A System for Semantic Integration of Geologic Maps via Ontologies*

Kai Lin and **Bertram Ludäscher**
San Diego Supercomputer Center
University of California at San Diego
9500 Gilman Drive, La Jolla, CA 92093-0505
{klin, ludaesch}@sdsc.edu

Abstract

This paper describes a prototype system for registering geologic data sets through ontologies to assist in integrating and querying heterogeneous geologic data sets. The system consists of three components: an *ontology repository*, the *data set registration*, and *ontology-aware applications*. User-defined ontologies in OWL are saved and used by the system. Each data set must be registered before it becomes available, and the registration semi-automatically generates a mapping from data sets to ontologies. The mapping between data sets and ontologies are used by applications to explore and extract information from the data set.

Introduction

It is widely recognized that ontologies (Guarino 1998) play a central role in modern GIS systems (Fonseca & Egenhofer 1999) and other scientific data integration systems (Ludäscher, Gupta, & Martone 2003). A GIS system often needs to integrate heterogeneous data sets with a unified logical view. However, integrating and using these data sets can be very difficult. This is primarily due to the fact that each data set uses different schema and semantics. In general, data heterogeneity can be divided into three categories (Sheth 1998; Schuster 2000): syntactic heterogeneity, structural heterogeneity and semantic heterogeneity. Syntactic and structural transformation approaches (e.g., database mediation) can be used to adequately handle the first two kinds of heterogeneities, however, they are not adequate for resolving semantic differences. The use of ontologies is considered a possible solution of the semantic heterogeneity problem (Wache *et al.* 2001).

In the GEON project (GEON 2003) we are developing an interoperability framework and system that allows a data provider to register a data set with one or more “mediation ontologies” and subsequently query the different data sets in a uniform fashion. The capability of querying the mediated data sets is significantly improved when all available data sets are registered in this way: Heterogeneous source vocabularies are made compatible via the ontologies, and multiple conceptual dimensions become queryable simultaneously. Our test cases include integration of different geologic maps from several state geological surveys. Within the system a user registers a geologic map using interactive tools against previously defined ontologies for *geologic age* (Poling 1997) and *rock types* (Gillespie *et al.* 1999; Struik *et al.* 2002). Some details of our system are described in the next section. We also report some initial experiences of using the system to build an ontology-enabled map integrator (OMI).

Design

In this section, we discuss the prototype system developed in the GEON project for exploring and integrating geologic data sets. The system consists of three components: the *ontology repository*, the *data set registration*, and *ontology-aware applications*.

Ontology Repository

The system accepts and saves user-defined ontologies in OWL DL, the description logic variant of the Web Ontology Language OWL (W3C-Consortium 2003). The following is a fragment of the rock *genesis* classification ontology in the ontology repository for OMI, in which a class `Metamorphic` is defined:

```
......
<owl:Class rdf:ID="Metamorphic" />
......
```

The (chemical) *composition* classification ontology is defined based on the genesis ontology. The following shows a fragment of the composition ontology:

```
......
<owl:Ontology>
<owl:imports rdf:resource=
    "http://www.geongrid.org/genesis" />
</owl:Ontology>

<owl:Class rdf:ID="Marble">
<rdfs:subClassOf rdf:resource="#Calcium"/>
<rdfs:subClassOf rdf:resource=
    "http://www.geongrid.org/genesis#Metamorphic"/>
</owl:Class>
......
```

A new class `Marble` is declared to be a subclass of the composition class `Calcium` and the genesis class

*This work supported in part by the National Science Foundation (NSF) under grants No. ITR 0225673 (GEON) and No. ITR 0225676 (SEEK).

`Metamorphic`. Additional ontologies for rock *fabric* and *texture* have been derived from a multi-hierarchical classification scheme (Struik *et al.* 2002).

In general, any web-accessible ontology having a URI can be imported and used by the system. If a user-defined ontology imports another ontology which is not in the system repository, the system will download the imported ontology and use the local copy whenever the remote ontology is not available. The system also provides a basic navigation tool to browse the ontologies in the system repository.

