DataNet Full Proposal: DataSpace Project Summary

PI: Stuart Madnick (MIT); Co-PI's: Hal Abelson (MIT), Ed DeLong (MIT), John Gabrieli (MIT), Jerrold Grochow (MIT), MacKenzie Smith (MIT); Senior Personnel: Timothy Berners-Lee (MIT), John Erickson (HP Labs), Alon Halevy (Google Labs), Geneva Henry (Rice University), Mei Hsu (HP Labs), David Karger (MIT), Michele Kimpton (DSpace Foundation), Thomas Malone (MIT), Joe Pato (HP Labs), Terry Reese (OSU), Michael Siegel (MIT), Stephen Todd (EMC), Tyler Walters (Georgia Tech), Danny Weitzner (MIT), John Wilbanks (Science Commons), Wei Lee Woon (MIST, Abu Dhabi).

Summary and Vision: Web technology brought tremendous efficiency gains for commerce, yet the world of scientific research has failed to fully leverage all its capabilities. As a result scientists duplicate research and miss opportunities for discovery, collaboration and translation of research into public goods. The DataSpace Project will bring these gains to science by providing a dramatically new approach to data management and long-term curation that accommodates multiple, heterogeneous data from a variety of distributed locations, and supports research across diverse disciplines and modalities, enabling investigators to easily access and aggregate data of known quality and provenance. It will build on proven technology and business models while bringing to bear the best research capabilities of MIT and nine partner organizations. To encourage sustainability and collaboration, organizations producing research data will be able to take responsibility for long-term stewardship of their data as part of a global network with only modest investment and expertise.

Intellectual Merit: the DataSpace Project implements this model by

- Working closely with, and actively supporting and being guided by, scientists and data in multiple scientific research domains to insure *relevance*, and *interoperable, interdisciplinary* solutions to data access and management. Initial domains are life science (e.g. neuroscience, including informatics and imaging data) and environmental science (e.g. oceanography, metagenomics). Additional domains (e.g. high energy physics, plant biology) will be included later to verify interdisciplinarity and interoperability.
- Building on our experience with existing *distributed infrastructure* for digital archiving and long-term preservation, but extending it to support the scale and complexity of research data.
- Developing an effective *sustainability plan* based on our successful experience with the DSpace open source digital archive platform and its current user base of about 500 research-generating organizations world-wide and the extensive business model experience at the MIT Sloan School of Management.
- Leveraging a large number of MIT and external collaborators who have *extensive and proven expertise* in key areas: many fields of scientific research, data integration, quality and interoperability, representation and visualization, database and storage technology, federated system design and policy management, long-term preservation, public outreach and education, and business planning and management for organizations.

Broader Impact: the DataSpace Project will impact science and society in multiple ways, including:

- Scientists have noted that *important advances* are limited by the current diversity of data formats, management and access policies, tools for visualization and use, preservation strategies, inability to accomplish multi-disciplinary multiple platform integration, and inability to easily extract needed subsets from enormous data collections. The DataSpace infrastructure will *support that diversity* and *empower researchers* to effectively utilize all relevant data resources.
- To ensure *continual availability* of scientific data to all communities, DataSpace will be highly distributed to *manage the risk of failure* either in technology (by supporting extensibility) or organization (for example, control by a small number of entities, or an inefficient monopoly caused by large-scale centralization). Technology is still evolving rapidly and innovation is occurring in many sectors and countries, so the cyberinfrastructure must be designed to leverage that situation.
- We strongly support the vision of *Open Access* to all information, while recognizing that there are sometimes necessary constraints that must be honored. Making universal, meaningful access is a high priority in the expectation that it will lead to *unexpected advances* in science and engineering.
- The DataSpace Project will produce *innovation* in several areas fundamental to data archiving infrastructure: data interoperability and visualization frameworks, large-scale database and storage infrastructure, legal and policy frameworks, broad public access, and new techniques to lower the cost of data to benefit scientific endeavors and society in general.

DataNet Full Proposal – DataSpace Project Description

I. Introduction

Many aspects of society have been transformed by the Web – the large-scale, sustainable cyberinfrastructure that includes the Internet, the World Wide Web and related technologies, and the rich ecosystem of services and social practices that have grown in that environment (Amazon, Google, eBay, Yahoo, Facebook, and innumerable other Web-based services). This cyberinfrastructure has transformed business and many other industries over the past decade not only because of its highly distributed, flexible architectures and protocols, but also because a legal and policy framework and a new set of social norms have emerged. So far, science has not seen the same transformative benefit from this infrastructure.

The proposed DataSpace Project has the potential to transform science and ultimately other data-intensive industries by creating an easy-to-use technology platform for curating and publishing data in standard formats (i.e. a 'Data Web Server') that enable a new generation of data services. As with the current Web, every data-producing organization, from individual research labs to international research institutes, will be able to publish their own data and develop locally-sustainable strategies for data curation. Starting with a group of premiere research universities and their libraries, DataSpace will create practical models for localized support services to archive research data more efficiently and effectively (i.e. addressing the "last mile" problem of support to researchers), and we will develop templates for long-term sustainability of data curation by research institutions.

Different domains (and sub-domains) of science, engineering, social science and humanities research require different data formats, management and access policies, tools for visualization and use, and preservation strategies. Scientists note that important advances are constrained by the inability to find and access relevant data from prior research, to accomplish multi-disciplinary integration, and to extract subsets from enormous data collections. The DataSpace infrastructure will be designed to support that data diversity and empower any researcher to effectively utilize all data resources across multiple domains. In addition, the DataSpace Project will produce innovations in several areas fundamental to data archiving to benefit science and society in general.

The goals of DataSpace are challenging but fortunately an exceptional and diverse world-wide team of experts have been assembled to meet the challenge. The core team consists of the Massachusetts Institute of Technology (MIT) and its nine research partners in this proposal: the DSpace Foundation, EMC, Georgia Tech, Google, HP Labs, the Science Commons, the Masdar Institute of Science and Technology (MIST), Oregon State University, and Rice University. These partners are already working on many aspects of data curation and interoperability issues central to the idea of the DataSpace cyberinfrastructure, such as:

- scientific research that makes extensive use of current cyberinfrastructure and exemplifies the opportunities and challenges faced by scientists and engineers in managing and leveraging data
- research into data quality, interoperability and representation standards
- development of the emerging Semantic Web
- designing platforms and services for data ingestion, curation, long-term preservation and reuse
- technology operations experience with network security, privacy and scalability
- development of policy and legal frameworks for data management (internal and cross-organization)
- innovation in data visualization, navigation and social networking technology
- a strong commitment to open access to research and teaching material
- research on and experience creating sustainable business models for distributed organizations

The DataSpace Project proposes to bring together these ongoing threads of work into a coherent program of research, development, operations and outreach. This will both define a *model data archive cyberinfrastructure* and serve as a *set of exemplars to implement the model* that scale well beyond what is currently achievable. The project will build on current expertise in data curation, interoperability and integration (e.g. the MIT Productivity from Information Technology research agenda and the DSpace open source digital archive project) to create a new infrastructure and a model for a distributed institutionally-based data archive program. The goal of the new infrastructure and associated service models are to be *locally sustainable, deployable by any organization that manages data, federated into a* *coherent whole, and scalable* over time to meet the needs of many types of research-generating organizations and scientific domains.

The resulting infrastructure (see Figure 1) should enable research organizations to *join a global data network of archives as easily as they can participate in the Web*. We will separate the concerns of physical data storage and low-level preservation, data curation and functional preservation, data access and visualization across different sectors, and belonging to the communities of practice that best serve each piece of the problem. The infrastructure will depend on a set of *practices, standards and protocols* that are *independent of particular hardware or software* systems, distribute and replicate the data, and support federated access to all the relevant communities of interest.

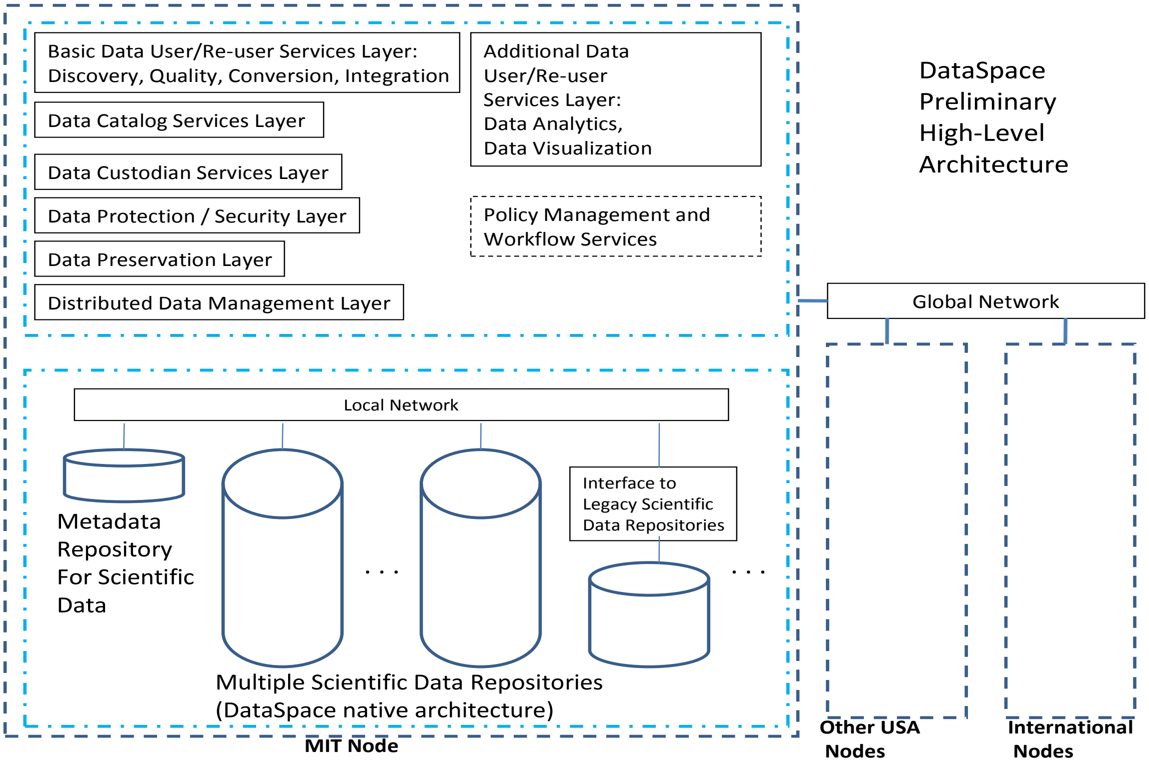

Basic Data User/Re-user Services Layer: Discovery, Quality, Conversion, Integration
Additional Data User/Re-user Services Layer: Data Analytics, Data Visualization
Data Catalog Services Layer
Data Custodian Services Layer
Data Protection / Security Layer
Policy Management and Workflow Services
Data Preservation Layer
Distributed Data Management Layer
DataSpace Preliminary High-Level Architecture
Global Network
Local Network
Interface to Legacy Scientific Data Repositories
Metadata Repository For Scientific Data
. . .
. . .
Multiple Scientific Data Repositories (DataSpace native architecture)
MIT Node
Other USA Nodes
International Nodes

Figure 1: DataSpace federated architecture

Building the science cyberinfrastructure for the twenty-first century will be an enormous undertaking requiring patience, determination, and rigor. The DataSpace Project has a team that can build a complete operational system. Once the model exists we have proposed an organization to extend it for decades to come, and because it is open and enables many forms of business, we can enable the evolution of a new ecosystem and incentive system for a cyberinfrastructure that will be transformative and sustained.

II. Scientific Research Challenges

While the DataSpace platform will be designed to support data curation and integration across *all* scientific and engineering domains, we recognize the central role of particular scientific research practices in designing data archiving and future usage strategies. To ensure that our work is driven by practice and will achieve its goal of improving scientific progress, we will work closely with senior research scientists in two scientific domains to inform our work initially: neuroscience and biological oceanography. Researchers in these two domains have identified data integration, interoperability, aggregation, and longevity as key problems, as well as the need to bring their data into modern technological infrastructure that will lower the cost of data processing and improve interdisciplinary data use. Beyond these two initial domains, we will quickly expand the project to additional domains (e.g. physics, materials science, ecology, social science and humanities) in later phases. The partner research institutions: Georgia Tech, MIST, MIT, OSU, and Rice University, have each identified additional scientists who are interested in the new service for their own domains. Again, the DataSpace infrastructure will support *any* type of data with associated metadata and tools for analysis, visualization, etc., but we will *begin* with scientists from the two selected domains who have significant experience at creating, collecting, managing, integrating, manipulating, mining, and visualizing data to inform architectural requirements and develop the collaboration and data sharing practices across the new organization.

The project will focus on defining standards for describing scientific data (i.e. “metadata”) consistent with the emerging Semantic Web. We will also investigate to what degree “raw” data formats in current use can be migrated into Semantic Web standards (e.g. RDF) or how they can be managed from DataSpace in their native environment. We do not assume that data must be converted into a new data format in order to become part of the DataSpace archive network – we will identify ways to include legacy data and as well as migration strategies that can implemented over many years (as we’ve seen with the migration of content into HTML and XML to be Web-enabled).

For data that can be brought into newer Web formats such as RDF, either by direct entry or by dynamically converting legacy relational databases into RDF via query conversion tools, we can immediately add value to the data by mapping ontologies onto it and adding new relationships to the data such as links to related research articles, or connections between data from the same physical locations. To accomplish we will leverage existing work such as the Object Reuse and Exchange (OAI-ORE) RDF ontology[1] for aggregating digital resources. These aggregations can be defined manually by expert data curators in the library working together with local scientists, or automatically via data mining techniques being developed by computer scientists. The relationships modeled in these aggregations can then be broadly federated (e.g. via the Concept Web Alliance[2]) to support improved search and reuse.

