White Paper on Knowledge Encapsulation
DataNet Federation Consortium

Reagan Moore

10 April 2013

The National Science Foundation has funded five DataNet projects with the intent of developing data cyberinfrastructure to support science and engineering projects. The projects are the DataNet Federation Consortium (DFC), Sustainable Environment through Actionable Data (SEAD), Terra Populus integrated data on population and environment (TerraPop), Data Observation Network for Earth (DataONE), and Data Conservancy. In this white paper, we discuss the evolution of data cyberinfrastructure from support for data and information to include support for knowledge and reproducible data-driven research.

The DataNet Federation Consortium originally had three main goals: 1) Enable collaborative research, 2) Build national data cyberinfrastructure, and 3) Enable student participation in research. These goals have now been augmented with an explicit focus on providing support for reproducible data-driven research. This extension of the goals is an outcome of interactions with Oceanography, Hydrology, and Engineering domains, and recognizes that disciplines need to manage not only data and information, but also knowledge. The knowledge can be captured in the processes used in analyzing the data to gather scientific results. The DFC supports three types of knowledge encapsulation processes: 1)scientific knowledge encapsulation that is needed for reproducible research, 2) interoperability knowledge encapsulation that is needed to manage access to community resources, and 3) management knowledge encapsulation that is needed to enforce policies on research collections.

Support for reproducible data-driven research can be interpreted as the encapsulation of knowledge through shared computer executable workflows that can be re-executed with minimal human support. The analysis procedures used by a researcher constitute knowledge that can be shared with another researcher. Support for interoperability with other community resources can be interpreted as the encapsulation of the required interaction protocol within standard functions (micro-services). The mechanisms for accessing external resources can be shared with other users. The knowledge needed to manage a collection can be interpreted as the set of controlling policies and procedures generated by a community consensus. The policies embody the intellectual management decisions of that community, and can be published and shared with the managers of other production data management systems.

For a single system to support all three types of knowledge, a description of knowledge encapsulation is needed, along with a description of the framework that

enables each community to apply their specific knowledge encapsulation procedures and rules. The experience of the DFC is that a sufficient basic knowledge encapsulation mechanism is a procedure that evaluates relationships. A sufficient basic framework for managing knowledge is a policy-based data management system that applies computer actionable rules.

In the rest of this status report, we will provide descriptions of the concepts underlying knowledge encapsulation, document how the DFC manages knowledge, and describe why these abilities are an essential component of national data cyberinfrastructure. The lessons learned in the DataNet Federation Consortium show that it is both reasonable and necessary to extend data management systems into knowledge management systems. The approaches used within the DFC also support the interoperability functions needed to build national data infrastructure through a loosely coupled federation of existing data management systems. The development of appropriate interoperability mechanisms has enabled rapid progress towards establishment of collaboration environments that extend across the DataNet Partner infrastructures.

Knowledge Characterization:

The DFC differentiates between data, information, and knowledge through the following computer-science based definitions:

- Data are the objects that are being managed, and consist of bits or bytes.
- Information is the application of a name to an object and the categorization/organization of a set of objects.
- Knowledge is the specification of a relationship between named objects through a transformational process.

This leads to the simple definition that information is the reification of knowledge, or more simply put, the evaluation of the knowledge relationships. The types of relationships that are applied in science domains include:

- Semantic / logical – relationships between names such as membership in a category
- Temporal / procedural – relationships based on causality or order of execution of procedures
- Spatial / structural – relationships based on location in space or position in a structure
- Functional / algorithmic – relationships based on application of a process or mathematical procedure
- Systemic / epistemological – relationships about the entire collection or about properties that are possessed by collections

A characteristic of the use of relationships to define knowledge is that the relationships have to be evaluated to determine whether or not the object in question satisfies the relationship. The application of knowledge is a dynamic process that requires the execution of a procedure that tests or applies the set of relationships. From this perspective, information is a static property that can be

stored as an attribute in a metadata catalog, data are the bits deposited in a storage system, and knowledge is the set of procedures that manipulate the date.

A system that implements knowledge management will need infrastructure to execute transformational processes that evaluate an allowed relationship. A system that enables information and knowledge management will need a stable repository to hold data, information and knowledge entities and their meta-descriptions. The DFC provides these capabilities through collaboration environments that manage the properties as virtual collections, independently of the physical storage system characteristics.