Ontology Mappings

Frequently a world can be modelled in several different ways. For instance, we have initially employed two different rock classifications, defined separately by the British Geological Survey (Gillespie *et al.* 1999) and a Canadian working group (Struik *et al.* 2002). Although these rock classifications have totally different class hierarchies, it is still possible to define mappings which translate between classes and properties in one classification and corresponding classes and properties in a second classification. The mappings between ontologies provide the possibility of combining and parameterizing and switching ontologies, and they are very useful in practice. More research on ontology mapping can be found in (Kalfoglou & Schorlemmer 2002; Bench-Capon & Malcolm 1999; Sleeman *et al.* 2002).

In the following, 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 O_A to O_B consists of a class mapping f and a property mapping g. The *class mapping* f is a partial function from the class set of O_A to the set of all derived classes in O_B, where a derived class is a class defined from other classes, for example, an intersection of two other classes. The class mapping should preserve the subclass (*isa*) relation, i.e., if A_1 and A_2 are classes in O_A and A_1 is a subclass of A_2, then $f(A_1)$ must be a subclass of $f(A_2)$ in O_B:

$$\begin{tikzcd}
\mathtt{A_1} \arrow[r, "isa"] \arrow[d, Rightarrow] & \mathtt{A_2} \arrow[d, Rightarrow] \\
\mathtt{f(A_1)} \arrow[r, "isa"'] & \mathtt{f(A_2)}
\end{tikzcd}$$

If f does not preserve the subclasses relation, then a query about individuals in $f(A_1)$ may incorrectly return some individuals in $f(A_2)$.

The *property mapping* g is a partial mapping from the property set of O_A to the set of all derived properties in O_B, and should satisfy the following condition: If p is a property between the classes A_1 and A_2 in O_A, then $g(p)$ is a property between the classes $f(A_1)$ and $f(A_2)$:

$$\begin{tikzcd}
\mathtt{A_1} \arrow[r, dashed, "p"] \arrow[d, Rightarrow] & \mathtt{A_2} \arrow[d, Rightarrow] \\
\mathtt{f(A_1)} \arrow[r, dashed, "g(p)"'] & \mathtt{f(A_2)}
\end{tikzcd}$$

Note that a class mapping and a property mapping induce a natural translation from the constraints of O_A to the constraints of O_B. For instance, if O_A has a constraint $|A| < 30$, i.e., the cardinality of A is less than 30, then this constraint can be translated into an O_B constraint $|f(A)| < 30$. An ontology mapping should satisfy the following condition: if a constraint c_a in O_A can be naturally translated into a constraint c_b in O_B by its class mapping and property mapping, then c_b is implied by O_B. In other words, this condition says that an ontology mapping should not introduce any new constraints into the target ontology. In fact, the requirement that a class mapping should preserve the subclass relation is a special case of this condition. The validation of an ontology mapping can be done when its underlying logic is decidable, as is the case, for instance, for description logics (Baader *et al.* 2003).

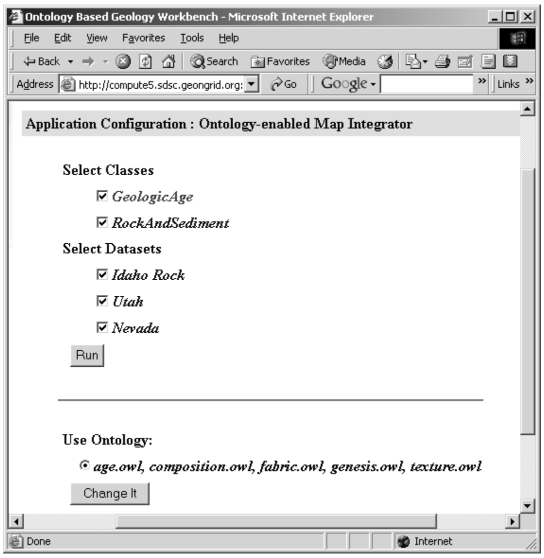
If there is an ontology mapping from O_A to O_B, then from any model M of O_B, a submodel can be extracted and naturally transformed into a submodel of O_A based on this ontology mapping. An ontology mapping is more general than a simple equivalence relation on classes and properties. In fact, it defines a structural translation from one ontology to another ontology. For example, let O_A contain a class `Person` and a data property `hasName` and O_B contain a class `Employee` and two data properties `name` and `id`, then a mapping sending `Person` to `Employee` and `hasName` to `name` is a valid general ontology mapping from O_A to O_B, so any model of O_B can be naturally transformed into a model of O_A via this mapping, that is, an employee is a person, but a person may be not 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_A has a property p defined between the classes A_1 and A_2 (i.e., $A_1 \xrightarrow{p} A_2$) and O_B has a property q between B_1 and B_2, and assume that A_1 and B_1 are conceptually the same. Combining O_A and O_B should give an ontology O containing a property p between a new concept A (of O, representing A_1 and B_1) and A_2, and a property q between A and B_2. In OWL, this can be done by using `equivalentClass` or `sameAs` tags. Theoretically this result is equivalent to the pushout (Barr & Wells 1990) of the following diagram:

$$\begin{tikzcd}
\mathtt{O'} \arrow[r, "\psi_1"] \arrow[d, "\psi_2"'] & \mathtt{O_A} \arrow[d, dashed] \\
\mathtt{O_B} \arrow[r, dashed] & \mathtt{O}
\end{tikzcd}$$

where O' is an ontology containing only one class A. The ontology mapping ψ_1 sends A to A_1, and ψ_2 sends A to B_1.

Note that parameterized ontologies and their instantiation can be implemented based on the same principle. A *parameterized ontology* is a pair (O', O_A) where the parameter ontology O' is included in the body ontology O_A, which means that there is an inclusion ontology mapping from O' to O_A. To instantiate this parameterized ontology with an actual ontology O_B, an ontology mapping from the formal ontology O' to O_B must be provided. The instantiation result is the


Ontology Based Geology Workbench - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://compute5.sdsc.geongrid.org:
Go
Google
Links
Done
Internet

Application Configuration : Ontology-enabled Map Integrator

Select Classes

☑ *GeologicAge*

☑ *RockAndSediment*

Select Datasets

☑ *Idaho Rock*

☑ *Utah*

☑ *Nevada*

Run

Use Ontology:

● *age.owl, composition.owl, fabric.owl, genesis.owl, texture.owl*

Change It

Figure 1: OMI interface showing the integrated ontology and a web form allowing users to change ontology.

pushout of the diagram above, i.e., the ontology O in the diagram. It is unique up to ontological isomorphisms.

Our system accepts ontology mappings in the form of OWL files. Currently, most axioms relating concepts between different ontologies simply use the `equivalentClass` tag. The following is a small fragment of an ad-hoc ontology mapping from the BGS rock classification (Gillespie *et al.* 1999) to the Canadian rock classification (Struik *et al.* 2002):[1]

```
 ......
<owl:Class rdf:about=
   "http://www.geongrid.org/br#FoidBearingMonzonite">
 <owl:equivalentClass rdf:resource=
   "http://www.geongrid.org/composition#FoidMonzonite"/>
</owl:Class>
 ......
```

Ontology mappings are used in navigation and query processing. If a data set is registered to an ontology O_A, and there is an ontology mapping from O_B to O_A, then users can choose both, ontology O_A and ontology O_B to query the data set. In the example above, a user can use BGS rock classification and/or the Canadian rock classification to send their queries to the system. If an another data set is registered to the ontology O_B, then the system that integrates two data sets still can use either ontology for accepting queries.

Figure 1 shows how different ontologies can be selected using the simple forms-based interface of our current prototype.

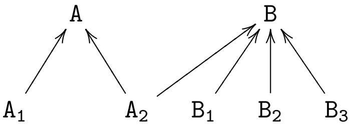
Data Set Registration

Each data set must be registered to an ontology before it is accessible from the system. In our current prototype only shapefiles (ESRI 1998) are acceptable for registering. Each shapefile consists of at least 3 files: `shapefile.shp`, `shapefile.shx`, and `shapefile.dbf`. The latter is the shapefile's feature attribute table stored in dBASE format. A shapefile can contain only one table. The other two files (`shapefile.shp` and `shapefile.shx`) contain information about feature geometry.

