This document is primarily of historical interest. It describes the work for a $40M project. The project was awarded in September 2011 at $8M over 5 years with $1M each of years 1 and 2.

Project Description: DataNet Federation Consortium Vision and Rationale

A new approach is needed to assure that the data generated today are available and useful tomorrow for the next generation of researchers. Long-term sustainability mechanisms are needed that deal with technological obsolescence and discipline and societal changes. The DataNet Federation Consortium (DFC) provides a new approach for implementing data management infrastructure that transcends technology, social networks, space, and time through federation-based sustainability models. To show the generality of the approach, DFC will federate data management systems from six science and engineering disciplines. DFC will support collections and services that span individual, institutional, regional, national and international repositories [1] and demonstrate collaborative research, education and outreach on shared collections.

DataNet Technology: The DFC federation technology is based on the integrated Rule-Oriented Data System (iRODS) [2], a second-generation data grid. The iRODS system provides a stable virtual data platform, based on policy-oriented data management, for federating multi-disciplinary collections across distributed heterogeneous resources. iRODS provides a rich interface that supports a range of client applications, from discipline-centric analysis and visualization tools to emerging social networking web applications. The iRODS system scales to 100s of millions of data objects in Petabyte storage systems and supports high-speed data transport [3]. Evaluations of the suitability of the iRODS data grid technology for managing large shared collections, digital libraries, preservation environments, and data processing pipelines have been conducted by six organizations: the Jet Propulsion laboratory (JPL) [4], the NASA National Center for Computational Sciences (NCCS) [5], the French National Library [6], the Large Synoptic Survey Telescope (LSST) [7], the Australian Research Collaboration Service (ARCS) [8], and AVETEC [9] for compliance with DoD High Performance Computing Centers requirements. These evaluations are published on the web and accessible at http://irods.org.

Sustainability Approach: Sustainability is the management of change, whether in technology, economic basis, access, or policy. Systems based on open source, modular, and service-oriented platforms such as iRODS are candidates for technological sustainability. Institutional commitments drive economic sustainability, either through cost models that demonstrate an economic incentive within a single institution, or through collaborations between institutions for joint custody of a collection. The ability to share collections requires support for the access mechanisms and enforcement of the management policies required by each institution. The DFC will develop a "Business-Model Classification for Data Collections" to study the effectiveness and utility of the use of federation as a sustainability model. The DFC will explore the transition from multi-tiered project management within the DFC to a multi-foundation business management model for sustainability of the DFC project itself.

The DFC will collaborate with the Data Conservancy led by Johns Hopkins University [11] on the use of Fedora Digital Library software [12] for preservation. The DFC will explore scalability extensions through integration of Fedora with the iRODS data grid [13]. The DFC will collaborate with the DataONE project at the University of New Mexico [14] on federation of the data infrastructure to enable data and service sharing between their environmental and ecologic data collections and DFC's oceanography and hydrology data collections.

Collaborative Science & Engineering: The initial research communities to be federated in DFC comprise six National Science Foundation funded projects in science and engineering. Each participating NSF project is a consortium of academic institutions. The consortia collectively involve more than 150 academic institutions, including eight minority-serving institutions. The consortia span six disciplines: Cognitive Science through the Temporal Dynamics of Learning Center (TDLC) [15]; Oceanography through the Ocean Observatories Initiative (OOI) [16]; Biology through the iPlant Collaborative (iPC) [17]; Hydro-Meteorological Sciences through the Consortium of Universities for Advancement of Hydrologic Science (CUAHSI) [18] and the Renaissance Computing Institute (RENCI) [19]; Social Science through the Odum Institute for Research in Social Science (Odum) [20]; and Engineering through the Cyber-Infrastructure-

Based Engineering Repositories for Undergraduates (CIBER-U) education initiative [21]. The DFC will demonstrate the wide applicability of the federation and sustainability approach by integrating data, applications, services and workflows from these diverse disciplines.

The motivation behind the DFC approach for transcending technology, social networks, space, and time is best understood by examining the research agenda and data management requirements from the collaborating NSF consortia. These consortia determine the requirements for the creation of generic data management infrastructure, and the types of data formats, processing services, management policies, assessment criteria, and long-term sustainability mechanisms that will promote their collaborative research goals.

Temporal Dynamics of Learning Center (TDLC): The TDLC is one of six NSF Science of Learning Centers (SLC) [22]. TDLC aims to achieve integrated understanding of the role of time and timing in learning, across multiple scales, brain systems, and social systems. The scientific goal is therefore to understand the temporal dynamics of learning, and to apply this understanding to improve educational practice [23]. Learning occurs at many levels: at the level of synapses and neurons; at the level of brain systems involved in memory and reward; at the level of complex motor behaviors; at the level of expertise learning; and finally, at the level of learning via social interactions between teachers and students. TDLC initiatives address such fundamental research questions as: How is temporal information about the world learned? How do the intrinsic temporal dynamic properties of brain cells and circuits facilitate and/or constrain learning? How can the temporal features of learning be used to enhance education? What are the best theoretical ways to conceive the temporal dynamics of learning in the brain and between brains?

Answering these questions cannot emerge from a single line of inquiry, so TDLC's research model is collaborative and interdisciplinary from the beginning. The center has created communities of scientists that cross disciplinary and institutional barriers in pursuit of these common research questions. Researchers in machine learning, psychology, cognitive science, neuroscience, molecular genetics, biophysics, mathematics, and education focus on these issues from multiple perspectives, synchronizing their research in parallel experiments in animals, people, and theoretical models. The center includes laboratories from 12 universities in the US, Canada, Australia, and UK. A significant challenge for collaborations among such geographically distributed scientists is sharing large quantities of data and stimuli quickly and easily, while carefully controlling access to only the collaborators permitted to view and manipulate the data.

Consider the following scenario: *Scott Makeig's group at UCSD has a local copy of EEG recordings made by David Sheinberg's lab at Brown, where monkeys learned to be experts at classifying certain visual objects. Scott's group applies their Independent Components Analysis tools to Sheinberg's data to remove artifacts and find the independent sources of the signals. Immediately, this new analysis becomes available to all authorized laboratories. The Data Grid now contains both the "raw" and processed EEG, along with the images that evoked the responses from the monkeys. Tim Curran, whose laboratory in Boulder is authorized to connect to the Data Grid, queries the iCAT catalog at UCSD. He had previously submitted his EEG data from humans trained on the same stimuli as Sheinberg's monkeys, which were analyzed by Makeig. Now Tim is able to compare directly the processed EEG components from the monkeys to the components evoked in humans learning to categorize the same stimuli. The system can line up the data according to the images used as stimuli. Local copies of the images do not exist on Tim's server, so they are automatically retrieved from the closest server with copies. Concurrently, Gary Cottrell's group at UCSD is able to process these same images using their machine vision algorithms, comparing the internal dynamics of their models with the neural dynamics of monkeys and humans. Eventually, the entire dataset, including neurophysiological measures, stimulus images, and models, will be shared with a broader scientific community.*

One initiative of the TDLC is to develop and deploy innovative technologies to support this kind of data sharing in the learning sciences, not only for the TDLC but in also partnership with the other NSF Science of Learning Centers [22]. The goal is to enable just-in-time sharing of neurophysiological data, motion-capture data, fMRI and electrophysiology data, and high-quality images and video across many laboratories. This requires easy, efficient, fault-tolerant transfer of hundreds of gigabytes, terabytes, and one day perhaps petabytes of data on a regular basis. Collaborators also need to be assured that shared data are seen only by those with permission dictated by human Institutional Research Board (IRB), HIPAA [24], and animal IACUC [25] protocols. And after a project or even TDLC ends, data need to be de-identified before sharing outside the immediate collaborative group, as dictated by IRB protocols. TDLC's challenge is technology for data sharing that includes speed, fault-tolerance, and sophisticated access control but at the same time is easy for scientists to install, maintain, and use on a regular basis.

