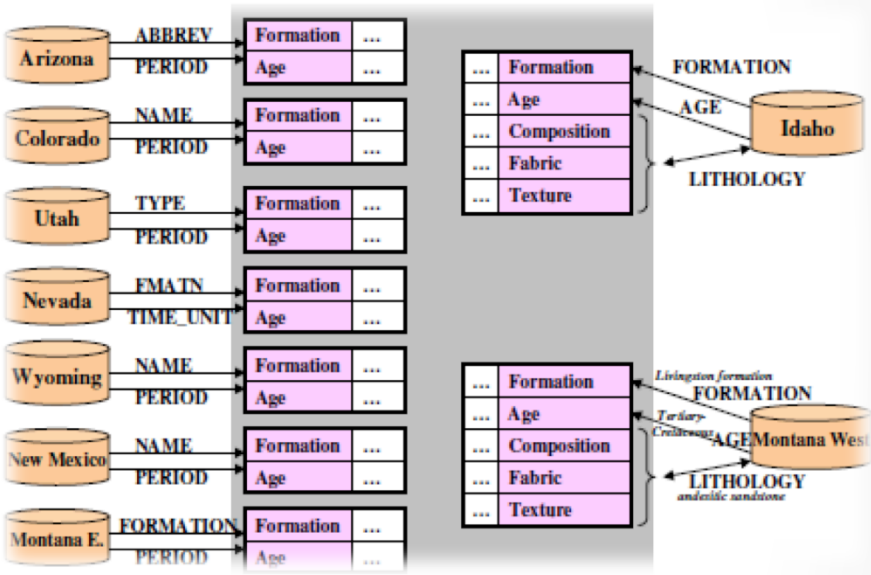
Data Integration

- From the Geological-Map Integration example:
- Allowing query expressions:

$$G(\text{"Arizona"}, AZ.Aid, Formation, Age, \ldots) \longleftarrow$$
$$Arizona(Aid, \ldots, ABBREV, \ldots, PERIOD, \ldots),$$
$$Formation = ABBREV, Age = PERIOD.$$

$$G(\text{"Nevada"}, NV.Aid, Formation, Age, \ldots) \longleftarrow$$
$$Nevada(Aid, \ldots, FMATN, \ldots, TIME\_UNIT, \ldots),$$
$$Formation = FMATN, Age = TIME\_UNIT.$$

Nine sources integrated into a single *global view*



- Shows the *global-as-view* approach
  - Global view ***G*** is filled with information from Arizona source by mapping the local
  - **(ABBREV** and **PERIOD** attributes) to the global (**Formation** and **Age** attributes)
- Spatial regions from the **AREA** column are identified by the **Aid** key attribute
- A unique prefix is used for each source to make the Aid attribute values unique across all sources. i.e (`AZ.Aid, NV.Aid`)

Data Integration

Semantic Mediation:

Based on Concept

- Knowledge-base extension:
  - Facilitate hard-to-relate data sources by going through shared ontologies and new types of concept-based queries against the data
- Ability to view the geologic-age ontology as a set of concepts
  - One for each geologic age organized as a hierarchy
  - Example:
    - A tree in which **children concepts** (*Devon, Cambrium*) are considered to be a subset of restricted set of ages described by the **parent concept** (*Paleozoic*).

- Ontology-enabled mediator:
  - Uses the information from the *geologic age ontology* to retrieve data that directly matches the search concept Paleozoic, and all data that matches any subconcept of *Paleozoic*, such as *Devon or Cambrium*