Our registration procedure takes the following 3 interactive steps to create the integration mapping from a shapefile to some ontologies:

1. **Ontology Selection:** The user is asked to select some class names from the OWL ontologies. Choosing a class name indicates that it might be able to (virtually) populate a corresponding class with instance object for each row based on some table columns. Such instance object can be considered as a value of a property of the polygon for this row. If a class is selected, its subclass are calculated and become automatically selected (see example below). So, to classify the objects generated for all rows, it is sufficient to choose one or several top level classes in the class hierarchy.

   The system can infer implied facts in the selected ontologies. As a consequence, users can provide minimum information to map their data into the ontologies. For example, suppose the classes `A` and `B` are two top level classes of two classifications:

   where an arrow represents an *isa* relation, i.e. a subclass relation. Consider a data provider who has data about objects that belong to the class `A`, she will register her data relative to `A` or its subclasses $\texttt{A}_1$ and $\texttt{A}_2$. Although $\texttt{A}_2$ is a subclass of `B`, it is sufficient to choose `A` as the target class to map data, ignoring `B`. The system can automatically use the fact that $\texttt{A}_2$ is a subclass of `B` to answer queries like "*find all instances of* `B`".

   The of ontology constraints are important for the data registration. For example, if the following constraint is declared:

   $$A_1 \cap A_2 = \emptyset \qquad A = A_1 \cup A_2$$

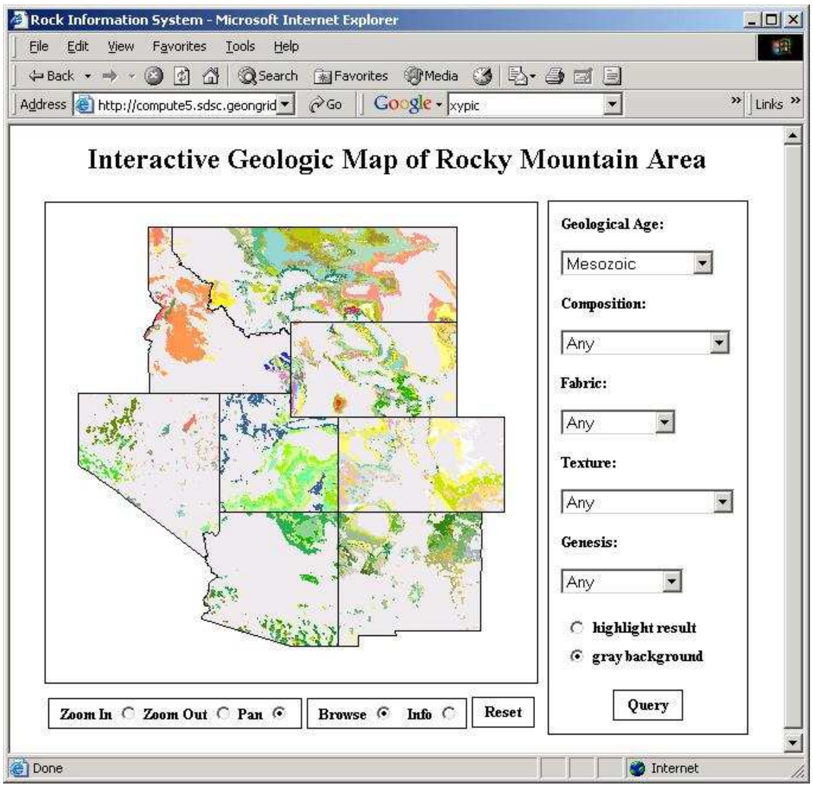
   which means $\texttt{A}_1$ and $\texttt{A}_2$ are disjoint subclasses of `A`, then it is illegal to populate $\texttt{A}_1$ and $\texttt{A}_2$ with instance objects from the same row. If we only have the constraint below:

   $$A = A_1 \cup A_2$$

   then each row of a table can generate instance objects of the class $\texttt{A}_1$ and $\texttt{A}_2$ at the same time.

2. **Mapping Data to Ontologies:** For each selected class, the user is asked to choose one or several columns to (virtually) populate the class. If a single column is chosen and

[1] The current mapping is ad-hoc, since it has been based mainly on syntactic matches – in particular, our immediate purpose was not to achieve an actual domain-specific semantic reconcilation, but rather to illustrate the capabilities of OWL as a concept definition language. For a more rigorous mapping, domain scientists need to be consulted.


