Proc. 3rd International Workshop on Semantic Web and Databases (SWDB'06), Atlanta, GA, April, 2006

A day in the life of a metamorphic petrologist

S. Adalı♯, B. Bouqata♯, A. Marcus♯, F. Spear♭, and B. Szymanski♯
♯Department of Computer Science & ♭Department of Earth and Enviromental Sciences
Rensselaer Polytechnic Institute
Email: {adalis, bouqab, marcua, spearf, szymansk}@rpi.edu

***Abstract*— In this paper, we describe the functionality of a toolkit for sharing and long-term use of different types of geological data sets across disciplines. Our tools allow users to describe the meaning of their data by attaching semantic information to it. The toolkit also makes use of the users' access patterns to learn how the data are used to further enhance the utility of data centric methods. These learned patterns are used in conjunction with the semantic data to help other users find common ways to navigate heterogenous collections and highlight interesting information. Our current prototype is being developed in close collaboration with the Metamorphic Petrology working group formed to facilitate sharing of data within this subdiscipline of geosciences as well as with other systems for sharing of geological data.**

I. Introduction

In this paper, we describe a set of tools that are being developed at Rensselaer Polytechnic Institute to facilitate the sharing and long-term use of different types of geological data sets across disciplines. These data sets include published data, geologic and topographic maps, satellite imagery, structural, seismic, geophysical, petrologic, geochemical and geochronologic data. The availability of a wide range of data sets about a particular geographic area will permit scientific questions to be addressed that are currently not feasible because of the difficulty in collating the necessary information. The tools will permit data sets collected at different laboratories to be readily accessible to colleagues at remote locations, and will facilitate and expedite multi-collaborator research efforts. The toolkit will provide global access to current data sets that can be utilized by researchers and students. Finally, the database will provide ways to organize and summarize data sets so that their usefulness will endure.

Making data sets widely accessible to different research groups and scientific fields is clearly a necessary first step in asking deeper and interesting questions in many scientific fields. Unfortunately, in many scientific fields where the data centric approach is relatively new, there is very little infrastructure for shared vocabularies and curated data. A major obstacle to sharing of data between different research groups and disciplines is the enormous amount of effort required to build a useful database from the ground up. For example, the PetDB database [5] initiated at Lamont has as its goal the cataloging of chemical analyses of rocks dredged from the ocean ridges. As of January, 2004, PetDB contained 785,000 chemical values for 8300 sample stations and 33,000 rock samples. Yet this database represents only a small fraction of the geochemical data available on rocks and serves a limited audience because it includes only chemical analyses keyed to specific samples.

Much geological data is highly visual, so a useful database must rely very heavily on images. Describing the content of images is a highly challenging task in itself. Developing a unified data model for this purpose is even harder as it is fairly common to find many exceptions to any model since each individual scientist might have a different naming and mapping scheme for his data sets. The continuous maintenance of the data remains to be a costly enterprise. A possible solution to the quandary of creating the database is to develop a system of tools that would (a) enable the researcher to use his/her own data in a way that adds value to the experience and (b) in the process of using the data, catalog the data in a way that would be useful to other researchers.

A key aspect of all geological data is that it has spatial significance: data sets refer to or were collected at a specific spot on or within the earth; data sets refer to a specific rock sample that were collected at a particular location; a chemical analysis of a mineral was collected on a specific location in a sample (e.g., the center or the rim of a grain); images were collected of particular parts of a sample and their spatial relation to other parts of the sample, other images collected on the sample, and chemical analyses collected on the sample. The spatial correspondances in the data of this form are critical to the interpretation of the data. Without this spatial relationship, most geological data sets have greatly reduced value.