The DFC will collaborate on development of institutional management policies and apply SLC research results in design and evaluation of tools for educational use of scientific data collections. This requires collaboration on development of IRB approval policies, development of appropriate context for each data set, and integration of data manipulation tools with DFC infrastructure to assemble reference collections for future researchers. Sustainability mechanisms will be explored through institutional commitments supporting education initiatives and local researchers.

Ocean Observatories Initiative: The OOI unifies observational data for oceanography research. Real-time sensor data streams from independent coastal, regional, and deep-sea sensor systems will be organized into collections analyzed to detect significant ocean events [26]. Data types include point measurements, time-series, LiDAR [27], and high-resolution video. Data are compared with prior observations and simulations of ocean state to actively control sensors and instruments. This will transform the conduct of oceanographic research [28]. Data streams will be assimilated into ocean models to interpolate the "now" state of the oceans and predict future states. The ability to respond rapidly to detections and ocean model variance will make it possible to study catastrophic events and provide better resolution of dynamic changes.

A Cyberinfrastructure Implementing Organization (OOI CI) is coordinating the development of the data management systems. Since multiple teams within the OOI control the sensors, social networking tools are needed to reach consensus on data formats, required representation information [29], analysis tools, and data sharing policies. Policy management tools are needed to enforce the consensus and support long-term preservation for a data life cycle at least to 2035. Climate changes slowly, with single cycles lasting as long as 25-30 years; a meaningful contribution can be made only if observations and derived knowledge are maintained over multiple cycles, including critical data and model provenance. Success in understanding climate and efforts at climate change mitigation depends on maintaining data for decades to centuries.

The OOI has completed its NSF Review Process and Final Design Review. Goals include open, near-real-time access (latencies of seconds) to ocean environmental data and use of event detection and ocean/atmosphere modeling. The OOI exploits previous experience with the Storage Resource Broker (SRB) [30-31] to use iRODS in the data and knowledge grid internal to the system architecture [32]. A unique component is the use of cloud computing and cloud storage caches to manage analysis of data streams. Data are cached at the appropriate location in the cloud to support real-time analysis, and archived for future comparisons in institutional repositories or Teragrid [33] storage resources. The iRODS data grid manages data distribution into the cloud storage, federation across storage repositories, and long-term preservation.

Work already conducted with cloud computing through Amazon's EC2 and S3 [34] for the data distribution network is unique within the proposed DFC. The OOI Cyberinfrastructure Implementing Organization also provides an interface to an allied Education and Public Engagement (EPE) Implementing Organization [35] tasked with developing OOI's interface with the non-oceanographic world. The OOI CI has also developed working relationships with NSF's NEON [36] and CUAHSI programs and NOAA's Integrated Ocean Observatories System (IOOS) [37].

The DFC and OOI will collaborate on development of policies for controlling sensors, federating sensor streams, applying data processing workflows, and archiving sensor and model data. The DFC will also promote interoperability between the OOI open, near-real-time data management system and regional projects at RENCI that support hurricane disaster planning. The DFC will collaborate with OOI on interfacing the iRODS data grid to the cloud to control caching of data near compute servers while managing a persistent archive of sensor data. The DFC will collaborate with the OOI on federating their data management system with institutional repositories, Teragrid archives, and other agency data collections. The DFC will also promote use of OOI data in education classes, based on OOI policies for re-use.

iPlant collaborative [38-40]: The mission of the iPlant Collaborative (iPC) is to support investigation of grand challenge questions in the plant sciences and to foster education and training for K-12 through graduate by accessing, developing, and managing digital dataset collections (both physical and virtual) and disseminating these resources. iPC will build a unified cyberinfrastructure to address these questions. The first grand challenge focuses on understanding the evolutionary relationships between plants by building and analyzing a plant tree of life. The second focuses on bridging the gap between genotypic information (genome sequence, genes, and chromosomal markers) and phenotypic information (traits and physiology) in the context of evolutionary and ecological relationships.

These grand challenges require virtual integration of a variety of datasets and types, including DNA sequences (e.g. GenBank [41]), protein structure (PDB [42]), gene expression patterns (NCBI GEO [43]), pathway networks (KEGG [44]), and external morphology (MorphBank [45]). The iPlant collaborative will be a federated system providing access to distributed data repositories. While several capabilities for iPlant can be supported by adapting and integrating existing tools and techniques, fundamental information representation and other computer science research problems need to be addressed. The ones of primary importance include: (1) Developing an information model to depict the semantics of plant sciences data and their inter-relationships; (2) Developing a framework for automatically deriving and annotating links among multiple plant sciences datasets and dynamically adding new links as discovered. (3) Resolving semantic conflicts among datasets. (4) Data Provenance to track, assess, and inform users about the quality of data, ensure proper attribution for data, and resolve data “ownership” issues. (5) Information Life Cycle Management to develop mechanisms and policies to decide what data to retain, remove, or archive and when. Processes and goals relevant beyond iPlant will be shared with the DFC federation to enable application and standardization across disciplines.

An important concern for the iPC is long-term sustainability and preservation of data, tools, services, and other resources. The iPC expects to develop robust economic approaches to ensure survival of iPC beyond the grant period, and will collaborate with DFC on sustainability research. The DFC will also collaborate with iPC on computer science research issues, provide generic data management infrastructure, and collaborate on application of DFC technology to iPC data challenges. Explicit tasks include development of data sharing agreements and replication policies for implementing multi-institution data repositories; development of an information model for provenance management and quality assessment; automation of administrative functions; and investigation of sustainability mechanisms for iPC digital holdings.

Hydro-Meteorological Sciences: Access to hydrological and meteorological data, both historical and real-time, is critical for scientific research on climate change and for forecasting a broad range of societal impacts from daily weather to water runoff. The fields of discipline include meteorology, hydrology and climate change. Changes to water resources impact human consumption, agriculture and manufacturing production; health effects from disease spread and heat stress; disasters such as hurricanes, flooding, landslides, and icing; environmental quality for air, water, and soil contamination; and socio-economics of urban and regional development and commodity trading [46]. However, untangling the complexity of these issues has been severely limited by the ability to locate, access, analyze, and disseminate the multiple datasets needed to understand the issues and underlying principles.

The DFC infrastructure builds a foundation for multidisciplinary research while exploring the approach for future mechanisms to maintain, integrate, and make hydrometeorological and geospatial datasets more transparent. Traditionally, weather and climate, and hydrological communities have segmented datasets into many, separate repositories with different schema and meta-data that are maintained by federal, state, and local government agencies, the private sector, and academic institutions. The DFC will develop strategies to virtually integrate the various repositories to enable cross-disciplinary research, and foster the development of context sensitive decision-support tools that can more fully utilize these datasets in practice.

