The GEON Portal: Accelerating Knowledge Discovery in the Geosciences

Ullas Nambiar
ubnambiar@cs.ucdavis.edu

Bertram Ludaescher
ludaesch@ucdavis.edu

Department of Computer Science
University of California, Davis
Davis, CA 95616

Kai Lin
klin@sdsc.edu

Chaitan Baru
baru@sdsc.edu

San Diego Supercomputer Center
University of California, San Diego
San Diego, CA 92093

ABSTRACT

Geoscience studies produce data from various observations, experiments, and simulations at an enormous rate. With proliferation of applications and data formats, the geoscience research community faces many challenges in effectively managing and sharing resources and in efficiently integrating and analyzing the data. In this paper, we discuss how this challenge is being addressed by the GEON Portal, a Web based distributed resource management system that provides integrated access to data and tools needed for knowledge discovery in the geosciences. Unlike previous data management efforts that were either data-driven or application-driven, the GEON Portal provides facilities for efficient sharing, discovery and integration of both data and services that use geoscience data. We identify the challenges involved in managing geonscientific resources and provide solutions that exploit the syntactic, semantic, temporal and spatial metadata associated with the resources. One of our goals is is to provide some insight into the challenges involved in providing a comprehensive scientific data management solution based on our experiences with geoscientific data.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Query formulation, Search process; H.3.5 [**Online Information Systems**]: Web-based services

Keywords

Geoinformatics, Data Integration, Semantic Search, Metadata

1. INTRODUCTION

Many scientific discoveries today are a result of collaborations between researchers sharing data and resources. But the extent of such collaborations is often limited to researchers working within close proximity. By allowing scientists to share data and tools (services) over the Web we can enable interactions between a larger group of researchers working on a common porblem. Unfortunately, scientific applications face unique problems that are not readily addressed by existing data management tools. Specifically, information sources often do not share a common terminology, have a variety of data representation formats and management architectures and exhibit complex relationships between data and tools used to analyze the data.

The GEOscience Network (GEON) [1] is focussed on solving the above problems in the realm of geosciences. The goal of GEON is to respond to the pressing need in the geosciences to interlink and share multi-disciplinary datasets to understand the complex dynamics of Earth systems. Creating a infrastructure to integrate, analyze, and model geoscientific data poses many challenges due to the extreme heterogeneity of geoscience data formats, storage and computing systems and, most importantly, the ubiquity of differing conventions, terminologies, and ontological frameworks across disciplines. Specifically, effective research in the realm of geosciences requires combining the information from a variety of subdisciplines. For example, concerns about climate change require an integrated understanding of stratigraphy, sea-level changes, fossil record, isotopes and tectonics [26]. Hence a scientist working on climate change must have access to data from a number of scientific processes. A single scientist or group would be unable to collect all the necessary information due to the size of data involved. Moreover, collecting, processing and storing geoscientific data is costly since methods of raw data capture such as photogrammetry, remote sensing etc; is very expensive. Thus, the expenses involved in collecting necessary information can become a barrier in the way of scientists exploring new directions of research.

In this paper, we discuss how this challenge is being addressed by the development of the GEON Portal [2]. GEON Portal is distinct from other efforts in scientific data management such as those described in [14, 9, 28] due to the:

1. *Resource registration strategy:* Our solution requires data and service providers to register their resources with the portal. Instead of explicitly mapping resources to each other as is done in mediation systems, we implicitly map the sources to common metadata framework by describing each resource using the 4tuple, *[Metadata descriptions, Ontology mappings, Spatial extent, Temporal extent]*. The 4tuple converts each resource into a point in a 4D space and thereby enables efficient discovery of resources using queries formulated over the 4D space. The ontology mappings become useful in overcoming heterogeneity in local schemas.
2. *Novel architecture:* The framework we have developed contains capabilities of both a *data warehouse* (data providers can store their datasets within the GEON network) and a *data mediation* system (users can design views spanning multiple distributed databases). Furthermore, by supporting integration of both data and services, our framework provides the unique capability to perform both data and application driven integration.

The GEON Portal is publicly accessible and currently contains more than *400 registered data sources, 600 services* and *20 ontologies*. At present there are over *750 registered users* of the portal, a significant number given that the data and tools provided are restricted to geosciences and several components are under early stages of development.

Organization: In Section 2, we provide the motivation behind this effort, and introduce *GEON Portal*, a distributed geoscientific data, services and tools management framework that we are building in Section 3. Section 3.1 describes the syntactic and semantic resource registration model supported by our portal. Resource registration is necessary for providing efficient resource discovery and integration services under the portal. Section 3.2 describes *GEONSearch*, the resource discovery component that allows users to find resources by searching over associated ontology concepts, spatial and temporal extents apart from the standard IR-style keyword search. Section 3.3 describes *GeoMed*, an ontology aware mediator that is being developed as part of the portal to enable integration of structured databases registered with the portal. Finally, we summarize our contributions in Section 4.

2. MOTIVATION

In order to answer new challenges facing the geoscience community, scientists are collaborating across discplines by sharing data and tools. Scientists not only share their applications and data in the raw form, but also those processed using products such as scientific workflows [11]. At first blush, one may argue that a possible solution would be to use a Geographic Information System. However, a number of GIS systems are available that have very similar architecture and functionality and for performance reasons require the data to be stored in proprietary formats. Therefore, a solution involving a GIS system would introduce the additional overhead of converting the data into a proprietary format before it can be shared. Moreover, in [3] it is shown that GIS systems themselves are not compatible with each other and there is growing need for providing a framework for integrating them.

We use the sample query given below to further motivate the need for and challenges involved in building a data management framework for geosciences.

