DIA: A Web Services-based Infrastructure for Semantic Integration in Geoinformatics

Zaki Malik, Abdelmounaam Rezgui, A. Krishna Sinha*, Kai Lin**, and Athman Bouguettaya

Department of Computer Science, Virginia Tech, VA. USA
*Department of Geosciences, Virginia Tech, VA. USA
** San Diego Supercomputer Center, UCSD, CA. USA
{zaki, rezgui, *pitlab, athman}@vt.edu, **klin@sdsc.edu

Abstract

We present DIA, a Web services-based infrastructure for the Discovery, Integration, and Analysis of geoscience data, tools, and services. DIA provides a collaborative environment where scientists can share their resources (e.g., geochemical data, filtering services, etc.) by registering them through well-defined ontologies. We have developed a planetary materials ontology in OWL for this purpose. The ontology is used by different geoscientists (using Web services) to explore, extract, and integrate information from different heterogeneous data sets. The DIA system is now in its final pre-release phase. It is currently made accessible to a few geoscientists for conducting usability analyses, and it will eventually be made available to the community at large through the geoscience portal (GEON) at the San Diego Supercomputer Center.

1. Introduction

Several research initiatives by geoscientists over the years have produced large amounts of data. However, the ability to find, access, and properly interpret these large data repositories has been very limited. Two main reasons for this lack of data sharing are the absence of publicly advertising research data, and the adoption of personal acronyms, notations, conventions, units, etc. by each research group, when producing data. This makes it difficult for other scientists to correctly understand the semantics of the produced data, and makes the automatic interpretation and integration of data simply infeasible. This leads to a poor ability in answering complex geoscientific questions.

We believe that to enable the sharing and integration of geosciences data on a global scale, ontology-based data registration and discovery is required. Hence, we have defined a “planetary materials” ontology for efficient, reliable, and accurate data sharing among geoscientists. The ontologically registered data can be used through different geoscientific tools to answer complex user queries. Thousands of such tools have been made available in the geoscientific literature, and a few have also been implemented. Up till now, most tool implementations have been private and similar to the data sharing problem, tool sharing has also been limited. This has resulted in redundant efforts for tool deployment. Geoscientists have been defining and using “stand-alone” tools, in segregated environments. The tools that are defined in one system are seldom used in the other. Even if the tools are shared in a public-setting, the interoperability of the different systems has been problematic. Therefore, geoscientists have mostly been “re-inventing the wheel” in terms of tool implementations. This has been a major impediment in taking geoinformatics forward. In many cases, geoscientists may neither be able to deploy, nor implement the necessary tools, due to lack of resources, knowledge, etc. Thus, many geoscientific queries remain unanswered.

We believe that the Web services initiative undertaken by the World Wide Web Consortium (W3C) can resolve the problem of tool sharing across different systems. Web services can provide a platform that facilitates cooperation among researchers and allows integration of tools. The participants of a Web services system do not have to worry about the operating system, development language environment or the component model used to create or access the services. The tools developed by geoscientists can simply be wrapped as Web services and made accessible to the scientific community at large. In Web services, only the input and the output parameters are defined and there is no need to know the exact details of the involved processes. This encourages application reusability and reduces development time, effort, and cost.

In this paper, we describe our approach towards organizing geoscientific data through well-formed ontologies, and the need for defining different tools as Web services. Specifically, we describe the DIA engine: a system for the Discovery, Integration, and Analysis of geoscience data. The DIA engine will be made available to the geoscience community through the GEON portal. GEON is a federally funded, multi-pronged project. As a collaboration among several institutions, projects, and institutions, GEON’s main objective is to provide...
cyberinfrastructure support for integrative geoscience research.

The paper is organized as follows. In Section 2, we define geoinformatics and show how geoscientists can benefit from this approach. In Section 3, we provide the details about the ontologies that we have defined. In Section 4, we introduce the DIA engine and show how Web services are used in the system. Section 5 provides some concluding remarks and direction for future work.