For example, RENCI is partnering with CUAHSI to utilize their Hydrologic Information System (HIS [47], providing access to point observations (e.g. streamflow, groundwater levels, snow levels, precipitation, etc.). These data are collected and maintained by organizations such as the USGS, the EPA, the NCDC/NOAA, and by scientific investigators that use the underlying CUAHSI-HIS Observations Data Model (ODM). All data served through CUAHSI-HIS are accessible through a web services API called WaterOneFlow offering standardization of the disparate data archives into a standard XML-based format.

While the CUAHSI-HIS provides unprecedented access to hydrologic time series data, it has been limited to point data. Understanding hydrologic systems, in particular when considering the effects of anthropogenic and climate change on water resources, requires a richer description of the environment. Many datasets necessary for completing the digital description of hydrologic systems for analysis and modeling purposes do exist – terrain data using LiDAR, land cover change data from classification of imagery data, hydrography data that represents lakes, streams and estuary boundaries, census data – however these data are not served by the CUAHSI-HIS and often remain isolated within individual data providers. Federating these datasets under DFC will enable scientists to gain a more complete understanding of the water cycle and address a variety of science questions related to hydrologic function across spatial and temporal scales.

For climate data a similar situation exists: national and worldwide climate data are housed at the NOAA National Climatic Data Center [48], regional data at the Southeast US. Regional Climate Center at UNC-CH, and state data at the North Carolina State Climate Office at NCSU. The DFC will support analysis across multiple geographic scales using data distributed in long-term archives with diverse types of data including point, area, imagery, model output, and technical summaries. These data are analyzed and disseminated for climatic studies, and fed into real-time products such as weather forecast models. The DFC will develop federation mechanisms to combine multi-dimensional data into a virtual space. This unique scaling of data will enable applications of climate research results within regional analyses. DFC will leverage a RENCI project with the NC Division of Emergency Management to assess impacts of sea-level rise on coastal and flood prone areas [49]. This project will ingest LIDAR, geospatial, and orthophotometry imagery from the State, combined with resource and geospatial datasets from NOAA, NASA, USACE, and USGS, and socio-economic data from a number of institutions. By leveraging on-going projects and products, DFC will facilitate understanding of issues of human health and agriculture with data from the US-EPA, USGS, USDA, DOT and DHHS federal agencies.

In another example, RENCI is partnering with NOAA at the National Climatic Data Center whose current data holdings of about eight petabytes are predicted to increase to 160 petabytes over the next decade. RENCI has actively partnered with NCDC to develop tools to access and use the massive archives and provide informative tools based on mining operations. RENCI is teaming with the NOAA HydroMeteorological testbed, initiating a five-year research-to-operations field measurements campaign in North Carolina starting in 2010. This program will develop from the ground up a data archive to support hydrological and meteorological research for the foreseeable future. Lastly, working with NOAA/NWS and the USEPA, DFC will develop strategies for creating data inputs for models and managing the large modeled outputs that can be generated from atmospheric meteorological and chemistry models.

Social Science Data Archives (SSDA): Social science researchers use datasets collected by researchers in sociology, political science, psychology, and city & regional planning. The many years of experience these disciplines have gained through working together has driven the development of multidisciplinary approaches to better understand the world. This ability to frame and engage complex social interactions and trends across a wide diversity of research questions is a tremendous asset [50]. The ability to seamlessly integrate data and to develop new methods of analysis is the key to future ground breaking social science discoveries.

Social science datasets will be integrated into the DFC through an expansion of the Odum Dataverse Network [51]. Building on previous work, Odum will develop policies and procedures that provide a bidirectional link between the DFC and Odum's social science data archive infrastructure. Social science data rely heavily on an extensive set of metadata that allows for the discovery, analysis, and preservation of studies and use the Data Documentation Initiative (DDI) as the metadata standard [52]. The expanded ability of the Dataverse Network to federate with the iRODS-based DFC will allow social science datasets to be managed with policy based rules that protect the authenticity, privacy, provenance, context, and integrity of social science datasets. The ability to preserve social science data seamlessly in an environment that has been tested and is proven to be scalable will be an important asset to the social science data community.

The Odum Institute has implemented a form of federation-based sustainability for long-term preservation. In addition to relying on multiple funding streams, Odum helped formed the Data Preservation Alliance for the Social Sciences (Data-PASS) [53]. Joint governance documents and MOUs protect the long-term interests of social science data preservation institutions through distributed storage and federated union metadata catalogs. Sustainability requires enlistment of a new institution in the event of an organization failure. Just as a multi-tiered funding approach smoothes out funding cycles, federation-based preservation environments provide even greater protection. The Odum Institute will provide replication resources and expertise for social science datasets in the DFC. Increasingly, efforts to study society and social frameworks cut across disciplines, and social science research is a natural place for multidisciplinary ventures.

CAD/CAM/CAE Engineering Archives: Recent reports from NSF, NAE and DoD have all lamented the fact that the "inter-disciplinary engineers" so desperately needed do not exist in adequate numbers to address today's national challenges [54]. There are not enough educational initiatives and programs to produce these new engineers, nor have the standard, stove-piped, curricula of engineering and computer science departments adapted to this need. Reports from the National Academies and elsewhere echo these concerns, documenting how disciplinary boundaries impede innovation, delaying converting new ideas into innovative products.

Traditional engineers are trained in specific disciplines, often finding it difficult to assimilate fundamental computing and information technology concepts needed in emerging areas of national need. *Furthermore, while information technology and computing is central to the creation of nearly all products and systems, the enterprise of developing interoperable digital representations and robust computational tools for design, modeling, simulation and analysis is still largely performed either by computer scientists (who do not adequately understand the engineering domains) or by engineers (who are inadequately trained in computer science).* Education and research in model building has been done in isolated fields—leading to advances in important, but relatively narrow, areas. The result is un-sharable representations, un-integrable systems and untested software tools of unknown accuracy and questionable reliability.

The DFC represents an opportunity to inter-connect existing education and training programs and create a truly national engineering informatics infrastructure that unites computer and information sciences with traditional engineering domains. Much like the transformation of "Science" to "eScience", the DFC is essential to the creation of an informatics-centered engineering. Engineering Informatics is the science of representation, simulation, archiving, and re-use of engineering knowledge in transformative ways. While other industries (e.g., financial, retail, digital entertainment) reap benefits and economies of scale from the information revolution, the manufacturing and engineering industries continue to lag behind.

Objectives in support of engineering under DFC will include support for engineering education. Previous NSF funding (Regli, OCI-0636273, OCI-0636235, SCI-0537125 and SCI-0537370), has built a substantial library of data, simulations and lecture materials to support engineering informatics and design education. Specific collections that will be migrated onto the DFC include: (1) virtual dissection laboratory; (2) bio-inspired robotics portal; (3) collaborative design exercises. The DFC will significantly expand the scale and scope of previous work under NSF OCI's CIBERU Program, enabling a national reach. The current cyber-collaboratory involves 32 faculty in 12 different disciplines at 9 universities: Penn State, Bucknell, Drexel, Northwestern, University of Missouri-Rolla, University at Buffalo, Virginia Tech, Sweet Briar College, and Norfolk State University (a minority-serving institution).