Rock Information System - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://compute5.sdsc.geongrid
Go
Google xypic
Links
Zoom In ○ Zoom Out ○ Pan ●
Browse ● Info ○
Reset
Query
Done
Internet

Interactive Geologic Map of Rocky Mountain Area

Geological Age: Mesozoic

Composition: Any

Fabric: Any

Texture: Any

Genesis: Any

○ highlight result

● gray background

Figure 2: OMI Interface showing the integrated geologic maps (left), and ontology-aware query forms (right).

the data set contains classification information, we found the following functions to be very useful in practice (here, we chose general examples instead of geological ones, for clarification):

(a) *All Matches:* If the name of a subclass of the target class is matched the value of the selected column, then assume an instance object of this subclass can be generated. This method can be used for a column containing multiple valid options, for example, “`student / housekeeper`”.

(b) *First Match:* If there are multiple matches by using the *All Matches* method, then take the first match to (virtually) create objects. This method can be used to select the begin point of an interval, for example, “`1960-1970`”.

(c) *Last Match:* Similar to the *First Match*, this method takes the last match. This method fit best if the column contains a sequence of narrowing down descriptions, for instance, “`java.util.Hashtable`”.

(d) *Manual Setting:* The user decides the classes to create instance objects.

The user can select one of these methods to populate the class. If multiple columns are selected, then the *Manual Match* method will be used to create (virtual) instances.

If the selected class has some properties defined on it, then for each property, the user is asked to select columns for populating it.

3. **Mismatch Resolution and Manual Setting:** The system prompts this step when one of following two cases occurs: 1) The manual setting is selected in the second step; 2) No match was found for some rows by the selected method other than the manual setting method in the second step. The user is then asked to select classes for these rows.

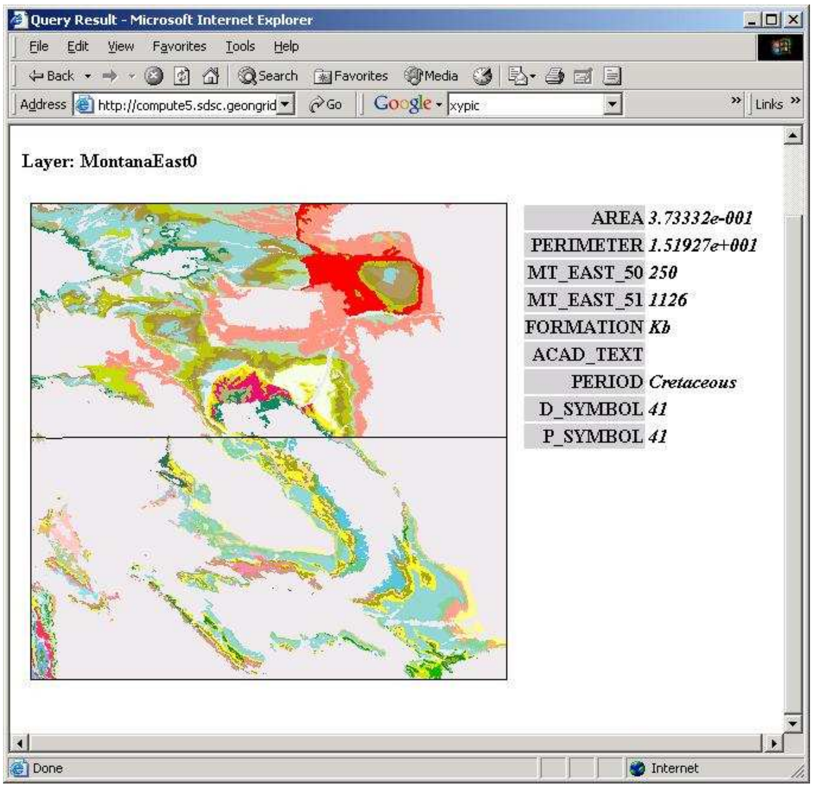
Ontology-Aware Applications

Any application that understands some ontologies in the system can be plugged into the system. There are two configurable options to launch the application: entities in the ontologies and data sets. Users can choose interesting concepts and data sets to start an application.