Sample Query: *Plot the gravity measurements near Rocky Mountains where the geologic age is Jurassic.* □

The first step in answering the query above is to identify the datasets of interest. Clearly, we need datasets containing *gravity measurements* in regions around Rocky Mountains and also their *geologic age*. Also required are maps of the region comprising Rocky Mountains. In fact, each state near the Rocky Mountain Range has its own geologic map that provides information about a portion of the range. It is obvious that the geologist would not be able to gather all the necessary data by performing experiments and so must rely on such data being shared by other researchers. However, merely publishing the datasets on the Web would not be helpful since the geologist would then be faced with the task of identifying relevant datasets from among the hundreds of sources on the Web – a task that would involve repetitive query formulation with minimal guarantee of finding the best resources. Thus, a geoscience data management system that allows *easy publishing* and *efficient discovery* of resources is needed to help the geoscientists. The popular keyword search techniques allow efficient searching of sources but require the contents to be searchable (for keywords). But many of the formats used to store geoscience data such as maps, images etc do not have searchable content making retrieving such sources impossible by using content based search. The same is also true for tools. GEON mitigates this problem by providing a comprehensive resource retrieval framework that associates additional user-given metadata to each resource. The metadata is represented as a *[Metadata descriptions, Ontology mappings, Spatial extent, Temporal extent]*. By formulating queries over the metadata, a user can then retrieve any resource that is registered under GEON.

Identifying relevant data sources is only the first step in solving our example query. The next step is to extract relevant information from the sources. However, sources often do not conform to a common format making data extraction a difficult process. In fact, many of the US state geologic maps are available in different formats and served in different projections. Moreover, some geoscientific sources refer to *geologic age* by a single attribute called *Period*, while others refer to it as *Geo_Unit_A*, *Time_Unit, Age,* or by a set of attributes *(Era, System, Series, Period)*. Apart from the schema differences, the actual values referencing *geologic age* may reflect different levels of detail in *geologic age* descriptions. For example, some states may use term *Quarternary* while others would refer to its subdivisions *Holocene* and *Pleistocene*. In such circumstances, asking for *Period=Jurassic* would return incomplete or empty results from some sources, and may even fail if the attribute *Period* is

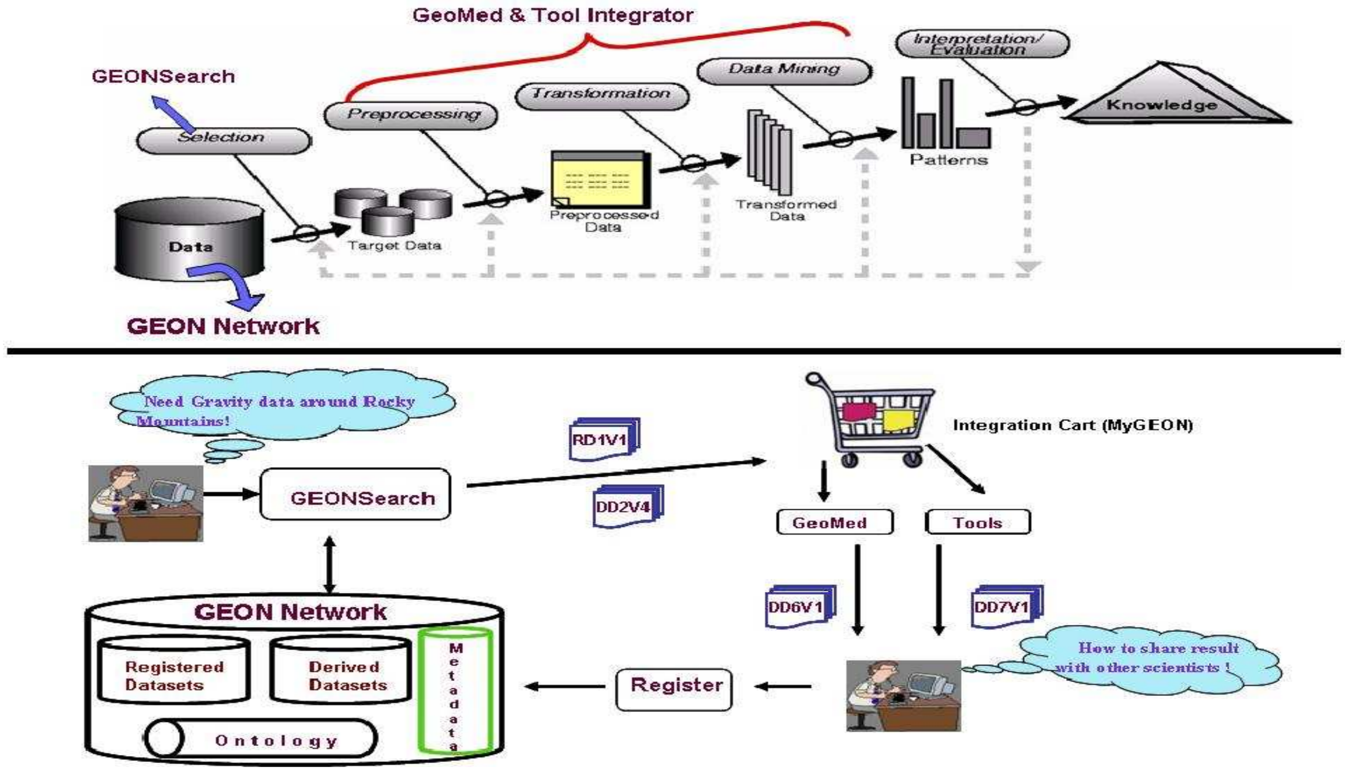


Figure 1: GEON Knowledge Discovery Framework

not present in the schema or if a superclass or subclass of *Jurassic* is used to represent *Jurassic*. Thus, the geoscientists querying the sources would have to resolve problems posed by heterogeneity in schema and data semantics before they can extract necessary information from the sources. Therefore, a second challenge before a geoscience data management system is that of providing a uniform interface over the sources. Obviously this requires resolving the *schematic* and *semantic heterogeneity* among the sources. The task of providing a common interface over sources has been the focus of research done in the context of data mediation (integration) systems [10, 5, 16, 12, 13]. Data integration systems combine multiple data sources such that they appear as a single (virtually) integrated source projecting a single global schema. In doing so, it shields the user from various heterogeneities arising from differences in data source types, data models and query capabilities among different sources. Therefore, a system managing geoscience data must incorporate a *query mediation system* that provides a uniform interface over the sources. GEON resolves the schematic heterogeneity problem providing a mediation system *GeoMed* that mitigates schematic differences between known commerical data formats by using corresponding APIs. To solve the semantic heterogeneity issue, GEON makes use of user-given mappings from schemas to ontologies that are part of the metadata 4tuple collected during resource registration.