A collaboration environment applies policies and procedures to control the properties of the collection. The choice of policies and procedures depends upon the type of data management application that is desired. For the Ocean Observatories Initiative, the goal is to automate archiving and replay of climate data records composed from real-time sensor data streams. The archived records are shared with both present and future researchers – requiring **digital curation for sensor and time-series data**. For the CIBER-U engineering design education course, the goal is the creation of digital library services that support analyses needed by the engineering community, including a format registry, format identification and transformation services, and controlled access. This is implemented as a **metadata rich repository and access environment** through a MediaWiki interface. The engineering digital library organizes the data, associates standard information attributes with each file, and enables re-use of data. For the Hydrology group, the goal is the automation of a hydrology watershed analysis, including the retrieval of data from remote repositories, the extraction of input parameters, and the execution of the hydrology analysis. This requires **management of a full-fledged scientific workflow eco-system** . The workflows are shared with other researchers, re-executed to verify results, and modified to test research hypotheses.

Knowledge Framework:

Each of these types of knowledge can be encapsulated in a basic function that is applied either within the collaboration environment or at a remote community resources. The collaboration environment serves as a mediator between an external community resource and a research environment, enabling input data sets to be assembled and shared for execution within a research environment. Depending on the computational complexity of the functions that are applied, either the function should be sent to the data source (low computational complexity), or the data should be sent to a compute engine (high computational complexity) through launching of an external workflow mechanism such as Keler, OpenFlow or Pegasus.

An implication is that the collaboration environment should support the virtualization of workflows, managing the properties associated with the workflow

independently of the execution environment. Within the DFC, procedures can be composed by chaining micro-services that can be executed on any compute platform. The I/O operations issued by a micro-service are automatically mapped to the I/O protocol required by the storage system where the data are located. The micro-services can be moved to the remote storage location, the processing can be applied to the data, and the resulting data set returned to the researcher.

The present data cyberinfrastructure is composed of a very wide variety of systems. To support interaction with existing infrastructure, interoperability mechanisms are needed that simplify access and encapsulate the knowledge required for interaction. The interoperability mechanisms can be characterized depending upon whether they support data, information, or knowledge manipulation. Within the DFC, the following interoperability mechanisms are applied in Phase I:

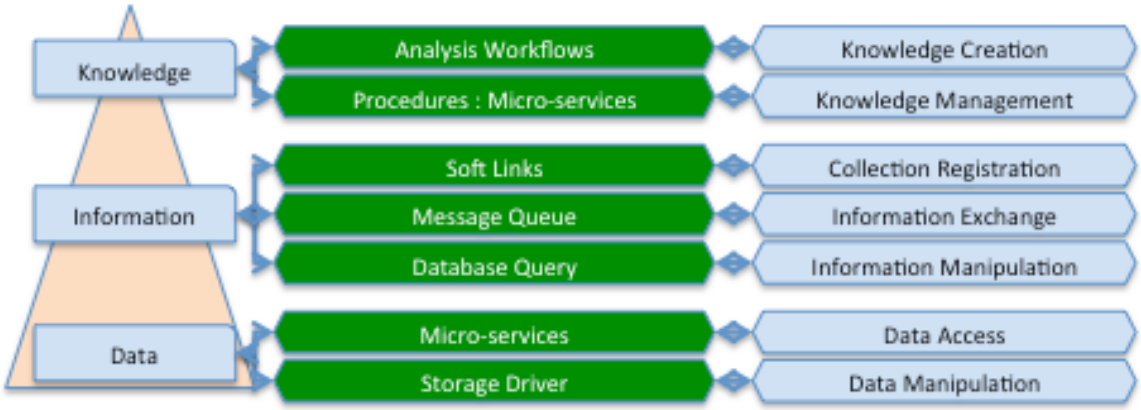

Figure 1. Types of interoperability mechanisms that encapsulate knowledge

The DFC supports both remote data access and remote data manipulation. For data manipulation, a storage driver executes partial I/O commands to support generation of data subsets at the remote storage location. For data access, a micro-service can be written that executes the protocol needed to access a remote repository. Thus data access is either through a micro-service or a storage driver.

At the information level, the DFC supports formation of collections through soft links. This registers the existence of a data set in remote repository into the DFC collaboration environment. In the registration process, additional operations can be performed when retrieving the data. Soft links are implemented through a micro-service. Information exchange is implemented as an asynchronous process through an internal messaging system within the data grid. Messages can be sent to a queue by one process, and read from the queue by another process. For interoperability between message queues, rules are used to check on the presence of messages and push them to an external message queue. For information manipulation, the DFC collaboration environment supports the execution of queries at remote databases. The queries can be registered into the collaboration environment. When the registered query is accessed, the system dynamically executes the query at the

remote database, and caches the result in the collaboration environment. The mechanisms for accessing the remote database can be implemented in a micro-service. Thus information access is either through a micro-service, or a rule, or a database driver (for access to the internal metadata catalog).

