Generating composite thematic maps from semantically-different collections of shapefiles and map services

Ghulam Memon, Ashraf Memon, Kai Lin, Ilya Zaslavsky, Chaitan Baru
(gmemon|amemon|klin|zaslavsk|baru)@sdsc.edu
San Diego Supercomputer Center

Abstract

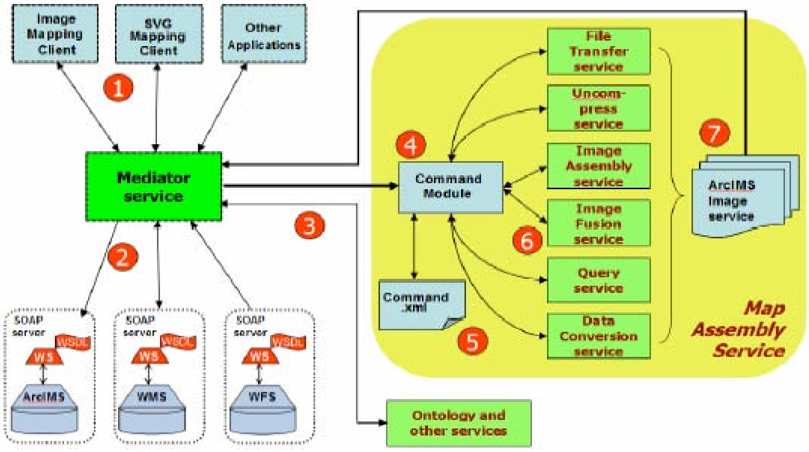
The Geosciences Network (GEON) project is a large-scale collaborative effort aimed at creating cyberinfrastructure for the Earth Sciences. GEON focuses on consistent representation, query access, integration and analysis of multiple semantically-different distributed spatial datasets. The paper will demonstrate knowledge-based integration of heterogeneous spatial data, stored mostly in shapefiles, and their online visualization as composite thematic maps.

Following design principles of Service-Oriented Architecture, we combined the Gridsphere portal framework with ArcIMS-based services. Shapefiles contributed by geoscience researchers and registered in the GEONgrid, are transferred on-demand into an "assembly" ArcIMS service. Once the service is instantiated via the Java connector API, it can be queried from the middleware or from map clients. To process user queries, the middleware consults dataset ontologies, creates a global map legend, and generates a composite thematic map with regard to this global legend. Color assignment is governed by ontology considerations. Finally, the paper proposes certain performance enhancements.