It is a task of the geoscientist to assimilate these data into a coherent picture/model/image of the properties/appearance of the Earth, and to interpret these data sets in the context of how the planet evolved. Because of the breadth of data that relate to any single geological problem, the scope of this effort is typically beyond the capabilities of any single researcher. Even when a researcher has the background to make use of diverse geological information (geoscientists are generally broadly trained even if they specialize in a narrow field of research), the effort to develop a command of the available data in related fields is often prohibitive. Collaborative efforts can bring scientists with different expertise together to work on a focused problem, but even these efforts are limited by accessibility and evaluation of data. To further confound the problem, much data collected in the geosciences is never published, the volume of such data being far too great to permit publication in traditional (i.e., paper) formats. Access to the large quantities of unpublished but high quality data

would result in greatly enhanced efficiency of effort and cost effectiveness by eliminating the need to recollect unpublished data. To this end, we are collaborating with a working group in the Metamorphic Petrology subdiscipline of geosciences to serve as a testbed for the collection and sharing of data. Our collaborators naturally interact with scientists from various subdisciplines of geosciences and efforts to collect data in these disciplines [2], [5].

II. Organization of Data

The problem of effective organization of data is not a new one. Many new tools for effective organization of one's desktop, for example the Haystack system [4], aim to solve this problem by storing relationships among the data and allowing the users to specify semantics of their data with the help of a data model or ontology. While this is a step in the right direction, there is a significant problem of scale. It is simply too hard to create tags for all the data to make them easy to query and find. There is also the question of finding and using the correct ontology for a large group of users. While it might be possible to agree on a series of general concepts for a specific scientific discipline, more general concepts tend to be more controversial in terms of their full meaning. It is hard to reach an agreement on the description and pertinent properties of many concepts as they are the subject of new research and hence are still being defined by new findings.

The complementary problem of creating tools to automatically find semantically related items and attaching meaningful classes to these groups poses other interesting problems. First of all, the data that our toolkit manages are very heterogenous in nature and are distributed to hundreds of files. It ranges from concrete physical objects (i.e. a rock) to large images or to a single number in a spreadsheet. Hence, it is hard to imagine an automated content-based tool that is able to deal with such a heterogeneity and provide highly accurate results. A large quantity of the data is created by ad-hoc analysis programs that do not create the necessary semantic information. In addition, a common process in the scientific inquiry is to conduct many different tests, examine the results, pose hypotheses and then test these hypotheses with many additional tests. Hence, the data with which we deal also contain many interesting results mingled with many dead-ends, tests that were later found incorrect, etc. In such a setting, it is very hard to continuously tag the data to indicate its importance and correctness. Yet, methods that monitor the user interactions with the data to arrive at conclusions is a good step in extracting this type of information. Still, we are then faced with the problem that the interactions with the data may be very sparse and may not provide any interesting insights.

One of the common methods in such a case is to process the tags or annotations in the *flickr*[3] style to extract similar content. We consider this a very important step towards solving some of the problems that we encounter. But, given that the data creation rate is very high compared to the size of the community that processes it, it is reasonable to assume that tags will not exist for all the information that is needed to solve the data organization problem. Furthermore, there is a need to attach some specific types of semantic information that will provide semantic context to the data such as the latitude and longitude of location from which data were obtained, data's type and relationship to other data objects. Hence, truly free-form tags that do not contain this information will have very limited use.

To this end, we aim to combine all of the above mentioned methods to solve our problem, semantic and free-form tagging are combined with automated tools that extract information from user interactions. We discuss this in detail in the next section.

III. The toolkit

Our toolkit is an interface that allows a single user to graphically annotate semantic links between files or file segments (such as a row from a spreadsheet) on his/her computer. The toolkit allows the user to organize her own data, share it with a collaborator or public, query existing data and upload the results for further processing. The current prototype of the toolkit is written in Java, Swing and uses XML/PDF for persistent storage of the semantic annotations. As explained in the following sections, the toolkit stores both implicit and explicit semantics as well as mined and fuzzy information. Currently, we are trying to assess how much of this semantic model can be exposed in RDF or OWL. This is a topic of ongoing research in the Semantic Web community [7].