Data Integration

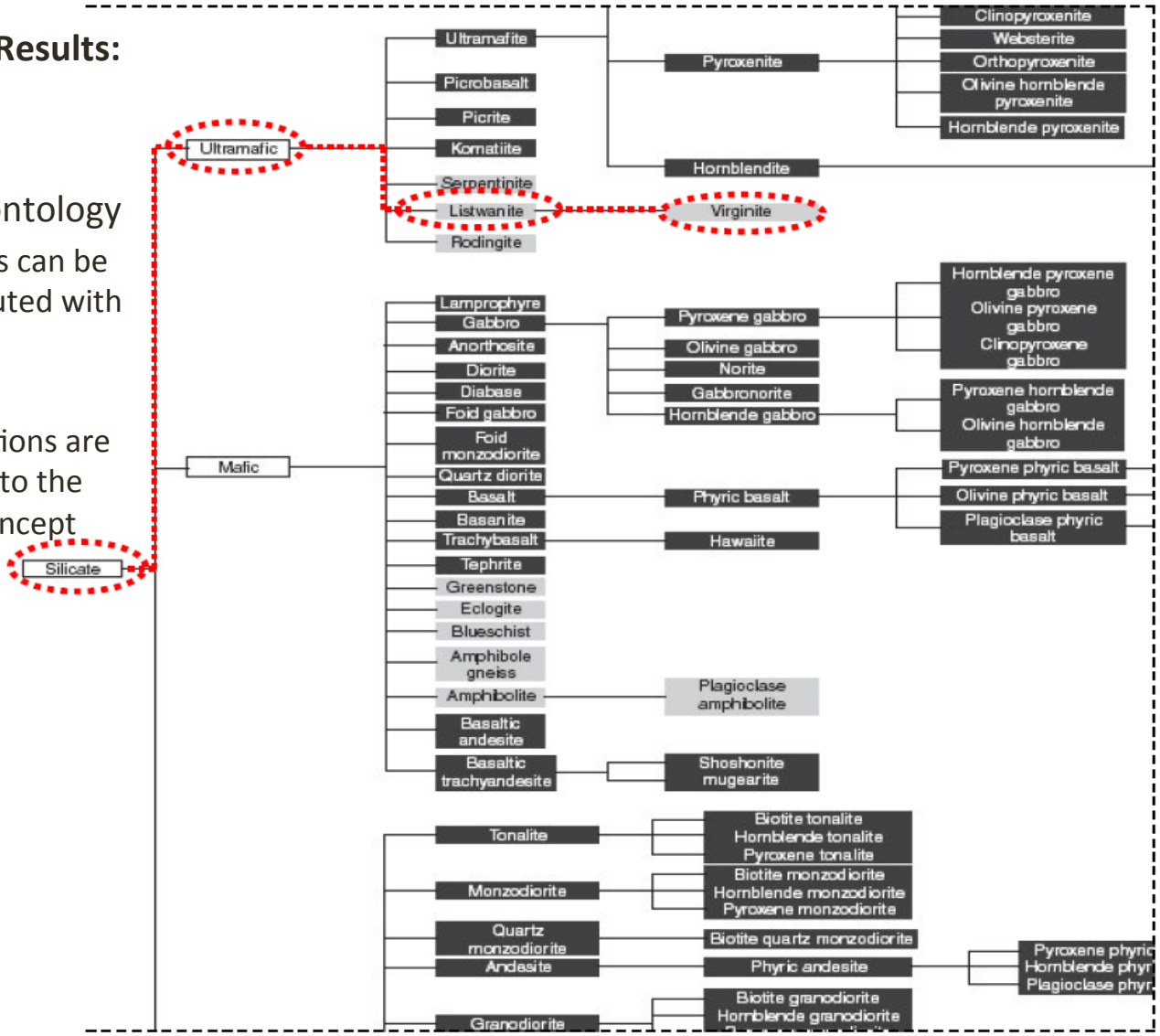
Semantic Mediation Results:

Concept-base query

Using this rock-type ontology

- New semantic queries can be formulated and executed with the prototype.
- Subconcept compositions are displayed as children to the right of the parent concept