Finally, as part of our exploration of realistic data archive service models, data ingested into DataSpace at the five initial sites will be tagged by library staff or other depositors for improved discoverability in local and Web-based search interfaces (e.g. local data catalogs, union catalogs such as the one in development by the TIB in Germany[3], and Google).

II.1 Domain: Neuroscience

In the past twenty years the life sciences – biology, biochemistry, psychology, etc. – have become a dominant branch of science. The quantity of data produced in these fields has seen a large increase, generating enormous quantities of genomic, proteomic, phenotype, computer-generated image, and many other types of data to describe biological phenomena, and we have seen a simultaneous growth in importance of computational approaches to life science research that leverages this data. Researchers in this domain have identified the challenge of data sharing and integration as a high priority, and not only sharing of primary data, which in the case of unstructured data such as images poses a great challenge, but also sharing of numerical models and methods used to build the models.

Neuroscience – a large and growing field within the life sciences – is an ideal candidate for DataSpace because of its pressing need for data archives, aggregation across institutions, large-scale integration, and interoperability across a wide variety of data types and time scales. Data used in neuroscience include those common to other branches of biology (e.g. genomics, proteomics, biochemistry, cell biology) but also include many distinctive data types: digital images (e.g. structural and functional MRIs, diffusion tensor images, microscopy images, high-throughput images and videos), electrophysiological activity data, behavioral, and in some cases clinical data (e.g. subject phenotype information and data from physical samples taken over time). Among neuroscientists there is a long history of quantitative modeling and computational approaches, and a broad consensus about the need to integrate data across multiple levels in order to understand the brain, the relationship between the brain and behavior, how the brain is modified by experience, the connection between genotype and phenotype (including the interaction between genetic and environmental influences), and the biological basis of brain disorders such as autism, schizophrenia and depression.

Computational neuroscience is a subfield of neuroscience that develops models to integrate complex experimental data in order to understand brain function. Although researchers in this field need access to

[1] Open Archives Initiative Object Reuse and Exchange (OAI-ORE) defines a standard for the description and exchange of aggregations of Web resources encoded in RDF. http://www.openarchives.org/ore/

[2] The Concept Web Alliance is a new international organization creating Semantic Web standards and federated services for concept-related “metadata” about data. http://conceptweballiance.org/

[3] STD-DOI is a shared registry of scientific research data that assigns DOIs and ISO 690 2 metadata for easier discovery and access. http://www.icdp-online.org/contenido/std-doi/front_content.php

a wide variety of experimental data in order to constrain and test their computational models, this data is not typically shared publicly or made easily available. Traditionally, neuroscientists have taken the “single lab” approach to research. Many fundamental aspects of brain function, such as the questions of how brains can perceive and navigate so robustly, how sensation and action interact, or how brain function relies on concerted neural activity across scales, remain unsolved due in part to this lack of data sharing.

Neuroscience requires a great degree of interdisciplinary collaboration among psychologists, biologists, computer scientists, chemists, and others who do not share a common vocabulary or understanding of each other’s data semantics. The data can range into hundreds of terabytes. For example, MRI technology is improving quickly and capable of generating more than five terabytes of raw data annually per machine. Searching across data types, integrating data for comparative analysis (e.g. combining MRIs with related gene sequence data), or applying novel data mining techniques is laborious. The trend in this field is away from single experiments and small studies toward larger scale, integrative studies that yield more reliable results. The importance of inter-institutional sharing is particularly acute where large sample sizes are needed to produce significant effects. One approach is to centralize data production in a single organization (e.g. the Allen Institute) but this is not viable in most cases due to high cost, cultural obstacles, lack of access to researchers, etc). Data are produced by many different groups at different institutions and the challenge is to collect and then aggregate data in useful ways.

Individual researchers in small to medium-sized neuroscience labs, and in biology, biochemistry and biological engineering in general, continue to run smaller experiments that produce large numbers of images and other data types that require active management for their future reuse. This is difficult for researchers now, given their dependence on disparate software and data formats supported by instrument makers (e.g. microscopes that produce specialized image formats). Given the difficulty of future reuse, many researchers lack motivation to archive their data effectively. Without support and best practices for data formats, metadata tags, storage conventions, and a real data management environment, enormous time is wasted and opportunities lost to leverage the data.

The DataSpace Project will draw on the deep expertise of senior scientists at MIT’s McGovern Institute for Brain Research, the Brain and Cognitive Sciences Department, and the Harvard-MIT Health Science and Technology Division, as well as scientists at Georgia Tech, OSU, and Rice University. These scientists will contribute their existing data, expertise, research goals, and prior experience with data integration, infrastructure and management practices. The project will provide specialist data curators to work with the researchers on data ingestion, description, organization, and management plans. Addressing the “last mile“ problem of access to researchers to support their contribution to and use of data archives has been notably missing from prior large-scale efforts in this field. Beyond the initial DataSpace parners, there are approximately 100,000 neuroscientists worldwide [WA08] so the potential impact of our work in just this initial domain is quite large. We will coordinate with DataSpace PIs, scientists at our partner institutions, members of our Advisory Board and other colleagues to reach out to these centers and individuals as DataSpace early adopters.

II.1.1 Example: Neuroscience Case Study

The idea of localization of function within the brain has been a cornerstone of neuroscience since the 19th century but the technology necessary for major progress has become available only in the past twenty years. The development of the imaging techniques of computerized tomography (CT) and magnetic resonance imaging made it possible to precisely locate damage in brain injured subjects. The measurement of the electrical signals on the scalp, known as electroencephalography (EEG) and arising from the synchronous firing of the neurons in response to a stimulus, opened up new possibilities in studying brain function in normal subjects. The development of the functional imaging modalities of positron emission tomography (PET), single photon emission computed tomography (SPECT), functional magnetic resonance imaging (fMRI), and magnetoencephalography (MEG) has led to a new era in the study of brain function.

At MIT, the Martinos Imaging Center is a collaboration among the McGovern Institute for Brain Research, the Brain and Cognitive Sciences Department, and the Harvard-MIT Division of Health Sciences and Technology (HST). The MIT Martinos Center opened in 2006 and provides one of the few places in the world where researchers can conduct comparative studies of the human brain and the brains of differing animal species. This case study describes the work of one of the DataSpace co-PIs, Professor John Gabrieli, the Grover Hermann Professor in Health Sciences and Technology and Cognitive Neuroscience and the director of the A.A. Martinos Imaging center at MIT. Professor Gabrieli’s research program aims to understand principles of brain organization that are consistent across individuals, and those that vary across people due to age, personality, and other dimensions of individuality [ACS*03, AOK*03, ATW*06, CZD*01, DCG*04, GGW*07, KDG04, MCE*04, OBG*02, OYS*07]. He is one of many researchers that rely on the Martinos Center to conduct his research.

II.1.2 Research Data

At the Martinos Imaging Center, about 160 researchers use a 3T Siemens Tim Trio 60 cm whole-body MR to perform functional magnetic resonance imaging (fMRI), logging more than 3500 hours last year alone. They also employ other brain measures as needed to address scientific questions, including diffusion tensor imaging (DTI), MRI structural volumes, and voxel-based morphometry (VBM). Once the fMRI scan is complete, the image files are converted from DICOM format files to NIfTI format files and stored by the individual researcher. Each image has approximately 170 fields of metadata associated with it, to describe the type of scanner used, subject information, "scan technique", and in some cases information about stimuli used and subject's behavioral performance. The Center currently generates approximately 5.4 Tb of data each year. Subsequent processing of scan data can generate another 6Tb of data annually per researcher. The volume of data grows linearly with the number of new MRI scanners acquired, and new Magnetoencephalography (MEG) scanners will generate significantly more data in the future.

DataSpace will curate the Martinos Center's image data and associated data and metadata as an initial test of the platform and set of services to researchers. In collaboration with Professor Gabrieli at MIT, Professor Randy Engle, Chair of the School of Psychology and director of the Center for Advanced Brain Imaging at Georgia Tech, will work with the DataSpace team on inter-institutional data management requirements and practices. This will provide an initial test of cross-institutional federation for search, aggregation and reuse of image sets from different studies.

II.1.3 Grand Challenges

Currently, there is no widely used system for distribution and sharing of brain imaging datasets across institutions, or across disciplines, reducing the chance for future re-analysis in light of new findings and imaging and analysis techniques. Many fundamental aspects of brain function, such as the questions of how brains can perceive and navigate so robustly, how sensation and action interact, or how brain function relies on concerted neural activity across scales, will remain unsolved until scientists begin to share data so that neuroscientists can combine expertise and develop "hybrid approaches" to their research that could extend and complement the traditional "single lab" approach. While data sharing is not the only barrier to progress, without it progress will be inevitably limited. For example, studies on human variation (i.e. intelligence, cognitive traits etc.) that depend on pooled data are limited statistically by sample size, and it is difficult for any one research center to recruit large numbers of subjects. Geneticists routinely study thousands of people pooled across multiple sites (e.g. for whole genome association studies) and neuroscientists will need to adopt this approach if they are to understand the neural basis of human variability.

While past efforts to create public databases in neuroscience had limited success, recent efforts like the BIRN and XNAT[4] projects have begun to show progress in defining the requirements for data interoperability and reuse. DataSpace can build on these prior efforts to take them forward with a distributed architecture and localized service model that will assist researchers to leverage the archive infrastructure more effectively. In 2007, a neuroscience workshop explored the best strategies for data sharing [THM*08]. The workshop participants identified the development of generally applicable service techniques and a pragmatic focal approach for data sharing of particular types of data as two critical endeavors to advance data sharing in neuroscience. They outlined several services that a data sharing infrastructure should provide including ontologies, monitoring of data uses, an expandable infrastructure, services for both contributors and users of data, teaching tools, and relevant challenges and competitions. Our project will build on these recommendations and will work with Gabrieli and his colleagues to develop immediate, practical data curation services in DataSpace.

II.2 Domain: Biological Oceanography

[4] The Biomedical Informatics Research Network (BIRN) is a large-scale NIH-funded project to create a centralized data repository including neuroscience data. The eXtensible Neuroimaging Toolkit (XNAT) was developed at Washington University in St Louis.

Microbial life has been integral to life on Earth for over 3.5 billion years. Microbes have evolved to be the fundamental engines that drive the cycles of energy and matter on Earth, past and present. Scientists in the field of biological oceanography conduct research in marine ecology by studying relationships among aquatic organisms and their interactions with the environments of the oceans or lakes.

Biological oceanography is an important component of the global climate system since ocean microbes play critical roles in ecosystem dynamics and biochemical cycles, but many of the feedbacks between marine biogeochemistry and climate are only poorly understood. The next major step in this field involves incorporating the information from environmental genomics, targeted process studies, and the systems observing the oceans into numerical models. This will improve predictions of the ocean's response to environmental perturbations, including climate change. Integration of genetics, populations, and ecosystems is the next "great challenge" of this field, involving the integration of marine metagenomic data with rich oceanographic data, process studies and numerical ocean and climate models. Marine metagenomics researchers need the ability to link gene sequences to the associated oceanographic data from where samples were taken. This will allow scientists from across several disciplines to incorporate their data into numerical models to improve predictions of the ocean's response to environmental perturbations, such as climate change.

Biological Oceanography is another ideal candidate for DataSpace, since it requires curating and preserving irreplaceable observational data that cannot be fully exploited at present, as well as combing data from many projects and many disciplines. DataSpace will provide these scientists with the functionality needed to bring this data together, and this interdisciplinary research will potentially lead to a new generation of more realistic oceanographic simulations including improved climate change projections.

II.2.1 ***Example: Biological Oceanography Case Study***

A new, but rapidly growing segment of biological oceanography is the field of marine metagenomics. Unlike traditional microbiology and microbial genome sequencing studies that rely on cultivated cultures, marine metagenomics draws on genetic material recovered directly from environmental samples. Metagenomic data has enabled scientists across disciplines, e.g. biological engineering, genomics, environmental engineering, etc., to begin to explore and model the relationship between marine microbes and things like climate change and the ocean's carbon cycle.

This case study is based on the work of another of the DataSpace co-PIs, Dr. Ed DeLong, a Biological Oceanographer at MIT and a Professor in the Departments of Civil and Environment Engineering and Biological Engineering, whose research focus is marine metagenomics. The overall goal of his research is to better describe and exploit the genetic, biochemical, and metabolic potential that is contained in the natural microbial world. The central focus is on marine systems, due to the fundamental environmental significance of the oceans, as well their suitability for enabling development of new technologies, methods, and theory for assessing the gene and genomic content of natural microbial communities. This is done without cultivation, quantitatively comparing gene content from different microbial communities based on environmental variables, and developing predictive models that relate community gene content to environmental process. Currently the lab is applying contemporary genomic technologies to dissect complex microbial assemblages. Biotic processes that occur within natural microbial communities are diverse and complex, much of this complexity is encoded in the nature, identity, structure, and dynamics of interacting genomes in situ. This genomic information can now be rapidly and generically extracted from the genomes of co-occurring microbes in natural habitats using standard genomic technologies [BAK*00, BSH*02, BSS*01, DPM*06, FMM*06, FST*08, HPP*04, KDK01, OHH*01, RKD08].

DeLong is currently involved in several large-scale data-rich collaborative projects, such as the Hawaii Ocean Time-Series Project, Monterey Bay Microbial Observatory and the Eastern South Pacific Oxygen Minimum Zone Project. We will coordinate with our DataSpace PIs, scientists at our partner institutions, members of our Advisory Board and other colleagues to reach out to these and other collaborative projects as DataSpace early adopters.