The toolkit revolves around the notion of a project that groups a number of related files, but the semantic information associated with files is shared among projects. The files related to a project may range from many different kinds of images, notes, paper references and emails to assertions about the findings. In this setting, there is a need to incorporate a simple data model about how data are related to each other. To this end, we have identified a small number of main classes of objects which which a petrologist deals routinely. We are currently revising this model with the help of our collaborators in the Metamorphic Petrology working group. Almost all the data collected centers around a sample which is a physical specimen like a rock sample collected from some geographic location. There is already a recognized need to identify samples uniquely [5] and an existing service [6] allows users to register their samples and obtain a unique SESAR sample number. We decided to adapt this number for identifying the sample to which that data belongs. Depending on a specific subdiscipline of geosciences, this sample is broken down to smaller pieces and is subjected to different types of analyses. Hence, our classes reflect these operations. One of the main classes of objects are thin sections of a sample, called a thin section class, that is subjected to various analyses.

The main aim of tagging is to identify and easily mark the main classes such as sample and thin sections with the names of files to which they belong. This allows us to group semantically related objects and describe their relationships

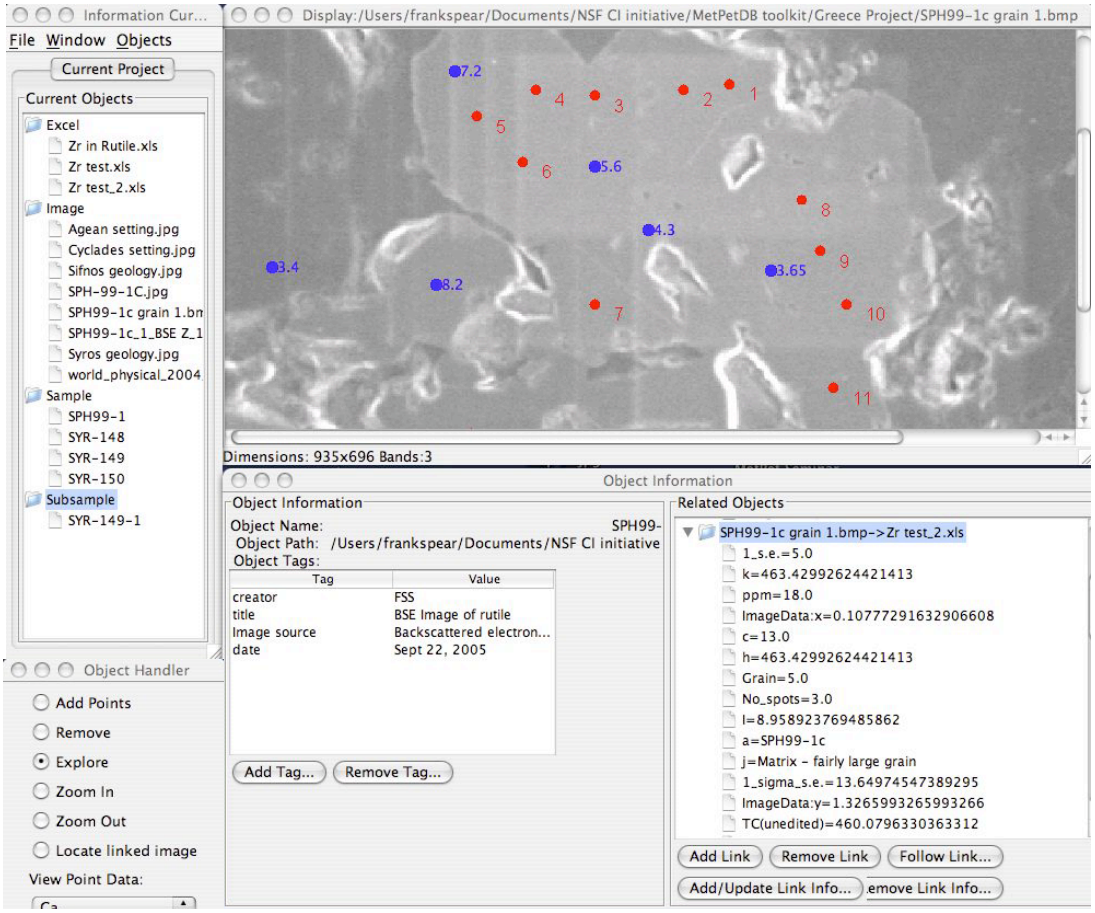

Fig. 1. The screen capture of the toolkit showing annotations on a thin slice

to each other. Some tag types for specific classes are predefined by the toolkit to provide a common starting point for understanding the data. Users are allowed to add new tags or freetext information to each object in the project. Three most important types of tags that are provided by the system are: *where* was it collected from; *when* was it collected and when was it recorded (internal to the system or extracted from files); and *who* collected it. We try to allow users to perform many tagging operations in bulk as much as possible with default values whenever possible. This information also allows the system to visualize information on many levels as long as the necessary relationships are provided by the user. The availability of visual information is of uttermost importance to the understanding of the data. Hence, our toolkit can show the availability of samples on a map; the availability or location of thin sections on a sample; images taken from thin sections and points where analyses are performed. The user is able to specify these relationships and use them to navigate through the data set.

