GEON: Toward a Cyberinfrastructure for the Geosciences—A Prototype for Geologic Map Integration via Domain Ontologies

By Bertram Ludäscher[1], Kai Lin[1], Boyan Brodaric[2], and Chaitan Baru[1]

[1]San Diego Supercomputer Center
University of California, San Diego
9500 Gilman Dr., MC 0505
La Jolla, CA 92093-0505
USA
Telephone: (858) 822-0864
e-mail: {ludaesch,klin,baru}@sdsc.edu

[2]Geological Survey of Canada
601 Booth Street
Ottawa, Ontario, K1A 0E8
Canada
Telephone (613) 992-3562
e-mail: brodaric@gsc.NRCan.gc.ca

ABSTRACT

When trying to combine different geologic maps, a number of interoperability challenges need to be overcome. We first provide an overview of those challenges and then briefly describe a mediator architecture devised to overcome them. We then focus on the problem of providing integrated access to a set of geologic maps from different state geological surveys, by defining a global view on the different local source schemas. Next we address the problem of content heterogeneity by defining an "integration ontology" to which the various local data content are mapped. The integration ontology consists of various sub-ontologies, such as one for geologic age (Poling, 1997), and several others relating to rock classification (chemical composition, texture, fabric, and genesis). The latter are derived from a recent proposal for rock classification (Struik and others, 2002). Based on the integration ontology, the prototype allows the user novel "concept-based" access and querying capabilities across the different geologic maps. This system is being embedded in the service-oriented data grid infrastructure under development in GEON.

INTRODUCTION

The Geoscience Network (GEON) is a collaborative NSF/ITR (National Sciences Foundation / Information Technology Research) project to create "cyberinfrastructure" for the Geosciences (GEON, 2003). GEON addresses the need in the geosciences to interlink and share multidisciplinary data sets in ways that allow researchers improved data access, data integration, and ultimately the construction of "scientific workflows" to combine data integration and analytical steps in a more seamless manner than is currently possible. The ultimate goal of GEON is to bring together heterogeneous scientific data and information to facilitate new ways of information integration and knowledge discovery.

The information technology (IT) research and development areas of GEON include data modeling and integration, grid systems development, and visualization. In the geoscientific component of GEON a set of scientific questions are centered around two test beds, the Rocky Mountain Region (uplift of Colorado Plateau), and the Mid-Atlantic Region (crustal (terrane) evolution). Data integration and mediation IT efforts are driven by and applied to the domain scientists' needs as exemplified by those test beds.

In this paper we describe one of the initial data integration efforts of GEON, the ontology-enabled map integration (OMI) system. Specifically, we describe the current prototype, which allows the user to query geologic maps provided by different state geologic surveys. The remainder of the paper is structured as follows: In the next section, we briefly recall the various levels of heterogeneity that present a challenge to data integration in general, followed by an overview of an extended mediator architecture that is used to overcome the interoperability challenges. Then, we elaborate on the ontology-enabled map

integration prototype. Finally, we conclude and outline some future plans.

DATA INTEGRATION: OVERVIEW AND MEDIATOR ARCHITECTURE

Levels of Interoperability and Standards

When trying to share distributed, heterogeneous data, a number of technical challenges must be overcome; these can be roughly classified as described below (see also Sheth (1998) for a historical perspective). Consider, for example, two systems having data sets that should be made interoperable (Figure 1). One can employ standards and technologies to overcome the various kinds of heterogeneities, and to facilitate interoperability at different levels. At the systems level in Figure 1, we may find different operating systems (Linux, MS Windows, MacOS, …), different data transport protocols (FTP or HTTP, which are built on top of a stack of internet protocols called TCP/IP), or higher-level protocols for discovery and interoperation of web services. Differences in system platforms and operating systems are usually overcome by standardizing protocols for data transport and remote service execution. For the latter, for example, one can employ web service descriptions (WSDL, 2001), which specify the input and output parameters of a web service. System level interoperability for GEON can also be achieved at the grid service level. Grid services extend the basic web service infrastructure and include additional features such as user authentication for secure data access. Apart from the generic issues of data access, transport, and remote execution, there are also a number of application specific system level issues, e.g., the choice and architecture of the mapping technology for rendering spatial information (server-side, client-side, mixed).