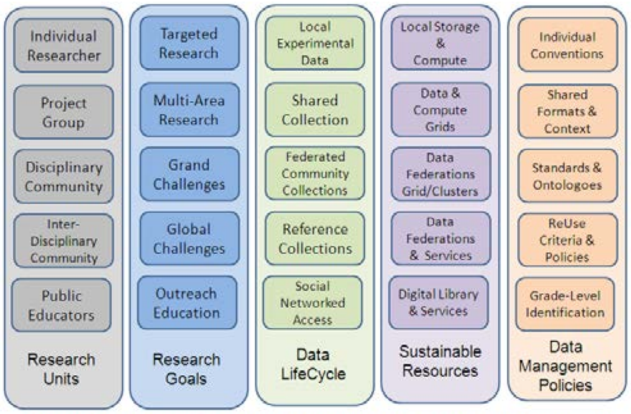
The DFC will establish cyberinfrastructure for the engineering equivalent of the Visible Human Project, complete, anatomically detailed, three-dimensional representations of the product components. Engineering dissection enables self-discovery and analysis of complex systems, enhancing engineering intuition [55]. The proposed cyberinfrastructure for engineering dissection will house data and educational materials and activities to support both physical and virtual dissection of engineered products and systems, links to hardware for reverse engineering (e.g., 3-D scanners, coordinate measuring machines), material testing (e.g., Brinell hardness and tensile testing machines), etc. This cyberinfrastructure will overcome several major deterrents to "hard-copy" engineering dissection: (1) start-up and maintenance costs, (2) space for disassembly and storage, and (3) preparation of educational materials and activities. It will also enable access to more complex products (e.g., refrigerators, automobiles) through virtual dissection. Engineering dissection is expensive (maintenance costs alone average $1,000/year at Buffalo for a class of 200 students to $5,000 per year at Virginia Tech for 1,200 students).

The National Design Repository [56] is a digital library of Computer-Aided Design (CAD) models and engineering data from a variety of domains, consisting of over 55,000 CAD models and assemblies [57]. Developed and maintained by Regli and his students, the Repository currently serves about 1,000 users a month. The NDF uses W3C and Semantic Web standards for engineering applications and enforces security boundaries. While the above narrative talks about "CAD/CAM" data, these facts are true in general for all forms of 3D data, including Architecture-Engineering-Construction (AEC) as well as the 3D models created by archeologists and game designers. DFC will be a vehicle that brings multiple 3D data sets together from Engineering and Science disciplines and that shares common tools to provide deeper insight across the disciplines.

The diversity of applications and services needed by the participating disciplines, and the heterogeneity of data resources in these disciplines, bring challenges in providing a seamless integrated system – infrastructure, policy management, administrative organization and sustainability models. The DFC federation framework provides a vehicle that can help not only federate data and services across disciplines and societal variations but also across technological evolution and fiscal sustainability for current and future usage.

DFC Federation Approach: Each collaborating science and engineering discipline has an objective or purpose that drives their initial choice of data to assemble in a collection. During the life cycle of their data, the purpose may evolve from managing data for a specific researcher, to formation of a shared collection to foster collaboration in a project, to publication in a digital library for use in a wider research domain, to formation of reference collections as the authoritative resource for evaluating future research, to federations with data collections from other disciplines to explore new research initiatives. Each change in the data life cycle corresponds to use of the collection by a broader community, and an associated evolution of management policies to meet the requirements of the expanded set of users. The context that entails the provenance, description, authenticity, integrity, and use of the data must encompass a broader range of knowledge to ensure that the larger community of users will be able to correctly apply and interpret the data. Standards to define semantics, data formats, and analysis tools require a consensus by each new community. Long-term data management requires evolution of management policies to address requirements of an expanding user community.

There is a strong correlation between data life cycle, broadening support for collection use by multiple communities, evolution of data management policies, and sustainability. Figure 1 lists these correlations, which drive the requirements for national scale infrastructure. The horizontal rows represent the stages of the data life cycle. The vertical columns denote the research units that generate the data collections, the motivating research goals, the storage resources used, and types of management policies that may control properties of the shared collection. Building generic infrastructure requires fundamental computer science research into the social network principles that underlie formation of research collections, their associated management policies, assessment criteria, and organizing principles [58].



Correlation between Social Aspects of Data Life Cycle, Sustainability and Management

Figure 1. Data Life Cycle Relationship to Social Networks

The DFC addresses these research challenges through the creation of six communities of practice: 1) Science and Engineering; 2) Facilities & Operations Center; 3) Data Cyberinfrastructure Technology and Research; 4) Policy and Standards; 5) Institutions and Sustainability; and 6) Outreach and Education. These communities of practice are social networks that provide coordination points for seeking input from external groups, for promoting the findings of the DFC, and for extending collaborations to new communities. The communities of practice will lead the formation of a social consensus on the policies and procedures for managing the data life cycle. The DFC will build social networks that integrate the findings of each community of practice, resolve these findings into a consistent set of data grid policies for managing the data life cycle, and demonstrate long-term sustainability through federation across institutional support commitments and storage facility providers. The DFC approach is to build on federation of policy-based data management systems that enable re-use of data collections by multiple communities, ensuring that future research can build upon today’s research results [59].

Science and Engineering Community of Practice: William Regli will lead this activity. Based on prior experience with applications of data grid technology, the DFC recognizes that the set of operations that need to be performed for the management of data are quite similar, but the specific data management policies and data manipulation processes are domain specific. In the iRODS data grid, the policies are translated into iRODS rules that chain together micro-services (micro-services are procedures, scripts or web-services that perform a well-defined operation). More than 170 micro-services are provided in iRODS for generic data management operations. One goal of DFC is to build standard policy sets (which may require some domain-specific micro-services to be coded/scripted or integrated) that can be modified for use within each of the science and engineering domains by collaborating institutions. This will provide six sets of policies and procedures that can be used to manage the data life cycle. These in turn will be analyzed to define generic infrastructure that is common across both science and engineering disciplines. The DFC will build reusable infrastructure for which the effort required to support a new community can be minimized. This working group also tracks the extensions that are needed to the generic DFC technology to support each of the domains. Each science and engineering domain has identified a liaison for interaction with the DFC.

The federation of multiple disciplines with a common infrastructure will enable cross-disciplinary research. While each DFC national consortium already involves multiple sub-disciplines, additional interactions will occur such as the proposed RENCI integration of hydrological data with oceanographic data to track climate change impacts [60]. Statistical models developed in Odum will be applied to the DFC itself to track formation of management policy consensus. iPlant researchers can use the hydrology data to develop new models for agricultural sustainability in global climate change; engineering models and 3-D visualization tools and CAD/CAM applications can be used to design or dissect complex sensor systems and ROVs. Temporal learning researchers may study how oceanographic researchers develop models and tools for time-series analysis, leading to new models of visualization tools to help rapid learning of temporal concepts. An exciting aspect of the DFC is that federation concepts can be applied to research collaborations, data sharing, and long-term preservation.

Facilities and Operations Center: Alan Blatecky and Sheau-Yen Chen will lead this activity. Given the wide diversity of purposes driving the development of data cyberinfrastructure and the diversity of existing resources that need to be linked to support the research initiatives, the DFC will rely upon data grid federation technology [61]. The DFC approach is feasible because data grids have been successfully applied for the last ten years at all scales from personal collections to internationally shared data collections, on data collections that range in size from a thousand files and a few gigabytes of data to hundreds of millions of files and petabytes of data [62].