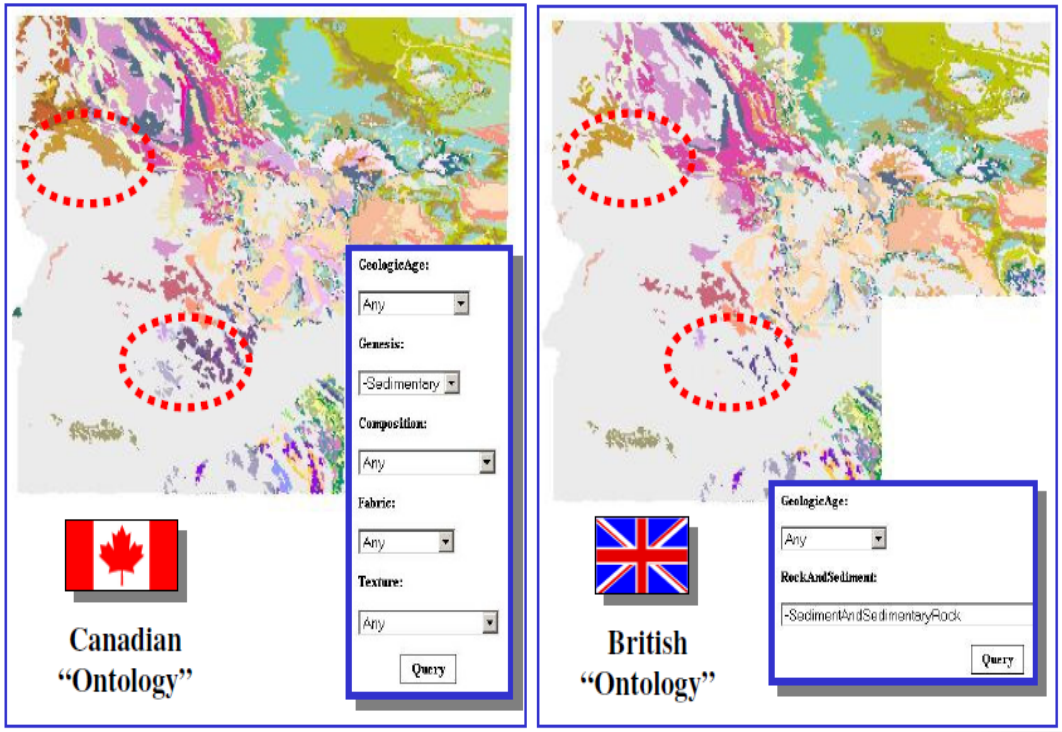
Canadian system for classification hierarchy of rock types



Data Integration

Semantic Mediation Results:

- Achieved by semantically registering existing data sources to one or more ontologies
- Extends the mediator approach to include expert knowledge for linking *hard-to-relate sources*



Semantic Query Results:

Left:

- Lithology information provided by two of the nine state geologic maps (Idaho and Montana West) linked to concepts in the composition, fabric, and texture hierarchies

Right:

- Result of a semantic query for Sedimentary rocks is displayed. Queries can be executed for
  - Composition (Silicate)
  - Fabric (Planar)
  - Texture (Crystalline)

Data Integration

Associating data objects in the sources with concepts defined in a previously registered ontology

Semantic Data Registration:

Needed to facilitate data integration

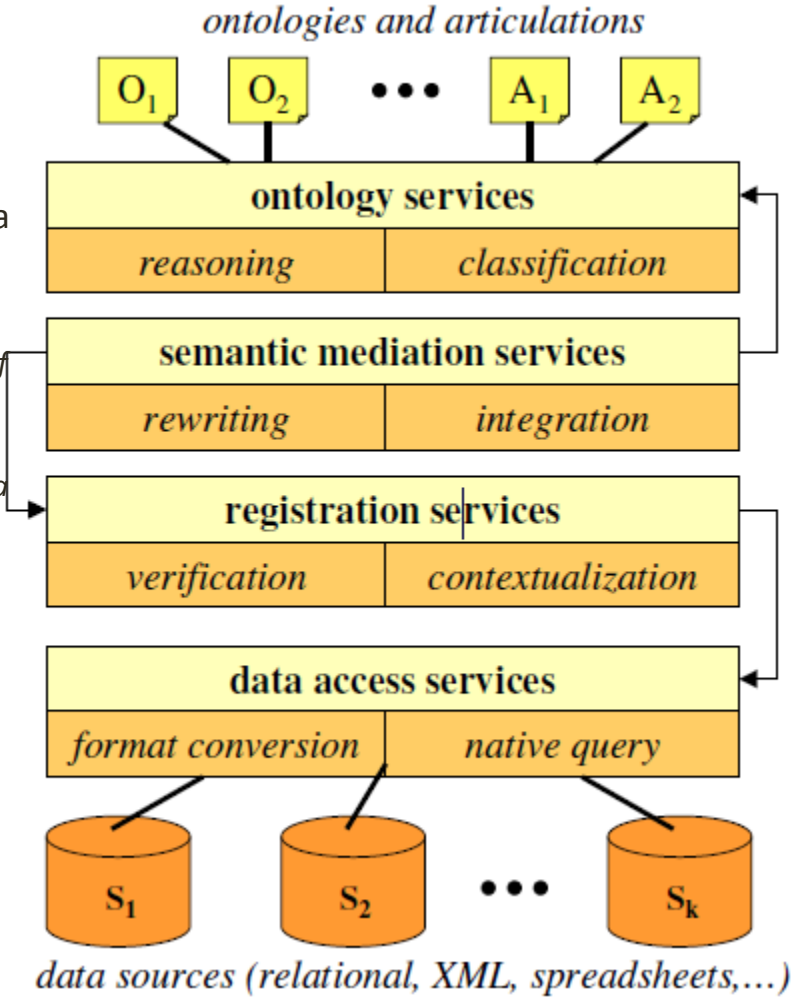
Proposed Framework

- Register the associations between data objects in a database *D* and a target ontology *O*
- *Example:*
- *Let k = id(D_k) and j = id(O_j) be unique recourses identifiers of D_k and O_j respectively*
- *The semantic data registration of D_k to O_j is given by constraints Ψ_{kj} where each $\psi \in \Psi_{kj}$ is a constraint formula over $\Sigma_D \cup \Sigma_O$*

Semantic mediation formula:

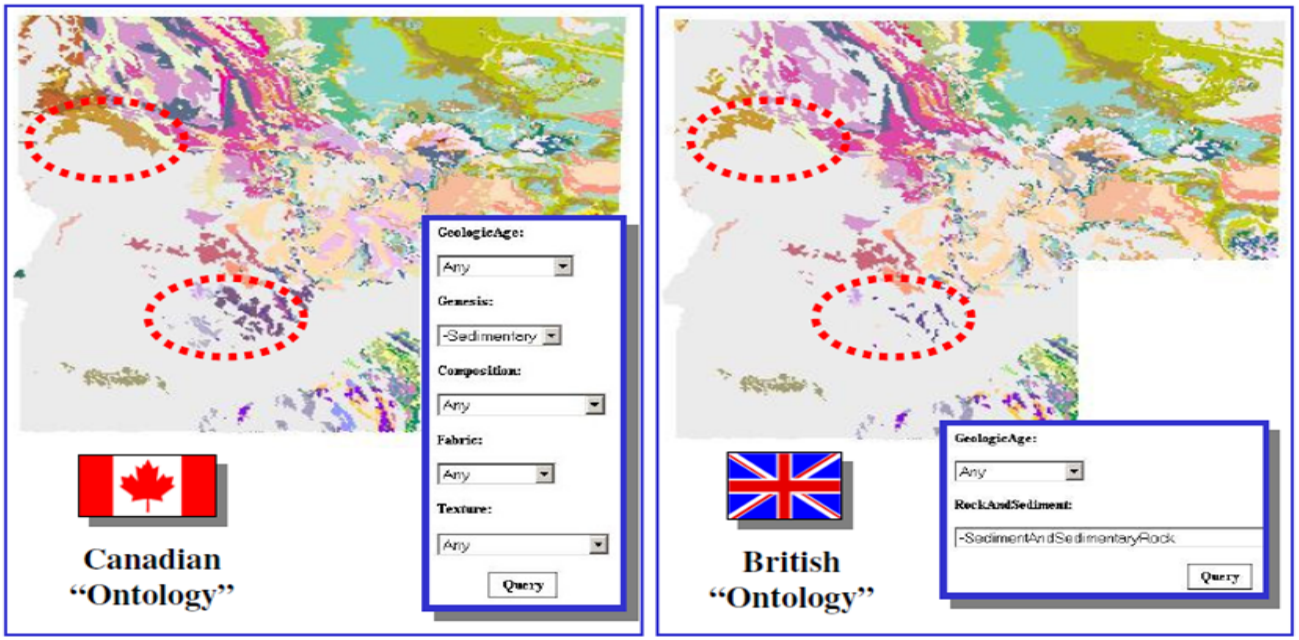
$$\psi = \forall x \forall y : j.\mathtt{D}(x) \wedge j.\mathtt{R}(x,y) \leftarrow \exists z : k.\mathtt{P}(x,y,z) \wedge k.\mathtt{Q}(y,z)$$

