



Web Service-based Distributed Data Mining in the Context of GEOSS

Liping Di

Center for Spatial Information Science and Systems (CSISS)
George Mason University
6301 Ivy Lane, Suite 620
Greenbelt, MD 20770
ldi@gmu.edu

Presented at
Workshop on Innovative Data Mining Techniques in Support of GEOSS
August 31, 2009-September 2, 2009
Sinaia, Romania



Contents

- GEOSS Background
 - GEO
 - GEOSS
 - GEOSS Architecture
 - GEOSS Core Capabilities
- Web Services
 - Feature of Web Services
- Implementation of Data mining algorithms as web services
 - How to implement data mining as web services (data mining services)
- Data mining in the context of GEOSS
 - Make data mining service available through GEOSS (mining algorithms)
 - The process of data mining
 - The mining workflow
 - Sharing of mining knowledge
- System Implementation
 - Design
 - Some major consideration
 - Data traffic
 - Security
- Summary



Introduction

- Data mining
 - A computer process of discovering unknown knowledge from massive data.
 - Data mining involves in massive data, high-performance computers, and mining algorithms.
- Earth observation
 - Earth observations, mainly through remote sensing, have generated huge volumes of geospatial data.
 - Vital for scientific research, socio-economic activities, and military missions.
 - Those massive data collections contain unknown knowledge about the status of the earth system, functions and interactions of its components.
 - Data mining is an important tool for discovering the knowledge
- About this talk
 - Will not talk individual mining algorithms.
 - Concentrate on how to bring the data, algorithms, and computing facilitate together to facilitate the data mining in the context of Global Earth Observation System of Systems (GEOSS).



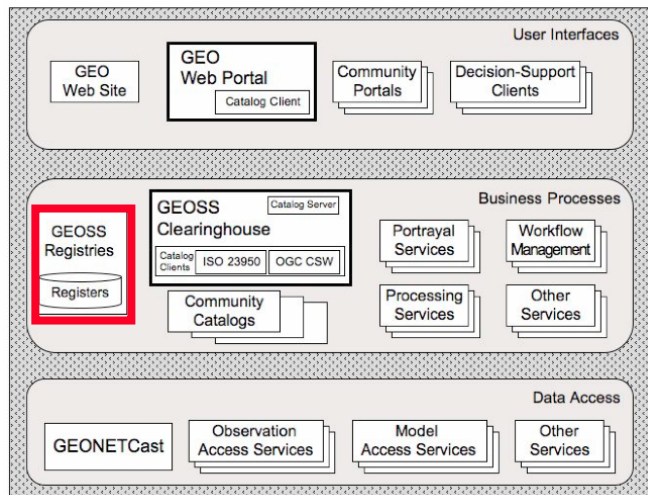
GEOSS Architecture

- GEO Architecture and Data Committee (ADC) designs the GEOSS architecture
- Some major features of GEOSS architecture
 - A system distributed around the world
 - Adopt the service-oriented architecture (SOA).
 - Web services as the major methods of implementation.
 - Standards and specifications developed by the Open Geospatial Consortium (OGC) and ISO TC 211 as the major interoperability standards.
 - Components and services are contributed by participating countries and organizations.
 - Machine-to-Machine interoperability as the major form of interoperability
- ADC defines a set of core components as the GEOSS common infrastructure
 - Enable the operation of the contributed components and services as a consistent system.
 - Core components: Component and Service Registry (CSR), Standard and Interoperability Registry, Clearinghouse, Portals.
 - GEOSS Common Infrastructure is implemented through the GEOSS architecture implementation pilots.

