

The Digital Library Frontier of Knowledge Generation

NSF Workshop on Post-DL Research Directions

Reagan W. Moore, San Diego Supercomputer Center, moore@sdsc.edu

The integration of digital libraries, data grids, and persistent archives is actively underway. While each community focuses on a different aspect of data management (data publication, data sharing, and data preservation) common software infrastructure is emerging. This process of integration is exemplified in the Library of Congress National Digital Information Infrastructure and Preservation Program, the National Archives and Records Administration persistent archive prototypes based on data grids, and the use of data grids to build persistent archives for the National Science Foundation National Science Digital Library. In the process, the importance of the information and knowledge content that is contained within data collections is being recognized. The worth of a massive scientific data collection is in the ability to identify and name physical relationships present within the data, and then apply the discovered relationships to the scientific processes that were used to collect or simulate the data. The ability to automate the feedback of knowledge from data collections to scientific applications is needed to increase the rate of scientific progress.

The generation, management, and access of knowledge relationships are the major challenges facing the information technology community. During the past five years, the computational science community has generated massive data collections of observational and simulation data. The extraction of knowledge from these large collections is becoming increasingly difficult as the size of the collections increases. The creation of domain specific collections that are tens of terabytes in size is now common, and many disciplines are planning the creation of collections that are hundreds of terabytes to petabytes in size. At the same time, the complexity of the concept spaces used to organize relationships for a given scientific discipline is increasing. The generation of knowledge from data collections is now the equivalent of a grand challenge that requires the use of the most powerful systems available.

Current attempts to automate the generation of knowledge proceed in multiple steps:

- Labeling of the information content by the identification of features within collections
- Characterization of knowledge by the identification of relationships between labeled features (semantic/logical, temporal/procedural, spatial/structural, functional/algorithmic)

- Organization of knowledge as relationships between semantic terms in ontologies or concept spaces
- Generation of logical, structural, and procedural rules from the relationships
- Application of the rules to discover the existence of knowledge relationships in new data collections

These steps appear to be possible on current Teraflop-capable systems. The original data bits are annotated with attributes that provide semantic meaning based on an underlying data model. Relationships between the attributes and between attribute values are used to define knowledge. Knowledge generation on current data collections can consume both the compute and I/O handling power of the entire Teragrid (for an analysis of a 10-TB collection within an hour).

A major challenge confronting current approaches to knowledge identification is that the meaning of semantic terms depends upon the associated context. Also, the choice of semantic terms for labeling features within data collections is not unique. Based upon experiences in the National Science Education Digital Library, semantic labels are expressed within a context that is formed by the aggregation of associated semantic terms and relationships. The identification of the presence of a named feature is done within the context defined by the cluster of related semantic terms and relationships used by a scientific discipline. Thus the extraction of knowledge from collections can be thought of as an all by all comparison of selected clusters of semantically labeled features that obey relational constraints. Each arbitrary cluster may be compared against all other possible clusters to determine whether a meaningful relationship exists across all digital entities within a collection. The result is knowledge as understood by that community. Petaflop capable computers are needed to generate knowledge that will be considered relevant across all communities within a scientific discipline.

This view of the evolution of data management technology is built upon successes within the data grid community for the management of data. The development of fundamental abstraction mechanisms have made it possible to create collections that are distributed across multiple administration domains, while managing technology evolution. At least five key functionalities or transparencies that simplify the complexity of accessing data in distributed heterogeneous systems are used within current data grids:

- Name transparency – The ability to identify a desired digital entity without knowing its name, typically accomplished by queries on descriptive metadata that are mapped onto a logical

name space. The logical name space is used to provide infrastructure independent names and global persistent identifiers for digital entities that are registered into a collection.

- Location transparency – The ability to retrieve a digital entity without knowing where it is stored, typically accomplished by mapping from the global persistent identifiers in the logical name space to a physical storage location and physical file name. For data that is stored under the ownership of a data grid, the administrative attributes for storage location and physical file name can be self-consistently updated every time the digital entity is moved and access controls can be associated with the logical name.
- Platform implementation transparency – The ability to retrieve a digital entity from arbitrary types of storage systems, typically accomplished through use of a storage repository abstraction. The data grid maps from the protocols needed to interact with the storage systems to the operations defined by the storage repository abstraction. Every time a new type of storage system is added to a data grid, a new driver is added to the data grid to support the new storage specific protocol. A similar platform implementation transparency is needed for accessing the information repository in which the metadata attributes are stored that are mapped onto the logical name space. An information repository abstraction is defined for the set of operations needed to manipulate a collection in an information repository, or database.
- Access transparency – The ability to use a preferred access mechanism to manipulate data stored within the data grid. An access abstraction is defined for the set of operations and services that will be performed within the data grid. Particular access mechanisms are then implemented on top of the access abstraction, making it possible to support access mechanisms preferred by each scientific discipline.
- Encoding standard transparency – The ability to display and manipulate a digital entity long after the originating application is gone. This requires understanding the associated data model and encoding standard. Transformative migrations are the conversion of an encoding standard to a non-proprietary, published format that is accessible to multiple applications. Transformative migrations are used to track the evolution of infrastructure independent encoding standards.