The Ontology-enabled Map Integrator (OMI)

As a part of the GEON, we try to build a system for integrating geologic maps from different geologic surveys. The objective is to integrate available geologic data sets to provide a web-based interactive geologic map for finding the location where rock has a specified geologic age or composition or fabric or texture or genesis property or any combination of these properties. The current prototype uses the open source MapServer (University of Minnesota 2003) to implement the required standard GIS functions; an integration with ESRI’s commercial technology is planned as well.

Five ontologies are submitted to the system: `GeologicAge`, `Genesis`, `Texture`, `Fabric` and `Composition` (Struik *et al.* 2002) (Harland *et al.*

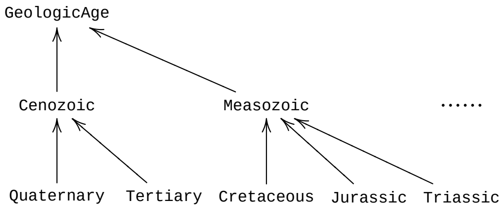

Query Result - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://compute5.sdsc.geongrid Go
Google xypic
Links
Done
Internet

Layer: MontanaEast0

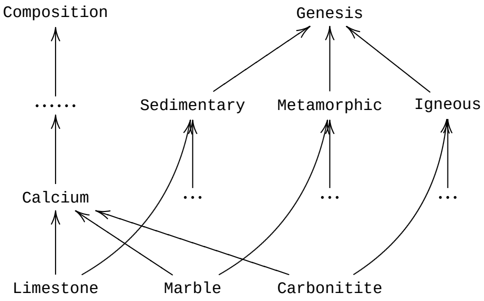
AREA	*3.73332e-001*
PERIMETER	*1.51927e+001*
MT_EAST_50	*250*
MT_EAST_51	*1126*
FORMATION	*Kb*
ACAD_TEXT	
PERIOD	*Cretaceous*
D_SYMBOL	*41*
P_SYMBOL	*41*

Figure 3: Snapshot of two adjunct geologic map (left), and some raw data (right), obtained by selecting a point of interest.

1989). All these ontologies represent the hierarchies of some classifications. The geologic age ontology can be simply described as the tree below:



The other ontologies demonstrate the similar tree structures. Furthermore, each subclass of composition or texture or fabric class may be a subclass of three genesis classes `Igneous`, `Sedimentary` and `Metamorphic`. The diagram below shows a part of composition classification, which says that calcium and limestone rock is also sedimentary rock, whereas marble rock is also metamorphic rock.



In one of our test case, nine data sets are attached to the ontologies above; they contain the rock age information in the states Arizona, Colorado, Idaho, Montana East, Montana West, Nevada, New Mexico, Utah and Wyoming. Additionally Idaho and Montana West data sets also provide rock type information. The following is the table schema of Arizona data set:

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

where the column `PERIOD` gives geologic age information. No rock classification information is provided. To register Arizona data set, we select `GeologicAge` class at the first step, then select the column `PERIOD` and the *All Matches* method at the second step. The system scans the data in the column `PERIOD` and found two unmatched terms: `Water` and `Algonkian`, we choose `Ignore` to omit these rows.

Other data sets have different schemas. For example, the Idaho data set has the following schema:

```
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)
```

Geologic ages can be found in the column AGE, whereas LITHOLOGY contains information about rock composition, texture, fabric and genesis. All these data sets are registered with the similar procedure as above.

Figure 2 shows the interface of the application for automatic map integration, where the map is the result of querying rock with the age Mesozoic. If Info is chosen, clicking on the map will return the raw data associated with the clicked polygon. Figure 3 shows the result after zooming in and selecting Info and clicking on a polygon. It took several days to integrate these data sets by hands before we built the system. With this data registration tool, a similar system for mid-atlantic area has been done in less than half a hour.

The ontology RockAndSediment derived from BGS rock classification and an ontology mapping from the RockAndSediment to the union of Genesis, Texture, Fabric and Composition are also submitted to the system. Then users are able to choose ontologies to send their queries.

Conclusions

An ontology based prototype for geologic map integration and its data set registration procedure have been discussed in this paper. The experiments we have done show that ontology based approaches are promising for scientific data integration and navigation. The novel design of the system makes it extremely easy to dynamically plug-in new data sets. Many problems are open for future research, for instance, what is the best way to organize the ontology and ontology mapping repositories, and what kind of reasoning services over those ontologies and mappings are needed. Also we plan to extend our system to be able to register database tables as well as XML documents in the future, and make it part of a generic framework for semantic registration of scientific data (Bowers & Ludäscher 2003).