Even though we are primarily motivated by the need in geosciences, the challenges faced and the solutions developed are pervasive over scientific data in general. In fact, solutions developed are being adopted by the CUAHSI Hydrologic Information System project (see http://www.cuahsi.org/his/) to setup a portal. Below we summarize the essential features a geoscience data management system should have:

- User controlled mechanism for publishing data and tools.
- Provision to account for changes in content of a published dataset.
- Rich search capabilities for efficient resource discovery.
- Enable seamless querying over multiple databases.
- Support remote invocation of tools for processing data.
- Fault tolerant and secure.

Research at GEON has focussed on responding to above mentioned challenges and has resulted in development of the GEON Portal - an infrastructure that supports discovery and integration of data and tools.

3. GEON PORTAL FRAMEWORK

The GEON Portal is being developed to enable geoscientists to discover new knowledge by bringing together data from various subdisciplines of geoscience. In accordance with the GEON policy of *reuse over rewrite (reinvent)*, we designed the portal based on the GridSphere portal framework[1]. Secure signon capabilities were

[1] Available at http://www.gridsphere.org/gridsphere/gridsphere

added by using GAMA[2] - a credential management solution built for web portals. The top half of Figure 1 borrowed from [6] is a flow representing the sequence of computational tasks involved in deriving knowledge from data. The lower half of Figure 1 depicts how we support the same in GEON. A scientist can select the datasets (and tools) of interest from the GEON Network by using the resource discovery component *GEONSearch*. The selected resources are then stored in a personlized workbench called *MyGEON* from where the scientist can then integrate the resources by placing them in the *integration cart*. For structured databases a uniform query interface is provided by the mediator *GeoMed*. The *Map Integration* [20] component uses mapping services to allow users to integrate (overlap) maps. The new datasets generated after integration is made available under MyGEON as a new resource created by the scientist. This new dataset containing knowledge extracted from multiple reources can in turn be made publicly available by the scientist if desired. The GEONSearch, MyGEON, GeoMed and Map Integration systems are components of the GEON Portal and can be accessed publicly over the Web. For enabling a personalized workspace for each user, we require interested users to register to the system.

A key inital step in the discovery process above is availability of the datasets. Only datasets that have been registered via the GEON Portal are available to the scientists from the portal. The GEON Portal supports two models of data storage – *local hosting* and *remote hosting*.

Local hosting: In this model, a copy of the dataset being registered is stored in the GEON repository. All subsequent references to the dataset from within the portal refer to the registered copy. No format transformation is done during this process. This mode of registration is useful in alleviating the problems of data management that a scientist would face. The GEON repository consists of a *Storage Request Broker (SRB)*[3] and a *PostgreSQL database*[4] that contains the GEON system catalog. The SRB is a middleware system that brings efficient archival storge capabilities to GEON - an essential capability for local hosting. A locally hosted dataset must have one of the following formats - *ASCII (simple text files), Excel datasheets, ESRI Shapefiles, GMT Raster data, GeoTIFF, Ontology (OWL files), PDF, NetCDF files and Tools (executables)*.

Remote hosting: Some resources registered via the portal may not be amenable to centralized storage particularly if they have complex schemas, are large in size and may be updated often. Relational databases are an example of such datasets. Since most commercial database implementations can be accessed remotely, GEON Portal allows such resources to be registered by only providing a remote connection URI and access rights. Only a copy of the source schema is made but no content is moved to the GEON repository. At query time, the portal establishes a connection to the resource using the URI and extracts the necessary content dynamically by forwarding the user query. An example of a remotely hosted resource currently accesible through the GEON Portal is the Paleobiology Database (PBDB)[5] that is hosted at University of California Santa Barbara and maintained by a internationl group of paleobiological researchers. Another category of resources that are registered as remotely hosted resources are *Web services*.

All resource registrations, irrespective of the model used, require the user to provide four types of meta information namely *syntactic Metadata, Ontology mappings, Spatial extent* and *Temporal extent*. The syntactic metadata captured by GEON is a subset of the ADN metadata framework[6]. The remaining metadata is optional but providing these when available will help in improving resource discovery and integration.

At GEON we must deal with two types of data source heterogeneities: *syntactic* and *semantic*. Syntactical heterogeneity arises when the underlying technology supporting data sources differs (e.g. web based interface, differing commercial databases, flat files etc). This kind of heterogeneity is hidden from the users by using corresponding *APIs*. Semantic heterogeneity may be due to differences in source ontologies and nomenclature or vocabulary used. The *Semantic Resource Registration* module of GEON portal provides a mechanism for resolving the semantic heterogeneity by allowing data providers to map the source schemas (attributes, values) to concepts and instances of an ontology. All sources mapped to a common ontology will then appear semantically homogeneous to the components of the portal.

In the following, we describe the resource registration, GEONSearch and GeoMed components of the portal. We describe the current implementations and list several new areas of research we are undertaking based on feedback we received from the users.

3.1 Resource Registration

The resource registration component of the GEON Portal consists of two subcomponents - (1) Syntactic registration and (2) Semantic registration. In the following, we elaborate upon them and then present details about current implementation of the registration system.

Syntactic Registration: The process of describing the physical schema of the sources alongwith the access methods and processing capabilities of the underlying database engine is called syntactic registration. For locally hosted data, this involves only providing the metadata. The access methods and capabilities are decided by the portal based on the format. For remotely hosted data, a URI that points to the resource must be provided alongwith all required metadata. If the resource is a database then GEON Portal will extract the schemas of the tables within the database. The registered relations are only visible to the user until they are explicitly shared by the registering user.

Semantic Registration: The process of describing mappings from source schema to an ontology is called Semantic Registration. There is no restriction on the number of ontologies to which a

[2] Details about GAMA are available at http://grid-devel.sdsc.edu/gama

[3] Available at http://www.npaci.edu/DICE/SRB/CurrentSRB/SRB.htm

[4] Available at http://www.postgresql.org/.