II.2.2 Research Data

DeLong's lab creates a combination of observational data (environmental conditions and oceanographic data describing where samples are taken) and experimental data (DNA sequences) using a high volume DNA sequencer (ROCHE 454 pyrosequencer) and a large computational cluster. The lab performs one pyrosequencing run per microbe sample. Each sample contains 100 megabase pairs (Mbp), equivalent to 500,000 DNA sequences. Each week, they perform 2-3 pyrosequencing runs which generate approximately 200 Megabytes of “raw data” (actual DNA sequences) or about 30-60 Gigabytes of raw data per year. Although the nature of the data is typical for this type of research, this represents a small number of DNA sequences, compared to other major labs that have multiple sequencers running samples throughout the year. They also import datasets posted by other labs on the National Center for Biotechnology Information (NCBI) GenBank database to use for comparative analysis (either comparing their data and the places it comes from to similar datasets or looking at the same gene sequence in different environments). To do this, they import and retain approximately 50 times the amount of data they generate so that the lab requires a storage capacity of 40-50 terabytes. The data is imported from GenBank for local processing because the formats GenBank supports do not meet the particular needs of metagenomics research so that the data must first be reformatted for reuse. The reformatted data cannot be resubmitted to GenBank, so the data must be archived locally for the lab and other researchers to reuse.

The metadata associated with their research is the oceanographic observational data from where the sample was taken, including: depth (m), temp of water (degrees C), salinity, chlorophyll concentration (micrograms/kg), biomass (micromoles/kg), dissolved oxygen concentration (micromoles/kg), oxygen (micromoles/kilogram), cell counts, and pigmentation information. Typically, once a frozen microbe sample arrives at the lab, it must be logged on a project website, along with the date that the sample was taken and the download of all of the metadata associated with the sample for local analysis.

Leveraging an ongoing collaboration with Professor DeLong at MIT, Professor Ricardo Letelier at Oregon State University (OSU) College of Oceanic and Atmospheric Sciences will work with the DataSpace team on inter-institutional data management requirements and practices for the Eastern South Pacific Oxygen Minimum Zone (ESP-OMZ) Project. Oxygen minimum zones (OMZs) are regions of the global ocean that present low dissolved oxygen concentrations (<22 µM) at intermediate depths (50-1000 m) due to a reduced ventilation and high respiration rates of the settling organic matter produced in the surface waters. These regions are considered to be an important sink for fixed nitrogen and are important sources of the greenhouse gases carbon dioxide (CO_2) and nitrous oxide (N_2O). The ESP-OMZ Project is a collaboration between the University of Concepcion in Chile, el Instituto del Mar del Perú, the Monterey Bay Aquarium Research Institute, Oregon State University and MIT.

A related effort in data archives for marine microbial ecology research is the CAMERA project at UCSD[5]. CAMERA is developing a centralized repository and database for genomic and related data in the field of marine ecology, and as part of the DataSpace outreach we will collaborate with the CAMERA team towards insuring full data interoperability across platforms, for automated deposit, search, and reuse. DeLong serves on the CAMERA science Advisory Board and we are working with the project’s director about an ongoing collaboration.

II.2.3 Grand Challenges

Marine metagenomic discoveries are occurring at a fast pace, challenging traditional paradigms. Although the National Center for Biotechnology Information’s (NCBI) GenBank database provides scientists in this field with the published gene sequences, the data is provided in formats unsuitable for further analysis with modern tools, and is not linked to the important oceanographic metadata.

Metagenomicists reuse data frequently, going back to look at previously unrecognized sequences of DNA data and applying new tools developed for a particular type of molecule. As an example, a recent breakthrough occurred when a type of rhodopsin derived from bacteria was discovered through the genomic analyses of naturally occurring marine bacterioplankton [BAK*00]. Rhodopsins are light-absorbing pigments that are formed when retinal (vitamin A aldehyde) binds together integral membrane proteins (opsins), and are currently known to belong to two distinct protein families: visual rhodopsins and archaeal rhodopsins. These two protein families showed no significant sequence similarity. In 2000 (when the results of this study were first published) no rhodopsin-like sequences had been reported in members of the domain *Bacteria*. By analyzing previously unknown parts of DNA, data researchers were able to demonstrate that archaeal-like rhodopsins are broadly distributed among different taxa, including

[5] CAMERA - Community Cyberinfrastructure for Advanced Marine Microbial Ecology Research and Analysis – is a Calit2 initiative led by Larry Smarr at UCSD with funding from the Moore Foundation [http://camera.calit2.net/]

members of the domain *Bacteria,* and that a previously unsuspected mode of bacterially mediated light-driven energy generation may commonly occur in oceanic surface waters worldwide. Since some relatives of the proteorhodopsin-containing bacteria use CO_2 as a carbon source, these results suggest the possibility of previously unknown phototrophic pathways that may influence the flux of carbon and energy in the ocean's photic zone worldwide. Many more important discoveries are expected if the data is archived so that it can be effectively analyzed with all the necessary metadata (i.e. oceanographic data). DataSpace will provide these scientists with the functionality needed to bring this information together. The DataSpace Project will work with DeLong, Letelier, and many other researchers to develop tools to search across project data (and data types) and integrate data for later analysis as required by their ongoing research collaborations.

II.3 Differences and Overlaps of Domains

The neuroscience and biological oceanography domains represent ideal initial domains. On the one hand, they involve different and diverse types of scientific data, thereby presenting DataSpace with challenges to demonstrate its ability to support that diversity. On the other hand, there are aspects of overlap for some of the data and research questions, which would demonstrate DataSpace's ability to support cross-disciplinary and multi-disciplinary research. Furthermore, they both demonstrate a key rationale for the NSF DataNet program: they pose many different challenges related to data expression, encoding, documentation, sharing, integrating, visualizing, and preserving; as a result, the use of disparate data makes it difficult to perform research that crosses sub-domains today. Also, these two domains are excellent test cases for DataSpace because it will build on existing collections and collaborations that span institutional boundaries, raising not only technical, but also social and legal challenges for inter-institutional data archiving and integration.

As the DataSpace platform matures (and as early as possible) we will bring in additional domains and researchers to continue testing the scalability and integrative features of the system, and its ability to support collaborative research across institutions. All of our institutional partners: Georgia Tech, MIST, MIT, OSU, and Rice University, will participate in recruiting additional data and working with the DataSpace team to improve the system and the service model we propose.

III. Implications for Scholarly Publishing

In the past, some have considered data sharing to be the responsibility of scholarly publishers since data are integral to the published papers that communicate the results of research, but as datasets have grown they can no longer be presented effectively in print. Some scientific publishers are taking steps to "enhance" publications to link to data and vice versa (e.g. BioMed Central, PLoS, and the Nature Publishing Group) but these efforts are fragmented and incomplete. Most publishers lack the technical expertise and business incentives to take on this responsibility and would benefit from more standardized access to the data, both in format and means of citation (e.g. persistent, Web-enabled identifiers and standards for data description and provenance). Approaches that use new data mining and other techniques to identify relationships between the literature and related data are needed. The DataSpace Project libraries will work with the scientific publishing community to advocate the use of our data encoding and metadata standards in publishing platforms that can support this enhanced functionality, and publishers are well positioned to promote use of the standards and best practices we develop since that the infrastructure necessary to enforce those practices will be readily available to any researcher.

Specifically, the DataSpace Project will investigate recent initiatives to assign persistent, unique Web-enabled identifiers (e.g. DOIs) to research datasets and create federated catalogs of data that are citable in published research papers. The DataSpace platform will assign such identifiers to each dataset (and in some cases each data entry) such that publishers can reasonably require authors to deposit data with their institution (if no other preferred and suitable archive is available) and cite it in their publications.

Additionally, our partners in the Science Commons are developing techniques to link research data to articles using advance text mining technology, and are working on standards for assigning persistent identifiers to data and other critical Web resources (e.g. authors) necessary to create a Web of open data.

IV. Public Outreach and Education: Harnessing Collective Intelligence to Advance Science

When scientific research data become widely available on the Web in reusable form, new technology for social networking can be applied to scientific problems that will engage the public's attention. As an example of this strategy, a new MIT research program – the Climate Collaboratorium – focuses on harnessing the collective problem-solving power of large numbers of people and massively shared computer simulation models to address the challenges posed by global climate change – which extends on the climate implications of the work on Biological Oceanography, described earlier. By creating *radically open computer models and data* - in the spirit of Wikipedia or Linux - far more people can see the models and data, understand their implications, and examine the consequences of alternative assumptions. This form of "citizen science" has the potential to allow for better collective understanding of the problem, consideration of more possibilities, more extensive analysis, and ultimately better solutions than are possible today

A core element of the project is to identify and make accessible datasets needed to do accurate modeling. For example, we often lack data gathered at the local level, which provide details not available in datasets gathered at the state or national level, such as locally measured coastal water conditions. Furthermore, many municipal and regional governmental organizations in the United States have done studies of ocean and microbial life, but such data is scattered amongst many agencies around the country, for the most part inaccessible to other researchers [M*03b, MC94, MCL*98, MCL*99, OM01].

For the DataSpace Project we will (1) identify datasets that can be useful for environmental modeling; (2) develop standardized data formats that make that data readily usable; (3) develop tools to convert existing data into the standardized formats so that these can be exploited by the Collaboratorium and similar projects, (4) expose research data to the public in ways that will engage their attention and contribution; (5) demonstrate the use of this tool in the MIT curriculum and publish the resulting materials open to the world through MIT OpenCourseWare. The Collaboratorium will be merely the first of a series of experiments to "open up" scientific data to the public as well as engage the public in the gathering and study of such data.

As another important form of outreach, in our initial scientific domain of microbial oceanography, the DataSpace Project will leverage the existing C-MORE Project based at the University of Hawaii. The Center for Microbial Oceanography: Research and Education (C-MORE) was established in August 2006 as a National Science Foundation-sponsored Science and Technology Center designed to facilitate a more comprehensive understanding of the biological and ecological diversity of marine micro-organisms. Its mission is to perform research and outreach to students and the public in this domain, and both MIT and OSU are currently key partner institutions. DataSpace will collaborate with C-MORE to provide data archiving support via the partner institutions, and to understand the project's requirements in support of teaching and public outreach.

V. Project Activities

V.1. Infrastructure Design, Development and Research Agenda

The infrastructure that will be designed for the DataSpace Project builds on existing infrastructure and extends it to Web-scale. We will create a basic operational system in the first year of the project, while doing the research necessary to evolve that infrastructure to further generations of design to allow for very large scale and broadly distributed network of archives.

- ***DSpace digital archive platform***: Originally created by HP Labs and MIT and released in 2002, DSpace is open source software designed to implement the Open Archives Information System (OAIS) reference model[6] for long-term digital archives. It is now used by more than 500 research institutions worldwide for access to and long-term archiving of research output in digital formats, primarily (but not exclusively) digital research publications. It includes functionality for data deposit, management, discovery, and preservation, and assigns globally unique, persistent identifiers to archived material. It facilitates "open" access but supports controlled access and embargoes where necessary.

Prior research on the DSpace platform related to long-term data curation includes work by HP Labs on federated, replicated archives; work by Cambridge University on scientific workflow integration (for crystallographic chemistry data deposit); work by MIT and the San Diego Supercomputer Center on policy frameworks for locally-controlled, globally-enforced distributed data curation; and several projects at MIT

[6] The Reference Model for an Open Archival Information System is available from http://public.ccsds.org/publications/archive/650x0b1.pdf

and other institutions in the DSpace community on digital preservation strategies for a variety of data formats [E07, MS07, S*05a, S*05b, SM07].

During the past year the DSpace open source software developer community has created a second generation platform known as DSpace 2.0. It is a complete reengineering of the original system designed to overcome earlier scalability and modularity limitations and will be released in July of 2009. The DSpace 2.0 platform provides simple APIs to a set of critical OAIS functional modules, and will provide the initial framework for a new DataSpace reference platform. While DSpace 2.0 will form a significant piece of the initial DataSpace platform, it will be integrated with other existing systems (e.g. the Fedora 3.0 open source software repository system, various existing user applications for search and retrieval, and existing storage platforms) to complete the first prototype of the platform. As our curation experience and research agenda evolves we anticipate annual or more frequent software releases of the DataSpace platform that integrate new features built from successful research prototypes. These will affect the curation layer, policy and security layers, user application layer, storage layer, and eventually every aspect of the platform required to support large-scale scientific data archiving.

Furthermore, a key design goal of the DataSpace Project is to achieve platform independence by defining the set of standards and protocols that can be implemented by *any data archiving system* so that a competitive market can emerge (similar to the marketplace for Web servers and browsers that exists today). We will identify a technical architecture and set of standards and protocols, building on the existing Web architecture; create a working *reference implementation*; develop an outreach plan that will encourage additional implementations of the architecture in the future, to meet the needs of, for example, other sectors or computing environments. All components of the DataSpace reference platform will be freely available as open source software, but over time, as the standards and protocols comprising the platform are published, we expect commercial and open source alternatives to appear.