At the syntactic level, we consider heterogeneities such as different data file formats, e.g. SHP (ESRI shape files) and DXF (AutoCAD drawings). XML, the Extensible Markup Language (XML, 2000) provides a simple and very flexible syntax for structuring many kinds of data and metadata to enable their exchange. Defining such a new structure in XML syntax can be done in different ways. For example, one can provide an XML Document Type Definition (DTD) or an XML Schema definition (XML Schema, 2001) to specify the allowed nesting structure and (in XML Schema) the data types of XML elements. This not only yields a data exchange syntax but also prescribes a schema for the exchanged data. Additional semantics such as domain specific integrity constraints have to be encoded by other means. The Resource Description Framework (RDF, 2003) can be seen as an XML dialect for encoding labeled, directed graphs, in particular ontologies (see below). For querying databases, query languages such as SQL (for relational databases) or XQuery (for XML databases) are used, each of which come with their own syntax for query expressions. Differences at the syntactic level are usually resolved either by adhering to a standard or by using format converters that can translate from one format to another.

At the schema level, heterogeneities can exist because

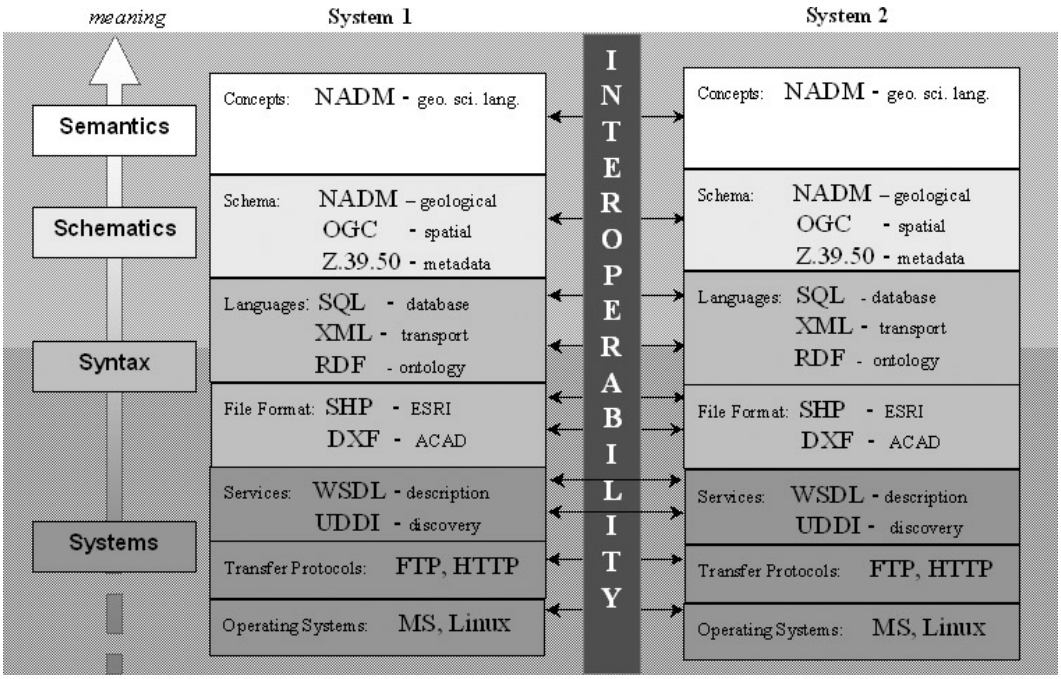

Figure 1. Levels of interoperability and standards.

the same (or at least similar) data can be represented using vastly different schema structures (even when the same file format or syntax is used). For example, two datasets may be organized in different ways across two relational databases, i.e., the table and column structure may be very different although the content (at the conceptual level) of the databases may be very similar. Similarly, for XML databases, different DTDs or XML Schemas can be used to describe the same data. To overcome schema level heterogeneities, we can again apply two approaches, schema standardization or schema transformation. An example of the former is Z39.50, which provides a syntax and schema for querying digital library collections. The Open GIS Consortium (OGC, 2003) and the Federal Geographic Data Committee (FGDC, 2003) provide a number of standards that cover syntactic and schema aspects, as well as controlled vocabularies. The North American Data Model (NADM, 2003) includes a comprehensive data model and schema suited for geologic maps and related information. For the latter, i.e., schema transformation, database query languages in general and XQuery in particular provide powerful means to express complex queries and transformations. Thus (XML) query languages play an important role in database mediators (see below).