[5] Available at http://paleodb.org/cgi-bin/bridge.pl

[6] Details at http://www.dlese.org/Metadata/adn-item/

resource can be mapped. Moreover, we allow partial mapping i.e. only parts of the source schema may to concepts in the ontology. Mappings even if partial are useful in resource discovery. Given that the users registering sources are most likely to be domain scientists, we cannot expect them to provide formal mappings. We therefore provide an easy-to-use interface for mapping the sources to the ontology. The internal representation of the mappings is done using the *Ontological Database Annotation Language (ODAL)* [19] a language with syntax close to OWL designed for generating partial mappings from database schema to ontologies represented using OWL. ODAL differs from OWL and its variants due its explicit support for databases. The mappings represented by ODAL can be formally represented by using the mapping signatures $\alpha : Schema \rightarrow Ontology$ described in [4].

While, the semantic registration is usually done from source to ontology, under GEON, we also wish to allow mappings between sources. The motivation there is to allow users to provide high-level information about semantic similarity of the content of two sources when available. Specifically, if the same user is registering multiple sources to a single ontology, then he may be able to provide information about whether the sources are *equivalent* or if a source is a specialization (subset) of another source (containment). Such mappings will help in optimizing queries issued over the ontology.

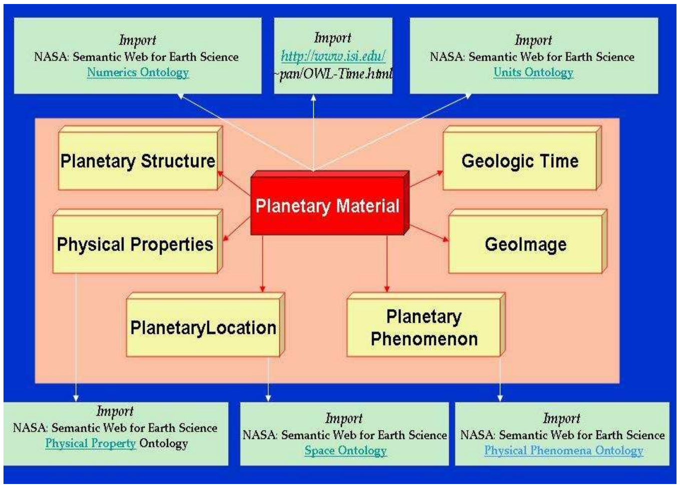

Import
NASA: Semantic Web for Earth Science
Numerics Ontology
Import
http://www.isi.edu/
~pan/OWL-Time.html
Import
NASA: Semantic Web for Earth Science
Units Ontology
Planetary Structure
Geologic Time
Planetary Material
Physical Properties
GeoImage
PlanetaryLocation
Planetary
Phenomenon
Import
NASA: Semantic Web for Earth Science
Physical Property Ontology
Import
NASA: Semantic Web for Earth Science
Space Ontology
Import
NASA: Semantic Web for Earth Science
Physical Phenomena Ontology

Figure 2: High Level Architecture of Planetary Ontology [27]

Current Implementation: The GEON Portal currently supports three levels of data registration, from most generic to most specific. They are

1. *Resource Registration:* In this mode the resource is registered with the GEON repository by providing the ADN metadata but is not associated with any ontology. While the task is simple, the lack of semantic metadata makes discovering and integrating such resources difficult.

2. *Item Level Registration:* This level of registration allows the datasets to be registered with one or more ontologies and higher level concepts within the ontology. However, the underlying schema is not mapped to the ontology.

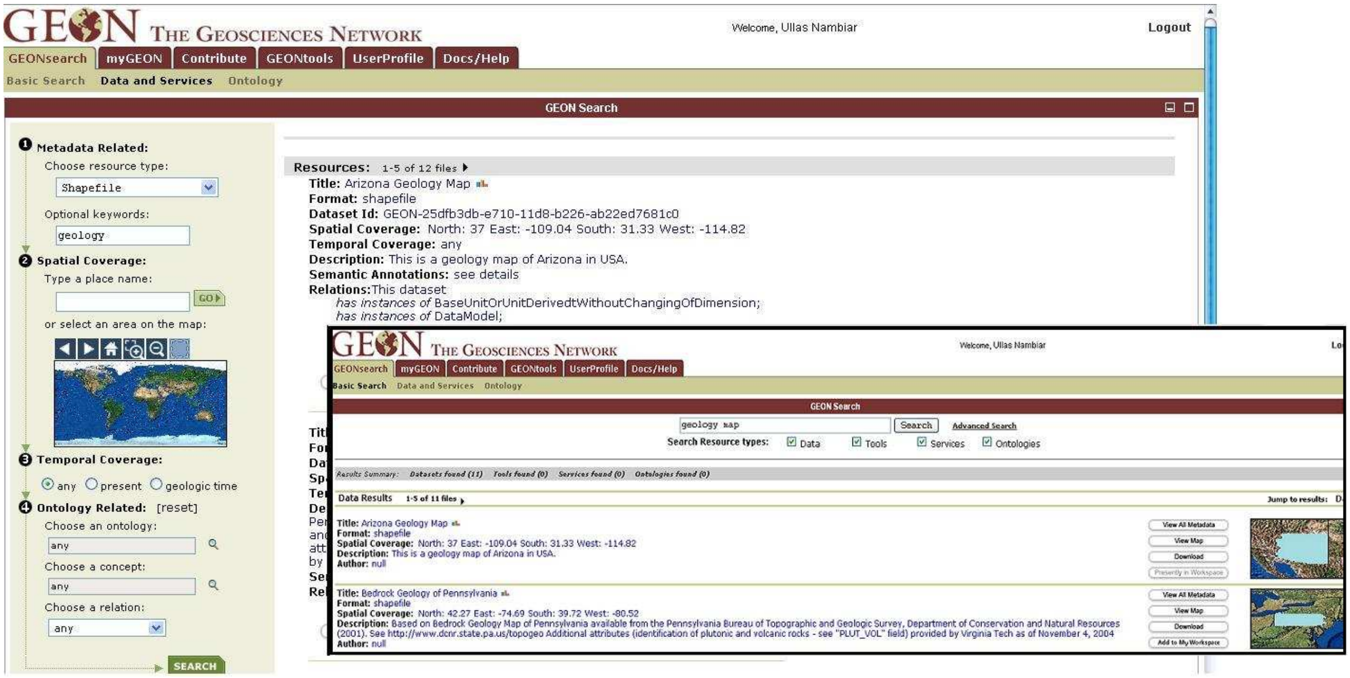
3. *Item Detail Level Registration:* This is the most detailed mode of data registration in which both attributes in the source schema and the values binding those attributes can be mapped to ontology concepts, thus allowing the resource to be queried using concepts instead of actual values. This mode of registration is most suitable for datasets built on top of relational databases. However, GEON also enables item detail level registration for Excel spreadsheets and maps in ESRI Shapefile format by internally mapping such datasets to PostreSQL tables. Registering data sources at the item detail level helps resolve the problem of semantic heterogeneity. For example, in [18], semantic heterogeneity issues among data from 8 state geology maps in Rocky Mountains is overcome by registering the maps at item detail level.