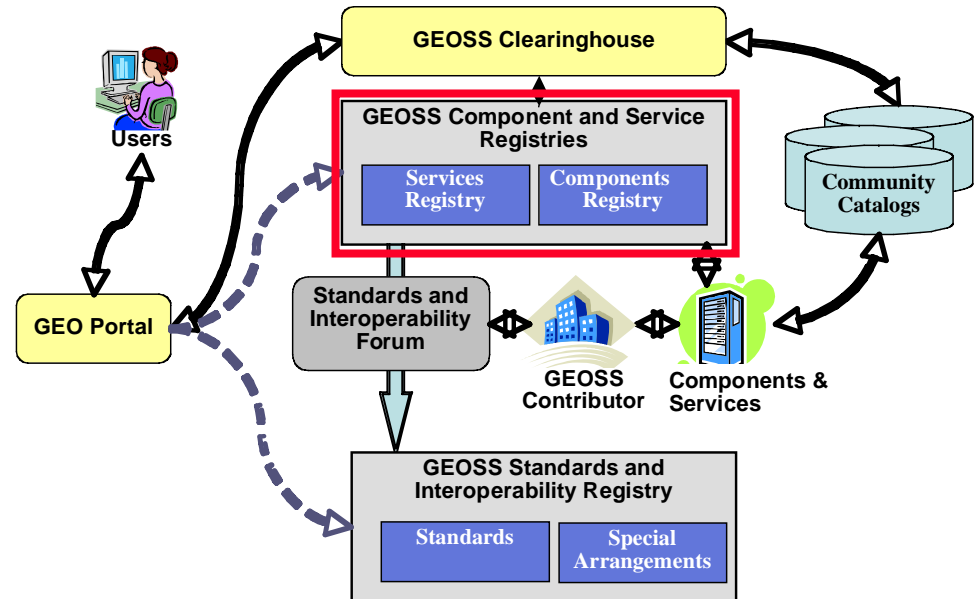


Implementation of GEOSS Architecture

- The Pilot aims to incorporate contributed components consistent with the GEOSS Architecture - using a GEO Web Portal and a GEOSS Clearinghouse search facility – to access services through GEOSS Interoperability Arrangements in support of the GEO Societal Benefit Areas.



GEOSS Architecture – Engineering Viewpoint



Interactions of the GEOSS Registries, Portal and Clearinghouse

Excerpted From Core Architecture Implementation Report (V0_9b) [George Percivall and Ingo Simonis]



GEOSS Components and Services

- **Component**
 - Part of GEOSS contributed by a GEO Member or Participating organization. Components expose service interfaces for providing access to earth observation-related functions and/or data.
- **Service**
 - Functionality provided by a component through its system interfaces. Services communicate primarily using structured messages, based on the Services Oriented Architecture (SOA) view of complex systems.



GEOS Component and Service Registry

- As a corner stone, the ***Component and Service Registry (CSR)*** includes mechanisms to:
 - register components and have them approved by the GEO Secretariat / Ad-hoc Registry Record Review Group
 - register service interfaces for specific component
 - associate service interfaces with GEOS-recognized standards or special arrangements for implementations using non-recognized approaches
- A taxonomy of standards types is also proposed to assist in the discovery and classification of GEOS service implementations in CSR Version 1.
- If you want to contribute your components and services to GEOS, you have to register them in CSR.



GEOSS Standard and Interoperability Registry

- Provide a centralized registry for registering
 - the GEOSS recognized interoperability standards and specifications.
 - the private interoperability arrangement.
- If you register your components and services in the CSR, you have to associate your components and services to GEOSS-recognized standards/specifications.
 - If no GEOSS-recognized standards/specifications is applicable, you have to register your interoperability interfaces as the special arrangement.
 - Hope the consumer of your components and services can understand your interoperability arrangement.



GEOSS Clearinghouse

- Standardized search across registered items and metadata catalogues to promote rapid access to
 - inventory-level information about offered components, services, and data.
 - other resource types, which may be registered with GEOSS catalogues, including, but not limited to, software applications, training materials and courses, web sites/portals, news feeds/RSS, models, and documents.



GEOS Portals

- Provide users a point of entry to access and utilize GEOS resources.
- Portals for individual societal benefit areas are being developed.
 - To serve individual community's needs.
- We can develop a data mining portal, or embed the data mining capability in the individual portals.