- ***Data protection and security***: While MIT and the DataSpace Project are committed to the principle of open access to research data wherever possible, we acknowledge that a lot of data cannot be shared for national security reasons, or to respect the privacy of data sources (e.g. HIPAA regulations for clinical data) [ADW01, BLH*06, SE02]. Furthermore some researchers need to limit access to their data for a period of time to protect their own research use of it. The DataSpace infrastructure will support this requirement, building on solid, mature, network security protocols, as well as defining federated security standards for data access and management using technology, such as Shibboleth. Partners at HP Labs have significant experience in this area and will contribute to our approach in DataSpace.
- ***Data discovery and data semantics***: Discovery of scientific research data is one of the largest problems facing successful deployment of a global cyberinfrastructure. Unlike text, which can be easily indexed by search engines, data are not immediately findable or usable without metadata to describe their content, structure and meaning. This metadata is often implicit in the databases and other technologies that house the data. In addition to providing metadata, it is important to fully understand the semantics of the data. Knowing that a piece of data is about temperature is one level of meaning, but knowing in which way temperature is recorded (e.g. ^{o}F, ^{o}C, or ^{o}K), is equally important. Domain specific ontologies will be used to represent this knowledge and, since DataSpace is intended to store data over decades, it is important that “temporal semantics” also be provided (e.g. in Russia, dates were recorded using the Gregorian calendar until 1918, subsequently dates used the Julian calendar.) For these reasons the DataSpace Project will work on metadata and ontology creation – both automatic and human-generated. This is an area of particular expertise in libraries and archives that we can leverage in collaboration with domain experts and the community at large [FMG04, GBMS99, LFG*98, LM96, LMS96a, LWG*98, Mad03, Mad95a, Mad96, MMS01a, MMS02b, MS91a, MS91b, MSG91, MSS01, MZ06, TM98, ZMS04a, ZMS04b, ZM06].
- ***Data quality****:* For data to be effectively reused, it must be fully understood and trusted. To accomplish this, extensive quality metadata on the metrics must be provided (such as data accuracy and precision) as well as provenance (such as the details of how, when, where, why, and by whom it was produced.) These have been areas of extensive research by members of the DataSpace team for almost two decades in, for example [MW89b, WMH89, MW90a, MW90b, MW90c, KMS95] and the affiliated MIT Total Data Quality Management (TDQM) project, and the MIT Information Quality (MITIQ) program. Some specific and important DataSpace research efforts, to be built on past accomplishments such as [WMK93], include the categorizing of quality metrics for scientific data and the incorporation of data

quality calculations, such as using provenance information [PM07, PM08] explicitly in DataSpace-supported scientific data analytics to greatly reduce the efforts of scientific researchers to reuse data correctly and effectively.

- ***Data analytics***: Current database technologies are optimized to provide fast access to "raw" data but have few provisions for efficient analytical processing – usually the key purpose for using scientific data. To address this, the DataSpace project will develop a framework which permits sources of data to be transparently combined with the required analytic models. Expected benefits include: (1) Combining the knowledge of storage and of visualization/analysis of data guarantees persistent access to both components. The system will have sufficient flexibility to allow data and algorithms situated in multiple, geographically distributed locations to be exploited, (2) users often need control over the form and consistency of the data – such as the type of normalization to be applied, units used and other reporting conventions. Further, the system should have the intelligence required to automatically conduct many of the basic pre-processing and consistency checks, and (3) the proposed framework will allow proprietary data and/or analytical methods to be protected. Providers of data will be able to specify the appropriate degrees of privacy attached to specific datasets, which would define the level of access granted to users. Similarly, providers of analytical tools or applications will be able to allow limited or subscription-based access without compromising their intellectual property rights.

Three related research activities under the DataSpace project, involving collaborating researchers from MIT, HP Labs, and Masdar Institute, have been identified that provide concrete situations in which different aspects of the above principles may be demonstrated:

- *Model calibration and mediation* for extracting and classifying patterns in complex data. In the example of predicting disease onset, drug companies may have data which can be used to calibrate or evaluate these models; but, as such data would often be considered commercially sensitive, a viable alternative might be to provide restricted usage of the data. For example, given information which characterizes the model, the system could provide gradient terms for updating model parameters, or accuracy estimates to facilitate model evaluation. Neither of these properties is commercially sensitive on its own, but both are of great use to researchers working on these models.
- *Operational Scientific Intelligence*. Business Intelligence (BI) has been an important area of research and commerce; there are clear parallels to "Scientific Intelligence" (SI). Particular areas of interest are real-time streaming and extraction analytics, operational data warehousing (to enable near-real time data integration and queries), and closed loop analytics. Distributing data storage and SI over multiple sites will enable high performance computing.
- *High-speed pre-processing and data consistency.* Efficient processing of petabyte-scale databases presents significant challenges, especially regarding the uncertainty in data. An example is in the analysis of satellite data, which often can be adversely affected by factors such as cloud cover and viewing angle. Researchers should be able to embed scientific requirements and conditions into the data. Such requirements could be "hard-limited" where data found to be violating specified consistency conditions would be flagged as such, or "soft-limited", where consistency or compliance scores could be used to assign some measure of data quality to be used in the processing.

- ***Data interoperability and integration***: Identifying standards for data encoding can make it easier to interoperate in a federated, distributed, interdisciplinary environment. This is the single biggest challenge, and greatest win, facing the DataSpace initiative. Storing and preserving data for the long term are only useful if that data can be interpreted in the future. Integrating research data (and metadata) within and across disciplines is enormously difficult today. The DataSpace Project will exploit emerging Semantic Web standards to achieve this data interoperability, while acknowledging that meaningful data integration will always require a large degree of knowledge about data semantics. Our organizational model includes library and scientific domain experts to provide this data curation service by leveraging local knowledge of data production methodologies and semantics, and combining this with international data ontology standards efforts [LMS96a, LMS96b, MMS01b, MMS02a, MMS02b, MMS03, SSR94, ZM06].

- ***Data conversion***: Although DataSpace will promote common data standards, heterogeneous data will persist due to its legacy heritage as well as to the particular needs of different domains and sub-domains of science. DataSpace will provide the means for automated generation of data conversion programs, based on extensions to MIT's Context Interchange (COIN) technology [BFG*97a, BGF*97b, FMG02a, GBMS99, MMS02b, ZMS08], which will facilitate data integration across multiple scientific data sources. Normally, multiple specific conversion programs must be written in order to convert data from each source into the appropriate format and semantics needed for processing, and each conversion program is likely to be different since there are usually multiple adjustments necessary. We refer to each type of adjustment as a “context modifier” and the collection of context modifiers for a source is referred to as the source’s “context.” Even when there are general-purpose data conversion utility programs available, someone must know all the source and processing contexts and the modifier adjustments necessary and specify them (a difficult, time-consuming, and error-prone process). In contrast, in [ZM06] an example is shown where there were 30 different data sources, each source having five different context modifiers. To allow all of these data sources to be used in any of the individual data sources’ contexts, up to 870 conversion programs would have to be created. The COIN project developed an automatic composition algorithm that utilized the semantics of the data sources (the “context” information) along with a library of basic conversion components that handled conversions within a single context modifier. Using this approach, in the worst case only 102 component conversions would be needed in the component conversion library. In many cases, the context modifier component conversions can be parameterized (e.g., a basic formula for converting among ^{o}F, ^{o}C, or ^{o}K). In the example in the paper, the actual number of component conversions needed was reduced from 870 to 8. Furthermore, since the construction of the conversion programs was automatic, it was accomplished without requiring any significant effort, or possible errors, by the scientific researchers or their staff. Although the basic theory behind context-based data conversion has been established and limited prototypes have been implemented, key research goals of DataSpace are to: (1) determine relevant context modifiers for broad categories of scientific data, (2) develop a library of conversion components, (3) adapt the context knowledge representation to fit within the overall metadata standards being developed for DataSpace, and (4) demonstrate the scalability and effectiveness of this approach in meeting the needs for large-scale scientific data users.

- ***Data analysis and visualization***: Like data encoding, long-term archiving and preservation of data are only useful if the tools to analyze and visualize the data also persist over time. Designing useful data processing tools is an active area of research itself – involving, for example, statistics, human/computer interaction, cognitive psychology, and systems modeling. The absence of well-developed methods for multivariate data analysis, data visualization, and spatial modeling of dynamic systems represents three major obstacles that currently limit use of image-derived data for systems modeling. In addition, we intend to develop a *data flow-based architecture* for high performance data-intensive analytics. Scientists can use data flow-based analytics engines, which may be provisioned as a cloud-based platform, to obtain high performance and highly scalable computation services that can be easily composed and applied to fresh or archived data sources, and have the results and the derivation processes be automatically preserved in DataSpace. Advanced data visualization techniques will be used not only to glean insight into the aggregate behavior of large quantities of data, but also used to allow scientists to navigate and drill down to data attributes of individual records. These techniques promise to elevate the data manipulation interfaces between DataSpace and the scientific user communities.

Scientists have produced many excellent visualization tools for their particular data and research questions, but these are usually created in an ad hoc manner that cannot be leveraged across domains or even different data sources [KHD*02, MCL*98]. The DataSpace Project will develop a technical architecture for implementing data visualization tools within a general framework as part of the DataSpace platform. Such a standards-based framework, building on well-defined, Web-based protocols, would support the creation and adaptation of analysis and visualization tools by domain experts that could then be tagged and archived for discovery and reuse in new contexts for unforeseen purposes. Even if such shared libraries did not emerge, a predictable framework for deploying new visualization tools over heterogeneous datasets could create new efficiencies for leveraging expensive visualization expertise.

- ***Data storage***: Given the size and growth rate of scientific research data, storage for a scalable DataSpace cyberinfrastructure will require a fundamentally new architecture. Storage services need not be co-located with curation and higher-order preservation activities, thus enabling curating organizations to use 3rd party storage service providers, with or without “persistence” services providing basic bit-level data preservation. Other industries (finance, insurance, commercial engineering, etc.) also face large-scale storage needs and are creating a vibrant commercial market in this area. The DataSpace Project will initially work with its partners EMC and HP Labs, and other storage providers as appropriate, to determine optimal storage solutions and near-term architectural options with commercial or non-profit service providers (e.g. Internet Archive, Akamai, Iron Mountain, Amazon, HP). In cases where the data are very large and requires data-intensive processing that can only be done locally, the data will need to reside within the organization. But in cases where the data are rarely used or not intensive to process, storing it remotely at lower cost would be desirable. The architecture will support the spectrum of storage possibilities with different qualities of service [BHM90, CH01, DHL01, DHL91, HS91, HX07, ZHF00].

- ***Legal Issues***: Infrastructure is not simply a technical issue. The law is a significant component of the network infrastructure for science. Copyrights, privacy rights, patent rights, data rights, and more all need to be managed. Data sharing and reuse are today plagued by varying practices and legal frameworks which makes data integration difficult to legally perform in many cases. The DataSpace Project will collaborate with the Science Commons to develop licenses and best practices for open, international data sharing, and promote those practices through outreach and the curation systems we deploy. The Science Commons has proposed a set of principles for open data usage and a protocol for implementing those principles, and plans to distribute an Open Data Mark and metadata for use on databases and data. However some data (as explained in the section on data security above) will not be possible to share, or only in limited contexts. These situations are straightforward examples of patent, privacy, copyright, and other intellectual property laws and conventions, and DataSpace intends to support these restrictions as well [E01a, E01b, E03, E06, EM04].

- ***Distributed policy management****:* Data integration across heterogeneous sources requires not only semantic integration, but also the identification and resolution of policy-based restrictions on dynamically-integrated data. For example, with policy metadata developed by the Science Commons, copyright information presented in Semantic Web formats with policies can be evaluated and acted upon in a machine-assisted manner. Project staff will develop extensions to the basic repository architecture that will enable the policy-based definition and management of repository federations. Our unique approach will permit peers within federations to share and interpret policy requirements that are intrinsic to membership in those federations. In particular, our framework will enable new members to expose their requirements to existing members while enabling new members to fulfill outstanding requirements for existing members [WHB*05, WHBL*05, ZMS01].

The approach to policy-based repository federation focuses on multi-party "covenants" that are built in simple, declarative fashion and may express policies governing replication, preservation, and various automated content management tasks. We will develop a policy model and framework for the management of federations and nodes which will include an extensible set of operations that nodes agree to perform under sets of conditions. At the repository level the policy model will be implemented as a component of a configurable repository. We will further develop a service node which can consume policies from all federation members and most appropriately assign tasks and agreements between them.

- ***Archives Internet domain***: Since trust and quality are such critical aspects of data management for science, we propose to explore the creation of a new top-level Internet domain – .arc – that would be granted by an accreditation process and guarantee certain functions and/or policies. This would allow both producers and consumers of scientific research data to understand an archive's data curation intent and reliability, and potentially data quality. Should such a domain prove useful, the accrediting body could be a governance body for the future federation of DataSpace archives (similar to the DSpace Foundation) or a dedicated purpose accreditation body, independent of DataSpace.

- ***Workflows for scientific research and archives***: Data curation is a holistic process that begins with data creation and is ongoing throughout the data's entire lifecycle. Successful data capture, including useful documentation, must be integrated into the research workflow of each domain. Extensive work is needed to identify or develop protocols to support simple integration of curation and preservation activities with other parts of the data lifecycle, and particular data creation. For example, much research data is currently generated from particular instruments in proprietary and custom formats defined by the instrument maker. Identifying data format standards and working with instrument producers to support those standards across domains could lead to significantly lowered cost for data curation and long-term preservation, without adverse impact on the instrument producers' revenue models.

At the DataSpace partner institutions (e.g. Georgia Tech, Rice, OSU) we recognize the need for up-front data curation as part of a viable service model, provided by the research library and working directly with researchers to inform best practice for data capture (e.g. "good" formats for long-term preservation) and to work with researchers on data ingestion to the archive. This addresses the "last mile problem" of data archiving that has plagued many prior attempts to create successful, large-scale data archives. Without local support for scientists to help with data archiving it becomes a burden that is difficult to overcome. With appropriate planned support, barriers can be overcome in time.