Finally, at the semantic level, we consider issues such as differences in terminology, different classification schemes (such as for rock types), and differences in the definition of and constraints for the various concepts that are relevant to the data sets being integrated. The main approach for reconciling semantic heterogeneities is the use of agreed-upon ontologies, which in their simplest form provide a controlled vocabulary with more or less formal descriptions of the pertinent concepts. In more sophisticated forms ontologies include formalizations (often through logic formulas) of properties of concepts and "inter-dependencies" of concepts. A prominent emerging standard for ontologies is the Ontology Web Language (OWL, 2003), which comes in three increasingly expressive variants: OWL Lite, OWL DL, and OWL Full. OWL is also an interesting example of how several interoperability levels and standards may be intertwined: for example, OWL DL builds upon the RDF model and syntax which in turn is usually denoted in XML syntax.

Mediator Architecture

Database mediator systems can be used to provide uniform access to distributed heterogeneous data sets, and thereby overcome a number of the interoperability challenges mentioned above. Figure 2 depicts a typical mediator architecture, in which a number of local data sources are "wrapped" as XML sources and subsequently combined into an integrated global view G. Thus a client application or end user is provided with the illusion of

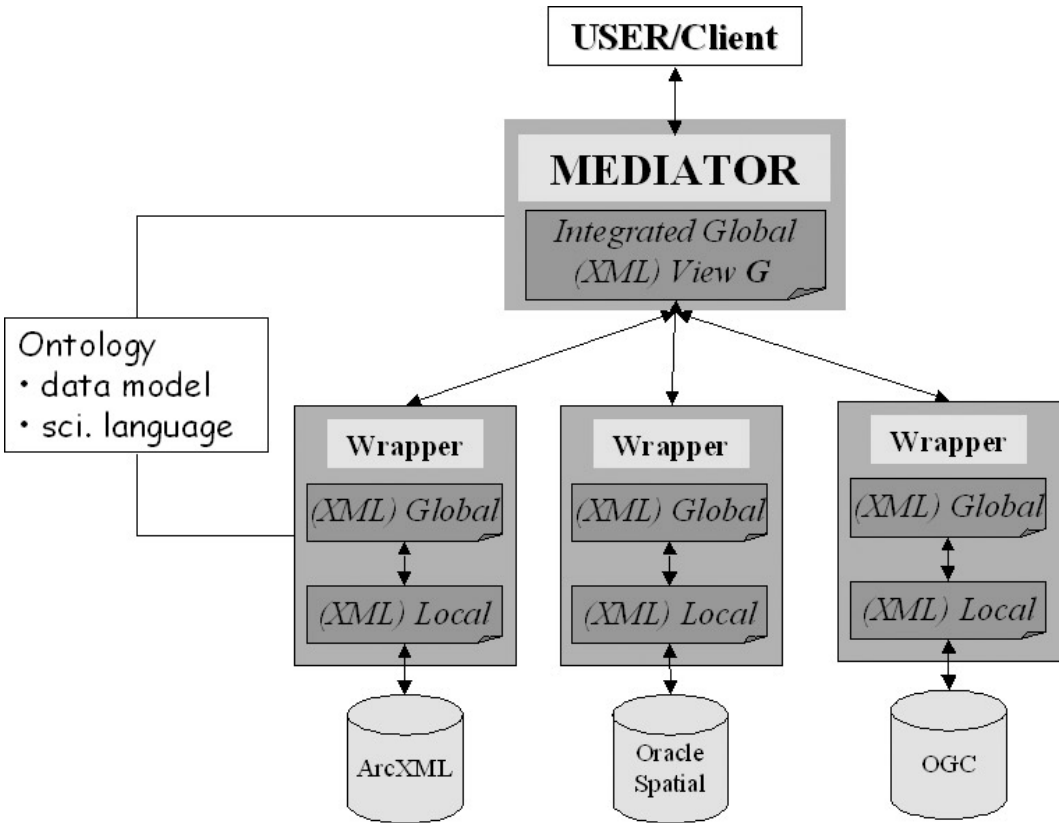

Figure 2. Extended mediator architecture.

querying a single, integrated (or global) database G.