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. A separate initiative to provide a single widely accepted geoscience ontology is also being supported under GEON. Figure 2 provides a high level architecture of the Planetary ontology [27] being developed as part of GEON. As depicted, this new ontology will be made up of several ontologies (both existing and new) for various sub-disciplines. Specifically, it will include concepts for *Planetary Materials (elements, isotopes, rocks and minerals), Planetary Structure, Planetary Location, Planetary Phenomenon, Physical Properties, Geologic Time and Geo Images*. Once developed, the ontology will be made available through the GOEN Portal and users can then provide mappings to this new ontology which will be comprehensive and if required will be extended to incoporate new ontologies. By moving to a single ontology we can avoid issues involved in reconciling differences in concept defintions arising due to multiple overlapping ontologies.

The datasets hosted by GEON are all generated as a result of a scientific inquiry such as measurement of seismic activity, study of geochemistry, mineralogy and igneous rock samples etc. Given that many of the datasets will be hosted locally by GEON, GEON can be considered as a virtual scientific database that is used to accession objects representing results of scientific inquiry. An accessioning system like GEON must provide the datasets with stable identifiers that can be included as references in publications resulting from new scientific inquiries that use such datasets. GEON provides each registered resource with a unique *GEON ID* that is based on the *UUID* standard defined[7] by OSF. However, the requirement for stable object identifiers can conflict with the tendency of scientific data to evolve over time. In fact, based on feedback we have received from users, we have identified following research tasks for improving data management under GEON:

Task 1: *Version Support and Provenance-* A resource registered under the portal may be modified at a later date. While portal fa-

[7] Definition available at http://www.ietf.org/rfc/rfc4122.txt



1 Metadata Related:
Choose resource type: Shapefile
Optional keywords: geology

2 Spatial Coverage:
Type a place name:
or select an area on the map:

3 Temporal Coverage:
● any ○ present ○ geologic time

4 Ontology Related: [reset]
Choose an ontology: any
Choose a concept: any
Choose a relation: any

Resources: 1-5 of 12 files
Title: Arizona Geology Map
Format: shapefile
Dataset Id: GEON-25dfb3db-e710-11d8-b226-ab22ed7681c0
Spatial Coverage: North: 37 East: -109.04 South: 31.33 West: -114.82
Temporal Coverage: any
Description: This is a geology map of Arizona in USA.
Semantic Annotations: see details
Relations: This dataset
has instances of BaseUnitOrUnitDerivedtWithoutChangingOfDimension;
has instances of DataModel;

geology map [Search] Advanced Search
Search Resource types: ☑ Data ☑ Tools ☑ Services ☑ Ontologies

Results Summary: Datasets found (11) Tools found (0) Services found (0) Ontologies found (0)

Data Results 1-5 of 11 files

Title: Arizona Geology Map
Format: shapefile
Spatial Coverage: North: 37 East: -109.04 South: 31.33 West: -114.82
Description: This is a geology map of Arizona in USA.
Author: null

Title: Bedrock Geology of Pennsylvania
Format: shapefile
Spatial Coverage: North: 42.27 East: -74.69 South: 39.72 West: -80.52
Description: Based on Bedrock Geology Map of Pennsylvania available from the Pennsylvania Bureau of Topographic and Geologic Survey, Department of Conservation and Natural Resources (2001). See http://www.dcnr.state.pa.us/topogeo Additional attributes (identification of plutonic and volcanic rocks - see "PLUT_VOL" field) provided by Virginia Tech as of November 4, 2004
Author: null