2. Geoinformatics

Man has been studying the earth and the solar system for centuries. Over the years, the study of several phenomena as eclipses, tides, volcanism or earthquakes, etc. has proven to be challenging. The inability of scientists to create frameworks that provide integrative solutions has been the prime impediment in this regard. Nowadays, the need for multidisciplinary observations and their integration is well-understood. However, we still conduct our science in similar ways to our forefathers. For instance, we make our observations both on the ground as well as through remote sensing techniques and store our information in computers, but still find it difficult to achieve an integrative understanding of our objectives. It poses a fundamental problem in seeking a more robust knowledge about the earth and the solar system. The prominent reasons for this are: (1) vast volume of data which individuals cannot assimilate, (2) inability to discover heterogeneous data, (3) variability in format of data even within a single discipline, (4) focus on narrowly defined science questions that cannot be easily expanded to a higher and integrative level of understanding of how and why the earth works as it does, and (5) changing the societal needs that influence support for research and education. To make novel geoscience discoveries, we will require the rapid availability of data and information that encompass a variety of temporal and spatial scales, as well as the capability to integrate heterogeneous data sets and tools to analyze them.

Geoinformatics can be thought as the field in which geoscientists and computer scientists are working together to provide the means to address a variety of complex scientific questions using advanced information technologies. Thus, as a concept, geoinformatics is rooted in providing rapid access to both academic and agency hosted databases, with tools and services, so that the best integrative approaches in (earth) science research and education can take place. As an initial step towards realizing the goal of geoinformatics, we have defined the planetary materials ontology that covers a variety of concepts, minerals, elements, isotopes, etc. for geoscience data. Moreover, we have built a prototype system (DIA) as proof of concept, that shares and uses a variety of tools to answer complex geoscience questions. In the following, we provide details of the tools and technologies used in constructing the DIA engine.

3. Semantic integration through ontologies

As the World Wide Web evolves from a set of isolated application systems to a network of interacting disparate systems, the need to represent the semantics of the exchanged information (so that it could be automatically understood) is becoming a necessity. This is where ontologies can play an integral role. Ontologies can aid in providing machine processable semantics of the information communicated between heterogeneous systems. Prior research regarding data heterogeneity has led researchers to divide data heterogeneity into three main categories, i.e., syntactic, structural, and semantic heterogeneity [1]. Data transformation approaches as database mediation can be used to handle syntactic and structural heterogeneity. However, semantic heterogeneity cannot be resolved using such techniques. Therefore, ontologies provide a viable solution for the problem of semantic heterogeneity.

Ontologies have gained popularity in various research communities such as e-commerce, language processing, knowledge engineering, and information integration. An ontology may be defined as a formal and explicit specification of a shared conceptualization [3]. The abstraction of concepts in a particular domain is known as “conceptualization.” “Shared” refers to the consensus of the relevant domain. Since the development of ontologies is a cooperative process that involves a number of entities (often not co-located), all entities (e.g., businesses, researchers, government agencies) have to agree on the concepts and definitions within that ontology. The word “explicit” in the above definition means that the concepts used in an ontology and the constraints on their use are explicitly defined. Moreover, “formal” means that the defined ontology should be described using a well-defined model or language. This enables the ontology to be machine understandable. In other words, an ontology may be defined as a set of knowledge terms, including the vocabulary, the semantic interconnections, and some simple rules of inference and logic for some particular topic [2].

3.1. Ontologies for Geoscience

To enable the sharing and integration of data on a global scale, we (with the support of fellow researchers) have introduced the idea of ontology-based data registration and discovery in geosciences. Our goal in
Defining ontologies is to provide an organizational structure for classifying data that can be discovered automatically by computers. These high-level ontologies and data level ontologies contained within, allow geologists to discover databases as well as datasets within databases using geoscience related concepts instead of simple keyword-based search. This is made possible as ontologies allow the registration of databases at different levels of granularity. For example, the hierarchical concept for rocks: [Rock -> Igneous Rock -> Geochemistry of Igneous Rock -> element abundance] provides the conceptual framework for registering and indexing a database at increasing levels of detail. This forms the basis for developing the ontologies that represent more explicit statements about the data.

As mentioned earlier, the use of these ontologies, especially at the most detailed level, facilitates semantic integration of heterogeneous geologic datasets. We favor ontology based integration as it systematically resolves both syntactic and semantic heterogeneity, thus allowing integration of multiple distributed databases into a single virtual database. Unlike integration based on merging multiple schemas, ontologic registration relates the data to concepts, rather than the structure within a database [12].