1. Introduction

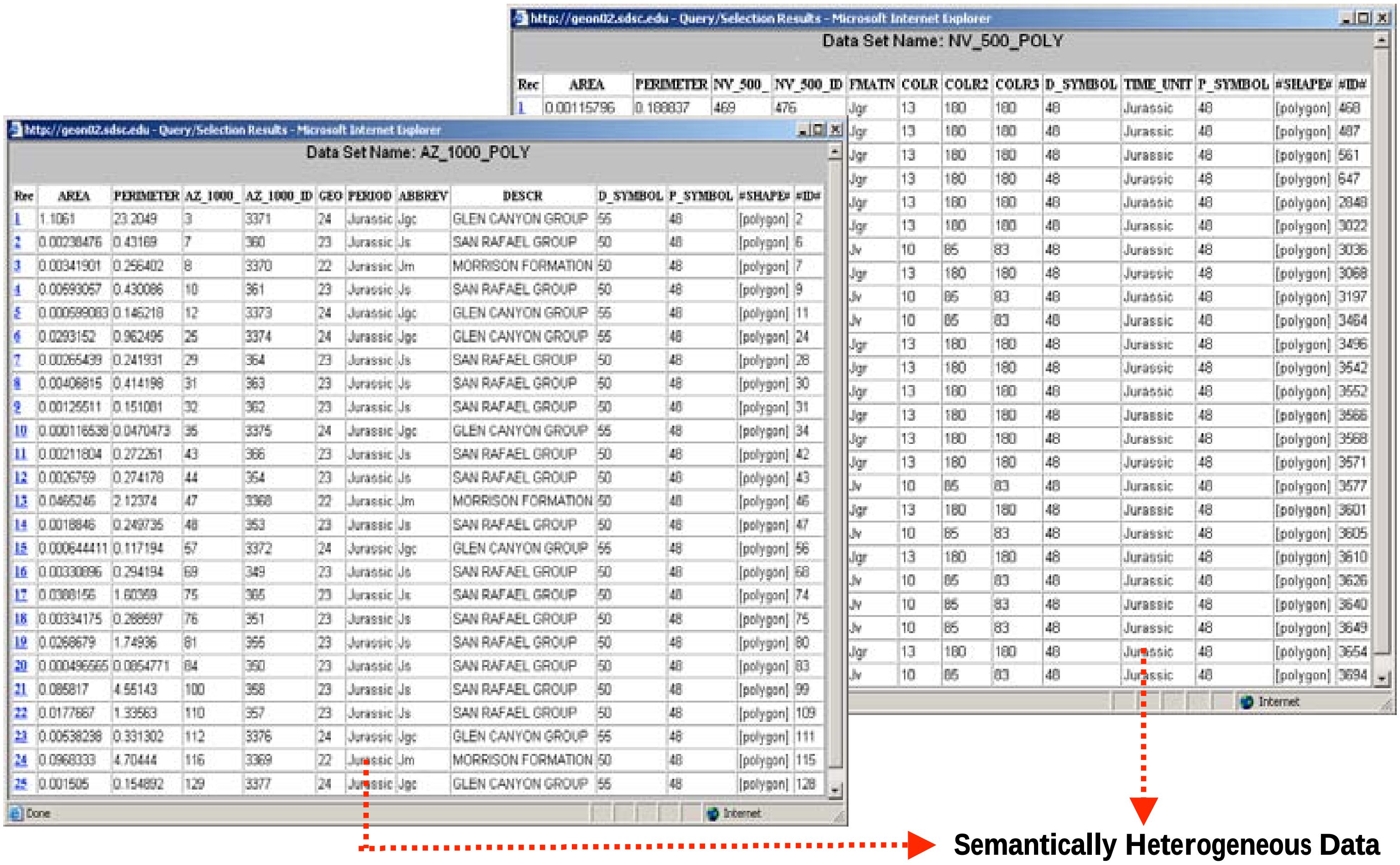
At its heart, GEON is a repository of heterogeneous data which can be integrated, queried, analyzed and visualized by geo-scientists with GEON-provided ontology enabled tools. Combining such data is one of the challenges that GEON faces, because of the inherent differences in data representation and formats.

Figure 1. Assembling a thematic map from heterogeneous sources, using mediator services.



Integration of spatial data from heterogeneous sources within GEON was initially explored in [1]. This paper described a spatial mediation system based on grid services middleware, where each service acted as a wrapper around a different data source to hide source differences and offer a uniform interface to the mediator client. The paper described spatial data service wrappers that converted each mediator request into a source-specific format, and merged the result fragments into a composite map service. The paper continued by describing a set of spatial mediation services for interrogating semantically diverse data sources.

Difference in the semantics of categories and terminology used by different sources represents a serious challenge even if other aspects of heterogeneity (structural, syntactic) are resolved. For example, while integrating geologic data from multiple sources, users may encounter that shapefiles for different areas contain the same data, such as geologic age, but the data belong under differently-named columns. Furthermore, different sources may apply different terms for the same geologic age (e.g. Jurr for Jurassic), or observe different levels of granularity in category description (e.g., Quaternary vs. Holocene)


http://geon02.sdsc.edu - Query/Selection Results - Microsoft Internet Explorer
Internet
http://geon02.sdsc.edu - Query/Selection Results - Microsoft Internet Explorer
Done
Internet