Figure 3: Basic and Advanced GEONSearch Interface

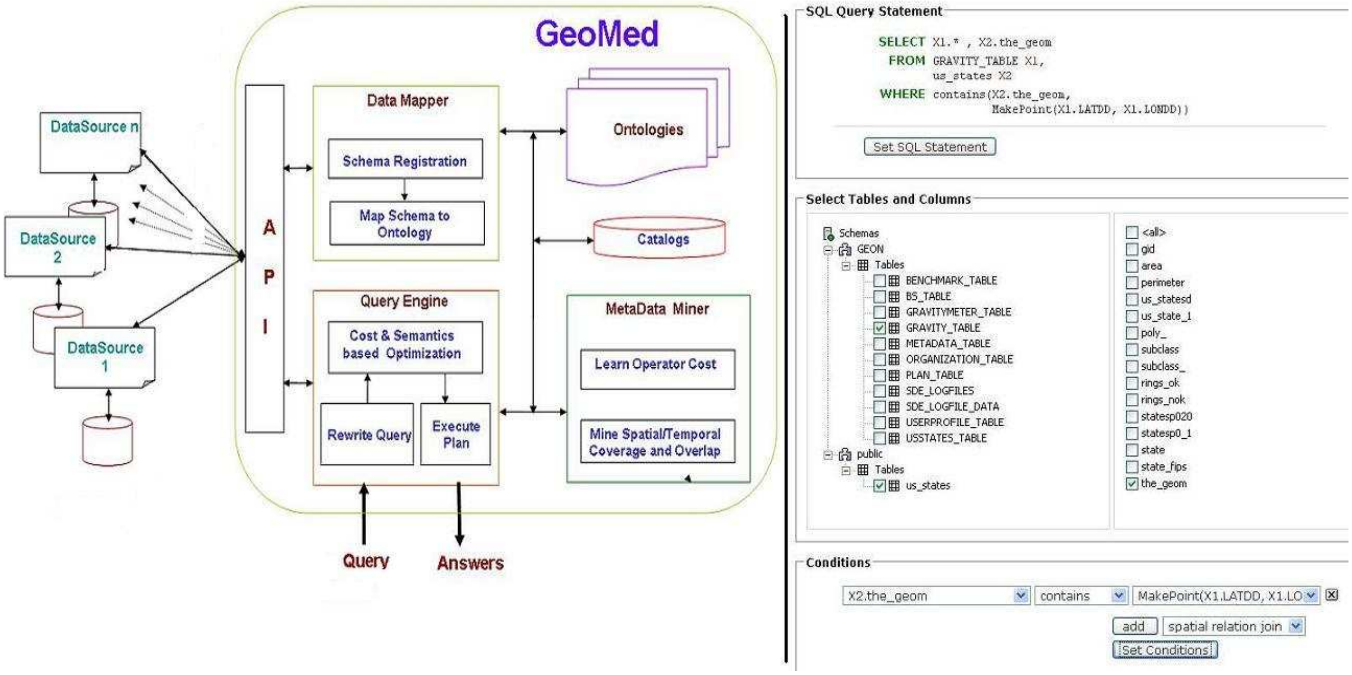
cilitates updating of registered data, it is not practical to replace existing content as it may have been used already in other scientific tasks. Therefore, GEON repository must support more than one version and each must be uniquely identified. Since content changes to remotely hosted data cannot be controlled by GEON, we are currently working on providing version support for locally hosted data. An immediate result of supporting versions is the need for *provenance management*. Even though much research in the realm of propagating provenance information has been done, there are no systems that can be used off-the-shelf. Morevoer, GEON faces unique challenges in that we wish to provide provenance information for a variety of data, tools and also processes that can be performed over the GEON Portal.

Task 2: *Automatic Change Detection/Reporting* - A user may want to see how much affect (improvement) the new version of an input dataset would have on the query/process she is interested in. Sometimes the new version might contain data that is not relevant to the process for which the user is using the dataset and so would not want to replace the dataset. But the user updating the input dataset would not be aware of all possible uses of the dataset and therefore is unlikely to provide information that would be satisfactory to all users of the dataset. For example, if a scientist only wants to plot gravity points in Davis, the fact that a new version of the gravity dataset contains samples from Sacramento is not useful to her and so she may not wish to use that dataset. Depending on the type of dataset being updated, the change detection may be simple or complex and will require further research.

3.2 Resource Discovery using GEONSearch

The GEON network together with the registered datasets can be seen as a intranet where as already motivated there is a need for supporting efficient resource discovery. Information retrieval (IR) is the science of searching for information in documents, searching for documents themselves, searching for metadata which describe documents, or searching within databases, whether relational stand-alone databases or hypertext networked databases such as the Internet or intranets, for text, sound, images or data. Automated information retrieval systems, popularly known as search engines have been built to extract information by formulating queries related either to the objects or their metadata. However, the resources registered to GEON have a variety of formats with many of them not suitable for content based search as is supported by search engines. Therefore, we designed GEONSearch, the retrieval engine for the GEON Portal to supports spatial, temporal, and concept-based search apart from the traditional IR-style keyword search.

Current Implementation: GEONSearch provides two search modes for extracting data from the GEON Portal - *Basic Search* and *Advanced Search*. Figure 3 shows the two modes provided under GEONSearch with Basic Search in the smaller pane on the right. The current implementaion of GEONSearch only provides the search capabilities over the metadata associated with the dataset. This decision is based primarily on the fact that content search is not feasible or not cost-effective over most of the resources currently registered in GEON. In the Basic Search mode users can issue keyword queries over the *Keywords, Title and Description* associated with each registered resource. The datasets whose metadata match the user query are returned in decreasing order of relevance where relevance is judged using traditional vector space models.



SQL Query Statement

```
SELECT X1.* , X2.the_geom
  FROM GRAVITY_TABLE X1,
       us_states X2
 WHERE contains(X2.the_geom,
              MakePoint(X1.LATDD, X1.LONDD))
```

Set SQL Statement

Select Tables and Columns

Schemas
- GEON
  - Tables
    - □ BENCHMARK_TABLE
    - □ BS_TABLE
    - □ GRAVITYMETER_TABLE
    - ☑ GRAVITY_TABLE
    - □ METADATA_TABLE
    - □ ORGANIZATION_TABLE
    - □ PLAN_TABLE
    - □ SDE_LOGFILES
    - □ SDE_LOGFILE_DATA
    - □ USERPROFILE_TABLE
    - □ USSTATES_TABLE
- public
  - Tables
    - ☑ us_states

□ <all>
□ gid
□ area
□ perimeter
□ us_statesd
□ us_state_1
□ poly_
□ subclass
□ subclass_
□ rings_ok
□ rings_nok
□ statesp020
□ statesp0_1
□ state
□ state_fips
☑ the_geom

Conditions

X2.the_geom | contains | MakePoint(X1.LATDD, X1.LO ☒

add | spatial relation join

Set Conditions

Figure 4: GeoMed Architecture and Query Interface

The Advanced Search mode allows users to search over the ontology mappings and spatial and temporal extents associated to each resource. Also users can issue keyword queries over a particular type of resource. To find resources using spatial information a user must provide a bounding box of region that might cover the spatial extent of a registered resource. A GUI is provided to help users determine a bounding box (see left panel of larger pane in Figure 3). Similarly users can provide a range of time to extract resources that may be extents within the range. Users can also look for objects that were semantically registered to GEON by selecting a concept available in registered ontologies and specifying the relation that the resource should have with the concept such as *has instances of, is related to, mentions, uses etc*. The relations are fixed by GEON and are same as those available for item level and item-detail level registration.