The DFC supports knowledge generation through the application of procedures to create new data products. The procedures are composed by chaining together micro-services into a workflow (micro-services are akin to actors in scientific workflow languages). The workflow can be registered into the collaboration environment. When the workflow is accessed, the data grid executes the workflow, and automatically saves the input and output files as well as any intermediate results that are desired. The workflow can be shared along with the input and output files. The workflow can be re-executed and the output results can be compared. The system has the capability to version the workflow executions. This makes it possible to encapsulate analyses within workflows and enable reproducible research. The knowledge represented by a community consensus on collection management properties can also be captured in rules that chain together micro-services. Both the knowledge needed to do an analysis and the knowledge needed to manage analysis results can be encapsulated within the DFC data grid.

These examples show that a policy-based data management system provides the basic framework needed to support application of procedures that encapsulate some form of knowledge.

DataNet Interoperability:

We can characterize DataNet interoperability mechanisms as types of encapsulated knowledge that capture interaction methods. For the current DataNet Partners, multiple types of interoperability mechanisms are needed: web service invocation for data access or deposition; message queue synchronization for depositing statistical information in triple stores; and native protocols for accessing storage repositories.

The interoperability mechanisms are implemented within the DFC collaboration environment. The collaboration environment constitutes middleware that unifies access between traditional research environments and community resources such as data repositories, information catalogs, and analysis services. Depending upon the type of procedure that is applied at the remote site, different interoperability mechanisms may be needed. The DFC policy-based data management environment is capable of implementing all of the interoperability mechanisms needed to link the DataNet Partner's cyberinfrastructure.

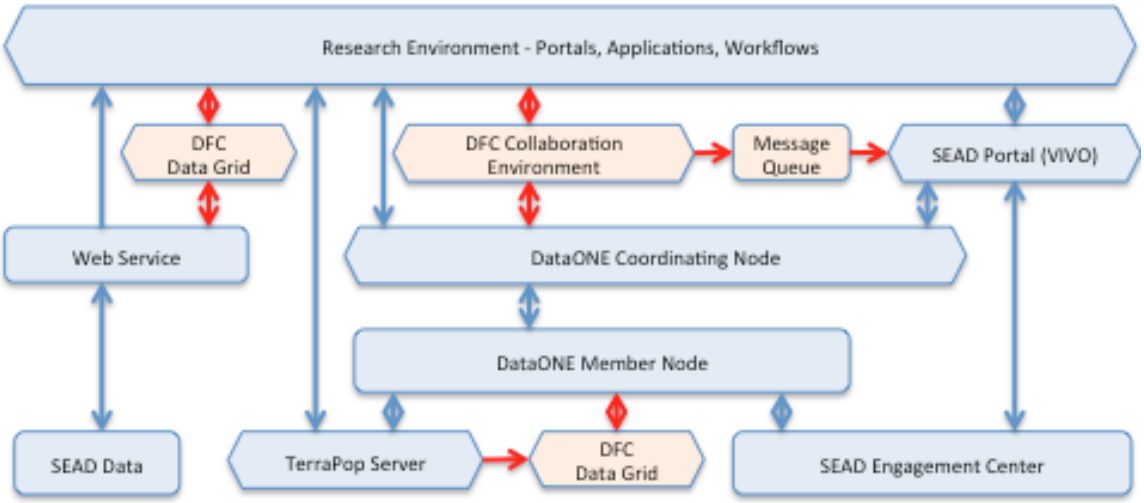

Figure 2. DataNet Interoperability Interfaces

An even more extensive set of interoperability mechanisms were tested by the NSF EarthCube Layered Architecture concept award. The DFC provided access to the DFC federation hub for the demonstration of these additional interoperability mechanisms. Interoperability demonstrations performed on the DFC federation hub included workflow interoperability (NCSA Cyberintegrator, Kepler), interoperability through brokers (GeoBrain) for accessing federal repositories, and interoperability through web services for accessing the CUAHSI HIS repository. The GeoBrain broker was also used within the DFC for access to Digital Elevation Maps for the RHESSys eco-hydrology analysis. In each case, the required mechanism was implemented as a micro-service that supported the appropriate external repository access protocol.

