

Prospectus for Digital Library Working Group
Reagan W. Moore
San Diego Supercomputer Center

There are two contrasting views of the world:

- how to impose order such that similar data collections are federated (digital library)
- how to impose discovery such that "un-related" data collections can be accessed through the same query (semantic grid).

The former is sought by the digital library and archivist communities to assert control over data collections. The latter is sought by web browsers to improve discovery of possibly relevant services.

The integration of grid technology with digital libraries, preservation environments, and data sharing environments (data grids) requires a better understanding of the semantics used by each community. In particular, digital libraries are an excellent source of semantic knowledge about the collections that are being accessed by grid applications. Coupling the semantics managed by digital libraries with the semantics used to control processing by grids implies interoperability across the semantics used by each system.

Two levels of semantics can be considered:

- the language used by scientists to describe the physical quantities they observe. Examples include the Uniform Content Descriptors developed by the astronomy community in the International Virtual Observatory Alliance
- the consistency constraints imposed on state information managed within a data grid. Note that the relationships between operations on data and updates to state information form an ontology. One would like to be able to dynamically change the set of consistency requirements without having to rewrite software. This requires a standard set of names for operations and for the state information created by the operations, and well-defined relationships that are implemented by the consistency constraints.

There is a tight coupling between knowledge (relationships imposed between semantic terms) and the semantics used by a community. One can invert this coupling and assert that any single semantic term is the result of the application of a set of inference rules or relationships. We can either describe the relationships that are satisfied when we use a term, or assume that within our community the semantic term correctly identifies the presumed relationships.

The set of relationships that underlie semantics are used in the grid community to impose consistency constraints on distributed data management. Two related questions can be asked about the integration of data management systems (digital libraries) and grid technology:

What essential capabilities are required for organizing distributed data into shared collections? Examples of such systems now are digital libraries, data grids, and persistent archives. The reason for managing distributed data is different in each case:

- distributed data source
- distributed data sink
- improve reliability
- improve scalability

- mitigate risk of data loss

Each driving reason can be the source of different semantic terms.

What capabilities should be provided by data management systems to simplify access? It is interesting to note that the individual data management systems can now be federated (controlled sharing of 5 name spaces used for data management - resources, file names, user names, metadata, and constraints). How can each of these name spaces be extended from a control environment to a discovery environment?