The DFC will build a sustainable data management infrastructure by federating existing storage systems provided by institutional repositories, regional data grids, national data grids, and international data grids. The DFC will serve as federation hub for data grids established for separate research initiatives. The DFC will federate the Carolina Digital Repository (reference collections used in research and education at UNC), ACCRE [63], the RENCI data grid (federating storage systems at engagement centers across North Carolina), the TUCASI data grid (federating storage resources between RENCI, UNC, Duke, and NCSU), state repositories (the Distributed Custodial Archives Preservation Environment) [64], the NSF Teragrid, NSF observatories and national scale research projects, federal agencies (National Climatic Data Center), and international data grids with similar science and engineering projects (UK e-Science data grid, SHAMAN preservation project [65], ARCS collaboration service). Data management policies will be negotiated with each institution for the acceptable use and storage of data. This requires the installation of DFC data infrastructure at the collaborating storage providers. The DFC will demonstrate federation policies that manage collections that may be distributed across local institutions, the Teragrid, and federal agency repositories and emerging architectures such as Amazon cloud computing, S3 storage system, and Google's MapReduce computing framework.

An operations center managed by RENCI will track the status of the federation, respond to queries on resource availability, and collaborate with each institution on the management of the infrastructure. Sheau-Yen Chen has nine years of experience in the management of data grids, and will lead the installation, management and operation of the iRODS technology. Replication capabilities of the iRODS data grid will be used to ensure copies exist at remote locations to minimize risk of data loss. The remote execution capabilities of iRODS will be used for remote data filtering and subsetting to minimize the amount of data sent over networks. The set of policies that govern use and storage of data at each institution will be organized into generic usage agreements from which other institutions can derive their local policy. The DFC will grow the sustainable data management infrastructure by federating with storage systems from other institutions and research projects and negotiating joint usage agreements.

Alan Blatecky will lead interactions with other grid communities, seeking opportunities to federate the DFC technology with the Teragrid, the Earth System Grid [66], the Open Science Grid [67], EGEE [68], and workflow systems such as Kepler [69].

Data Cyberinfrastructure Technology and Research Community of Practice: Arcot Rajasekar leads this social network, integrating the software technologies needed to support the science and engineering domains and provide technological sustainability. The architecture of DFC is a federation of multiple cyberinfrastructure nodes. Each CI node, or *DFC platform*, consists of a set of distributed services coordinated by a data grid system. DFC Platforms federate to form the DFC Cyber-Infrastructure for the DataNet community as a whole. The DFC federation allows communities to share data and services. The DFC Federation architecture is shown in Figure 2 and components of the DFC Platform in Figure 3. The DFC Platform is a service-oriented architecture providing technological sustainability in a scalable and extensible solution. At the heart of the DFC platform is the integrated Rule Oriented Data System [70] developed by the Data Intensive Cyber Environments Center at the University of North Carolina and the University of California, San Diego. iRODS federates distributed and heterogeneous data resources into a single logical file system (called the collection hierarchy) and provides a modular interface to integrate new client-side applications as well as server-side data and compute resources. iRODS also acts as a third-party mediator providing authentication, authorization and auditing, optimized data movement protocols and rich support for metadata at multiple levels of data collections. iRODS also has a built-in distributed rule-engine. Administrators and collection owners can encode policies as rules for managing their data collections. These policies can be applied to realize operational functions such as:

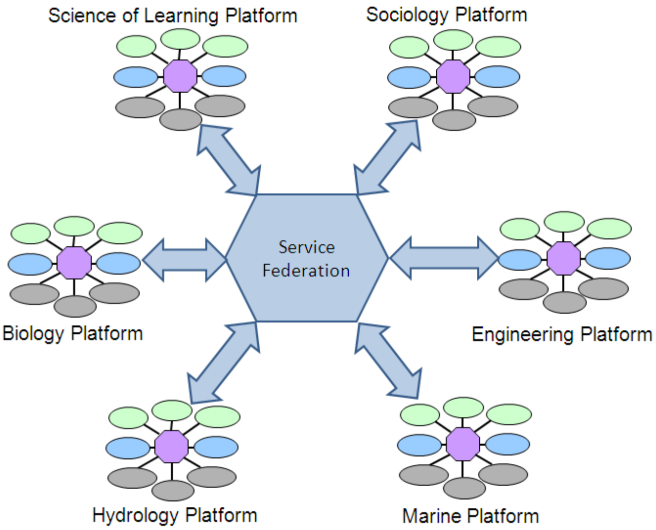


Figure 2. DFC Federation Architecture

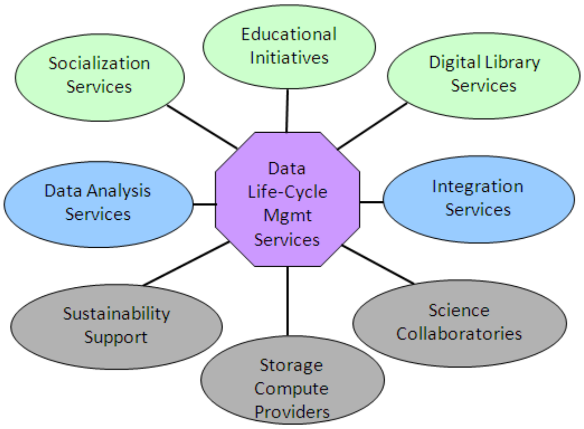


Figure 3. DFC Platform

- *data accession workflows* (operations such as integrity and format/style checks, metadata extraction, pre-processing, relationship association with other data, replication, format translation that need to be performed on ingestion of a new data object),
- *archival processes* (long-term management and assessment policies [71] such as data migration, obsolescence control, provenance checking, periodic corruption checking, disposition enforcement, creation of vendor-independent Archival Information Packages),
- *publication and dissemination processes* (associating metadata, indexing for search and browse, creating multi-formats for ease of access, distributing and replicating for load balancing, access control checking, auditing, notification),
- *analysis, synthesis and access processing* (operations such as discipline-specific analysis, integration of multiple data sets, capture of derived products with their provenance).

The rule engine in iRODS provides a way to customize the iRODS data grid to meet the demands of each discipline and also encode trust policies for sharing data across disciplines.

Positive evaluations of the iRODS by independent agencies have shown the applicability of iRODS in a wide variety of situations [72]. The DFC technology has been chosen by JPL for managing academic access to Planetary Data System products, NASA NCCS for archiving and managing the portal to a cache of simulation output, LSST for managing up to 150 petabytes of data distributed between Chile and the US, ARCS for building shared collections between Australian academic institutions, and the French National Library for integration with Fedora as a digital library. In addition, iRODS is used in the NARA Transcontinental Persistent Archive Prototype (TPAP) [73], the Carolina Digital Repository (CDR) institutional repository at the University of North Carolina at Chapel Hill (UNC-CH), and the RENCI regional data grid linking sites across North Carolina. The DFC will implement a long-term data preservation environment that manages collections replicated across institutional repositories, regional data grids, national research projects, federal repositories, and international collaborations. Letters of collaboration demonstrate a wide range of interest in the DFC approach. These external applications of iRODS technology will bring additional data into the DFC.