The source wrappers not only provide a uniform syntax, but also reconcile system aspects, e.g., by means of a unified data access and query protocol. In a conventional relational or XML-based mediator system, interoperability is facilitated at the structural level. That is, differences in schema can be reconciled by corresponding schema transformation as part of the view definitions for the global view G. However, terminological differences or other semantic differences are not adequately handled at the purely structural/XML level. To this end, source schema and contents can be registered to an ontology, which encodes additional "knowledge" about the registered concepts. As we shall see in the next section, by "ontology-enabling" the system in this way, one can evaluate high-level queries over concepts that are not directly in the source databases, yet indirectly linked via the ontology.

THE ONTOLOGY-ENABLED MAP INTEGRATION SYSTEM

Figure 3 shows an end-user's view of the current ontology-enabled map integration prototype (OMI, 2003). Nine different spatial data sources, i.e., geological maps from various state geological surveys, are made interoperable in such a way that the user can seamlessly query across the different data sets, and even view the data through "conceptual-level glasses". By this we mean that once a source has been "semantically registered" relative to an existing ontology (such as the rock classification system of Struik et al., 2002), we can use that ontology's conceptual entities and relations to query the data set.

On the left in Figure 3 we see a collection of geologic maps from a number of state geological surveys. The query forms on the right of the depicted windows allow the user to query for regions having a specified geologic age and/or rock type. On the upper right, the results of a conventional query with GEOLOGIC-AGE='Paleozoic' are shown. Note that the system finds only very few regions, since the information that the Paleozoic Era contains periods such as Permian and Carboniferous, etc., is not "known" to the system. In contrast, on the lower right, we have "turned on" a geologic age ontology (Poling, 1997), and a much larger set of data is now found. Here, we have used a technique called "concept expansion" that replaces a query term such as 'Paleozoic' by all suitable "sub-concepts" (here the Periods, etc., belonging to 'Paleozoic') in order to retrieve all relevant data.

The Use and Role of Ontologies

In information integration systems (Figure 2), ontologies can be used to provide information at the level of conceptual models and terminologies, thereby facilitating conceptual-level queries against sources, and resolving some of the semantic-level heterogeneities between them. In our system (Figure 3), the rock classification ontology and geologic age ontology are used as a global view for registering data sets and processing queries. When a data set is registered to an ontology, a mapping from the data set to the selected ontology is generated. The registration process consists of the following three steps:

1. Select classes in the system ontology repository to register this data set: e.g., select the time scale and/or rock classification systems to be used as the global ontologies into which the data structure and contents are to be mapped;
2. Select columns for each selected class for populating virtual objects in these classes: e.g., map the columns in the data set to specific ontologies, thus indicating that a column's contents (their range) are mapped to an ontology, such as mapping a lithology column onto a rock classification ontology;
3. Select the populating methods or populate manually: map the column contents to classes in the ontology manually or semi-automaitcally using word-matching or other provided techniques, e.g., map "granite" from a lithology column to "Granite" in the ontology.

However, before such mapping can occur the sources' local data schemas must first be registered. For example, in our implementation we used the following two schemas for the Arizona and Idaho data sets:

Arizona—(AREA, PERIMETER, AZ_1000_, AZ_1000_ID, GEO, PERIOD, ABBREV, DESCR, D_SYMBOL, P_SYMBOL)

Idaho—(AREA, PERIMETER, ID_500_, ID_500_ID, FORMATION, UNIT_NAME, ROCK_TYPE, ERA, SYSTEM, SERIES, LITH1, LITH2, LITH3, LITH4, LITH5, LITH6, LITH7, LITH8, LOCATION1, LOCATION2, COMMENTS, IDCARB, IDK, IDBASE, IDFAM, IDPHOS, IDSG, IDBATHAB, LITHA, LITH_FORM, PERIOD, D_SYMBOL, P_SYMBOL, LITH_MAJOR, LITH_MINOR, LITHOLOGY, AGE, IDLITH)

After these steps, wrappers are created for the registered data sets. Each wrapper uses the mappings between the data source and ontology to translate queries from the global ontology to the local schema, and also to translate content from the local schema to the global ontology. As explained above, the system can automatically use the subclass relation to expand concept queries when required.