The challenge for the next ten-years is that equivalent abstractions are needed for the manipulation of knowledge. In the case of data grids, the problem was relatively simple, in that digital entities were treated as bits that were moved between storage systems under the control of the logical name space. For the management of knowledge, the problem is much more difficult because knowledge relationships are pervasive throughout the infrastructure. The management of knowledge requires the

ability to manage relationships expressed not only within the digital entities, but also within the infrastructure that is used to manage the knowledge itself. There are multiple abstractions that are needed to manage each type of knowledge:

- Digital ontologies – The characterization and organization of the relationships inherent within a digital entity. For example, these relationships describe how to map bits into bytes, map bytes into integers or floating point numbers, map numbers into arrays, map arrays onto coordinate systems, assign physical variable names, assign time stamps, etc. The order of the application of the mappings is important.
- Application ontologies – The characterization and organization of the operations that can be performed upon the relationships present within digital ontologies. The preservation mechanisms of emulation and migration are a combination of these two ontologies. For a digital entity that is being preserved, an evolving set of application ontologies is seen over time. The operations that originally could be used to manipulate the digital entity can be identified and applied based upon the application ontology of a future application. For an application that is attempting to display archived material, an evolving set of digital ontologies is seen. The application can apply operations on the relationships present within the digital entity that the application knows how to manipulate. Since both the encoding format of digital entities and the operations supported by applications evolve over time, both ontologies are needed to preserve the ability to manipulate data collections. The same characterization can be used when considering the manipulation of digital entities created by different scientific disciplines, using different encoding formats and different analysis programs. Digital and application ontologies are essential for interoperability between scientific disciplines.
- Grid ontologies – The characterization of the consistency constraints between the services provided by a digital library, persistent archive, or data grid. Consistency constraints specify the order in which updates to the attributes that are mapped onto the logical name space need to be done. The addition of a new service requires a specification of the dependencies on the old services. A grid ontology is used to specify dynamic changes to the consistency constraints so that software services themselves can evolve. Since grid services can depend upon the structure of the digital entities, grid services may need to manipulate digital entities based upon the relationships described within the digital ontology. An example is the application of a grid service to extract descriptive metadata for the features present within a digital entity. Relationships between labeled features present within a digital entity may impose relationships between the grid services used to identify those features.

- Concept spaces – The ontology used to organize semantic relationships for terms defined within a scientific discipline. A challenge is that the application of a semantic label to a structure within a digital entity is itself governed by a logical/structural/temporal rule that can be described and managed. The assignment of semantic tags to digital entities for the creation of information is governed by rules that can be quantified as relationships within an ontology, managed as a concept space, and separately applied. The information content of a collection can be expressed as a set of processing rules that can be organized as relationships within a concept space. The processing rules in turn can be expressed as services within the grid.
- Knowledge repository abstraction – The characterization of the operations that can be performed on a knowledge repository for the manipulation of a concept space. Typical operations are the mapping from frame-based descriptions of relationships to logic or procedural rules that can be applied to test for the presence of the relationship, the addition of new relationships to a concept space, the query of a concept space for the existence of a defined relationship, etc.

The process of generating knowledge relationships about collections of digital entities requires an infrastructure that has embedded knowledge about services, storage repositories, and even the digital entities. As long as the ontologies associated with the scientific discipline can be kept disjoint from the ontologies needed to manage the processing infrastructure, the development of the required infrastructure is feasible and can be used by multiple disciplines. Present day approaches, however, interlink the infrastructure services with the structure and knowledge of the digital entities that are being manipulated. An example is the embedding of the knowledge needed to extract a pressure array for a digital entity directly within the data management system. Changes in the digital entity structure would need to be correlated with changes in the data management system. The concern is that each scientific discipline is then forced to develop their own knowledge generation system for analyzing their digital entities, and that progress within disciplines is slowed.

We can avoid the problem of discipline specific infrastructure by building a knowledge generation system that maps from relationships organized within a discipline concept space to rules that can be applied by an inference engine, to the set of operations that will be performed by the application or grid service, to the characterization of the digital entity provided by the digital ontology. The infrastructure components can be generic, with separate characterizations for the interactions between services within the knowledge grid and for the interactions of the grid with knowledge repositories. The goal is to build a knowledge generation infrastructure that uses abstractions to characterize the relationships

present in both digital entities and analysis applications, such that the same infrastructure can be used for all scientific disciplines.

Multiple demands are being placed upon the infrastructure design for the integration of digital libraries, data grids, persistent archives, and knowledge generation environments. The integration will build upon concepts from the persistent archive community related to abstractions for the characterization of digital entities and presentation applications, concepts from the data grid community related to abstractions for storage and information repositories, and concepts from the digital library community on characterization of information and semantic interoperability. The integration of these concepts into scalable technology usable by all scientific disciplines is the next major challenge for the digital library community.