- ***Business models for archives and distributed organizations***: The model for sustainability of the DataSpace archives network will depend on its distributed, multi-purpose nature, so that the cost is not borne by a single institution or sponsor alone. Institutions that support research and other industries that share many of the DataNet's infrastructure needs can and should bear the costs of the infrastructure, much like the Internet and the Web. We assert that research-generating organizations already have some infrastructure in place to manage research output (particularly in written form, with the enormous book and journal collections support by research university libraries and archives) and are under increasing pressure to provide similar support for other types of research output such as data. There are existing examples where data archiving is done within research institutions – social science research datasets, GIS datasets, and storage services for small, personal datasets – but few institutions have developed general-purpose data archiving infrastructure to date. Some research domains have existing high-quality, large-scale, often international data archives in place (e.g. NCBI, NCAR) but most disciplines lack such infrastructure, and even where it exists, integrating data across domains is too difficult. We believe that the architecture we propose to define can be implemented at many scales – from a single lab, to an entire institution, to a national or international archive – and the costs will be distributed – making it affordable and sustainable for all. For the existing DSpace federation that cost is primarily borne by research universities and their libraries and IT departments. For new DataSpace federation we expect a similar pattern to emerge initially, but to be more widely distributed across smaller and larger organizations over time [Mad87, MOW89, MS02, MW88b, MWC*04, ZMS01, M*03a, M03*b, M04a, M04b, MC94, MCL*99, ML98, MYB87, OM01].

- ***International consensus***: A key requirement for the long-term sustainability of DataSpace is a governance mechanism to evolve and gain consensus support for technical and operational standards. As the global body responsible for setting Web standards and defining the future architecture of the Web, the World Wide Web Consortium (W3C) offers expertise and an existing institution that can build consensus on DataSpace-derived standards [BL99, BLC*99, BLC96]. W3C could lead the standardization effort alone or in partnership with other standards groups in the library, publishing and archival communities. In addition, the current DSpace federation is a global organization of hundreds of research universities that has evolved a governance model for collective action and software maintenance that might be leveraged for this project.

V.2. Data Management Operations

The DataSpace Project is built on the assumption that research-producing organizations, such as universities, have ultimate responsibility for managing and preserving the data their researchers generate and depend on. This is an extension of the concept of "institutional repositories" of which DSpace is a leading example that has achieved widespread adoption by research organizations over the past five years. These repositories place responsibility for digital archiving within institutions – often the library, IT department, or a combination of both – but they are not presently designed to manage large-scale scientific research data, and are not easily federated to create subject-based archives. So the concept of institutional responsibility for data management is becoming established, but the means to accomplish that concept for more complex forms of data is still emerging.

The DataSpace project will develop a prototype service plan for institutionally-based data archiving operations. These prototype service models will both demonstrate the feasibility of institutionally-based data archives, and document the associated costs and skills necessary to run them. Within institutions we expect to see a range of DataSpace nodes from supported, enterprise archives to locally-managed archives in large research units. Some institutions may choose to outsource archiving operation or run them in consortia -- all of these models have been employed for institutional repositories. The architecture will support all of these modes of operation as well as explore third party service providers (e.g. for persistent storage provision).

The service models prototyped by the initial DataSpace partners (MIT, Rice, Georgia Tech, OSU, Masdar) will investigate roles and responsibilities across an institution, including the library, IT department, research labs, etc. Each partner institution will build a prototype operation involving the library and including staff with scientific domain expertise, data curation expertise, and IT expertise.

Science partners will advise project staff on types of data, priorities for data management, opportunities for improved data integration, and realistic methods of outreach to local scientists working with data from small lab experiments to global research collaborations. Each domain of science has different conventions for data production, management, and integration, and there is a wide variety of archiving practice across a given institution, from highly normative (e.g. major research organizations like the McGovern Institutes) to ad hoc (e.g. a biologist working in a personal lab environment). The DataSpace service model needs to account for the range of scientific norms and practices while respecting current working practices and discipline conventions. The model must also offer direct benefit back to researchers to motivate them to adopt new metadata formats and (over time) shift workflows the new infrastructure available to them. The services models and operations thus developed will be described and promoted as a model to other research institutions by our partners in the DSpace Foundation.

Finally, the DataSpace Project recognizes that in many cases there will be a need for data archives that are not institutionally supported – i.e. that are run directly by a lab, department, or center, by a consortium of institutions, or by a scholarly society, a publisher, or any other part of the scientific community, so, we expect to see a large number of different types of archives to emerge. The DataSpace architecture and governance model developed for the future network of DataSpace nodes will be implementable by any archive as long as it conforms to the standards we establish. This follows the general principles of Internet and Web development, where anyone can play, but there are social layers built on top of the technology to create virtual communities with known practices and procedures to define trust and quality within that community.

V.3. Work Plan Structure

The DataSpace Project will proceed along four parallel tracks over the five-year period:

- *Research and Development*: Identification and analysis of scientific datasets from initial domains, Development of initial platform (year 1), Annual demonstrators of research activities and new platform releases
- *Exemplar Operations*: Ingest of initial scientific datasets, Creation of new operational infrastructure, Documentation of cost and service models for DataSpace operations
- *Business Development*: Definition of cost models for DataSpace operations at MIT, Rice, Georgia Tech, OSU, and Masdar. Experimentation with business model options (see Appendix A1)
- *Education and Outreach*: Documentation of service models for DataSpace operations, Outreach to DSpace community (500+ research institutions worldwide), Outreach to existing data archive operations, Outreach to Web community and data managers from other domains

In each of the four tracks there will be annual deliverables to distribute project output as quickly as possible. The initial outputs include research findings, architectural designs, recommended standards and protocols, open source software implementations, exemplar service models and cost models, and identified third-party service providers.

V.4. Data Management Community Outreach

V.4.1. International Relations

The DataSpace Project team has many existing international ties that it can leverage for outreach and coordination with related efforts in other countries. Among our partners, EMC, Google, HP Labs, the DSpace Foundation and the Science Commons are global organizations with representation throughout the world, including developing countries. The W3C is an international organization that coordinates the standards and protocols underlying the Web.

Several members of the team are part of the Web Science Research Initiative, a new joint program with the University of Southampton in the UK (also the source of the EPrints open source digital repository platform and significant research in digital repositories and preservation, Semantic Web tools and technology, and Open Access promotion and support models). WSRI coordinates international research around all aspect of the Web including sociology and economics, biology and law, etc., as well as its technical impact. Furthermore, our advisory board includes representatives of major international research organizations such as ERCIM and CODATA, as well as related cyberinfrastructure initiatives in the UK, Europe, Australia, Canada and elsewhere. Through the Advisory Board (initially of the project, but ultimately of the new federation of archives) we will insure continuous input from other parts of the world and coordinate standards and activities.

Since DataSpace is a distributed architecture, like the Web, it will establish a global virtual organization. We are already in discussions with research universities emerging throughout the world to join the DataSpace Project. As an example, the newly established Masdar Institute of Science and Technology (MIST) in the United Arab Emirates will be an unfunded partner organization and is proposing to establish one of our first international nodes as well as collaborate on some of the research agenda.

V.4.2. Educational Outreach

Use of science and engineering data in educational contexts is growing as active and lab-based learning methods become more accepted. While it is out of scope for this project to work on changing the science and engineering curriculum to make more use of data, we will actively support cases where this happens by making the data citable and reusable, and documenting how the data were used in teaching MIT courses via the OpenCourseWare (OCW) program [A08], which publishes the materials of every MIT course as freely available content on the Web. OCW sites include archived data and related documentation (e.g. visualization tools, problem sets using the data), and these are also archived into MIT's DSpace archive. To gain broader adoption of this practice we will also liaise with popular open source software course management systems (e.g. Sakai, Moodle) to insure ease of incorporating available data and related tools into course materials within those platforms. In addition, we will explore ways, in conjunction with MIT's Center for Collective Intelligence, to encourage and empower society at large to become more actively engaged in scientific investigation and the use of scientific data.

VI. Assessment Plan

DataSpace will track progress along 3 major dimensions for each quarter of the project.

- *Data Recruitment*: volume of data ingested (TB's and distinct datasets), number of data formats ingested, number of researchers contributing data, number of data attributes per dataset (i.e. metadata)
- *Platform adoption*: number of software downloads, number of software installations, number of institutions "federating" data, number of software implementations of DataSpace standards (i.e. beyond the DataSpace reference implementation of the system), number of contributions of software from outside the project team (i.e. measurement of open source community development).
- *Archive Usage*: for DataSpace archives at MIT, Georgia Tech, MIST, Rice, and OSU initially, number of unique visits to the archive, number of searches, downloads and views of data.

Metrics are an important measure of progress for key goals (e.g. broad adoption of the platform) but there are other important, less measurable, dimensions that also determine long-term success. The DataSpace Project will create both a high level advisory board (detailed elsewhere) and a DataSpace Business Development Management Team (DBDMT) (see Sustainability Plan in Appendix A1). The responsibility of these boards will be to help the project develop both realistic operations plans and business models for ongoing sustainability.

The DBDMT will use the capabilities provided by Sloan School faculty and researchers to assess DataSpace operations, performance and sustainability. Examples of practices proposed include:

1. Utilizing system performance data capture tools in conjunction with statistical analysis to examine performance and use.
2. Developing and utilizing models along with the data capture described in (1) to determine choices in existing implementations and promising directions for future development.
3. Use of customer surveys and assessment tools to obtain user perspective. Using these results to improve DataSpace utility and operations.
4. Interaction with Entrepreneurship and Innovation initiatives to assess business models, customer base expansion and add-on markets.

With these and other assessment practices we plan to provide business models for sustainability, feedback to R&D and Operations, and Guidance for outreach (training and education).

VII. DataSpace Project Organizational Structure

It is our belief that research-generating organizations can and will play an active role in curating the data their researchers generate. Once the costs of doing so are in line with other enterprise operations (e.g. Internet connectivity, Web site management, Library and Course Management Systems) we believe these organizations will accept this responsibility, including ongoing financial support for their archiving activities. Using DataSpace Project partners as initial exemplars and partnering with the DSpace Foundation for outreach to other major research organizations worldwide, we will test operational and business models to demonstrate the technical and economic viability of a distributed, federated model of data archives.

The archive network that emerges from the DataSpace Project will be highly decentralized, flexible, and multi-modal. It will be light-weight and low-cost to lower barriers to adoption by any organization that generates research data. A new organization will serve to coordinate these distributed activities, similar to the role of ICANN. Since much of the DataSpace network's interoperability will build on existing infrastructure and standards, the added coordination costs of the curation activities can be kept low.

VII.1. Organization

The DataSpace Project's organizational structure will iterate research and development phases with operations and expansion phases over the five year period.

- Research and Development phase

In the R&D phases we will work with a group of specific individuals (i.e. the PI, co-PIs, and the senior personnel) to develop the initial DataSpace platform (see Figure 2); to work on the range of research problems identified in the activities section; to deploy and test exemplar infrastructure and operations (at MIT, Rice, Georgia Tech, OSU, and Masdar); and to ingest initial data collections and provide the improvements to data integration and interoperability that our research will enable. During this phase the work will be done in as broadly inclusive a manner as possible, using well-known techniques for virtual community building (e.g. public mailing lists and virtual meetings, wikis for documentation, open comment periods). Outcomes of this phase beyond the platform and research results are working data collections and operational service models that can be promoted within the partner organizations and outside the project to organizations worldwide.

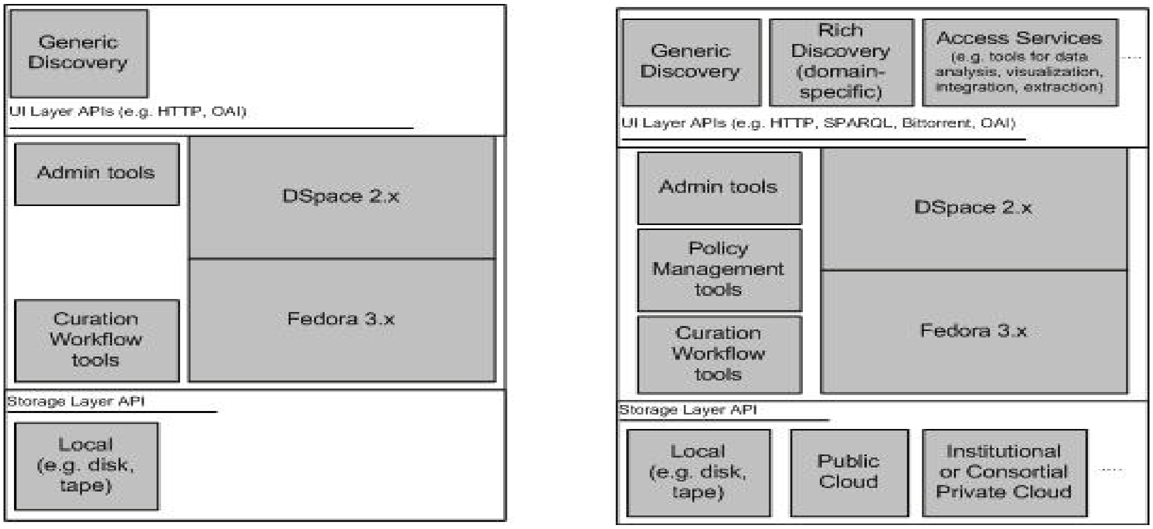


Figure 2: Initial DataSpace v0.1 Architecture on left (Year 1), v1.0 on right (Years 2-5)

Within MIT, the project team consists of members from the computer science and management research community, the Libraries and Archives, the main campus IT organization, and senior researchers and IT experts from the scientific domains of initial interest (neuroscience and biological oceanography). The project will hire a dedicated project director to coordinate the activities of the partners, both at MIT and externally, and to liaise with the other NSF DataNet partners and the global data archiving community.

External to MIT we have expert policy partners (the **Science Commons**), corporate technology research partners (**EMC, HP Labs** and **Google Labs**), and data archiving institutional partners (**Rice, Georgia Tech, Oregon State University** and the **DSpace Foundation**) and initial international DataSpace node (**Masdar**) These partnerships will be managed by MIT, the Project Director and the

Management Board, and are described in detail in Appendix A5. These are the project's formal partners but we believe in (and have extensive experience with) inclusive processes that will accommodate any sector, any organization or individual who wishes to participate.