IV. Data Use and Sharing

In our model, the use of data is a natural part of the data. We record implicitly the time data are added and accessed by users. While each piece of data originates from a specific user, each user is allowed to add new annotations to the data at will. In general, we expect two modes of data sharing. A collaborative project involves sharing of data among a group of scientists where each user works on the data on their own computer by adding objects or annotations, and then share it with others in the group using a check-in/check-out procedure. Data in this project are visible only to the members of the group. The second type of sharing is when a group or an individual decides to share the findings with the general public. In this case, the same type of information is made available, but everybody has access to it. Currently, we are developing a centralized database server to facilitate the sharing of data. However, as the system grows, we are going to move to a distributed environment. An important aspect of the database server is reliability. Given data are available to public, they must always remain in the public domain and it must be possible to cite them unambiguously. In this context, data sharing is not different than publishing the data.

The querying and navigation of this type of data poses many challenging questions that we will discuss in this section. Typed and free form semantic tags allow users to locate data easily, but the answers to such queries may be too large to be digested easily. In general, even finding data within one's own computer is becoming a challenging problem.

For a project with many similar tests, navigating the objects can be a problem even for the owner of data. When the data is shared with another person, the problem becomes exponentially harder. Even though semantic information helps the users view data with respect to some structure, in most cases, this structure is not sufficient to understand where to start and how to effectively navigate the data. Clearly, more effective semantic summaries may be needed to point out what is important in the project. As the data is shared between users, this summary must be expanded to highlight what is new in the project.

We intend to use the navigation patterns to help us with this problem. We expect that objects accessed at the same time tend to be related to each other semantically. As these access patterns are repeated, our belief that they are semantically related also becomes stronger. We can analyze access patterns for specific objects, but the amount of data that we obtain about each individual object may be very sparse and hence may not contain many frequent patterns. In this case, semantic groups and groups of objects sharing frequent tags provide us with a meaningful abstraction for conducting this analysis.