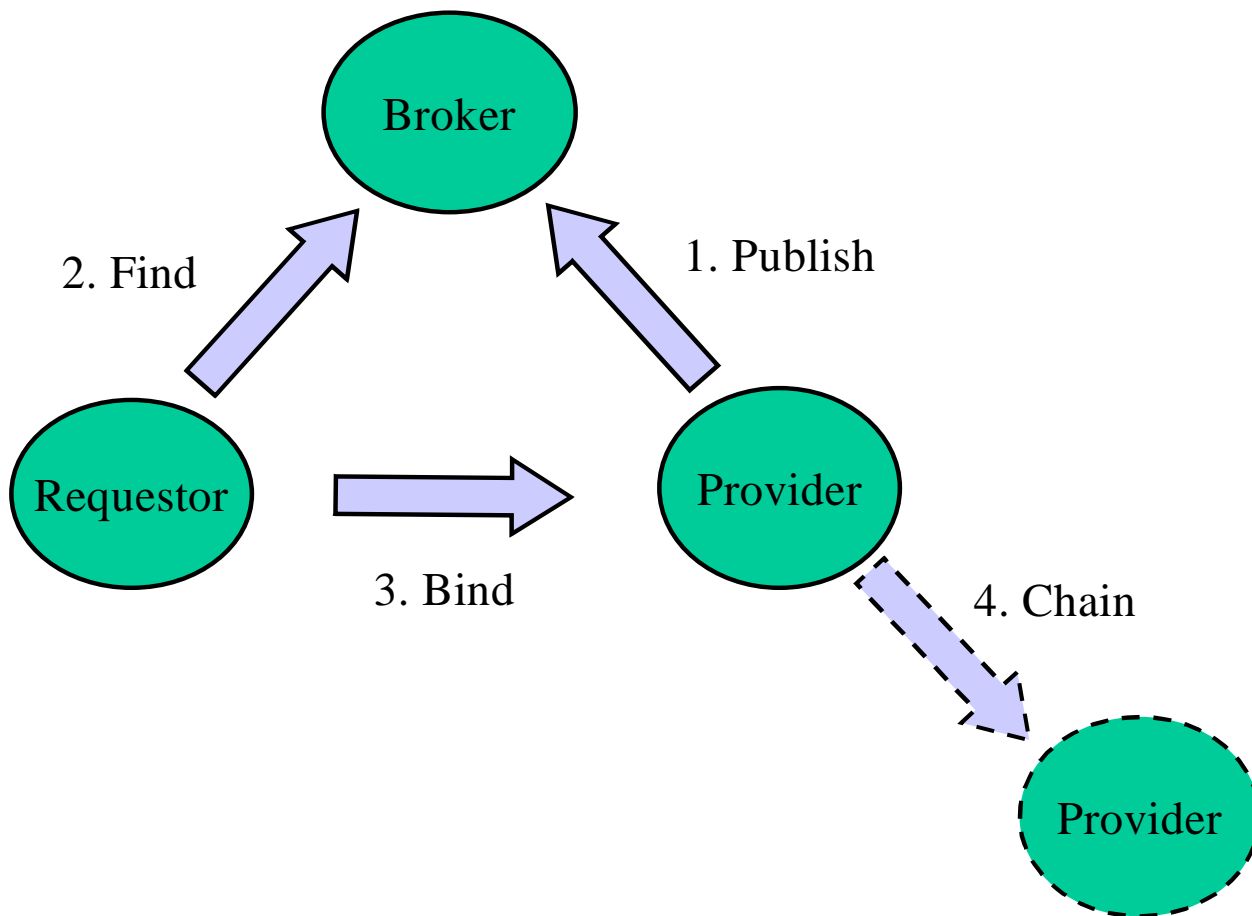


Geospatial Web Services

- Web Services are self-contained, self-describing, modular applications that can be published, located, and dynamically invoked across the Web.
- Web services perform functions, which can be anything from simple requests to complicated business processes.
- Once a Web service is deployed, other applications (and other Web services) can discover and invoke the deployed service.
- Geospatial Web services are the web services that process geospatial data and information