The social networking paradigm emerging from non-traditional usage of the web is playing an increasing role in scientific collaborations. The DFC proposes to explore such novel usage through integration with DFC cyberinfrastructure. Enabled through Web 2.0 [74], multiple social networking systems have emerged, including community building (e.g. LinkedIn, Facebook), collaboration wikis (e.g. Wikiomics, openwetware), content sharing (e.g. YouTube, Flickr), collaborative tagging (e.g. Delicious, Digg), blogging (e.g. Twitter), scalable data sharing/movement services (e.g. BitTorrent), internet services (RSS, Wiki, podcasts, mashups, etc.) and other innovative services in data sharing (eBay, matchmaking, personal spidering agents). Most are seldom used for scientific collaboration and when used are applied in particular projects/groups or narrow sub-disciplines. Harnessing social Web networking services as standardized consortium-guided mechanisms in the DFC will provide a means for innovative access and expand the reach into newer communities and usage models. This extended outreach will provide impetus to federation-based economic sustainability of the collections in DFC.

The DFC will provide an important extension of these tools through *long-term preservation of the contents exchanged through these social networking sites*. The DFC will manage scientific social networking. Capturing and preserving the “data” exchanged through these tools will allow provenance tracking by future scientists. The DFC will enable automated extraction of provenance information, saving the context in which social consensus is created. Such novel usage models will increase federation sustainability, transforming traditional passive data repositories into active and interactive “live” community libraries. An example is the integration of preservation workflow services with TV broadcast production pipelines performed as part of an NSF-funded project for DIGGARCH [75]. In this project, a preservation workflow was integrated with the video production of the TV series “Conversations with History” from UCTV. All data and provenance information were captured including multiple formats of the video, transcripts, emails and blogs exchanged by the show host, and web pages used by the site promoting the TV series. Similar mechanisms can capture human interactions performed through socialization services within scientific exchanges, and derive provenance information.

iRODS provides interoperability mechanisms for integration with other data management systems. A current demonstration of federating multiple data management solutions involves collaborations funded outside the DFC, including integration of the merged Fedora digital library middleware and DSpace digital library [76], the Dataverse Network [51], the LStore logistical networking file system [77], and LOCKSS preservation environment [78] with the Storage Resource Broker [79] and iRODS data grids. Associated research issues include development of interfaces for federation of structured information resources, federation policies across shared data collections, and integration with workflow systems for data intensive processing of collections.

DataNet Policy Research: CISE research relevant to preservation policies will be conducted to enable federating a diverse set of data collections from multiple domains. Since users from the DFC partnerships are domain scientists, tools that can ease definition and description of rules and policies will be of great help. The DFC proposes fundamental research on rule generation interfaces and properties of rule-based systems. The main work in this area will be led by Prof. Chitta Baral from Arizona State University who has developed tools for bio-medical databases and conducted research on logic programming and answer set systems. He will work with Dr. Jorge Lobo, who has been leading work on network policy definitions at IBM Watson and Prof. Arcot Rajasekar who developed the rule engine for iRODS.

Policy Framework and Rule Base Optimization Research: There is a clear relationship between the expressiveness of a policy language and the ability to analyze its properties and impact on system behavior. Without sufficient expressiveness, a policy language may not be able to regulate complex system behavior, apply across heterogeneous components, or apply to systems involving frequent changes. Without analysis, much of the benefit of using policy-based techniques and declarative policy languages may be lost. Arguably, the lack of effective analysis tools accounts in part for the lack of wider adoption of policy-based techniques. The first part of the research is based on specification and verification of the goals and intentions of specific rule bases using temporal logic outside security. Lobo introduced in [80] a policy framework for the description and analysis of history-based security policies. In the rule-based system used in iRODS, many of the micro-services performed by the rule engine are situated in time. Indeed, if the analyses of data are time-consuming, and parallel processing is the norm in performing these services, it becomes important that the purpose of the rules be formally specified. The run-time execution of these rules needs to satisfy the purpose. DFC will build on earlier work [81-84] on expressing goals of agents using temporal logic to express the purpose of a specific rule base. The specification language will need to have temporal constructs as well as aggregation constructs. The temporal constructs such as “eventually,” “always in the future,” “next time,” and “until” will be needed to express the evolution of the data as the rules are executed. The aggregation constructs will be needed to express consistency aspects of the database. Extending specification of iRODS rules in temporal logic will be a research topic addressed as part of the proposed work.

Extending the previous work by Baral and Lobo [85-88], the DFC also proposes to develop tools that will help in the verification of the correctness of a rules base with respect to a formally specified purpose. The tools will use the language of answer set programming [89-91] which has constructs that allow enumeration of work trajectories and can express constraints built up from temporal and aggregation operators. The possible initial states of the database and the various possible evolutions can be expressed as dictated by the rule base and then verified that the rule base satisfies the formally specified purpose by encoding the purpose as constraints and checking that the resulting answer set program is consistent.

The second part of this research will be in reasoning and evaluation of the correctness and consistency of rule-based systems. The language of answer set programming (ASP can be used for meta- reasoning on other rule-based systems. Baral plans to use ASP to check the correctness of data processing workflow and to check the consistency of rule bases. In this research DFC will exploit the `knowledge representation’ ability of ASP and its ability to emulate `description logic’ to reason about ontologies as well as knowledge bases. A third area of research that DFC proposes in rule-based systems is in automatic generation of rules. In certain domains and for certain goals it may be easier for a domain user, who is not an expert in writing rules, to specify what they want rather than write the rules that will achieve what they want. DFC plans to provide them with interfaces with which they can specify what they want. Tools will be developed that will automatically generate the desired rules. In this DFC will build on earlier work on automatic construction of rule-based policies with respect to given agent goals [92].

The last research topic relevant to the DFC is in the area of policy negotiation tools and protocols. If two institutions have different policies and they want to work together how do they reconcile policies? Some of the work in this area [93-95] is done in the framework of negotiation. DFC proposes to see how this can be applied for combining data management policies. Baral has worked extensively in the area of combining databases [96-98] and will try to apply similar research in policy negotiation. Lobo also has extensive experience in policy combination algorithms for privacy and security policies [99-102].

Policy and Standards Community of Practice: Helen Tibbo will lead this activity. The digital library and information science communities are defining policies and procedures needed for access and sustainability, as well as standards for descriptive metadata [103]. While the DFC expects a core set of procedures and management policies to apply across all data life cycle stages for controlling administrative tasks, assessment criteria, and display behaviors, each stage has unique policies controlling the collection. The DFC will create a reference implementation for the core set for each stage of the data life cycle that can be used by other communities as a starter kit for building their unique management system. The DFC will collaborate with the Open Grid Forum [104], the new DuraSpace ™ Organization formed from DSpace Foundation and Fedora Commons, the DICE Foundation [105], and Sun [106] on the development of reference implementations for data grids, digital libraries, and preservation environments. Explicit analysis of the policies and procedures required by each DFC science and engineering domain will be performed to identify the generic reference implementations. Four graduate students in the School of Information and Library Science will serve as liaisons with the science and engineering domains to help assemble the policies. These policies will be compared with those from the standards community, such as the ISO Mission Operations Information Management System repository assessment criteria [107]. DFC members will participate on digital library conference program committees (JCDL, ECDL), on preservation conferences (Helen Tibbo is the new president of the Society of America Archivists), on the ISO MOIMS-rac standards committee, on the IEEE Mass Storage Symposium, the yearly Supercomputer Conference, and present papers on the results obtained by application of the DFC technology.