Data Set Name: NV_500_POLY

Rec	AREA	PERIMETER	NV_500_	NV_500_ID	FMATN	COLR	COLR2	COLR3	D_SYMBOL	TIME_UNIT	P_SYMBOL	#SHAPE#	#ID#
1	0.00115796	0.188837	469	476	Jgr	13	180	180	48	Jurassic	48	[polygon]	468
					Jgr	13	180	180	48	Jurassic	48	[polygon]	487
					Jgr	13	180	180	48	Jurassic	48	[polygon]	561
					Jgr	13	180	180	48	Jurassic	48	[polygon]	647
					Jgr	13	180	180	48	Jurassic	48	[polygon]	2848
					Jgr	13	180	180	48	Jurassic	48	[polygon]	3022
					Jv	10	85	83	48	Jurassic	48	[polygon]	3036
					Jgr	13	180	180	48	Jurassic	48	[polygon]	3068
					Jv	10	85	83	48	Jurassic	48	[polygon]	3197
					Jv	10	85	83	48	Jurassic	48	[polygon]	3464
					Jgr	13	180	180	48	Jurassic	48	[polygon]	3496
					Jgr	13	180	180	48	Jurassic	48	[polygon]	3542
					Jgr	13	180	180	48	Jurassic	48	[polygon]	3552
					Jgr	13	180	180	48	Jurassic	48	[polygon]	3566
					Jgr	13	180	180	48	Jurassic	48	[polygon]	3568
					Jgr	13	180	180	48	Jurassic	48	[polygon]	3571
					Jv	10	85	83	48	Jurassic	48	[polygon]	3577
					Jgr	13	180	180	48	Jurassic	48	[polygon]	3601
					Jv	10	85	83	48	Jurassic	48	[polygon]	3605
					Jgr	13	180	180	48	Jurassic	48	[polygon]	3610
					Jv	10	85	83	48	Jurassic	48	[polygon]	3626
					Jv	10	85	83	48	Jurassic	48	[polygon]	3640
					Jv	10	85	83	48	Jurassic	48	[polygon]	3649
					Jgr	13	180	180	48	Jurassic	48	[polygon]	3654
					Jv	10	85	83	48	Jurassic	48	[polygon]	3694

Data Set Name: AZ_1000_POLY

Rec	AREA	PERIMETER	AZ_1000_	AZ_1000_ID	GEO	PERIOD	ABBREV	DESCR	D_SYMBOL	P_SYMBOL	#SHAPE#	#ID#
1	1.1061	23.2049	3	3371	24	Jurassic	Jgc	GLEN CANYON GROUP	55	48	[polygon]	2
2	0.00238476	0.43169	7	360	23	Jurassic	Js	SAN RAFAEL GROUP	50	48	[polygon]	6
3	0.00341901	0.256402	8	3370	22	Jurassic	Jm	MORRISON FORMATION	50	48	[polygon]	7
4	0.00593057	0.430086	10	361	23	Jurassic	Js	SAN RAFAEL GROUP	50	48	[polygon]	9
5	0.000599083	0.146218	12	3373	24	Jurassic	Jgc	GLEN CANYON GROUP	55	48	[polygon]	11
6	0.0293152	0.962495	25	3374	24	Jurassic	Jgc	GLEN CANYON GROUP	55	48	[polygon]	24
7	0.00265439	0.241931	29	364	23	Jurassic	Js	SAN RAFAEL GROUP	50	48	[polygon]	28
8	0.00406815	0.414198	31	363	23	Jurassic	Js	SAN RAFAEL GROUP	50	48	[polygon]	30
9	0.00125511	0.151081	32	362	23	Jurassic	Js	SAN RAFAEL GROUP	50	48	[polygon]	31
10	0.000116538	0.0470473	35	3375	24	Jurassic	Jgc	GLEN CANYON GROUP	55	48	[polygon]	34
11	0.00211804	0.272261	43	366	23	Jurassic	Js	SAN RAFAEL GROUP	50	48	[polygon]	42
12	0.0026759	0.274178	44	354	23	Jurassic	Js	SAN RAFAEL GROUP	50	48	[polygon]	43
13	0.0465246	2.12374	47	3368	22	Jurassic	Jm	MORRISON FORMATION	50	48	[polygon]	46
14	0.0018846	0.249735	48	353	23	Jurassic	Js	SAN RAFAEL GROUP	50	48	[polygon]	47
15	0.000644411	0.117194	57	3372	24	Jurassic	Jgc	GLEN CANYON GROUP	55	48	[polygon]	56
16	0.00330896	0.294194	69	349	23	Jurassic	Js	SAN RAFAEL GROUP	50	48	[polygon]	68
17	0.0388156	1.60359	75	365	23	Jurassic	Js	SAN RAFAEL GROUP	50	48	[polygon]	74
18	0.00334175	0.288597	76	351	23	Jurassic	Js	SAN RAFAEL GROUP	50	48	[polygon]	75
19	0.0268679	1.74936	81	355	23	Jurassic	Js	SAN RAFAEL GROUP	50	48	[polygon]	80
20	0.000496565	0.0854771	84	350	23	Jurassic	Js	SAN RAFAEL GROUP	50	48	[polygon]	83
21	0.085817	4.55143	100	358	23	Jurassic	Js	SAN RAFAEL GROUP	50	48	[polygon]	99
22	0.0177667	1.33563	110	357	23	Jurassic	Js	SAN RAFAEL GROUP	50	48	[polygon]	109
23	0.00538238	0.331302	112	3376	24	Jurassic	Jgc	GLEN CANYON GROUP	55	48	[polygon]	111
24	0.0968333	4.70444	116	3369	22	Jurassic	Jm	MORRISON FORMATION	50	48	[polygon]	115
25	0.001505	0.154892	129	3377	24	Jurassic	Jgc	GLEN CANYON GROUP	55	48	[polygon]	128