Figure 1 shows the high-level representation of the Planetary Ontology. For example, the package “Planetary Material” can be used to represent the nature (physical, chemical) of substances and their properties. This figure also shows the utilization of imported and inherited properties from additional packages, e.g., Physical Properties, Location, and Planetary Structure, to more fully define the concept of Planetary Materials. Existing ontologies available from SWEET ontology library [4] such as Numeric, Time, and Units are also used as they provide common concepts which are useful for the development of data level ontologies.

### 3.2. DIA’s Ontology Repository

As part of the GEON project, DIA uses the ontology support functions of GEON [12]. The ontology repository defined to facilitate data exploration and integration, accepts and saves the defined ontologies in OWL DL, the description logic variant of OWL [5]. Apart from browsing options, the ontology repository at GEON, provides an option to combine, parameterize, and switch between different ontologies. This is a useful practice as mappings can be defined that translate between classes and properties in one classification and corresponding classes and properties in another classification [6, 7]. In describing the GEON ontology mappings below, we assume that all ontologies are formalized in the same logic, i.e., we consider mappings between different ontologies (not between different logics).

An **ontology mapping** from an ontology $\Omega_\alpha$ to $\Omega_\beta$ is defined as having a class mapping $f$ and a property mapping $g$. Class mapping is a partial function from the class set of $\Omega_\alpha$ to the set of all derived classes in $\Omega_\beta$. A derived class is defined from other classes, e.g. an intersection of two classes. The mapping should preserve the subclass ($is a$) relation, i.e., $\alpha_1$ and $\alpha_2$ are classes in $\Omega_\alpha$ and $\alpha_1$ is a subclass of $\alpha_2$, then $f(\alpha_1)$ must be a subclass of $f(\alpha_2)$ in $\Omega_\beta$:

![Figure 1. The planetary ontology](image)

If $f$ does not preserve the subclasses relation, then a query about individuals in $f(\alpha_1)$ may incorrectly return some individuals in $f(\alpha_2)$ [12].

The property mapping $g$ is a partial mapping from the property set of $\Omega_\alpha$ to the set of all derived properties in $\Omega_\beta$, and should satisfy the following condition: If $p$ is a property between the classes $\alpha_1$ and $\alpha_2$ in $\Omega_\alpha$, then $g(p)$ is a property between the classes $f(\alpha_1)$ and $f(\alpha_2)$:

![Diagram of property mapping](image)

Note that a class mapping and a property mapping induce a natural translation from the constraints of $\Omega_\alpha$ to the constraints of $\Omega_\beta$. For instance, if $\Omega_\alpha$ has a constraint
ontology OA, which means that there is an inclusion
where the parameter ontology \( \Omega \) is included in the body
of any model
valid general ontology mapping from
been done by using \( e \) and a property
model of \( \Omega \alpha \) for the same concept. For example, suppose ontology
ontologies may contain similar but different definitions
but a person may be not an employee.

Data properties
of the following:
Theoretically this result is equivalent to the pushout [8]
this condition [12].
should preserve the subclass relation is a special case of
this condition [12].

If there is an ontology mapping from \( O \alpha \) to \( O \beta \), then
from any model \( M \) of \( O \beta \), a sub-model can be extracted
and naturally transformed into a sub-model of \( O \alpha \) based
on this ontology mapping. An ontology mapping is
more general than a simple equivalence relation on
classes and properties. In fact, it defines a structural
translation from one ontology to another. For example,
let \( O \alpha \) contain a class \( \text{Person} \) and a data property
\( \text{name} \) and \( O \beta \) contain a class \( \text{Employee} \) and two
data properties \( \text{name} \) and \( \text{id} \), then a mapping sending
\( \text{Person} \) to \( \text{Employee} \) and \( \text{name} \) to \( \text{name} \) is a
valid general ontology mapping from \( O \alpha \) to \( O \beta \), so any
model of \( O \beta \) can be naturally transformed into a model of
\( O \alpha \) via this mapping, that is, an employee is a person,
but a person may not be an employee.