Institutions and Sustainability Community of Practice: Richard Marciano will lead this activity. Sustainability comprises mechanisms that ensure a data collection will remain usable and understandable over time. The DFC addresses four aspects: Economic sustainability; Technological sustainability; Policy sustainability; and Access sustainability. The DFC relies on federation across multiple institutions to ensure sustainability, using multiple sets of policies that govern the data life cycle to support re-use of data collections by broader communities, and to develop required access mechanisms. Collections used by a single community will remain viable as long as that community can demonstrate an economic incentive for keeping the data. If multiple communities can validate an economic incentive for maintaining a collection, the risk of losing support for a data collection is minimized.

Technology Sustainability: For long-term viability, the underlying generic infrastructure needs to federate both existing and future data management systems. The technology sustainability in the DFC design is embodied in the two main concepts of the underlying open source iRODS data grid technology and its extensions: 1) infrastructure independence – iRODS middleware provides a uniform and modular interface to integrate new applications and data resources and acts as a third-party mediator providing data grid functionality for authentication, authorization and auditing, optimized data movement protocols and rich support for metadata for collection hierarchies. 2) policy-driven services – the iRODS architecture controls procedures through rules which enact policies. Each community, collection or data curation system can have its own policies for data life-cycle management and expose these policies as rules. This provides not only transparent policy definition, but also provides a means to compare policies and procedures. The DFC ensures the continued maintenance of the technology through multiple governing institutions: A DFC Cyber Infrastructure Foundation will guide technology sustainability and will

re

work with standards organizations, other DataNet partnerships and independent software development teams to define and tune the DFC's CI software stack and hardware requirements to adapt to the growing needs of the DFC collaborations. A DFC Data Collection Foundation will develop and apply business and fiscal models for managing and sustaining the scientific and engineering data collections and the underlying cyber infrastructure. Both institutions will promote open source development of the core infrastructure.

Additional organizations will be allied with these foundations. The new Data Intensive Cyber Environments Center at UNC [108] promotes development and application of cyberinfrastructure within the UNC system. The Data Intensive Cyberenvironments Foundation promotes open source development of the iRODS software. The Sun PASIG community is exploring integration of DFC technology with Fedora and DSpace for open source distribution integrated with the Sun open source operating system. By using multiple institutions that promote international use of the technology, the DFC can ensure sufficiently wide-spread use that the software will remain viable.

Economic Sustainability: The DFC believes economic models for long-term sustainability can be based on joint governance by multiple institutions, each of which may have a different driving usage model or purpose for a data collection. The DFC will develop generic infrastructure that allows multiple governance policies to be applied to copies of a collection. The original purpose under which the collection was formed defined the data format, semantic labels for identifying physical quantities, descriptive metadata, coordinate system, geometry, resolution, accuracy, calibration, and access services for data manipulation. The ability to transform from the original data context to a new data context for a new purpose can be implemented as a chain of iRODS procedures controlled by each institution's governance policies. If necessary, the procedures can create transformed data products that are also stored within the data grid. This makes it feasible to consider re-purposing of data as a long-term sustainability mechanism. Policies specified by an institutional community of practice can be implemented that manage the data context changes.

The DFC will examine the repurposing of collections as reference material. In collaboration with institutional repositories such as the UNC Carolina Digital Repository, the DFC will support the registration of collections from their original project repositories into institutional repositories. This requires the identification and enforcement of new management policies for the standards under which the reference collection is governed. The institution can then encourage use of the reference collection to support local research, to support education classes, and to enable assembly of collections that encompass a wider set of material across the entire discipline. The iRODS data grid explicitly supports control by multiple governing institutions through separate management policies, usage accounting information, audit trails, and application of separate procedures for transforming the data context for users from each institution.

The DFC will investigate models for recovering cost of storage, either by pre-paying the expected lifetime of the material, paying for use, or paying for subscription. DFC will also investigate federation models in which storage space is exchanged as a risk mitigation mechanism against data loss, or local collections are created to support research and education initiatives.

Policy Sustainability: The policies controlling each collection must evolve to support the purpose of the new constituency. Generating a new management policy is a difficult task. By building upon the reference policy implementations derived from the six NSF science and engineering domains, the DFC expects to simplify the process. The DFC will track how the policies governing an individual collection change over time, and use this information to understand how policies for other projects should also evolve.

Access Sustainability: A data collection that does not provide the appropriate analysis and display tools will not be used. A significant amount of effort is typically required to enable re-use of data. If the tools that are required to parse, manipulate, and display data are maintained with the collection, collection re-use can be effectively promoted. The DFC manages Access Sustainability by decoupling the access and storage protocols from the tools that do the data manipulation. The iRODS data grid provides the mechanisms to encapsulate data parsing and manipulation routines as micro-services that interact with a standard set of remote operations. The micro-services respond to standard actions invoked by a wide variety of clients. This means that the tools for parsing data can be migrated to new operating systems, and accessed by new clients without requiring modification to handle the new protocols. The DFC will work with each science and engineering domain to encapsulate the required tools as micro-services.

Outreach and Education Community of Practice: Marilyn Lombardi will lead this activity. Education is a social process that is facilitated by interactions between students and between students and teachers. The open-source Croquet-based Cobalt 3D browser and authoring toolkit [109] allows for the easy creation and deployment of avatar-mediated 3D visual development environments (“virtual worlds”). Within these deeply social and collaborative contexts, researchers and educators will be able to access, aggregate, and arrange the multi-media data resources (texts, images, simulations, animations, 3D models, etc.) available to them by virtue of the proposed data infrastructure. Integration of the Cobalt 3D visual development environment into the DFC will enhance the accessibility, the sustainability, and the educational value of the national data infrastructure [110]. The DFC will also explore integration of social networking technology to improve education by enabling interactions between students, and between students and teachers. The DFC will use SLC mechanisms to evaluate the success of the approach.

The DFC will federate educational initiatives and examine repurposing of data in educational tools, enabling student participation in research initiatives. One example is creation of a tool for a classification task, and then training of students in the tool’s use. The tool applies required context transformations for the classification to be meaningful. Associated management policies govern allowed manipulations, publication requirements, and permitted access. A consensus on management policies is required between the project providing the collection, the educator managing the course, and the students, which must be enforced on data access. The DFC will implement management policies in collaboration with the science and engineering disciplines.

The CIBER-U project, Cyber-Infrastructure-Based Engineering Repositories for Undergraduates, aims to realize a high-impact from application of cyberinfrastructure in engineering undergraduate curricula. Specifically, CIBER-U combines activities at three universities and undergraduate summer research experiences to enable product dissection activities. One key demonstration objective within the DFC is to create a ``Source Forge" focused on the domain of building shared engineering models. This will include tools for design collaboration and information sharing, as well as distance learning and educational materials. This repository will be made available over the Internet and provided for use by educators and researchers around the country and the world. Further, the DFC will work with partners in government and industry to rapidly transition new concepts on knowledge representation, standards, and software interoperability into ongoing efforts at the National Institute of Standards and Technology (NIST), United States Department of Energy (DoE), ISO and the W3C.