Interoperability Layers:

There are multiple layered components within national data cyberinfrastructure. Each layered component addresses some property such as authentication, or network access, or workflows. For each component, multiple technologies are currently used in the existing data management systems. Thus an interoperability mechanism is needed for each layer of the architecture.

Figure 3 shows a list of ten component layers, the interoperability mechanism that is used to support access, and the types of technologies that are in common use. The interoperability layer name is listed on the left, the interoperability mechanism is listed in the center, and the types of currently used technologies are listed on the right. The purpose of the figure is to identify the types of interoperability mechanisms that are needed for current infrastructure, and see whether they correspond to the interoperability categories of drivers, micro-services, and rules.

Note that the interoperability mechanism for authentication has evolved over the last twelve years. Previously, the Generic Security Service API (GSSAPI) provided the desired interoperability mechanisms. Now Pluggable Authentication Modules (PAM) are used to support interoperability across identity management systems and authentication systems.

Also note that the DataNet Partners are included in the categories as examples of external data repositories that can be accessed. In practice, the DataNet Partners also provide information catalogs and data analysis services. The interoperability interfaces were implemented through micro-services and also through rules.

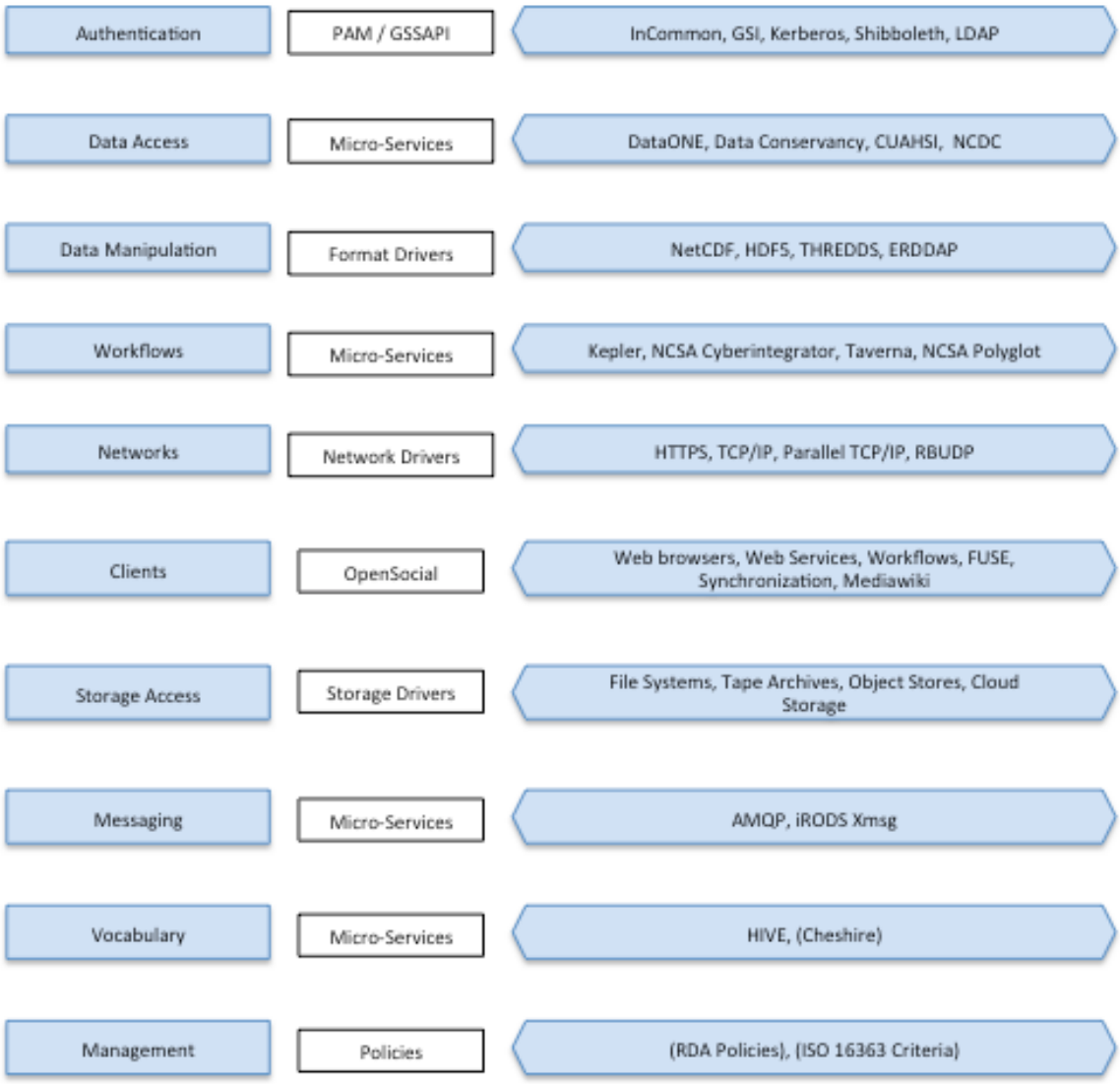

Figure 3. Interoperability layers needed for national data cyberinfrastructure

In each layer, either storage drivers / format drivers, or micro-services, or policies / rules are used to manage interoperability

Implementing National Data Cyberinfrastructure:

The set of tasks that are needed to implement national data cyberinfrastructure can be quantified through seven basic steps. The infrastructure needs to support the specific data, information, and knowledge needs of each discipline. The infrastructure needs to provide a generic framework for implementing each unique requirement, and for supporting the appropriate interoperability mechanisms. The framework is then applied through a loosely coupled federation architecture to link the existing data cyberinfrastructure resources with collaboration environments. These steps have been applied within the DFC through application to Oceanography, Hydrology, and Engineering projects. In years 3-5 of the DFC project, the steps will be applied to three additional domains including Cognitive Science, Plant Biology, and Social Science. The expectation is that the same approach will enable management of data, information, and knowledge for all science and engineering disciplines.

1. Build national data cyberinfrastructure prototype
   Support multiple science and engineering domains by loosely coupling their existing infrastructure with a collaboration environment
2. Develop generic interoperability framework
   Define the generic infrastructure needed for the national infrastructure to manage knowledge as well as data and information
3. Define interoperability mechanisms
   Support access across the disparate types of infrastructure in common use
4. Define domain specific extensions
   This is done at three levels: technical interoperability, project level policy, and end user usage requirements