In general, combining several ontologies can not be
done by a simple union operation, because these
ontologies may contain similar but different definitions
for the same concept. For example, suppose ontology
\( O \alpha \) has a property \( p \) defined between the classes \( \alpha 1 \) and
\( \alpha 2 \) (i.e., \( \alpha 1 \rightarrow p \alpha 2 \)) and \( O \beta \) has a property \( q \)
between \( \beta 1 \) and \( \beta 2 \), and assume that \( \alpha 1 \) and \( \beta 1 \) are
contceptually the same. Combining \( O \alpha \) and \( O \beta \) should
give an ontology \( O \) containing a property \( p \) between a
new concept \( \alpha \) (of \( O \), representing \( \alpha 1 \) and \( \beta 1 \) and \( \alpha 2 \),
and a property \( q \) between \( \alpha \) and \( \beta 2 \). In OWL, this can be
done by using \( \text{equivalentClass} \) or \( \text{sameAs} \) tags. Theoretically this result is equivalent to the pushout [8]
of the following:

\[
\text{\alpha 1} \overset{p}{\rightarrow} \text{\alpha 2} \quad \text{\beta 1} \overset{q}{\rightarrow} \text{\beta 2}
\]
\[
\downarrow \quad \downarrow
\]
\[
\alpha \quad \beta
\]

where \( \hat{O} \) is an ontology containing only one class \( \alpha \). The
ontology mapping \( \psi 1 \) sends \( \alpha \) to \( \alpha 1 \), and \( \psi 2 \) sends \( \alpha \) to
\( \beta 1 \).

Note that parameterized ontologies and their
instantiation can be implemented based on the same
principle. A parameterized ontology is a pair \( (O, O \alpha) \)
where the parameter ontology \( \hat{O} \) is included in the body
ontology \( O \alpha \), which means that there is an inclusion
ontology mapping from \( \hat{O} \) to \( O \alpha \). To instantiate this
parameterized ontology with an actual ontology \( O \alpha \), an
ontology mapping from the formal ontology \( \hat{O} \) to \( O \beta \)
must be provided. The instantiation result is the pushout
of the diagram above, i.e., the ontology \( O \) in the
diagram. It is unique up to ontological isomorphisms.
The system can infer implied facts in the selected
ontologies. Ontology mappings are used in navigation
and query processing. If a data set is registered to an
ontology \( O \alpha \), and there is an ontology mapping from \( O \beta \)
to \( O \alpha \), then users can choose both, ontology \( O \alpha \) and
ontology \( O \beta \) to query the data set.

4. The DIA engine

The primary objective of constructing the DIA
engine is to build a service-oriented computational
infrastructure that enables scientists to Discover,
Integrate, and Analyze earth science data. The DIA
engine is designed to provide earth scientists the
capability to better understand the relationships between
the observed geologic records and the complex
processes that have shaped them over the years.

DIA comprises of three main phases, namely,
discovery, integration, and analysis. Data discovery
enables the users to retrieve the distributed data sets,
which are located at multiple sites that are pertinent to the
research task at hand. Data integration enables users to
query various data sets along some common attributes
to extract previously unknown information called data
products. The data products that are generated can either
be used in their delivered form or used as input to the
data analysis phase. Data analysis may be used to verify
certain hypothesis or it may refine the data product with
further data discovery and integration. We believe that
such a cyber-infrastructure for the geosciences is a
requirement for both improved efficiency and trust for
online conduct of science.

To illustrate the different DIA phases, and the
science and technology components that constitute
geoinformatics, consider the following example query:
What is the distribution, U/Pb zircon ages and the
gravity expression of A-type plutons in Virginia? Are
these A-type igneous rocks in Virginia related to a hot
spot trace? Are the bodies tabular in 3-D geometry?