The DFC will create a national portal for DataNet-centric education and vocational advocacy. The DFC will contribute materials to sites such as http://www.tryengineering.com, http://www.engineergirl.org, and other websites aimed at outreach to future scientists and engineers. While distance-learning technologies do not replace the classroom experience, they do provide alternative ways in which to reach non-traditional students. The DFC will attempt to identify on-line delivery mechanisms for these modules and, more importantly, create a revenue mechanism in which all of the universities can participate. Further, the DFC will identify ways in which these e-learning revenues can be re-invested into Cyber-Infrastructure development.

The DFC will publish results in leading conferences across the whole spectrum of science and engineering disciplines, as well as in symposia for the key computing, information technology, database and rule areas central to the technical themes of this proposal. The DFC will also disseminate results to researchers and practitioners through government (NIST, NARA, etc) and industrial partners (Sun, IBM, etc). The DFC will work closely with them to ensure timely transition of developed technology and training materials. The tools developed in this project can have a significant influence on emerging international standards. This team is uniquely positioned to bring advances desperately needed by the engineering design community into the mainstream of W3C efforts, as well as influence ISO standards efforts.

The DFC will support summer schools on the DFC technology. Tutorials on application of the DFC infrastructure will be given at conferences and interested institutions each year. Education classes on the technology will be taught at UNC. An example is a course taught this year within the School of Information and Library Science at UNC on policy-based data management.

Timeline and Milestones: The DFC is a service-oriented system providing multiple federation layers for effective data sharing. DFC will deliver a standard architecture, *the DFC Platform*, usable for extensible and scalable data sharing and preservation. The primary DFC software activity will to integrate and develop policies and services for scientific domains. The DFC will apply the policies in a production data grid that is used to sustain reference collections through federation of storage resources across multiple institutions. The DFC will function as a hub that imposes preservation management policies on collections registered from participating institutions. The DFC platform will provide two core functionalities: 1) Policy-driven middleware that will provide extensible data life-cycle management and sustainability services; and 2) A suite of social network services that increase the use of scientific data. The DFC plans to support requirements from other research communities such as the astronomy community (Hubble Legacy Archive [111], National Virtual Observatory [112], Large Synoptic Survey Telescope [113]), and the seismic community (Southern California Earthquake Center [114], USArray [115]). The DFC goal is to build generic infrastructure that supports all disciplines.

The DFC will focus on nine sets of activities that constitute yearly deliverables:

Year 1: Implementation of a national preservation environment. A sustainable data management infrastructure will be based on federation of institutional repositories, regional data grids, national grid infrastructure, and federal repositories. An operations group will be established to respond to questions and provide data grid support.

Year 1: Analysis of application requirements and installation of policy-driven middleware that will provide extensible data life-cycle management and sustainability services. The DFC will build additional domain-specific services each year to parse and manipulate the data formats of the science and engineering disciplines. Integration of multiple social networking tools will be started in the first year and will continue throughout the project.

Year 2: Development of reference implementations for management policies, context manipulation procedures, and data grid frameworks. In collaboration with the Sun Microsystems Preservation and Archiving Special Interest Group, a reference implementation for preservation environments is planned that will be distributed as open source software. Reference implementations will also be designed for data sharing, digital libraries, data processing systems and data visualization systems. To support data processing pipelines, services that integrate data management with workflow systems such as Kepler and distributed visualization systems such as Vista [116] will be developed.

Year 2: Development of interoperability mechanisms between data management solutions to enable participation by all projects in national infrastructure. This includes support for testbeds for evaluation of new technologies from DataNet Partners and from international collaborations. Services for digital libraries will be developed that support large-scale publication, indexing, and curation that are relevant to the science and engineering disciplines.

Year 3: Development of generic services needed for the manipulation of structured information. This requires development of the procedures that allow re-use of data through mapping to a new context. An existing collaboration with the European Union SHAMAN project is integrating the Cheshire on-line catalog technology [117] with the iRODS data grid to enable the manipulation of descriptive metadata. The SHAMAN project is also integrating the

Multivalent parsing tools [118] for composing multi-level displays of text and data. DFC will build on extensions to these systems that are planned by SHAMAN for manipulating scientific data that are based on the Data Format Description Language technology developed at NCSA.

Year 3: Development of a suite of social network services that increase the use of scientific data in education. The integration with Croquet and SLC technologies will be supported by the DFC. Management polices will be developed for deposition, acquisition, access control, integrity, trustworthiness and privacy (including constraints such as HIPAA), replication, transformation, retention, curation, discovery, access, disposition, data interoperability, and standards and institutional policy enforcement. These policies govern the usage model for the collections, define the collection assessment criteria, and define the expectations of the user community.

Year 4: Automation of assessment criteria validation. By defining appropriate rules that are periodically executed, the DFC will validate assertions about each collection. Evaluations will include usage, performance, reliability, and trustworthiness. The DFC will collaborate with the ISO community on implementation of the ISO MOIMS-rac repository assessment criteria, and on specification of representation information that defines the data context.

Year 5: Consensus building within communities of practice for standard data management policies, standard procedures for manipulating data, and models for economic sustainability. The DFC will demonstrate federated environments in which data collections are repurposed for use by multiple institutions, and governed by separate management policies. Models for economic and technological sustainability developed and deployed through the project will be assessed for their utility. The DFC will validate social networking tools for effectiveness in broadening access.

Year 5: Development of sustainability models for economic, technological, policy, and access support. The DFC will exploit institutional federation models, interoperability mechanisms between cyberinfrastructures, and data access standards to ensure the ability to migrate collections between resources.

In each case, the development of the original technology is already being funded. iRODS development is funded through the NSF Software Development for Cyberinfrastructure program, and by NARA for preservation research in the Transcontinental Persistent Archive Prototype [47]. The DFC will collaborate with the original software developers to ensure incorporation of new technology in reference implementations. DFC liaisons from each community of practice and application domain will facilitate use of the integrated software framework in research and education. The creation of reference implementations for management policies at each phase of the data life cycle and implementation of a sustainable preservation environment based on federation of institutions will be unique contributions of the DFC.

<Section on Organizational Structure and Personnel deleted for this version.>

Broader Impact: The DataNet federation will create a social context for creating extended collaborations, support the integration of infrastructure that enables data sharing between communities, and support the federation of storage facilities. The DFC will implement a preservation facility that organizes distributed data into shared collections for a project, federates shared collections across projects, federates shared collections across disciplines, enables transition of data through all phases of the data life cycle, promotes educational use of research data through social networking technology, enables publication of scientific data, enables preservation of scientific data, enables analysis across collections, and enables student interactions with “Live” data. DFC will enable the use of national scale collections by collaborating diversity partners and minority serving institutions. The partnership is truly national in scale, spanning partners on both coasts and six disciplines, the entire data life cycle, technologies from digital libraries to supercomputers, and six communities of practice.

The DFC will demonstrate strong incentives for additional projects to participate through support for: Interoperability with other data management systems to enable data sharing; Life cycle management, enabling publication, preservation, and re-use of data; Education links to enable classroom use of active research collections and development of new collections, through social networking and virtual world technology; Broadening of participation in data access through support for management policies tuned to each community of practice; Development of novel tools and services in support of science and engineering education; and Transparent access to facilities linked by DFC, based on local resource allocation.

The DFC will operate a production data preservation facility through use of data grid federation technology. Federation-based sustainability mechanisms will be prototyped that promote long-term preservation through use of multiple institutions, by repurposing data collections for new uses. The software developed as part of this project will be distributed freely as open source.