- Operations and Expansion phases

The second phase (which overlaps with the first) involves deploying systems that conform to the architecture and standards defined in phase one, and defining service models for institutionally-based data archives. Those systems and service models will serve as exemplars for a new, global virtual organization, currently referred to as the DataSpace federation that will be continuously sustainable beyond the life of the project. The DSpace Foundation and the DataSpace Project's institutional partners will lead the outreach efforts to the current DSpace federation, to the general research community, and to other organizations and sectors. This is similar to the model implemented by the World Wide Web and other informal communities that form around a set of technological standards and social conventions. The costs of this virtual organization will be similarly distributed, borne by each organization that contributes data to the network.

Each DataSpace Project partner has cyberinfrastructure capabilities that the project will leverage (see Appendix A3). One of the deliverables of this phase will be to template cyberinfrastructure requirements for various the service models deployed. For example, we would identify hardware, network and storage requirements for different-sized archives (e.g. below a petabyte, an exabyte, and beyond), and minimum requirements for federated operations at Web-scale. We do not assume that organizations will run their data archiving operations entirely in house and will explore commercial and non-profit offerings for various parts of the architecture (e.g. persistent storage providers, file format migration services, and "cloud" computing and storage services).

We will evaluate the possibly of new organizations to support specific defined activities (e.g. certifying archives for the .arc domain, maintaining standards that are unique to data archives, providing community governance and technology planning) similar to the role of the Internet Corporation for Assigned Names and Numbers (ICANN) in coordinating and assigning domain names. The structure and business model of such new organization(s) will be determined during phase two of the DataSpace Project as one of its deliverables, based on what we learn about the needs of the larger virtual organization. Governance of such virtual organizations must be flexible and responsive so we will not require or proscribe any particular governance model a priori.

Because every organization has data that it manages and archives for various business reasons, we believe that the DataSpace infrastructure will not be specific to scientific research data and can be adopted and maintained collectively with other sectors of the economy in much the same way as the Internet and the Web are today. Other domains of practice – finance, transportation, defense, health care, etc. – all face similar problems of integrating and leveraging data over long time frames, and should be involved in the creation and support of the data Web that this project proposes to build for science. The DataSpace Federation will include such non-science members, and the governance model will be designed with that assumption. Enthusiastic support for DataSpace from other sectors (e.g., financial services) has already been evidenced.

VII. 2. Management Structure

To ensure the effective management and evolution of the DataSpace Project, we will create a term-based Management Board and Advisory Board as well as a continuous DataSpace Federation virtual organization and governance body.

- *Project Management Board* will consist of the Project Director, PIs, and Senior Researchers, and will meet at least once a month to review progress and set near-term and long-term goals and plans. The management board is to ensure good collaboration and communication between and among the domain experts and the DataSpace researchers and operations teams. A Project Management Executive Committee, consisting of the PI's and Project Director, will make day-to-day management decisions.
- *Project Advisory Board* will consist of a diverse ensemble of knowledgeable leaders from throughout the world and diverse industries (including corporate as well as scientific) to ensure that "best practices" are understood and adopted. The Advisory Board will meet at least twice a year and will be available for advice on an ongoing basis. Besides providing insights to the Management Board, the Advisory Board will be one of the mechanisms to achieve outreach to communities that are potential users and supporters of DataSpace. A number of notable individuals have agreed to join the Advisory Board (see Appendix A5).

- *DataSpace Federation* is a virtual organization that will initially consist of the early adopters of the DataSpace technology and will evolve, over time, to be the sponsoring/coordinating body for the sustainment of DataSpace, much like the DSpace Federation.

VII.3. Diversity and Inclusiveness

The DataSpace Project is committed to diversity and inclusiveness, and will engage in a range of activities to support these goals, including:

- Project staff and research assistants will be retained following MIT's and our partner institutions' standard equal opportunity hiring policies.
- Infrastructure we design or build will comply with the W3C's accessibility guidelines, and we will apply similar standards to the research data where possible.
- The Project's working practices will be public, open and transparent to encourage the broadest possible participation across communities and without regard to race, color, religion, sex, age, national origin, sexual orientation, gender identity or expression, disability, or veteran status.
- The Project's software deliverables will be entirely open source, and the community we will build to maintain and enhance that software will be similarly open and inclusive (as is currently the case for the DSpace software).
- In the spirit of DataSpace's goal to make scientific data widely available, DataSpace will enable research groups at small or modestly funded institutions to engage in research not currently feasible.

The Project will support and enhance MIT's Interphase and Minority Introduction to Engineering and Science (MITES) programs and similar programs at partner universities. Interphase is a rigorous summer residential, academic program for admitted freshmen in their transition to MIT, who especially benefit from academic enrichment and support and is especially designed to foster high achievement and content mastery for underrepresented minorities (African American, Mexican American, Hispanic/Latino and Native American). MITES is a rigorous academic enrichment program for promising high school juniors interested in studying and exploring careers in science, engineering, and entrepreneurship. In the summer before their senior year, participants tackle advanced academic challenges, develop the skills necessary to achieve success in an increasingly globalized economy, and forge relationships with individuals from diverse racial, ethnic, cultural, and socioeconomic backgrounds. We will work with Interphase and MITES to incorporate access to and utilization of DataSpace into their curriculum – to highlight the ways that scientific endeavors can be leveraged through such extensive repositories of scientific data.

VII.4. Other NSF DataNet Partners

MIT recognizes NSF's desire to build a network among the DataNet projects. Our project is highly collaborative by its very design and intends to build a distributed network of partner organizations in a new virtual organization. To that end, we have communicated with the PIs of the Johns Hopkins University-led DataNet proposal and confirmed a mutual intention to work collaboratively, building on our existing relationship and activities. We have also communicated with PIs at the University of Washington and the University of North Carolina to collaborate and coordinate our activities if we are selected. Areas of collaboration would minimally include standards for a range of identifiers (e.g. URIs for dataset, data elements within datasets, metadata tags, data formats, and interoperability protocols across systems), and standards for data representation to enable interoperability across platforms and scientific domains.

VIII. Results from Prior NSF Research

The DataSpace project builds on many prior research projects as well as operations experience at MIT and our partner organizations. This prior work is described in detail in Appendix A6 and the papers listed in the References section. Prior work by the PI and co-PIs specifically relevant to the DataSpace Project and funded by the NSF in the past five years comprises:

- Integrating Data Management with Data Grids (NARA Collection on Persistent Archives, 2004-2005). Smith, NSF award no. 10168632-007, $174,859. To integrate DSpace with data grids, specifically SDSC's Storage Resource Broker (SRB), to provide long-term access and preservation of digital material.

- Developing Scalable Data Management Infrastructure in a Data Grid-Enabled Digital Library System (NARA Collection on Persistent Archives, 2005-2007). Smith, NSF award no.10254712, $621,203.

Methods for data management policy definition and exchange across distributed grid-based systems, primarily DSpace and iRODS.

- Transparent Accountable Datamining Initiative (the TAMI Project, 2005-2008). Abelson, NSF award no. 0524481 $1,370,000. Created technical, legal, and policy foundations for transparency and accountability in large-scale aggregation and inferencing across heterogeneous information systems.

IX. Summary of Key Features of the DataSpace Proposal

Intellectual merit: develops a general distributed infrastructure for data archives, and a model for institutionally-based support services for digital archiving, long-term preservation, and open use.

Scientific progress: Data are initially from life and environmental sciences (e.g., neuroscience and biological oceanography) with later broadening to others (e.g., high energy physics, plant biology) to insure interdisciplinarity and interoperability of the platform design. The initial disciplines represent very distinct scientific domains but share some common data types, such as genomic data. The absence of coordinated storage facilities, encoding standards, cataloging, curation, and retrieval methodologies leads to needless duplication of effort while at the same time inhibiting information exchange and important discoveries.

Open Access: Supports a vision of open access tempered by certain demands of privacy, property rights, data rights, etc., and the legal and policy framework to support this.

Exemplars: Develops a set of exemplar data archives at leading research universities that can be deployed at any research organization, for any discipline, and federated to advance science. Addresses the range of data life-cycle issues across a wide variety of disciplines and a long time span.

Some Unique issues incorporated: Builds on the successful DSpace platform and open source model of development and maintenance; proposes the creation of a new top-level internet domain (".arc"), analogous to the role of ICANN and W3C; addresses 'temporal semantics' recognizing that in addition to changes in software and hardware, there will be change in the meaning of the semantics of the data.

Diverse and expert personnel: The project personnel, from nine geographically distributed organizations, represent a diversity of skills (including life scientists, environmental scientists, computer scientists, library scientists and business management experts). This addresses the need for experts in both technical as well as science areas. Many are uniquely qualified experts in their specific fields.

Risk mitigation: *Research risk*: One or more of the senior personnel have extensive experience and, at least preliminary, results in the each of the research areas proposed. *Operational risk and sustainability*: The risk of system failure is managed through a distributed design and federated approach to policy. For sustainability, takes the approach that research data-generating institutions should play an active role in curating their own data, and that cross-sector infrastructure adoption will spread costs and risk beyond research institutions.

Assessment: Recognizes the need to assess the project technically, socially, and economically. Addresses the need for platform independence by designing standards and protocols, and by actively encouraging cross-sector adoption.

Public/Private Partnership: Brings significant technology corporate expertise into the research enterprise to insure realistic, real-world, scalable solutions and proven innovation-to-market strategies to minimize project risk.

Expert Advisory Board: Ten recognized experts from all fields represented in the project (i.e. science, law, business, technology, libraries, and digital preservation) will advise the project over its lifespan.

Organizational Structure and Plan for Expansion/Evolution: Well thought out structure and effective use of the three boards: Advisory Board, Management Board, and DataSpace Business Development Management Team.

Broad impact: Can serve as a model for creation and maintenance of archival data in a variety of scientific communities involving a variety of partners. Provides new ways to make it easier to incorporate data into coursework, such as through Open CourseWare, and linking underlying data to published research articles. Potential for outreach to minority and pre-college students to enhance and broaden their resources. Potential to improve the use of science by governmental decision makers. Preserving science for the future can have huge benefits to society. DataSpace is capable of being a truly transformational project

DataNet Full Proposal –DataSpace Project
Appendix A1: Sustainability Plan

Overview and strategy

The proposed DataSpace research follows four significant tracks: (1) technology development, (2) business development, (3) application and operations, and (4) technology transfer, education, and outreach. The business development effort will be headed by faculty and researchers from MIT's Sloan School of Management who have successfully used internal and external resources to develop and sustain numerous infrastructures, organization and companies. Here we present our initial approach to sustainability along with some of the resources and examples of how they will be used to further validate, evolve and assure a sustainable DataSpace.

In many ways, our initial DataSpace Sustainability Plan draws on the Internet experience. In abbreviated and simplified manner, the pre-Internet period could be described as follows:

- Large organizations (both scientific and corporate) developed proprietary computer networks. They had the financial resources to do so and sustain them. But these networks, besides being very expensive, were awkward to use, maintain, and did not (except in very limited ways) connect to other networks – even internally many of their networks were not interconnected.
- Small and medium organization largely did not have resources for such networks – except in very limited ways. The costs – both in terms of hardware and software resources as well as personnel expertise required – were too high to be sustainable.

The Internet changed the situation for both types of organizations. The impact on small and medium organizations was dramatic. By creating a world-wide infrastructure whereby the base costs, such as R&D, standards development, hardware and software, and operation, were widely shared, the cost of using the Internet dropped to levels that any organization – including individual users – was easily within reach. Even the large organizations benefited because (a) they could utilize the cost savings for other purposes in their organization and (b) they were much better able to interconnect their separate internal networks as well as connect to other organizations in ways never practical before.

It is important to understand how the costs of the Internet are widely dispersed. For example, each organization does pay for: (a) the cost of its own computers and internal networks, plus (b) the cost of connecting to the Internet – whereby (b) *collectively* covers the cost of operating the world-wide Internet. In almost all cases, (b) is a very small fraction compared with (a), and is, thus, an easy to justify and sustain expense. That is the model that we envision for DataSpace. The pre-DataSpace period has many similarities.

- Large research organizations are able to operate their own scientific data repositories. But these are usually very specialized and expensive. Often there are multiple, largely separate, repositories within large research organizations with minimal inter-connections and limited ability for interoperability and integration.
- Small and medium research organizations only have very limited resources for data repositories, if at all.

The proposed DataSpace will benefit both types of organizations. By extensive sharing of the base costs, every organization will be able to participate. By each organization taking on most of the responsibility for the storage and management of its own data, there will not be any huge central cost burden for NSF or any one organization. Furthermore, much like the Internet, the modest central costs will be distributed across a broad ecosystem of DataSpace participants.

A premise of the DataSpace project is that research-generating institutions, including research universities, have a key role in archiving and actively managing the research data produced by their own faculty and researchers. While many researchers have created archives of data they rely on, either at their own institution or across institutions that they collaborate with, these archives have uncertain futures and require constant planning to sustain over time, with a constant risk of complete disappearance if lead researchers leave or retire. The proposed model would allow researchers' departments or institutions to take the individual researchers' data and federate it with their collaborators' data to create a virtual archive of the data desired by the particular research community using that data.

Another major advantage of this approach is the possibility for sustainability of the data itself over time, irrespective of the organization that originally archived the data. Archives will always need exit strategies, since the data curators and their organizations change over time. In the print environment this was straightforward, but digital data archives require more sophisticated exit strategies that allow for the data to be transferred along with other tools such as visualization software, documentation, and expertise needed to interpret it. Defining standards for data encoding and documentation has the potential to significantly reduce the cost of transferring data between archiving organizations (or replicating the data) to lower risk and improve data sustainability.

Finally, we will make an early and aggressive effort to identify and promote the development of added-value business opportunities and applications. This will be significant as much of the sustainment of networks is correlated with the value and use of these networks. Promoting an active business environment around our DataSpace infrastructure will sustain this level of use and continued development and provide more organizations that will contribute to its costs and development.