References

[Baader *et al.* 2003] Baader, F.; Calvanese, D.; McGuinness, D.; Nardi, D.; and Patel-Schneider, P. F., eds. 2003. *The Description Logic Handbook: Theory, Implementation, and Applications*. Cambridge University Press.

[Barr & Wells 1990] Barr, M., and Wells, C. 1990. *Category Theory for Computing Science*. Prentice-Hall.

[Bench-Capon & Malcolm 1999] Bench-Capon, T. J. M., and Malcolm, G. 1999. Formalising ontologies and their relations. In *Database and Expert Systems Applications*, 250–259.

[Bowers & Ludäscher 2003] 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)*.

[ESRI 1998] ESRI. 1998. Esri shapefile technical description. www.esri.com/library/whitepapers/pdfs/shapefile.pdf.

[Fonseca & Egenhofer 1999] Fonseca, F. T., and Egenhofer, M. J. 1999. Ontology-driven geographic information systems. In *ACM-GIS*, 14–19.

[GEON 2003] GEON. 2003. Geoscience network (geon) project.

[Gillespie *et al.* 1999] Gillespie, M. R.; Styles, M.; Robertson, S.; Hallsworth, C. R.; Knox, R. W. O.; Evans, C. D. R.; Irving, A. A. M.; Merritt, J. W.; Morigi, A. N.; and Northmore, K. J. 1999. Bgs rock classification scheme. British Geological Survey Research Reports. http://www.bgs.ac.uk/bgsrcs/home.html.

[Guarino 1998] Guarino, N. 1998. Formal ontology and information systems. In *Proceedings of the 1st International Conference on Formal Ontologies in Information Systems*. IOS Press. Trento, Italy.

[Harland *et al.* 1989] Harland, W. B.; Armstrong, R.; Cox, A.; Lorraine, C.; Smith, A.; and Smith, D. 1989. *A Geologic Time Scale 1989*. Cambridge University Press.

[Kalfoglou & Schorlemmer 2002] Kalfoglou, Y., and Schorlemmer, M. 2002. Information flow based ontology mapping. In *Proceedings of 1st International Conference on Ontologies, Databases and Applications of Semantics*. Springer. Irvine, CA, USA.

[Ludäscher, Gupta, & Martone 2003] Ludäscher, B.; Gupta, A.; and Martone, M. E. 2003. A model-based mediator system for scientific data management. In Critchlow, T., and Lacroix, Z., eds., *Bioinformatics: Managing Scientific Data*. Morgan Kaufmann.

[University of Minnesota 2003] University of Minnesota. 2003. Mapserver homepage. http://mapserver.gis.umn.edu/.

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

[Schuster 2000] Schuster, V. S. 2000. Ontologies for geographic information processing.

[Sheth 1998] Sheth, A. 1998. Changing focus on interoperability in information systems: From system, syntax, structure to semantics. In Goodchild, M.; Egenhofer, M.; Fegeas, R.; and Kottman, C., eds., *Interoperating Geographic Information Systems*. Kluwer. 5–30. http://lsdis.cs.uga.edu/lib/download/S98-changing.pdf.

[Sleeman *et al.* 2002] Sleeman, D.; Robertson, D.; Potter, S.; and Schorlemmer, M. 2002. Enabling services for distributed environments: Ontology extraction and knowledge-base characterisation. In *ECAI 2002 Workshop on Knowledge Transformation for the Semantic Web*.

[Struik *et al.* 2002] 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. http://www.nrcan.gc.ca/gsc/bookstore/free/cr_2002/D10.pdf.

[W3C-Consortium 2003] W3C-Consortium. 2003. Owl web ontology language reference. www.w3.org/TR/owl-ref/.

[Wache *et al.* 2001] Wache, H.; Vogele, T.; Visser, U.; Stuckenschmidt, U.; Schuster, H.; Neumann, G.; and Hubner, S. 2001. Ontology-based integration of information - a survey of existing approaches.