Note that although all system-registered ontologies can be considered as conceptual-level query mechanisms, the system can suggest suitable ontologies based on, first, the user's choice of data sets and, second, the sources' schema information.

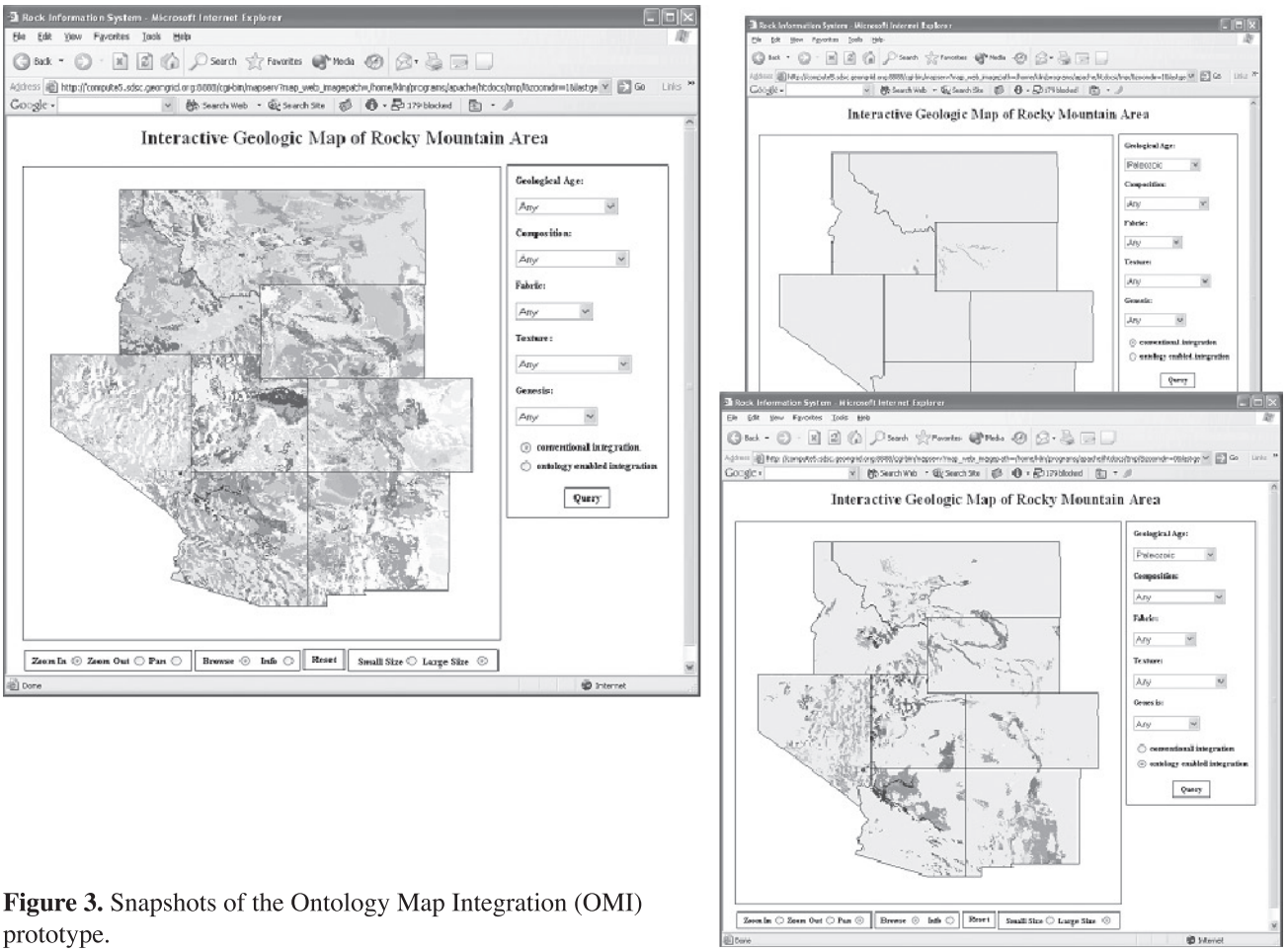

Figure 3. Snapshots of the Ontology Map Integration (OMI) prototype.