A number of steps are involved in answering
the above-mentioned query. The logical sequence of these
steps, as discovering data, identifying A-type bodies,
accessing their published ages, and deploying gravity
modeling tools after accessing gravity data have been
described by Sinha et.al. in [10]. We will trace these
steps through the DIA engine. Querying the DIA engine
taunts five phases: (i) resource registration, (ii) query
specification, (iii) data discovery, (iv) filtering and
integration, and (v) analysis.
4.1. Resource registration

DIA’s registration system allows geoscientists to register their data (which is normally in Excel files) to one or more ontologies, for the purpose of sharing. The goal is to allow researchers to associate one or more ontologies to their files so that a unique and definite meaning is associated with each column. Moreover, it will become possible to relate this column to columns with similar (or close) semantics in other files. Since GEON enables users to register their own ontologies as well [9], it is possible for them to use their own ontologies when registering Excel files. Discovery can be facilitated by resource registration at three levels:

(i) **Keywords-based registration:** Discovery of data resources (e.g., gravity, geologic maps, etc) requires registration through use of high level index terms. GEON has deployed extension of AGI Index terms. These extensions will be cross indexed to others such as GCMD and AGU.

(ii) **Ontological class-based registration:** Discovering item level databases requires registration at data level ontologies. In this case, a data set is registered to a data level ontology, e.g. bulk rock geochemistry, gravity database, etc.

(iii) **Item detail level registration:** Item detail level or fine-grain registration consists of associating a column in a database to specific concept or attribute of an ontology, 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 PostgreSQL tables. For example, a column in a geochemical database may be specified as representing SiO2 measurement. This level of registration is a requirement for semantic integration, i.e., the automatic processing (by tools) of shared data.

4.2. Query specification

In the DIA engine, the user query can be expressed in one of two ways: it can either use a text-based format or a menu-based format. The text-based format allows a user to query the entire database while the menu-based format lets the user select only specific items, which in turn queries only a subset of the data. The user does not need extensive knowledge of the querying techniques, models or keywords (which may be required in a text-based format). The task at hand can be completed with the help of a few “mouse-clicks,” and query results are definitely produced as long as the data required to answer the query is present, i.e., empty result sets are only returned in case of missing data. The user clicks through the different menus to “build” an exact query. Using our running example of distribution of A-Type plutons, Figure 2 shows what menus are involved in getting to the A-type filtering (Four top-level menus are navigated to get to the A-Type menu). Similarly, the “Tool Selection” layer in Figure 3 shows how menus are navigated and how Web services are used.

![Figure 2. Query specification through menus](image)

4.3. Data discovery

Currently, GEON supports four methods for data discovery. These methods differ by the type of their search key: by keywords, resource type, temporal attributes, or spatial attributes. A user specifies the type of search key and the system returns a list of documents (maps, Excel files, etc.) that have the entered Search key in their metadata [9]. The GEON discovery 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 [11] 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.

4.4. Filtering and integration

Data filtering is a process in which the DIA engine transforms a single raw data set into a data product. Data filtering may also take a data product as its input. Examples of data products include a map showing the A-Type bodies in the Mid-Atlantic region, an Excel file giving the ages of those A-Type bodies, a gravity database table spatially related to A-Type bodies saved as a contoured gravity map, etc. Data products used in data integration may be of two types: pre-packaged or created dynamically. Querying pre-packaged data is usually faster but is not flexible and provides little support for complex scientific discovery. Dynamically created data products may require on-the-fly integration
Figure 3. A layered bottom-up depiction of DIA phases from data to data products and extensive query processing, but enable far richer possibilities for scientific discovery. There are two main classes of integration:

*Intra-class integration* is a process whose input is two or more homogeneous data sets, i.e., registered to the same ontology. This process uses the common ontology to interpret all data sets and generate an integrated data product. *Inter-class integration* is a process whose input is two or more heterogeneous data sets (i.e., registered to different ontologies.) This process uses the appropriate ontology to interpret each data set. It uses an integration class to generate a data product out of two or more data sets. Figure 3 shows the integration between geospatial, geotemporal, and gravity gridding data products that are obtained after data filtering through a “bounding box” (used to limit the area of interest).

**4.5. Analysis**

Geoscientists can use the data products generated as a results of the above mentioned phases in hypothesis evaluation, i.e., to analyze the results. In Figure 3, using a space-time analysis and studying the generated 3-D model, a geoscientist can verify the hypothesis on A-Type bodies. Our running example query (*A-type rocks*) has major implications on tectonic models such as (1) a failed rift associated with a triple junction [13], (2) gravitational collapse of crystal regions over thickened by Grenville orogenesis [14], (3) flanks of an active within plate rift zone similar to the Red Sea region [15], and (4) as a continental plume track. Thus, facilitating rapid data /tool access clearly is a requirement as geoscientists engage in more complex queries.