Semantically Heterogeneous Data

Figure 2. Semantic Heterogeneity between attribute tables of Arizona and Nevada geology map sources

These semantic heterogeneity challenges for heterogeneous spatial data sources federated within GEON, were addressed in [2]. The paper focused on the use of ontologies in GEON to register and query disparate data. GEON supports 3 levels of data registration, from most generic to most specific:

1. Data Registration: In this mode data is registered with the GEON repository but is not associated with any ontology and thus cannot be queried intelligently.

2. Item Level Registration: This level of registration allows the data to be registered with one or more ontology concepts, without exposing its schema. Due to the unavailability of schema in this mode, data can be discovered when searched for particular ontology concepts but cannot support fine-grained queries i.e. queries on specific columns.
3. Item Detail Level Registration: This is the most detailed mode of data registration in which one or more data columns can be registered to an ontology and specific mappings of data values to ontology concepts can be defined, thus allowing the data to be queried by concepts instead of actual values.

Registering data sources at the item detail level allowed authors of [2] to resolve the problem of semantic heterogeneity for a given test case of integrating data for 8 state geology maps in Rocky Mountains.

In addition to data registration, GEON also supports ontology registration and thus gives users the flexibility to create their own ontology and associate their data with it. This approach makes GEON a central repository for geoscience ontologies, in addition to geoscience data.

This paper extends the geology map integration application described in [1] and [2] by converting it into a portlet (thus incorporating it with existing geongrid portal) and by extending it to incorporate any GEON-hosted shapefiles instead of only geology data. This paper will explain how and to what extent ArcIMS was used to support this integration and the workflow followed by the map integration portlet. We will also describe how the ontology engine was extended to support ontology enabled creation of thematic maps, including global legend generation and its use in map rendering.

2. Color Enabled Ontologies:

In this application, we consider combining multiple layer fragments retrieved from different spatial data sources. It is fairly easy to use ArcIMS' programming API to generate composite maps dynamically. But assigning proper colors to the maps requires *knowledge about* the spatial data, and *knowledge about* the coloring scheme. As GEON already maintains semantic descriptions of datasets (by associating data elements with concepts in one or more ontologies represented as OWL files), we will use the same framework for managing coloring rules.

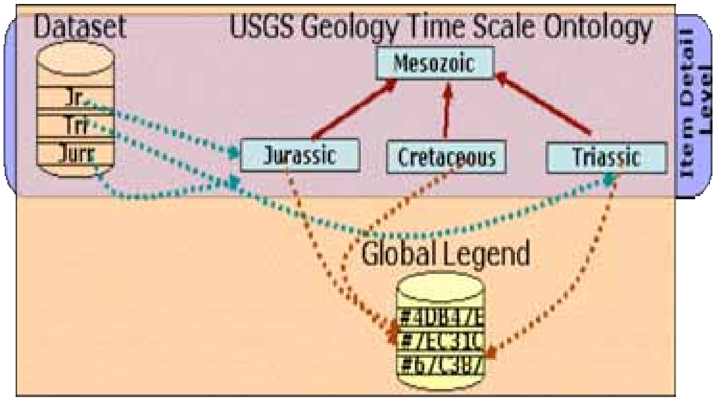


Figure 3. Item Detail Level Registration with colors

Specifically, we allow the ontology owner to provide a simple text file which maps each concept from the given ontology to an RGB color value. This information ultimately leads to mapping data values to proper colors, as shown in the figure 3. Once the global legend is generated, the associated colors, for example, can be included in ArcXML requests against individual sources which retrieve map fragments complying with a common legend. In a slightly different architecture, when spatial data sources return features rather then image fragments, the color scheme is applied differently. Both examples will be described in the next two sections:

3. Generating Integrated Thematic Maps:

As mentioned earlier, this section will describe how shapefile fragments are combined and rendered according to a global legend, which is generated in the process of querying semantically-different resources. The flow of the application is shown in figure 4. It includes the following main components:

1. *GEONSearch/Composition Portlet*: This portlet is one of the tools provided within the GEON portal (www.geongrid.org), which is used to search GEON repository of hosted and non-hosted datasets by dataset type (e.g. shapefile, WMS Service e.t.c.), subjects (a controlled vocabulary of terms), keywords, spatial extent, and associated ontology.
2. *GEON Metadata Catalogue*: This is a database which holds metadata for each dataset in addition to ADN [3] XML metadata schema.
3. *Storage Resource Broker* (SRB) [4]: SRB is data grid middleware developed at San Diego Supercomputer Center. It provides a uniform interface for multiple heterogeneous data resources, supporting common namespace, grid security and other critical attributes of a data grid solution. In our application, SRB is used to control access to GEON hosted datasets.
4. *Mapping Services*: These are web services which act as a wrapper around ArcXML[1] and Java Connector API[2], thus decoupling GEON portal from ArcIMS.
5. *Ontology Engine*: This component represents the ontology based query rewriting mechanism as described in [2].
6. *Ontology Service*: This is a web service wrapper around the ontology engine, responsible for

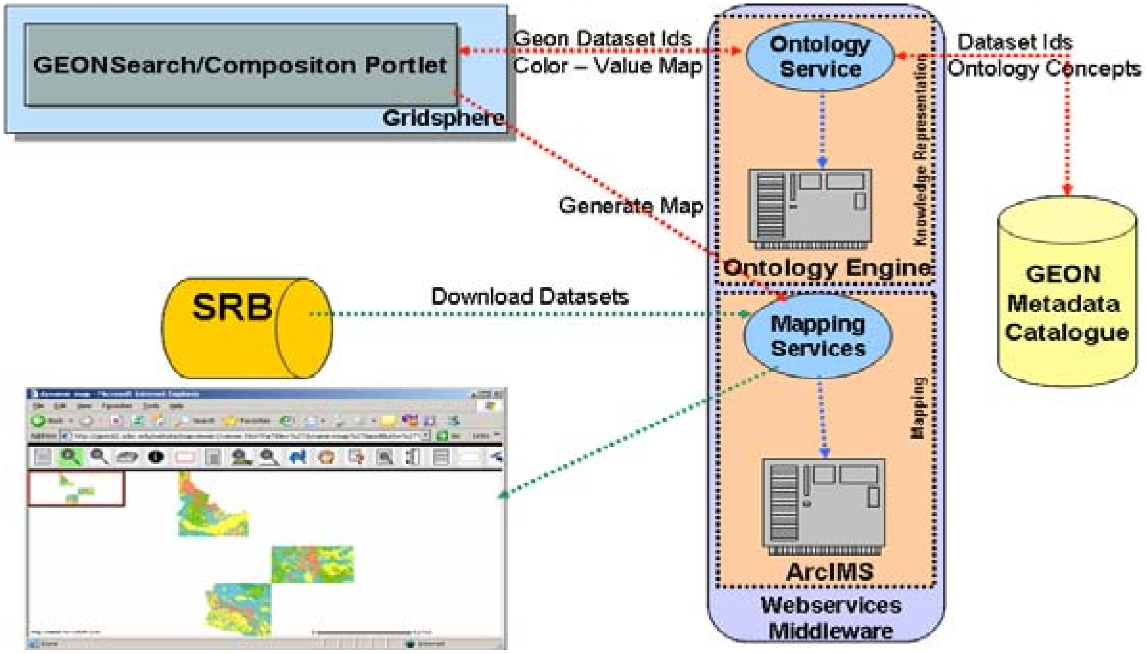


Figure 4. Architecture for Composite Thematic Map Generation

[1] ArcXML is an XML based language which is a standard way of communicating with any ArcIMS Image Service, though HTTP.