The Architecture of the Prototype

Figure 4 shows the system architecture: the system consists of a http server, a query mediator, and a map server (MapServer, 2003). When a request is received by the web server, the system generates a query against the global ontology, and then sends the query to the mediator; the mediator decomposes the query into several subqueries, and sends each of them to its target database. Then a mapfile is created based on these query results, and finally the map server renders a map according to the mapfile, and sends the map back to the user. Note that a map can contain several remote layers from other web map servers.

If a query, for example, asking to show rocks of Cenozoic age is received, the system takes the following steps to process the query:

1. Concept expansion: gets all the subclasses of the queried class (not shown in Figure 4).
2. Query rewriting: generates new queries to find formations against the two virtual tables by using the subclasses in the set found in step 1;
3. Map rendering: renders a map based on the query results of step 2 and predefined colors.

CONCLUSIONS AND FUTURE PLANS

We have described our current ontology-enabled prototype for integrating geologic maps from different sources. Syntactic and structural differences are overcome by traditional schema integration and database mediation techniques. In addition, "semantic mediation" and conceptual-level queries are supported by registering source data sets to domain ontologies such as for geologic age, rock type classification, etc.

On the systems side, we are adding commercial mapping technology (ESRI) in addition to the current open source technology (MapServer, 2003). Moreover, we will also "grid-enable" the application, i.e., use Grid standards for data access and querying. At the level of ontologies, we are working on a "3rd-party registration mechanism", that will allow the user to register a data set relative to one ontology (rock type ontology A), and then query the data set using another ontology (rock type ontology B); cf. (Lin and Ludäscher, 2003; Bowers and Ludäscher, 2003). This is only possible by having a "mediation engineer" devise a so-called articulation ontology that maps concepts between ontologies (such as A and B). We have already conducted preliminary studies in this