Service operations





Service Operations

- Publish – advertise (or remove) data and services to a broker (e.g., a registry, catalog or clearinghouse).
- Find – Service requestors and service brokers collaborate to perform the find operation. Service requestors describe the kinds of services they're looking for to the broker and the broker delivers the results that match the request.
- Bind – A service requestor and a service provider negotiates as appropriate so the requestor can access and invoke services of the provider.
- Chain – The chain operation binds a sequence of services.



Service Chaining

- A *Service Chain* is defined as: a sequence of services where, for each adjacent pair of services, occurrence of the first action is necessary for the occurrence of the second action.
- When services are chained, they are combined in a dependent series to achieve larger tasks.
- Three types of chaining defined in ISO 19119 and OGC:
 - User-defined (transparent) – the Human user defines and manages the chain.
 - Workflow-managed (translucent) – the Human user invokes a service that manages and controls the chain, where the user is aware of the individual services in the chain.
 - Aggregate (opaque) – the Human user invokes a service that carries out the chain, where the user has no awareness of the individual services in the chain.



Contribute Your Data Mining Algorithms to GEOSS

- Implement your data mining algorithms as web services
 - Follow the OGC and W3C web service interface standards
- Deploy your web services in a computer to make both the services and the computing power accessible by other GEOSS members.
 - If you don't want to share the computing facility, you can ask someone else to host the services.
- Register the mining services in CSR.
- It is expected the world-wide contributions of mining services to GEOSS.
 - Many different mining services
 - Multiple instances for a single type of service



The Data Mining Process

- The data mining involve in multiple consecutive steps
 - Find where the needed data are located.
 - Obtain the data
 - Preprocess the data
 - Apply mining algorithms to the preprocessed data
 - Analyze the final result.
- In the web service environment, each steps involve in multiple services
 - The process step can be chained together as a mining workflow.
- In the GEOSS context, the services and data can be located anywhere in the world
 - Discovered through the GEOSS CSR.