**4.6. DIA’s service-oriented approach**

The DIA engine is a Web-based, service-oriented system developed using a variety of technologies including: ESRI’s ArcGIS Server 9.1, Microsoft’s .NET framework, Web services, Java, and JNBridge 3. Users submit queries through the DIA’s Web-accessible
The DIA engine’s architecture and workflow (Figure 4). The engine translates these queries into a sequence of tasks that include: accessing map servers, discovering and accessing data sources, invoking Web services, filtering features, joining layers, and the graphic rendering of query results for visualization. The DIA engine also enables users to save their query history as well as export data products for future references. Since the DIA engine is developed along a service-oriented approach, key code modules are wrapped as Web services. This approach has two advantages. First, it makes the system readily extensible. As the geoscience community introduces new services, these could be integrated in the DIA engine as new functionalities. Second, services developed for the DIA may be used as building blocks to produce other systems.

Figure 4 depicts the service-oriented architecture of the DIA engine and also describes how this architecture supports the discovery, integration, and analysis of Earth science data. The DIA engine supports several querying modes (Geological map-based queries, Region-based queries, etc.) To answer the running example A-type query, the user first selects the option “Geological map-based queries” in the DIA’s main menu. The system then accesses a geological map server, gets a (default) geological map, and displays it to the user (Figure 4). This map enables the user to select the area of interest (i.e., Virginia). This may be done by selecting a bounding box or by selecting the entire state. In the latter case, the DIA engine accesses a gazetteer to determine the selected state’s latitude-longitude coordinate.

The user then uses DIA’s drop down menus to identify a computational filter (A-type filter in this case) to be applied to the data samples located in the selected area. The DIA system invokes a GEON Web service called GEONResources that provides functions for searching and getting the metadata information for GEON resources including all the registered databases that have samples located in the area of interest. In the case of the A-type query, the DIA invokes the service GEONResources with the keyword “GeoChemistry” as well as Virginia’s latitude-longitude coordinates. GEONResources then returns a set of database identifiers that correspond to (registered) data sets containing geochemical data for samples located in Virginia. For each returned database, the DIA system executes a two-step process. First, it builds a virtual query (expressed in SOQL, a language developed by GEON’s researchers at SDSC) that requests all the data (i.e., columns) that are necessary to apply the filter specified by the user. The DIA system then invokes a GEON Web service called SoqlToSql that translates this SOQL query into a SQL query. In the second step, DIA submits the SQL query to the GEON server that interacts with the actual database server, gets a record set containing the relevant data samples, and returns the data to the DIA engine. The DIA engine then inputs the received data to the A-type filter and displays the query results to the user. Note that the A-type filter tool is also a Web service that may be hosted other geoscientists. This facilitates reusability, as new filters do not need to be written by each individual group. Similarly, gravity information etc. is gathered and a data product is generated. The A-type bodies data product and gravity data product are then integrated to deliver the desired query result in a visual manner.

The DIA Engine is now in its final pre-release phase. Currently, the DIA engine is accessible through http://mapserver.geos.vt.edu/ATypeDemo at Virginia Tech. It will eventually be made available as a portlet in the GEON portal.
5. Conclusion

We have presented our approach for the semantic integration of data and tools for geoinformatics. The resulting system termed as the DIA engine, uses ontologies and Web services to organize, annotate, and define datasets and geoscience tools. We have defined several geoscience ontologies and have wrapped several geoscience tools as Web services. These tools and ontologies are available for public use through the GEON portal at the San Diego Supercomputer Center. The DIA engine is in its final pre-release phase, and will soon be made accessible to the geoscience community at large through GEON.

We expect that as the Semantic Web matures, and more geoscientists adopt this paradigm, a number of geoscience tools will be made accessible as Web services. This would require that similar to data management, Web services are also ontologically registered. Annotating Web services with semantics would ensure that appropriate tools (in form of Web services) are selected in an efficient and automatic manner for answering geoscience queries. In the future, we intend to define a service ontology for geosciences.

6. Acknowledgement

This work is supported by the US National Science Foundation award EAR 0225588 to A. K. Sinha and National Science Foundation grant number 0627469 to A. Bouguettaya.

7. References