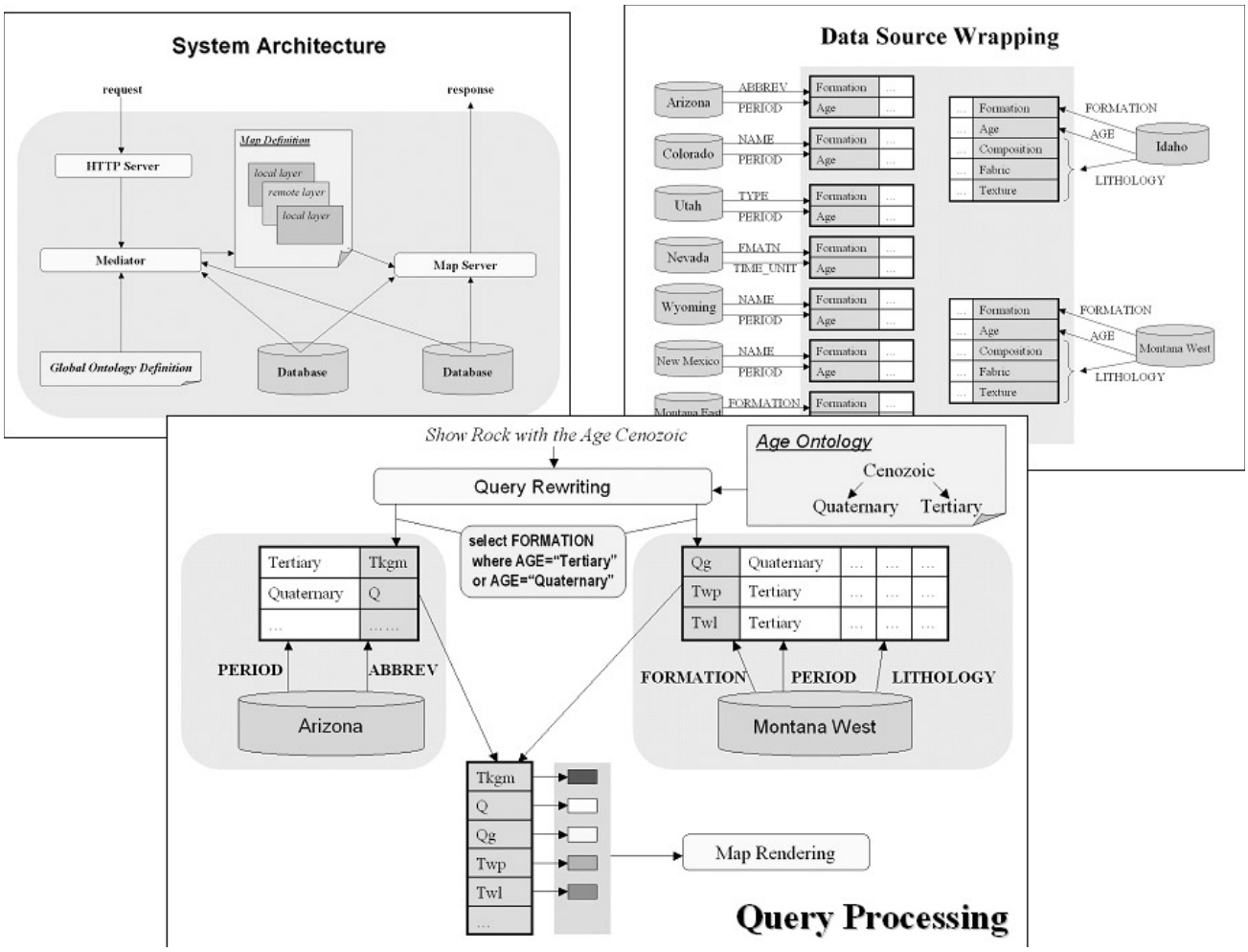

Figure 4. System architecture of the prototype.

direction, which employ OWL to map concepts between different ontologies (for example, the rock type classifications of the British Geological Survey and Struik and others, 2002). Ultimately we are interested in embedding applications such as the one described here into a GEON workflow environment that will allow the user to combine data integration steps (e.g., geologic map integration) with analytical steps (e.g., rock classification) to form a high-level "scientific workflow".

REFERENCES

Baader, F., Calvanese, D., McGuinness, D., Nardi, D., Patel-Schneider, P., 2003, The description logic handbook—Theory, implementation and applications, Cambridge University Press.

Bowers, S. and Ludäscher, B., 2003. Towards a generic framework for semantic registration of scientific data, in Semantic web technologies for searching and retrieving scientific data (SCISW), <http://www.sdsc.edu/~ludaesch/Paper/scisw03-seek.pdf>.

FGDC, 2003, Federal Geographic Data Committee, <http://www.fgdc.gov>.

GEON, 2003, The Geosciences Network, <http://www.geongrid.org>.

Lin, K., and Ludäscher, B., 2003, A system for semantic integration of geologic maps via ontologies, in Semantic web technologies for searching and retrieving scientific data (SCISW). <http://www.sdsc.edu/~ludaesch/Paper/scisw03-geon.pdf>.

MapServer, 2003, University of Minnesota, <http://mapserver.gis.umn.edu/>.

NADM, 2003, North American digital geologic map data model, <http://geology.usgs.gov/dm/steering/teams/design/>.

OGC, 2003, Open GIS Consortium, <http://www.opengis.org>.

OMI, 2003, Ontology-enabled map integration prototype, <http://kbis.sdsc.edu/GEON/map-integration.html>.

OWL, 2003, OWL Web ontology language reference, W3C candidate recommendation 18 August 2003, <http://www.w3.org/TR/owl-ref/>.

Poling, J., 1997, Geologic ages of earth history, <http://www.dinosauria.com/dml/history.htm>.

RDF, 2003, Resource Description Framework (RDF), <http://www.w3.org/RDF/>.

Sheth, A., 1998, Changing focus on interoperability in information systems: From system, syntax, structure to semantics, interoperating Geographic Information Systems, pp. 5–30, Kluwer.

Struik, L., Quat, M., Davenport, P., and Qkulitch, A., 2002, A preliminary scheme for multihierarchical rock classification for use with thematic computer-based query systems: Geological Survey of Canada, <http://www.nrcan.gc.ca/gsc/bookstore/free/cr_2002/D10.pdf>.

WSDL, 2001, Web Services Description Language (WSDL) 1.1, W3C Note 15 March 2001, <http://www.w3.org/TR/wsdl>.

XML, 2000, Extensible Markup Language (XML) 1.0 (Second Edition), W3C Recommendation 6 October 2000, < http://www.w3.org/TR/REC-xml>.

XML Schema, 2001, XML Schema Part 0: Primer, Part 1: Structures, Part 2: Datatypes, W3C Recommendation 2 May 2001, <http://www.w3.org/TR/xmlschema-0/>.