[2] Java Connector API is a programming interface, provided by ArcIMS, to support its dynamic operation

relaying requests to the ontology engine thus keeping it separate from the portal.

As shown in Figure 4, the workflow for generation of integrated thematic maps starts when the required shapefiles are discovered using the GEONSearch services. The GEONSearch portlet forwards unique GEON ids[3] for the discovered shapefiles to the ontology service. The service queries the metadata catalogue to return the ontology concepts used to annotate the dataset at registration time. If all selected datasets are registered at the item detail level, and annotated with a single common ontology, then a union of all returned concepts, including both children and parent concepts, is created. At the next step, if a mapping between concepts and symbology (in our case – polygon colors) is defined for the ontology, then the ontology service associates the actual data values corresponding to each given concept in every dataset, with its respective symbol (polygon color). Once the corresponding symbols are retrieved and assigned to geographic objects with particular values, the thematic map is sent back to the GEONSearch Portlet. In the example we consider here, the system discovers color values associated with the expanded query terms, and applies them to the polygon coverages (in ArcIMS lingo: constructing VALUEMAPRENDERER with EXACT element).

The portlet then sends the GEON id and the generated legend (the color-value map, in our case) for each dataset to mapping services, specifically to a ShapeToMap[4] service. This service uses GEON ids to download shapefiles from SRB and stores them on the local disk for a finite period of time (e.g. 6 hours). These cached datasets are registered with a garbage collector, which periodically removes expired datasets (datasets which have outlived their allocated time). Next, an ArcIMS Image Service configuration file is created referencing the locally cached datasets and the color-value map. Once instantiated, this ArcIMS service represents a transient composite thematic map service available for attribute and spatial querying from middleware or directly from various user clients.

The execution flow described above represents a best-case scenario. The following special cases may also occur:

1. None of the datasets is registered at the item detail level: In this case all the datasets are colored randomly.
2. One or more datasets are item detail level registered and atleast one of them is registered with more than one ontologies: In this scenario, the ontology service attempts to find a common ontology for as many datasets as possible. All the datasets, for which a common ontology could not be found, are colored according to their respective ontologies or randomly if they are not associated with any ontology.
3. No color-concept mapping is provided for an ontology: Under this condition, all the datasets which are associated with the given ontology are colored randomly.

[3] In GEON we use randomly-generated unique ids, called GEON ids, to identify all hosted as well as non-hosted datasets. These ids are also used to identify metadata and data itself in SRB.

[4] ShapeToMap is a web service which utilizes Java Connector API to generate dynamic map services from shapefiles.

4. Ontology based Map Integration:

This section will describe how semantically heterogeneous datasets can be queried to generate map feature fragments (as opposed to map images for the entire dataset), which can be combined and visualized using the semantics-aware map assembly services described above. The approach was initially proposed in [2] to demonstrate the usefulness of ontologies in resolving semantic heterogeneity. Here we will describe an extension of the original system to all GEON-hosted shapefile collections.

The system architecture is shown in Figure 5. The following components were added to the original system:

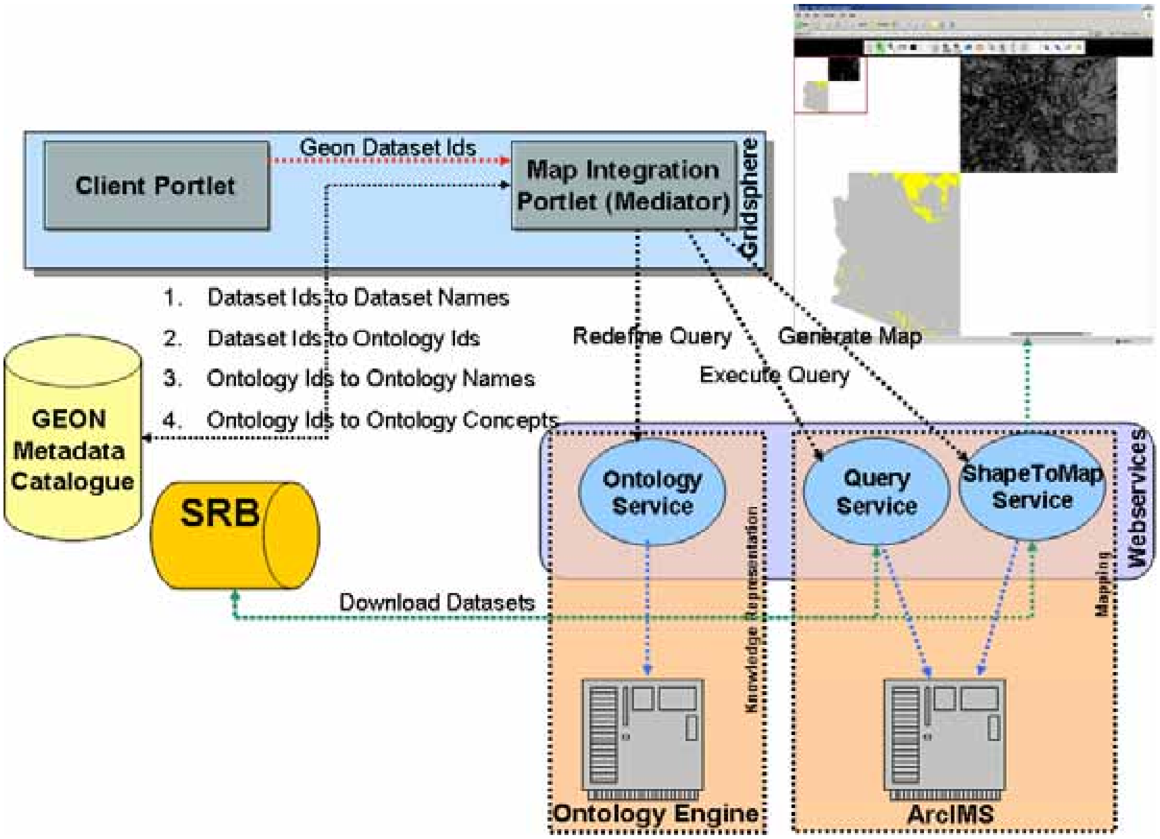


Figure 5. Architecture for Query based Map Integration

1. *Map Integration Portlet*: This portlet displays dynamically generated query interface for the composite map application, manages communication with the user, and serves as a gateway to other services.
2. *Query Service*. This is an additional web service from the mapping services suite, which facilitates querying shapefiles and generation of subsets of data.

As before, a map generation request is initiated when the required sources of shapefiles are discovered through the GEONSearch portlet (Client Portlet). The search portlet sends GEON ids of the selected datasets to a Map Integration Portlet. This portlet resolves GEON ids to ontology concepts and generates a dynamic user interface for data query. If any of the selected datasets are not registered at the item detail level then they are classified as background layers i.e. they are included in the map but cannot be queried. Next, the user chooses the concept(s) for which each dataset will be queried. The portlet, then sends queryable (item detail level registered) GEON ids and selected concepts to the ontology service for query expansion i.e. resolution of concepts to column names and actual data values for each dataset. For each dataset, the ontology service may send any of the following three responses:

1. Dataset is not registered with the selected ontology.
2. Dataset is registered with the selected ontology but does not contain any occurrences of the selected concept(s).
3. Dataset is registered with the selected ontology and contains occurrences of the selected concept(s). For all such datasets the ontology service maps concepts into data values and sends them back to the portlet, along with the column name in which they may occur.

At the next step, the portlet sends GEON ids, data values and column names for all the datasets, which fall in the third category, to the Query Service. The Query Service downloads each dataset from SRB and creates a composite ArcIMS Image Service configuration file, and instantiates the service. The created ArcIMS service supports GET_EXTRACT[5] requests that allow the query service to request shapefile fragments given the selected column names and data values. Finally, the Query Service sends HTTP URLs of all generated zipped shapefiles to the portlet. After that, the portlet sends all original data ids and URLs to the ShapeToMap Service for map generation. In case of geologic polygon coverages, the composite legend is generated using the following rules:

1. All datasets which fall in category 1 are colored dark grey, which symbolizes that these datasets could not be queried as they were not registered with the selected ontology.
2. All datasets which fall in category 2 are colored light grey, which shows that these datasets were searched but no results were found.
3. All datasets which fall in category 3 are colored yellow and light grey. In this combination, yellow color marks those fragments of data which actually contains queried features. Light grey, on the other hand, is used to symbolize the remaining dataset, which were queried but returned empty results.