The GEONSearch system has been well received by users. While many users are satisfied with the available techniques, some have requested a more powerful system. An ongoing task [23] is to move towards a system that not only retrieves answers satisfying a query but can also suggest related answers that may not satisfy the original query completely.

Task 3: *Recommender System* -Effective retrieval of datasets from GEONSearch will require users to construct keyword queries that clearly identifies their need - a task that is often difficult even for experts. By focussing on exact keyword matches GEONSearch would miss to provide datasets that are relevant to the user query but whose metadata do not match with the query. For example, if users search for *California*, they will not find relevant documents mentioning only *San Diego*. Similar problems could arise when searching over spatial and temporal extents of the resources. The problem also appears for concept based search provided by GEONSearch since the current framework would only return the datasets that are mapped using a user specified relation. Thus even in the presence of advanced search capabilities, users may have to formulate multiple queries to find relevant data. A system that can recommend datasets and tools that are similar to those which are retrieved by user query would greatly help users. The similar resources may not satisfy the user query but actually be the resource the user inteded to look for when formulating the query. Thus, by providing similar resources we may be able to reduce the time users spend in extracting resources. Hence, we are focussing on providing similar results apart from the relevant ones currently given by GEONSearch. The similarity estimation will be done using background knowledge available from GEON ontologies, dataset-to-ontology mappings, past usage information etc. We plan to leverage our experiences with supporting similarity search over autonomous databases [22] to help support similarity search in GEONSearch.

3.3 Integrating databases using GeoMed

Integrating the databases registered under GEON is challenging as they are registered mostly as remotely hosted sources making them autonomous. Moreover depending on the underlying database engine (DB2, PostgreSQL, Oracle, SQL Server) they may have differing formats and processing capabilities. Moreover, data from different sources may have differing semantics, e.g. data can be defined under different geospatial models. Thus, effective integration in this scenario warrants resolving both the syntactic and semantic heterogeneity among the sources. As explained earlier in Section 3.1, by registering at item detail level both types of heterogeneities can be resolved. The mediation system we have built as part of GEON, *GeoMed* relies on the sources being *"aware of"* the integration and providing item detail level registration if they want the resource to be effectively integrated. There are two predominant techniques to map the database (local) schema to the the global schema [7]. In the *global-as-view (GAV)* model, the global schema is defined as a view over local schemas. In contrast, in the *local-as-view (LAV)* model, a global schema is defined first by modeling the application domain. Then the source schemas and their data objects are defined as views over the global schema. For query evaluation, the rules have to be *inverted* or *folded* [17, 12].

Fixing a unique global schema that would limit the potential types of queries users can pose over the system. Moreover, such a schema may not be acceptable to the users. A likely solution then is to use an *ontology* that is widely accepted in the geographic domain as the *global schema*. The user queries can be posed on the ontology, which the mediator can translate to queries over the local schemas. In recent years, this approach has been adopted by several researchers to overcome the heterogeneity problem in distributed processing environments [8, 15, 25, 21]. Hence, GeoMed also uses ontologies to resolve the schema heterogeneity problem. In essence, the *ontology* becomes the *global schema* and the semantic mappings given by the data provider (user registering the database) become the rules necessary for rewriting a query on the ontology to a query on the underlying source.

Current Implementation: The GeoMed system as illustrated in the left half of Figure 4 consists of four main components - *Data Mapper*, *Ontologies*, *Query Engine* and *Metadata Miner*. The Data Mapper provides tools for syntactic and semantic registration of the databases. Data Mapper is part of the resource registration module. The schema information and mapping rules are stored in the *Catalogs*. Mapping the source schemas to ontologies provides us with a mechanism to easily overcome the schema heterogeneities that will arise when multiple sources are registered to GeoMed. Query rewriting will be done by the *Query Engine*. The rewritten query will be optimized by the Query Engine before it is executed. Efficient query optimization necessitates access to a number of statistics like selectivity of the query, cardinality of the source, response time of the source given the query, response time of a source for given spatial operations etc. Given a source, the number of feasible queries on the source is exponential in the number of attributes and their binding values. Since, much of the statistics required for optimization are query specific, we cannot assume these to be provided by the data provider. Moreover, the statistics will change over time as additional data is added to the source thereby requiring periodic updates. *Data Miner* will learn and update the various statistics required for optimizing the query plans. The statistics will also be stored in the *Catalogs*.

We support two types of data integration under GeoMed depending whether the user of the system reconciles the schema differences. The techniques are:

- *User driven integration:* In this model, the user is allowed to select the databases she wishes to integrate and thereby decide the global view that she wishes to have. Details of how to access the databases etc are still hidden from the user. The right half of Figure 4 shows the query interface that supports such integration. The query language supported is SQL since it is the common language supported by all commercial relational databases. In database parlance, we can equate this mode to supporting formulation of join queries over a federated database. No ontology mappings are used in this case and so the user must do resolve the schema heterogeniety if any. Obviously this mode is targeted for users who are comfortable with issuing SQL queries.
- *Ontology driven integration:* As the name suggests, this model allows users to formulate queries over the ontology. These queries will in turn be pushed down to the databases that map to the ontology. The system is currently under development as there are several challenges involved, primary among them being the lack of a single ontology for GEON. Other challenges include, the need for an easy to use querying model given the large size of the ontology, efficient source selection in the presence of overlapping sources and collecting statistics for supporting cost-based optimization.

Based on our observations about the content distribution of datasets and their usage we have identified the following challenge:

Task 4: *Accounting for Content Overlap* – Mutliple sources registered under GEON Portal provide overlapping content. Calling all sources relevant to a would be time consuming and costly due the potential number of duplicates. Irrespective of the integration model used, it is necessary to identify the best subset of sources that can answer a query. Source specific metadata such as coverage of each source and the degree to which it overlaps with other sources is required to identify the best subset order for answering the query. Such statistics are not usually available given the autonomous nature of the sources. We intend to leverage our experience [24, 13] in mining source statistics over autonomous Web sources to solve this problem.