To understand and model interesting patterns in the data, we will use a two step method as given in [1]. We first feed access patterns to a data mining engine CSPADE, a sequence mining algorithm [8] which selects paterns with a frequency higher than a given minimum support. Each pattern is a sub-sequence of the form $A\, before(x)\, B$ where B occurs after A with at most x elements in between, also called a *gap* of x elements. The patterns and their frequecy are then used to construct a generalized form of Hidden Markov Models called VOGUE (Variable Order-Gap for Unstructured Elements) HMM [1] that describes the most common patterns in the data. This models allows us to construct a context dependent map. For each object that the user visits, we can find the next set of possible objects ordered with respect to how probable it is that the user might select them during the next data access. This gives us a somewhat hierarchical view of the data with possible cycles. It is possible to use this type of information in a number of ways in the system. First of all, it allows the user to visualize the most common patterns of use or the latest common patterns of use based on the dataset used to construct the VOGUE HMM. A specific project may have multiple patterns of use at the same granularity corresponding to different research activities or problems being investigated. It allows the system to detect these by finding divergence from known access patterns. As a result, a given data set may have multiple views corresponding to different interpretations of the data. Each view may be ordered with respect to the use frequency allowing users to visualize common data models. As the data are shared, we expect more interesting data to be tagged by a larger group of users hence allowing the access models to become more and more specific. This will allow us to develop a special type of HMM, *coupled VOGUE HMM* with information at multiple level of granularity, where some states in the original HMM are replaced by another HMM that reflects the new level of granularity. For example, in the original HMM, a state may correspond to samples from a specific region. This state could be expanded to an HMM that reflects the specific mineral properties of the samples for this region. More detailed observations emitting from the states of the original HMM could be added to include other relevant information for the types of mineral analysis that could be performed. Hence, semantic information allows us to improve the effectiveness of the machine learning module by allowing the learning process to be employed at different levels of granularity. We are in the process of integrating these functionality to our prototype.

V. Conclusions

In this paper, we described a toolkit for organization and sharing of data sets between geosciences. The functionality currently is centered around the subdiscipline of Metamorphic Petrology that our collaborators are coming from. However, we are coordinating our efforts with other efforts to provide semantics to geological data. We believe that our methodology for extracting use patterns from data using data semantics will be useful in all of these efforts. At the same time, our methods will benefit greatly from the annotated data being created by these efforts.

One of the main goals of our approach is to reduce the cost of semantic tagging. Since we expect users to tag data in a way that makes sense to them, developing automated tools to help them is an even harder problem. One possible approach to solving this problem is to find models that predict to which classes an object may belong based on existing information. To this end, HMMs provide us with a new kind of information by summarizing access patterns at different levels of granularity for a specific user or groups of users. This information describes the way the objects are used (object A is followed by object B) and provides a way to understanding how problems are formulated and researched. Making use of this information in a predictive fashion provides many interesting possibilities that we intend to explore. For example, is it possible for the system to propose which tests to run for a specific data set? Is it possible to determine possible dead-ends before even trying them? Can these learned patterns be used to educate students in how to ask questions and examine them? Hence, the examination of use patterns provides us with many interesting possiblities that we intend to explore.

References

[1] B. Bouqata, *Using Data Mining to Improve HMM Estimation and Complexity*, Ph.D thesis, Rensselaer Polytechnic Insitute, 2005.

[2] *GEON: Cyberinfrastructure for the Geosciences*, www.geongrid.org.

[3] *Flickr online photo management and sharing application*, www.flickr.com.

[4] D. R. Karger, K. Bakshi, D. Huynh, D. Quan, and V. Sinha. *Haystack: A General Purpose Information Management Tool for End Users of Semistructured Data* in Proceedings of CIDR 2005.

[5] *PETDB: Petrological Database of the Ocean Floor*, www.petdb.org.

[6] *SESAR - System for Earth Sample Registration*, www.geosamples.org

[7] A. Sheth, C. Ramakrishnan, and C. Thomas, *Semantics for the Semantic Web: The Implicit, the Format and the Powerful*, International Journal on Semantic Web & Information Systems, 1(1), pp.1–18, Jan-March, 2005.

[8] M. J. Zaki. *Spade: An efficient algorithm for mining frequent sequences*. Machine Learning Journal, special issue on Unsupervised Learning (Doug Fisher, ed.), 42(1/2):31.60, Jan/Feb 2001.