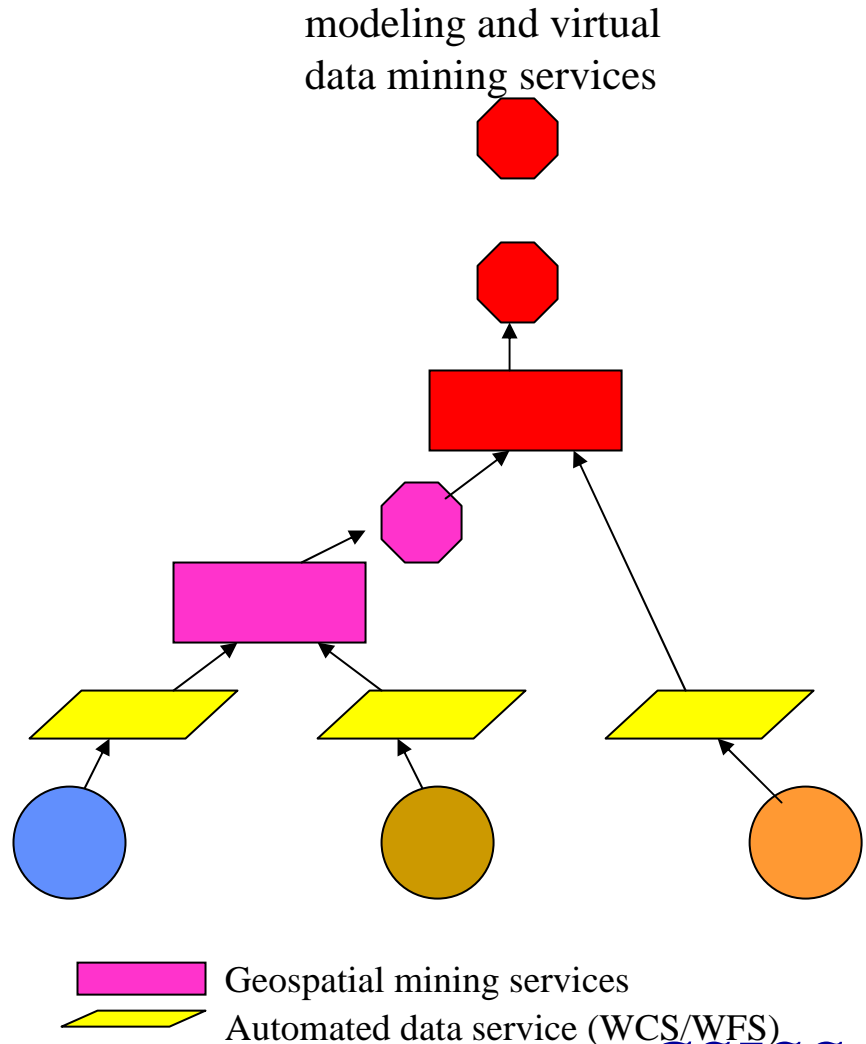
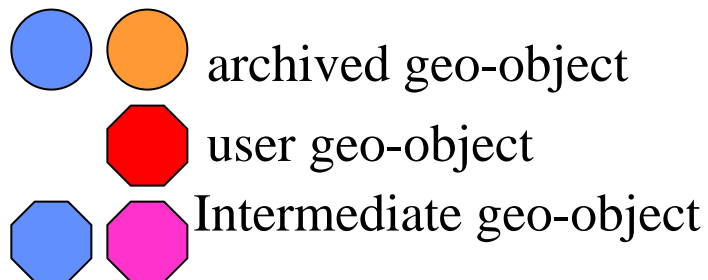


Mining Workflow and Mining Process Model

- The mining workflow contains the instances of data and instances of mining services
- Mining process model—Tell conceptually steps to take to do a specific type of mining.
- A mining process model represents the knowledge for conducting a specific type of data mining.
- Geo-tree – The conceptual/graphic representation of a mining process model.
 - Two types of nodes in a geo-tree: Mining process node and geo-object (data) node.
 - The mining process node is a geospatial service type.
 - The geo-object node contains a geo-object type.
- By defining mining process models at abstract level
 - Allow the development of general model that can use for generating unlimited numbers of instances.
 - Make domain experts easier to create models by eliminating the needs to details on product specifications.