- Shows ontology ***O***'s concept ***D*** and its role ***R*** can be populated using certain tuples ***P***(x,y,z) from ***D***
- When you combined the constraint and articulation $\psi \wedge \varphi$
- Allows data object from D_k to link to concepts like $i.\mathtt{C}(x)$ in ontology O_i



Data Integration

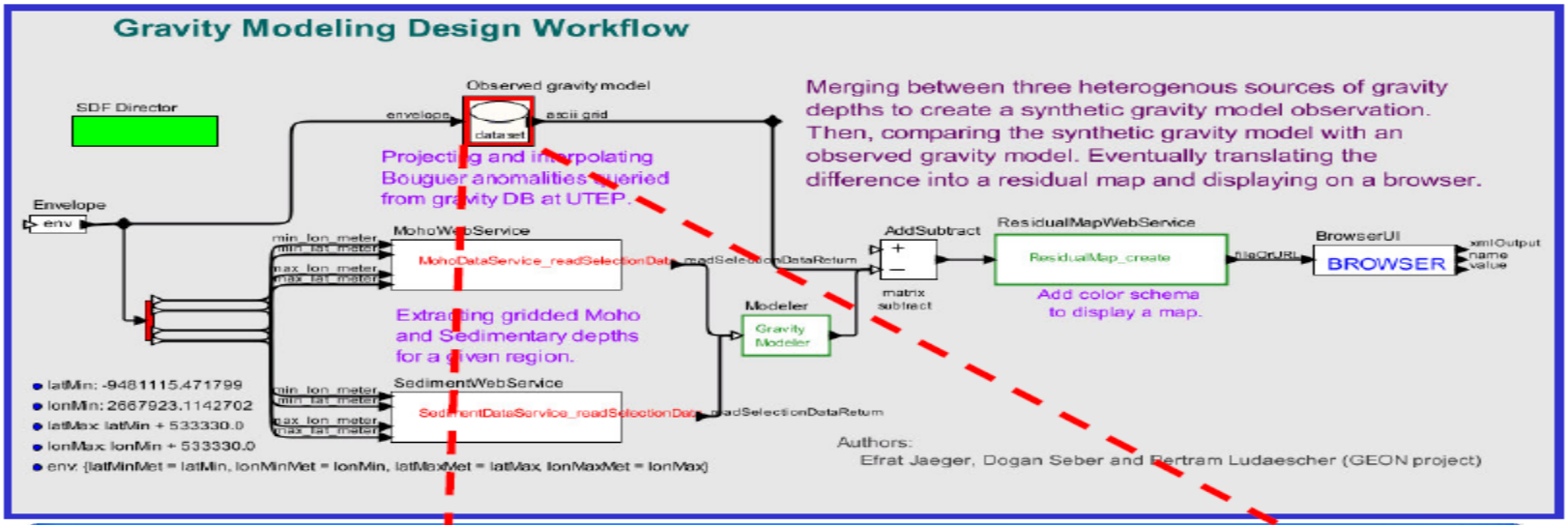
Results of Semantic Data Registration

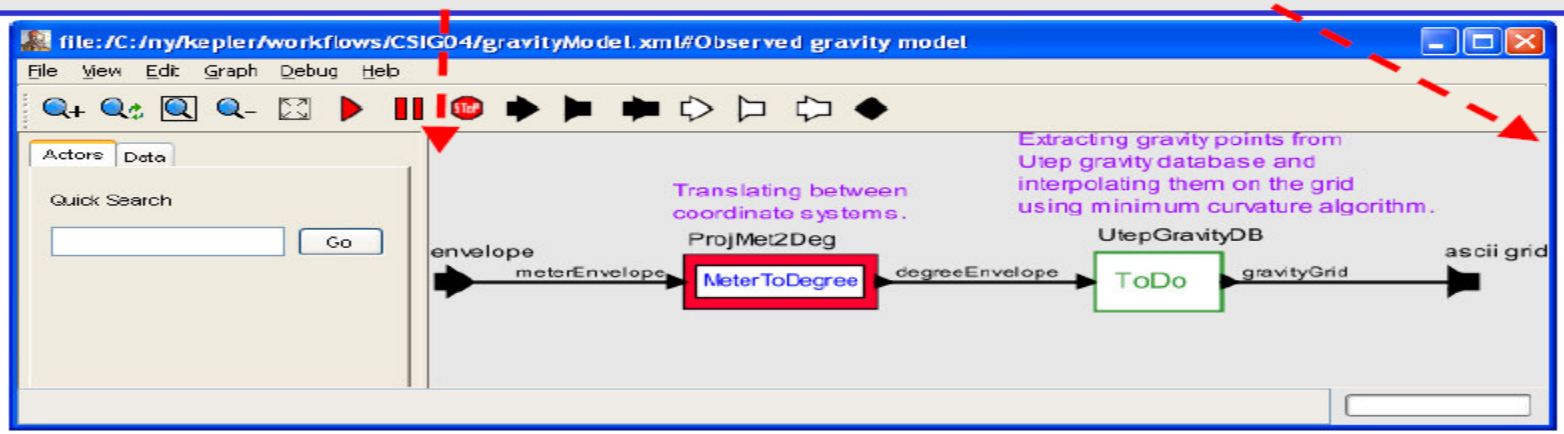


- Key Point:
  - Application of articulation axioms is that O_j and O_i can be used to query and view data from D_k through the conceptual view provided by O_i
- Geological Map Example:
  - Enables the ability to query the geological maps using British rock classification view despite the fact that the geologic map database was originally registered to the Canadian System

Scientific Workflows

- Involves components that has not yet been implemented
- Advantage:
  - Allow the user to go from a conceptual design workflow to an executable version by replacing the design components with implemented ones as they become available





Scientific Workflows

Semantic Workflow Extensions

Ontology-Driven Data Transformation

Implementation of ***Semantic Type*** ensures that the data is connected meaningfully:

- Structural type check guarantees the actors connections have compatible data types