DataSpace resources and organizational structures for economic and technological sustainability

Although we believe the basic strategy described will be successful, it is important to facilitate the success of DataSpace's economic and technological sustainability[1]. In this regard, we propose to utilize a number of resources and organizational structures.

It is our intention that research organizations can join the DataSpace global data network of archives as easily as they can join the Web. We will separate the concerns of physical data storage and low-level preservation, data curation and functional preservation, data access and visualization across different sectors and belonging to the community of practice that best serves each piece of the problem. The infrastructure will depend on a set of practices, standards and protocols that are independent of particular hardware or systems. We propose to create a new ecosystem and incentive system for a science infrastructure that will be transformative and sustained.

Organization Structures

International consensus: A key requirement for the long-term sustainability of DataSpace is an institutional mechanism to evolve and gain consensus support for the DataSpace technical and operational standards. The participants in our DataSpace proposal have extensive and successful experiences in developing and nurturing such global sustainable organizations. Some examples are:

World Wide Web Consortium (W3C): As the global body responsible for setting Web standards and defining the future architecture of the Web, the W3C offers expertise and an existing institution that can build consensus on DataSpace-derived standards. W3C could lead the standardization effort alone or in partnership with other standards groups in the library, publishing and archival communities. Timothy Berners-Lee, head of the W3C, is one the DataSpace senior personnel.

DSpace: DSpace is the open source software designed to implement the Open Archives Information System (OAIS) reference model for long-term digital archives. It is in use by approximately 500 research institutions worldwide for access to and long-term archiving of research output. It includes functionality for data deposit, management, discovery, and preservation, and assigns globally unique, persistent identifiers to archived material. It facilitates "open" access but supports controlled access and embargoes where necessary. Michele Kimpton, head of the DSpace Foundation, is one of the DataSpace senior personnel.

DataSpace Federation: As part of the DataSpace effort, we propose the formation of the DataSpace Federation. This will initially consist of the early adopters of the DataSpace technology and will evolve to be the sponsoring/coordinating body for the sustainment of DataSpace, much like the DSpace Federation. This virtual organization may be created as a separate legal entity or these activities may be combined with those of an existing, complementary organization with an established business model.

Research on Virtual Organization for DataSpace: In addition to the efforts and experience of the DataSpace senior personnel mentioned above (and other DataSpace senior personnel), we will also build on and conduct research on the possibility of new types of virtual organization infrastructures to support

[1] Note: There is significant overlap between economic and technological sustainability.

the success and sustainability of DataSpace with the assistance of key experts, such as Prof Wanda Orlikowski, an authority on virtual organizations and a member of our Advisory Board.

Business models for DataSpace archives and distributed organizations

The model for sustainability of DataSpace will depend on its distributed, multi-purpose nature, so that the cost is widely distributed and is not borne by scientific researchers alone. All the institutions that support research, as well as other industries that share many of the DataNet's infrastructure needs (e.g., financial services, healthcare), will only bear incremental costs of the infrastructure, much like the Internet. We propose to facilitate this process in several ways.

Other industries with common interests: Although the predecessors to today's Internet, ARPAnet and NSFnet, were primarily for government and research purposes, today much of the cost of the Internet is borne by diverse industries, ranging from manufacturing companies to financial services to online retail stores, etc. – and the scientific community is able to leverage on those common investments. We envisage that many industries have data, both scientific and non-scientific (such as financial data and sales data), that must be archived and processed over time – common needs and interests to DataSpace. To facilitate and accelerate this process, we will engage people from such industries. For example, Dan Schutzer, head of the Financial Services Technical Consortium (FSTC), is a member of our Advisory Board.

DataSpace infrastructure sustainability: Although most of the early research on the technologies that underlie the Internet was conducted by universities and funded by government agencies, the vast majority is now conducted by the corporate sector – by companies ranging from Microsoft (e.g., browsers) to Google (search), to Cisco (e.g., routers) to Akamai (e.g., data caching.) These companies included those that existed prior to the Internet but extended their scope onto the Internet (e.g., Microsoft), those whose business was dramatically accelerated by the Internet (e.g., Cisco Systems), and new organizations that were created as a result of the needs of the Internet (e.g., Google and Akamai.) We propose to help and encourage all of these types of organizations to take on various aspects of the long-term sustainability, especially the technical sustainability, of DataSpace. Part of the strategy for achieving this is to follow well-known open source software development practices to encourage early investment by other industries such as those mentioned in the previous section. The DataSpace team has extensive experience creating broad-based open source software communities that can be leveraged for DataSpace. The initial software will be a reference implementation of the DataSpace architecture and we expect to emerge with other implementations over the course of the project, perhaps optimized for particular data needs (e.g. large individual datasets vs. large numbers of small heterogeneous datasets, or highly structured data vs. unstructured data like images). The project's reference implementation will be deployed and tested initially by three of DataSpace external partners – Georgia Tech, Rice University, and Oregon State University – each of whom have been active participants in the DSpace community and have experience capturing and managing research data in their own environment. Also, Masdar, our initial international DataSpace node, will help to extend the impact and visibility internationally. During that time the DSpace Foundation will do extensive outreach to the DSpace community about DataSpace and develop training and testing opportunities for other institutions that are interested in deploying DataSpace data archives at their own institutions. This process worked well to build a community for the DSpace software in the research library community, and has worked well for other software systems in higher education.

Business plan development: MIT and its DataSpace partners have extensive experience and resources to facilitate the development of such business plans to produce viable technology sustainment. Some of the key resources include:

- The **MIT 100K Entrepreneurship Competition** (http://www.100k.mit.edu) is a leading business plan competition. The competition was founded in 1990 to encourage students and researchers in the MIT community to act on their talent, ideas and energy to produce tomorrow's leading firms. Entirely student-managed, the competition has produced hundreds of successful ventures. For example, the business plan for Akamai was a finalist one year. The success of this effort has led to affiliated competitions at many other universities.
- The **MIT Entrepreneurship Center** (http://mitsloan.mit.edu/faculty/research/entrepreneurship.php) has the mission to train and develop leaders who will make high-tech ventures successful. The interdisciplinary

center nurtures new ideas, novel approaches, and advanced technologies, while fostering continued competitiveness, success, and national and global prosperity. A related activity is the MIT Entrepreneurship Society (http://entrepreneurship.mit.edu/esociety.php) which has established an entrepreneurial support network among MIT faculty, staff, students, alums, and enthusiastic participants in MIT-related ventures. Members based in Boston, New York, and Silicon Valley meet and network regularly to advise and support each others' careers and ventures. All members pledge to donate between 3% and 25% of the shares of their new ventures to MIT, providing a stream of funds to support intellectual and material contributions. Prof. Ed Roberts, the founder of the MIT Entrepreneurship Center, is a member of the DataSpace Advisory Board.

- **MIT Entrepreneurship & Innovation Program** (E&I) (see http://entrepreneurship.mit.edu/E_and_I.php) is a part of the MIT Sloan MBA Program. The program focuses on launching and developing emerging technology companies. The Entrepreneurship Lab (http://entrepreneurship.mit.edu/elab.php) is a key part of the E&I Program. It is a semester-long course in which students work one day a week in a start-up company. Interdisciplinary teams of MBAs and engineering students are charged with helping to solve a real-world problem – such as the further development and sustainment of DataSpace technologies. Assignments range from conducting market research for a pre-IPO company, to participating in the creation of a marketing plan, to helping a software company develop high-level customer profiles. Prof. Ed Roberts, a member of the DataSpace Advisory Board, is the faculty head of the E&I program and Prof. Stuart Madnick, PI of the DataSpace initiative, has been a member of the E&I faculty advisory group.

The DataSpace Business Development Management Team (DBDMT)

DataSpace faculty and researchers, largely drawn from the MIT Sloan School and headed by Dr. Siegel, will form the DataSpace Business Development Management Team (DBDMT) and will work closely with the DataSpace Advisory Board and other relevant advisors to develop the DataSpace Sustainment business plan. The DBDMT will use internal (e.g., MIT Entrepreneurship Center and others described above) and external resources (e.g., DSpace federation). A few examples of how these resources can be used include:

(1) Promote a DataSpace viral social network. These networks have been actively utilized and examined at Sloan in research and development projects and have been very successfully used in the development of the DSpace federation. We plan to actively pursue this approach to DataSpace, building communities of interest, collecting ideas, and actively engaging key participants.

(2) Assist members of the DSpace federation to transition to DataSpace, become members of the DataSpace federation, and draw on their support for the concept of digital repositories to bring in new members and support.

(3) Fully utilize the resources and support of the DataSpace Advisory Board to promote DataSpace broadly – both nationally and internationally, academic and commercial – so that it gains a universal acceptance comparable to the Internet and Web and ensured continual development through both non-profit and commercial interests.

(4) Develop educational, training, and high-visibility opportunities such as workshops, conference presentations, and other programs attracting participants, research and further building communities.

(5) Actively engage in Entrepreneurship Competitions, such as the 100K competition (both at MIT and/or other universities), to assist in the development of the long-term DataSpace business model and also to increase the visible and acceptability of DataSpace.

(6) Utilize the active innovation and entrepreneurship culture and project-based activities of the Entrepreneurship and Innovation program to promote student class projects, theses, and other graduate student programs to support developments that contribute to our Sustainability Plan.

(7) Incorporate modeling techniques (such as the use of System Dynamics) to understand and improve DataSpace evolutionary processes while avoiding unintended and undesirable consequences.

(8) Involve MIT Marketing students to develop approaches to promote usage through class projects.

The DBDMT active engagement in the business development process, through efforts listed above and others, and the effective usage and deployment of DataSpace will ensure its long-term sustainability.

DataNet Full Proposal –DataSpace Project
Appendix A2: Management Plan

DataSpace Project Management

The DataSpace Project will be led by the PI and co-PIs (see Appendix A4 on Key Personnel and the Biographical Sketches) and managed by a dedicated Project Director to be hired at the commencement of the project. The Project Director will be recruited by, and report to, the Project's PI, and will be a research staff member of the MIT Sloan School of Management (with a dotted line report to the Sloan School's Finance and Administration department). To hire the Project Director we will establish a search committee comprised of the PI and co-PIs, the Director of the MIT Libraries, and an HR representative from the Sloan School of Management. The committee will follow standard MIT recruitment policies and procedures for the position, including conformance with FLSA and EEO policies, and following MIT's standard affirmative action plan.

The DataSpace Project Director will be responsible for the day-to-day business operations and administration of the overall organization, including management of infrastructure requirements, administrative staff, production of management reports; and oversight of the financial operation of the Project, organization of events, and similar non-technical activities. The Project Director will also have primary responsibility for fundraising, public relations, and distributing information on the results of the research and development undertaken. The Project Director will spend a considerable time meeting with the various constituents of the Project, especially the partner organizations, and will have primary responsibility for the finance and operations of the Project.

The Project Director will be a highly experienced project manager but is not required to be an expert in the research or operational domains of the DataSpace Project (i.e. technical architecture, scientific research data management, data interoperability, or long-term data curation). That expertise will come from the PIs and the Senior Personnel of the Project so that the Project Director will have primarily a coordinating and managerial role. The Project Director will also have prior experience with creating new organizations and developing business plans.

To assist the PIs and the Project Director in ensuring the effective management and evolution of the DataSpace Project, we will create a term-based Management Board and Advisory Board, as well as a long-lasting continuous DataSpace federation virtual organization and governance body.

- *Project Management Board*: The Project's management board will consist of representation by the PIs and Senior Personnel from each of the organizations in the proposal, and will meet at least monthly to review progress and set near-term and long-term goals and plans for the Project. A key goal of the management board is to ensure good collaboration and communication between and among the domain experts and the DataSpace researchers and operations teams. Since the Senior Personnel are not all co-located at MIT, we will establish means to meet virtually, and will communicate regularly via network-based infrastructure (see the discussion of communication management below).

- *Project Advisory Board*: To assist the Project's leadership, we will also establish an Advisory Board consisting of a diverse ensemble of knowledgeable leaders from throughout the world and different industries (including corporate as well as scientific) to ensure that "best practices" are understood and adopted. The Advisory Board will meet at least twice a year and will be available for advice on an ongoing basis. Besides providing insights to the Management Board, the Advisory Board will be one of the mechanisms to achieve outreach to communities that are potential users and supporters of DataSpace.

There will be a minimum of ten members of the advisory board, and a maximum of fifteen. Members will include representatives from higher education, scientific research (life sciences and environmental sciences initially), law and public policy, finance, management and entrepreneurship, technology, informatics, and digital curation/preservation. A number of notable individuals have agreed to join the Project's Advisory Board (listed in Appendix A5) and to serve as unfunded collaborators.

Project Research Management

A large component of the DataSpace Project's work involves conducting research on the many aspects of scalable, sustainable data management and archiving that are not yet well understood. These are enumerated in the Project Description section on activities and the DataSpace research agenda.

These research activities will be conducted at MIT, EMC, Google, HP Labs, the Science Commons, and the Masdar Institute of Science and Technology. The DataSpace research activities will be overseen by the Project's PI and coordinated by the Project Director. The Project will establish collaboration mechanisms early on, via the Project website, wiki, and other such tools. We anticipate regular meetings of the research personnel at MIT and including external partners via remote connections and the Project Director will collect documented research findings on a monthly basis.

Oversight and Accountability Mechanisms

Oversight and accountability for the DataSpace project are the primary responsibilities of the PIs, but a key measure of and method for evaluating their success will come from the scientific domain experts. "Challenge problems" (i.e., valuable scientific research activities, consistent with the DataSpace goals, which either cannot be done or that are extremely difficult to accomplish with currently available technologies) will be defined in consultation with the co-PIs from MIT, colleagues at our partner institutions, and project advisors. Progress towards overcoming these challenges will be measured, as well as adjustments and additions. The success of DataSpace in overcoming these challenges will not only provide accountability but will also assist in the development of the business case for DataSpace to the broader scientific community.