Geo-object, Geo-tree, Virtual mining products, Geospatial mining models



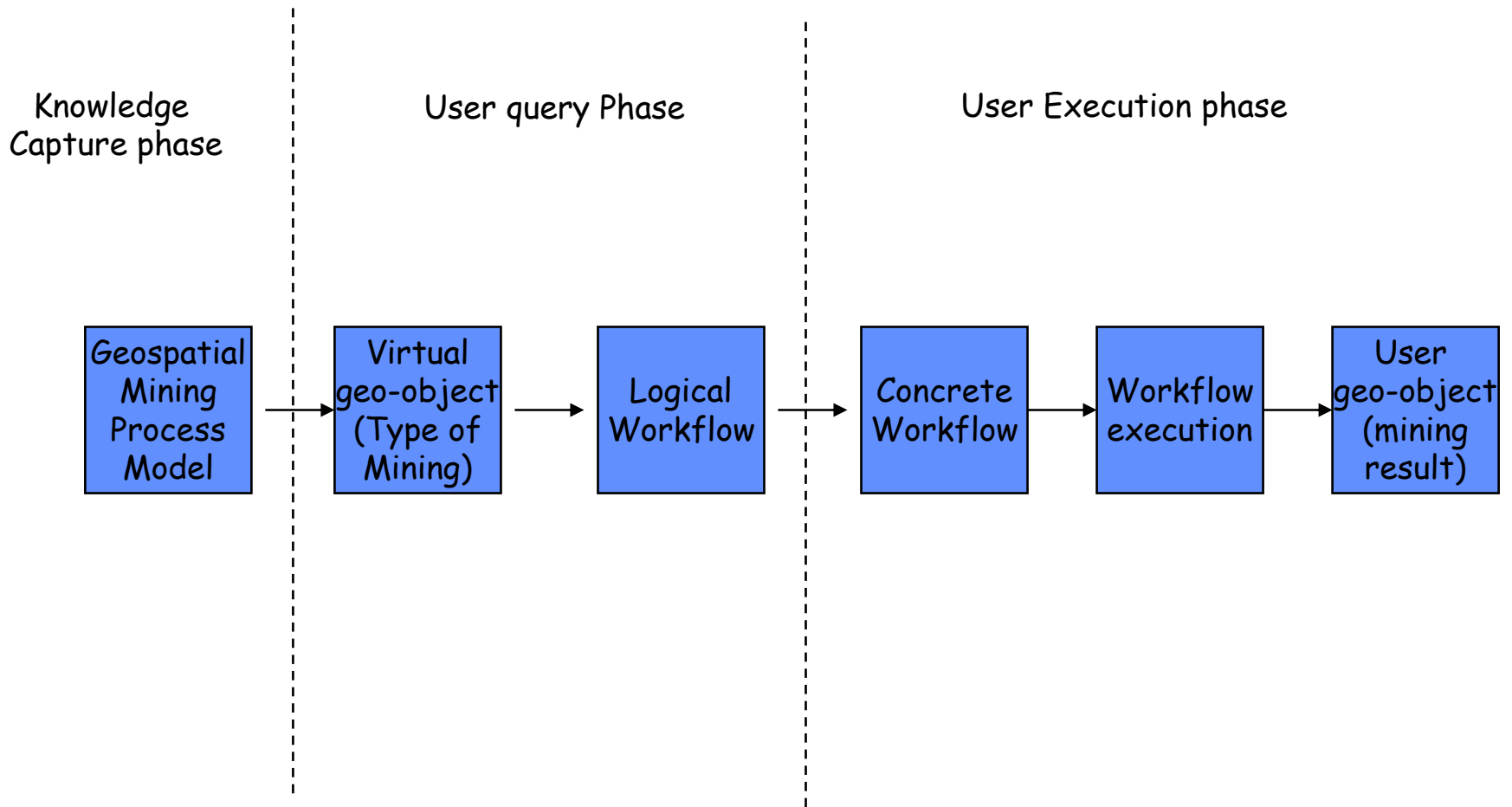


Virtual geo-objects and virtual geo-object type

- A virtual geo-object is a geo-object (the mining result) that:
 - not exist in a geospatial information system (e.g., GEOSS)
 - The system knows how to create it on-demand through mining.
- Virtual geo-object type—abstract of virtual geo-objects that share the same features.
 - All geo-object nodes in a geo-tree, except for those nodes for the archive geo-objects, are virtual geo-object types;
 - The root node in a geo-tree is a virtual geo-object type the model can generate.



From geospatial mining model to user geo-object





Knowledge capture: the construction of mining process models

- Two catalogs are essential
 - geo-object types
 - Mining service types
- Save the model in a formal way
 - Use the modified/simplified BEPL to store the geo-tree
 - Keep geo-trees in the geo-tree library
 - Catalog the geo-trees in the geo-object catalog.
- Two ways to create geospatial models
 - Domain experts create models and share with others.
 - Easy to construct a model through a graphic model construction client.
 - Automatically creation of model.
 - Require the system to have domain knowledge and AI capabilities



Knowledge capture: Expert Creation of Geospatial Mining Models

- An expert user knows the thought process to do specific type of data mining from lower-level inputs step-by-step (the logical mining process modeling)
 - With help of a good user interface and the availability of mining service ontologies and data ontologies, the expert can construct abstract mining process models interactively.
 - The expert-created model can be incorporated into the system as a part of the mining capability the GEOSS can provide.
- Advantages
 - Allow the mining capability of GEOSS to grow with time.
 - Allow the sharing of mining process knowledge.