The ShapeToMap Service retrieves datasets from SRB and the ArcIMS server's output directory where the generated results are deposited. Next, using the color scheme described above, it creates an ArcIMS Image Service that references the locally cached datasets. A GET_MAP request against this newly generated service returns a tri-color map which represents query results graphically.

The portlet also allows users to download the generated data fragments, which then can be registered with the GEON system as regular datasets and shared with the community. If

[5] GET_EXTRACT is an ArcXML request, which is used to return shapefiles through an ArcIMS image service interface. This request results in creation of zipped shapefiles which are subsets of original datasets. These results are stored in ArcIMS output folder and thus are accessible through HTTP.[5]

downloading and registering a dataset to the GEON system is not desired, the user can still work with it by adding it to his/her personal workspace.

5. Future Work:

We plan to enhance the current implementation in the following ways:

1. *Query Optimization*: The procedures described above are fairly time-consuming. In order to prevent repeated execution of map assembly steps in response to the same or similar queries, query results will be saved in SRB and recorded in a Query Tracking database. This database will keep track of queried GEON ids, the concepts for which they were queried and references to generated result sets. Every subsequent query will be checked against the Query Tracking database. If a query identical or reducible to a previous query is requested, then the corresponding resultset will be downloaded from SRB and returned to the client. If the requested query contains concepts which are children of a previous query, then the corresponding resultset will be queried for child concepts, thus avoiding re-querying original resources. The complete query execution machinery will be only invoked if none of the above conditions is valid.

2. *Search by spatial extent*: As mentioned earlier, in our current implementation users are required to pre-select datasets in the GEONSearch portlet in order to view composite thematic maps or otherwise integrate the data. For this application we plan to expose a WMS world map service to enable dataset discovery by spatial extent in addition to text-based search.

6. Acknowledgement:

Partial support under US National Science Foundation grant #0205049 "ITR: GEON: The Geosciences Network: A Research Project to Develop Cyberinfrastructure for the Geosciences" is gratefully acknowledged

7. References:

[1] Zaslavsky, I., Memon, A. GEON: Assembling Maps on Demand From Heterogeneous Grid Sources, in ESRI User Conference, San Diego, California, 2004

[2] Lin, K., Ludäscher, B. A System for Semantic Integration of Geologic Maps via Ontologies, in ESRI User Conference, San Diego, California, 2004

[3] ADN Metadata Schema, http://www.dlese.org/Metadata/adn-item/index.htm

[4] SRB, Storage Resource Broker, http://www.npaci.edu/DICE/SRB/

[5] ArcXML Programmer's Reference Guide 9.0, http://downloads.esri.com/support/documentation/ims_/ArcXML9/Support_files/arcxmlguide.htm

Authors Information

Ghulam Memon
San Diego Supercomputer Center, University of California San Diego
9500 Gilman Drive, La Jolla, CA 92093-0505
(858) 822-3609
gmemon@sdsc.edu

Ilya Zaslavsky
San Diego Supercomputer Center, University of California San Diego
9500 Gilman Drive, La Jolla, CA 92093-0505
(858) 534-8342
zaslavsk@sdsc.edu

Ashraf Memon
San Diego Supercomputer Center, University of California San Diego
9500 Gilman Drive, La Jolla, CA 92093-0505
(858) 822-0017
amemon@sdsc.edu

Kai Lin
San Diego Supercomputer Center, University of California San Diego
9500 Gilman Drive, La Jolla, CA 92093-0505
(858) 822-3649
klin@sdsc.edu

Chaitan Baru
San Diego Supercomputer Center, University of California San Diego
9500 Gilman Drive, La Jolla, CA 92093-0505
(858) 822-5035
baru@sdsc.edu