4. CONCLUDING REMARKS

This paper summarizes some lessons learned in building a resource management system, the GEON Portal, that addresses effective integration of multi-disciplinary data and tools in geosciences community. GEON Portal allows both local and remote hosting of data and tools, provides easy interface to markup resources using ontologies and spatiotemporal metadata, provides efficient resource discovery through GEONSearch and a semantics aware mediator, GEOMed for integrating databases. The portal isolates the users from the complexity of using distributed heterogeneous resources. It does so by providing interactive assistance to associate datasets to ontologies, and then to formulate integrated queries that are automatically mapped to the underlying databases.

We described several ongoing research tasks that were undertaken based on feedback received from registered users of the portal. To the best of our knowledge, there is no scientific data management system similar to GEON Portal although the need for such system is becoming critical in other domains. The challenges we faced are not limited to geosciences and therefore our solutions can be easily used to satisfy resource management needs in other scientific domains.

ACKNOWLEDGEMENTS

We thank GEON team members - particularly, Dogan Seber, Ashraf Memon, Ghulam Memon and Choonhan Youn of San Diego Supercomputer Center; for helping us understand various facets of the GEON portal.

5. REFERENCES

[1] GEON - Cyberinfrastructure for Geosciences. *http://www.geongrid.org*.

[2] GEONGrid Portal. *https://portal.geongrid.org:8443/gridsphere/gridsphere*.

[3] O. Boucelma. Experiences in Building a Geographic Integration System. *IIWeb*, 2004.

[4] S. Bowers and B. Ludaescher. A Calculus for Propagating Semantic Annotations through Scientific Workflow Queries. *EDBT'06 Post-Conference Workshop on Query Languages and Query Processing*, 2006.

[5] S. Chawathe, H. Garcia-Molina, J. Hammer, K. Ireland, Y. Papakonstantinou, J. Ullman, and J. Widom. The TSIMMIS project: Integration of Heterogeneous Information Sources. *In Proceedings of the 100th Anniversary Meeting, Information Processing Society of Japan, Tokyo, Japan*, pages 7–18, October 1994.

[6] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors. *Advances in Knowledge Discovery and Data Mining.* AAAI/MIT Press, 1996.

[7] D. Florescu, A. Levy, and A. Mendelzon. Database Techniques for the World-Wide Web: A Survey. *SIGMOD Record, 27(3)*, September 1998.

[8] J. Grant, J. Gryz, J. Minker, and L. Raschid. Semantic Query Optimization for Object Databases. *ICDE*, November 1997.

[9] L. Haas, B. Eckman, P. Kodali, E. Lin, J. Rice, and P. Schwarz. DiscoveryLink. *Bioinformatics: Managing Scientific Data*, 2003.

[10] L. Haas, D. Kossmann, E. Wimmers, and J. Yang. Optimizing Queries Across Diverse Data Sources. *In proceedings of VLDB*, 1997.

[11] E. Jaeger, I. Altintas, J. Zhang, B. Ludaescher, D. Pennington, and W. Michener. A Scientific Workflow Approach to Distributed Geospatial Data Processing using Web Services. *SSDBM*, 2005.

[12] S. Kambhampati, E. Lambrecht, U. Nambiar, Z. Nie, and G. Senthil. Optimizing Recursive Information Gathering Plans in EMERAC. *Journal of Intelligent Information Systems, Volume 22, Issue 2*, March 2004.

[13] S. Kambhampati, U. Nambiar, Z. Nie, and S. Vaddi. Havasu: A Multi-Objective, Adaptive Query Processing Framework for Web Data Integration. *ASU CSE TR-02-005*, April, 2002.

[14] Z. Lacorix, O. Boucelma, and M. Essid. The Biological Integration System. *WIDM*, 2003.

[15] L. Lakshmanan and R. Missaoui. On Semantic Query Optimization in Deductive Databases. *ICDE*, 1992.

[16] A. Levy, A. Rajaraman, and J. Ordille. Querying Heterogeneous Information Sources Using Source Descriptions. *In proceedings of VLDB, Bombay, India*, 1996.

[17] A. Levy, A. Rajaraman, and J. Ordille. Querying Heterogeneous Information Sources using Source Descriptions. *In proceedings of VLDB, Bombay, India.*, pages 251–262, 1996.

[18] K. Lin and B. Ludaescher. A System for Semantic Integration of Geologic Maps via Ontologies. *ESRI User Conference*, 2004.

[19] K. Lin, B. Ludaescher, and C. Baru. Ontological database annotation language. Technical report, San Diego Supercomputer Center, 2005.

[20] G. Memon, A. Memon, K. Lin, I. Zaslavsky, and C. Baru. Generating composite thematic maps from semantically-different collections of shapefiles and map services. *ESRI*, 2005.

[21] E. Mena, V. Kashyap, A. Sheth, and A. Illarramendi. OBSERVER: An approach for query processing in global information processing systems based on interoperation across pre-existing ontologies. *Conference on Cooperative Information Systems, 41: 14-25*, 1996.

[22] U. Nambiar and S. Kambhampati. Answering Imprecise Queries over Autonomous Web Databases. *ICDE*, 2006.

[23] U. Nambiar, B. Ludaescher, G. Memon, and D. Seber. GEONSearch: From Searching to Recommending. *GeoInformatics*, 2006.

[24] Z. Nie, U. Nambiar, S. Vaddi, and S. Kambhampati. Mining Coverage Statistics for Websource Selection in a Mediator. *In proceedings of CIKM, McLean, Virginia*, November 2002.

[25] N. Paton, R. Stevens, P. Baker, C. Goble, S. Bechhofer, and A. Brass. Query processing in TAMBIS bioinformatics source integration system. *Statistical and Scientific Database Management*, 1999.

[26] A. Sinha, editor. *GEOINFORMATICS: Data to Knowledge*. The Geological Society of America, 2006.

[27] A. Sinha, K. Lin, R. Raskin, and C. Barnes. Cyberinfrastructure for the geosciences: ontology based discovery and integration. *GeoInformatics*, 2006.

[28] R. Tuchinda, S. Thakkar, Y. Gil, and E. Deelman. Artemis: Integrating Scientific Data on the Grid. *Innovative Applications of Artificial Intelligence*, 2004.