Knowledge Capture: From Geo-tree to virtual geo-object

- The root node of a geo-tree is a virtual geo-object type
- When user/client requests a geo-object, user will provide a description of the geo-object they want;
- If the type of the user geo-object matches with virtual geo-object type in a geo-tree,
 - The geo-tree is selected;
 - The root node is instantiated with the descriptions provided by the clients.
 - The root node now becomes a virtual user geo-object.
- The next step is to determine if the virtual user-geo-object can be materialized on the fly.



User query phase: Logical Instantiation of Geo-Tree

- Check if a virtual user geo-object can be materialized by instantiating the whole geo-tree.
 - Push the description of the virtual user geo-object down to each node of geo-tree (e.g., spatial coverage, format, etc);
 - Discover instance of service and geo-objects through searching both service instance catalog and geo-object instance catalog;
 - If an archive geo-object is found as the input of a process, then the push down will be stop for the branch of this tree.
- The logical instantiation will not create an actual workflow, but conceptually, it creates a logical workflow.
- If a geo-tree can be instantiated logically, the virtual user geo-object can be materialized.
 - A logical ID will be created and return to client to indicate the user-requested geo-object is found in the system.
 - The logical ID will be used by the client/user to request for the geo-object.



User execution phase: Physical Instantiation of Geo-Tree

- When client requests a user geo-object, the geo-object ID will indicate if the user geo-object is virtual.
- If the geo-object is virtual, the geo-tree associated with the virtual geo-object will be instantiated to create a concrete workflow.
 - A workflow language will be used to encode the workflow.
 - The workflow is executable in a workflow execution engine.



User execution phase: Creation of user geo-object

- The workflow engine will execute the workflow and generate the user geo-object (the mining result).
- The user geo-object will be return to user/client.
- The above mentioned steps reflects two stage processes:
 - User query
 - User execution
- The two stages can also be merged into one stage process that when a user query can meet, the resulted user-object will be pushed back to user automatically without user initiation of the retrieval.



Implementations

- The schema discussed above is being implemented in CSISS as a prototype GEOSS Data Mining Facility (GDMF)
- GDMF needs
 - Mining service instances (provided by worldwide contributors)
 - Earth observation data (provided by worldwide contributors)
 - Computing facilities (provided by CSISS and worldwide contributors)
 - GEOSS CSR, SIR, Clearinghouse (GEOSS common infrastructure)
 - Geospatial workflow engine (BPELPower—CSISS)
 - GeoBrain instantiation service (CSISS)
 - GEOSS data mining portal— being developed
 - Mining process model construction
 - GEOSS resource access
 - Mining execution and result analysis
 - Mining service ontology— need to be developed by the community
 - Data ontology – being developed by NASA and ISO. Mining portion may be needed to be enhanced.



Some Implementation Consideration

- Some major considerations:
 - Data traffic
 - Security
- Data traffic
 - Data mining involves in large volume of data
 - Pushing the large volume of data through the network is not practical in the GEOSS scenario
 - Push mining services to the data sources – not common because of security reason and incompatibility of executable.
 - It is desirable that data mining services are deployed as close to data sources as possible.
- Security
 - Authorization and authentication
 - Try to use OpenID to enforce the security



Conclusions

- This presentation discussed the schema for the web-service-based distributed data mining within the context of GEOSS.
 - Current technologies, interoperability standards, and network infrastructure allow building a web-service-based distributed data mining facility on GEOSS infrastructure.
 - Significant development efforts are needed.
- GEOSS Data Mining Facility built on such a schema are flexible and scalable and can provide much better mining services to the user community than traditional local mining system.
 - Allowing worldwide users to utilize the worldwide geospatial resources available under GEOSS.
 - Allowing the worldwide contribution and sharing of geospatial data mining resources.