5. Support federation of community resources
   Link the community resources to a collaboration environment through knowledge encapsulation.
6. Manage a community based collection life cycle
   Track evolution of management policies as a collection is re-purposed from a local project, to a shared collection, to a digital library, to a processing pipeline, and to an archived reference collection.
7. Manage cyberinfrastructure evolution
   This is based on the management of knowledge, as well as data and information.

The above steps constitute a core set of tasks that can be applied to enable a new science and engineering domain to manage data, information, and knowledge. The domain is loosely federated with other disciplines through collaboration environments that manage knowledge encapsulation. The knowledge required to interact with a domain's data can be captured in procedures, and used by researchers from another domain to extract desired data subsets. This forms the basis for interdisciplinary research. Without some form of knowledge encapsulation, data re-use across domains is a labor intensive process that slows

down research advances. With knowledge encapsulation, analysis processes can be automated, re-executed on demand, and compared with prior research results to better understand new phenomena.

Summary:

A collaboration environment provides the knowledge encapsulation needed for the sharing of data. The collaboration environment enables common semantics, shared collections that span multiple storage systems, and a single sign-on environment for authentication. Collaboration environments constitute middleware that encapsulate the knowledge generated by a community consensus for management of shared data. Thus collaboration environments are a limited form of knowledge environments.

The DFC has explored the extension of collaboration environments into scientific knowledge environments and interoperability knowledge environments. Scientific knowledge is captured in research workflows, and interoperability knowledge is captured in the interoperability mechanisms needed to support loosely coupled federations of community resources. The result is a general knowledge environment that enables reproducible data driven research.

The general framework that supports knowledge environments is a policy-based data management system that registers data, information, and knowledge workflows. Specific knowledge functions are encapsulated within drivers, micro-services, and policies. The DFC uses the iRODS integrated Rule-Oriented Data System as the basis for implementing knowledge environments for Oceanography, Hydrology, and Engineering. Based on the success with these disciplines, a similar approach can be used for additional NSF funded research projects.

The DFC technology is capable of implementing the wide variety of data management applications in current use, from simple collection formation for a project, to organization of distributed data into a data grid for building shareable collection, to management of a digital library for publishing data, to provision of a data processing pipeline for automating generation of derived data products, to preservation of records in an archive. Each application applies the generic mechanisms (policies, micro-services, drivers) to encapsulate community knowledge. Through policy evolution, it is possible to migrate a set of data through all of these stages of a community-driven collection life cycle.