- There is no means to check whether the connections are potentially meaningful
  - For example:
    - A connection between an actor outputting a velocity vector and one that takes as input a force vector may be structurally valid, but not semantically

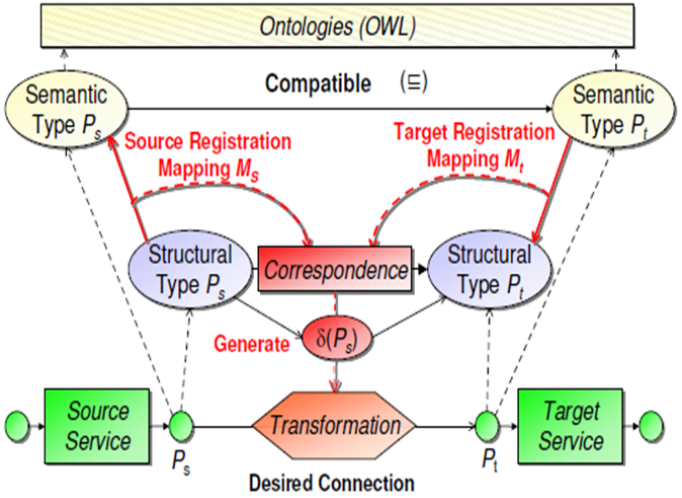
Scientific Workflows

Semantic Workflow Extensions

Ps is not compatible with **Pt** $P_s \not\preceq P_t$

```
<rinfo>
    <age>...</age>
    <ccomp>...</ccomp>
    <text>...</text>
    <fab>...</fab>
</rinfo>
```

```
<properties>
    <lithology>... </lithology>
    <geoage>...</geoage>
</properties>
```



Needs source registration to assist:

Source registration mapping **Ms** = by the rule $q_s \rightsquigarrow E_s \in M_s$

/rinfo/age $\rightsquigarrow O$.geologic_age
/rinfo/ccomp $\rightsquigarrow O$.lithology.composition
/rinfo/text $\rightsquigarrow O$.lithology.texture
/rinfo/fab $\rightsquigarrow O$.lithology.fabric

Target registration mapping **Mt** = by the rule $q_t \rightsquigarrow E_t \in M_t$

/properties/lithology $\rightsquigarrow O$.lithology
/properties/geoage $\rightsquigarrow O$.geologic_age

Correspondence mapping **Mst** = by the rule $q_s \rightsquigarrow q_t$

/rinfo/age $\rightsquigarrow$ /properties/geoage

Questions