Project Operations Management

The vision of the DataSpace Project is to create a new, highly distributed virtual organization consisting of any and all research-generating organizations that need to manage data produced or consumed by their members. Research universities, national research laboratories and libraries, and private research companies are all examples of organizations that have responsibility for research data and should be able to manage and curate that data locally or via contract with third party service providers. Because of this vision the DataSpace Project does not define a single operational model or service model that will fit all cases. We will instead provide the necessary infrastructure (architecture, standards, and reference implementation) to enable any organization to take on data curation activities, and will define a small set of operations exemplars to demonstrate and document possible service models and associated costs models. Exemplars will include functional service models for organizations that do not choose to establish their own. With these exemplars established, the Project will perform outreach to the research community to promote the vision and encourage further adoption.

As a key project deliverable, the exemplar operations will be built and documented at MIT, Rice University, Georgia Tech, Oregon State University, and Masdar. At MIT, the operation will be managed jointly by the Information Services and Technology (IS&T) department and the Libraries. At Rice University, Georgia Tech, and Oregon State University the operations will be established and managed by the Libraries. Masdar is so new that the management structure has not yet been determined and we will work with them to determine the best options. To serve as diverse examples each of these operations will be developed independently, exploiting existing relationships within each institution that best serves the needs and capabilities of the institution and its members. For example, at some institutions the DataSpace operation may be managed entirely within the Library, or entirely within the existing IT department or data center, or by some combination of both (as will be the case at MIT). Some institutions will form consortia to set up their central operation – either with one member hosting the service for the group, or by a sub-federation of nodes that interoperate. Other institutions will outsource the hosting of the entire system (or parts of the system, such as the storage subsystem) to third party commercial or non-profit providers. The Project will identify such third party providers and make them known to potential DataSpace adopters as part of its coordination and outreach activities. Finally, some institutions may choose not to provide a central service for data archiving, but rather rely on individual departments or labs to run their own local archives. Since the architecture will be distributed and federated, one of our design objectives will be to support this scaling of operations from an individual researcher to a consortium of large institutions or even centrally managed national data archives, as proposed for Masdar.

Virtual Organization Management

While we believe that a scalable, sustainable data archiving platform must be distributed and autonomous – not centrally provided or managed – we recognize the need for a governance structure for this virtual organization, currently referred to as the DataSpace federation. So another key deliverable of the Project is the definition of a governance plan for the future. In addition to the PIs, several of the

DataSpace Partners will contribute to that plan, including the DSpace Foundation (which has established just such a governance model for the current community of DSpace-adopting organizations) and the W3C (which currently provides technical governance for the Web community). The governance body we establish will be lightweight, and will not directly provide operations or services (at least initially) in order to encourage the growth of services within research-generating institutions and by third party service providers. Examples of the activities that might be found appropriate for the governance body are:

- Establishing and maintaining technical standards for the DataSpace system architecture, data models, communication and security protocols.
- Establishing and maintaining standard ontologies for data description (i.e. metadata) about different types of research data, or for the data itself where appropriate.
- Establishing and maintaining policies for data archiving activities that member organizations could choose to adopt (e.g. drawn from the Trustworthy Repositories Audit & Certification: Criteria and Checklist produced by the Research Libraries Group and the US National Archives and Records Administration).
- Creation of an official Archives Internet domain (.arc) that could be granted to data archives based on some accreditation process.
- Certification or identification of trusted third party service providers such as persistent storage providers, data migration or reformatting services, or system hosting services.
- Provision of training on system implementation, customization or operation; on data management policies and practices; on service and cost models.

Although we have identified important and promising activities for the DataSpace federation above, the detailed list as well as the internal structure, governance, and operation of the DataSpace federation are important research issues to be resolved in the course of the project.

Education and Outreach Management

Outreach to the community of potential DataSpace platform adopters will be performed by all members of the DataSpace Project team, including its advisory board, but will be the particular focus of the DSpace Foundation. The Foundation will hire a technical advocate/outreach coordinator to perform education and outreach activities to the current DSpace community and beyond. The advocate will be retained in years three through five of the Project to create awareness of the DataSpace work, promote and be the technical advocate for the new platform, develop migration strategies for current archives, organize training and education opportunities, and provide feedback from platform adopters to the DataSpace Project team. This will be done via existing user group meetings, conferences, seminars, and newsletter and journal articles across diverse sectors. We will initially focus on the scientific research community and its current organizations (e.g. research universities and government research labs) but will broaden the outreach to include other sectors as opportunities are identified by the Project team and the Advisory Board.

Communication Management

The DataSpace Project is characterized by being geographically distributed, and broadly inclusive. While some of the Project's personnel are located at MIT, there are partners at other universities and research centers, both in the United States and other regions of the world (e.g. the Masdar Institute of Science and Technology in Abu Dhabi). Because of this, and because of our commitment to building a successful, distributed virtual community and organization as an outcome of the Project, we will invest significant effort into designing and implementing an effective, open virtual communications infrastructure.

A number of meetings and other communications will be conducted virtually via network-based infrastructure (e.g. a Project Web site, wiki, mailing lists, IRC, and telephone or video conference calls) with records kept and made public to encourage interest and participation in the Project by outside organizations and individuals. The Project PIs and Senior Personnel are all very experienced at working with globally distributed, virtual project teams, and can demonstrate effective success with this approach. For example, the working practices of the W3C – a highly distributed, virtual organization – are well established and successful, and can serve as a model for the DataSpace Project's efforts. In addition, each organization will have a designated point of contact to simplify coordination.

DataNet Full Proposal –DataSpace Project
Appendix A3 – Cyberinfrastructure Capabilities

The DataSpace Project builds on the idea that many research data-generating organizations have the infrastructural capability of running local data archives already, but lack a set of agreed upon standards, protocols, and readily-available technologies and support services to create a distributed, federated "data Web". So, this project will create a set of exemplar implementations of a new data archiving system that leverages existing infrastructure and builds it out in ways that promote data interoperability and long-term access. MIT, Georgia Tech, Masdar Institute of Science and Technology, Oregon State University, and Rice University will each implement a DataSpace node to test the infrastructure and develop a service and cost model optimized to the particular institution. These nodes will be federated to test our vision for shared data collections, and then we will broaden the federation to include additional institutions and other types of data-producing organizations. Federated services may include data discovery (e.g. search), delivery, integration, analysis and visualization.

Because of this distributed model, we cannot specify the exact cyberinfrastructure capabilities of every potential DataSpace site, and there should be a range of capabilities and costs supported by the architecture. We also assume that sites can leverage emerging "cloud" infrastructure to supplement the local cyberinfrastructure capabilities, so that detailing current capabilities is misleading. At present, we can only describe the core capabilities available at MIT and the partner organizations.

I. MIT cyberinfrastructure capabilities

Production hardware and software proposed will be developed and operated by staff based at the MIT Libraries and MIT Information Services and Technology (IS&T) department (MIT's central information services organization), in collaboration with our partner organizations. DataSpace hardware and software for MIT use will be housed in MIT IS&T facilities under the management of project personnel.

- **Infrastructure**

The DataSpace system proposed will be able to operate in a distributed fashion within an organization such as MIT (i.e. processing nodes may be established at each participating database site), although it is likely that one or more "control nodes" will be designated and several processing nodes will also be operated centrally. The initial control nodes will consist of a number of high speed servers configured in a Linux cluster (exact number and specifications to be determined at the time of purchase based on the best available technology at that time). MIT's IS&T department currently operates over 100 such clusters from manufacturers such as Sun, Dell, Apple, and Hewlett-Packard ranging in size from several computers up to hundreds of computers. Limited storage (approximately 1 TB) will be needed for local processing at the control node, and storage will be maintained for some research databases for testing purposes. MIT IS&T will also be able to host processing nodes for individual research projects on a contract basis should individual faculty at MIT or elsewhere desire this type of "co-location" service.

In addition to the infrastructure resources managed by MIT's IS&T and to be used for the DataSpace "control nodes" and "co-location" storage services, most of the collaborating research groups (i.e., the Martinos Imaging Center at the McGovern Institute of MIT and the Center for Advanced Brain Imaging at Georgia Tech) have computing, storage, and networks resources of their own. Since the funding and operation of these facilities are pre-existing and they have been partly described in the Project Description section, we will not elaborate upon them here – other than to note that they will be incorporated into the federated architecture of DataSpace.

- **Networking and access**

Each computer in the DataSpace central node cluster will have two or more network connections to the internal network of the MIT data center. The data center's internal network allows configuration of multiple sub-nets each operating at 1Gb/second connecting with MIT's backbone network currently operating at 10Gb/second. [Planned upgrades during the life of the NSF contract will raise these speeds to 10Gb/second and 40Gb/second respectively.] Each sub-net can additionally be configured with several “virtual networks” should our research indicate that isolation of network traffic between or among different servers is desirable.

The internal MIT network connects to the Internet through multiple paths including Internet 2, three different commercial Internet service providers (ISPs), through a common network ring among several other universities in the Boston area, and via MIT’s optical network to a major Internet interconnection point in New York City. Aggregate capacity from MIT to the Internet is 15Gb/second, with regular upgrades planned to ensure that average load is never more than 10% of capacity.

During the life of this contract, it will be possible to have a dedicated connection from the DataSpace control node at MIT to Internet 2 and other external networks at speeds of 10-40Gb/second. These speeds will allow direct data transfers of sizable databases as well as all control information to manage a distributed worldwide network.

- **Service, fail-over, and archiving**

All hardware will be housed in MIT’s advanced data center with redundant power and cooling designed to meet specifications for Tier II (N+1) standards of the Up-time Institute. MIT also maintains long-term leased space in a commercial Tier III data center (expected site availability of 99.982%) where redundant DataSpace equipment can also be located. In addition, we expect to locate reduced capacity redundant hardware and software at another MIT-leased site at least 100 miles distant from the primary site.

Software configuration will allow dynamic designation of additional control nodes so that control nodes can be designated at non-MIT sites. Control data will be replicated in real-time with a target replication delay of no more than two seconds among all control nodes. Fail-over will be based on standard “heart-beat” control protocols with transfer of control in less than two seconds and with no expected loss of control data. Control data will be archived on secondary (disk) and tertiary (tape) storage for analytical purposes.

- **Hardware and software replacement/development cycles**

Current hardware technology suggests a standard 3-4 year replacement cycle for computer hardware and a 4-5 year replacement cycle for networking hardware. The DataSpace proposal allows for one replacement cycle during the initial five year contract term. While we can generally predict the direction of hardware for this period of time, we cannot specifically know whether these replacement cycles will continue past the initial five year contract. We will re-evaluate replacement cycles in applying for contract renewal, with the possibility of shorter replacement in order to achieve effective cost-benefit for operation of DataSpace.

Software development and replacement cycles vary considerably more than do hardware replacement cycles. The overall “service oriented´ architecture of the DataSpace system is designed to allow incremental replacement of components as new or better modules become available. Since most of the software used in DataSpace is expected to consist of open-source components (either developed and contributed to the open source community by the DataSpace efforts or from other sources), and since there are very active communities working in the areas of data storage and control, it is expected that we will have a fairly active replacement cycle as well. New releases of DataSpace processing and control node software will designate specific standards and recommended modules complying with those standards.

- **Computer and information science and engineering capabilities of project personnel**

Key project personnel include world-class computer and information science and software engineering researchers and experienced software and system architects and engineers, as elaborated in their biographical sketches. Dedicated architects and engineers will be hired for the DataSpace Project to work with the senior personnel from within the Libraries and the IS&T organization to develop the initial reference implementation of the new DataSpace architecture. The DataSpace project will additionally use the existing services of IS&T for development, operation, and other IT-related services. MIT IS&T consists of almost 350 people with experience as system programmers, network engineers, client support specialists, and other related IT functions. The average experience level of MIT IS&T personnel in areas related to the DataSpace project is over 10 years.

- **Strategy for staying at the leading edge of evolving technologies**

The modular service-oriented architecture of the DataSpace system is designed to allow rapid change and advancement of the system. Personnel on the DataSpace project, as well as experts in MIT IS&T, are constantly scanning the environment for significant advances that may have a bearing on DataSpace. This awareness of evolving technologies, as well as a software architecture designed to accommodate rapid evolution, will ensure that the DataSpace system provides leading edge data capabilities to researchers throughout the country and world.

II. DataSpace Partner Institutions cyberinfrastructure capabilities

All of the DataSpace partner institutions (Georgia Tech, EMC, HP Labs, Masdar Institute of Science and Technology, Oregon State University, Rice University) have extensive cyberinfrastructure capabilities that will be utilized in this effort. Due to time and space limitations, we will only describe some of the unique facilities being provided by EMC and HP Labs.

II.1 EMC cyberinfrastructure capabilities

EMC has extensive cyberinfrastrucrure capabilities. Of particular relevance to the DataSpace effort, EMC will provide access to on-line clusters of its EMC Centera data archiving platform. MIT will be able to test against these clusters, which support the industry standard XAM protocol. This will enable extensive evaluation and feedback to the DataSpace researchers on performance and also provide a reference point for testing XAM-based implementations of data management.

EMC also has a research test lab that prototypes advanced technologies. One of the recent projects from this lab is a virtual-machine-based version of the EMC Centera platform. The virtual-machine-based version provides greater flexibility in deployment and offers another reference point for testing data management protocols in DataSpace against a clustered computing environment. The facilities of this EMC test lab will be available to support DataSpace experimentation.

II.2 HP Labs cyberinfrastructure capabilities

HP Labs also has extensive cyberinfrastrucrure capabilities. Of particular relevance to the DataSpace effort, HP is a key partner in the Open Cirrus joint initiative -- sponsored by HP, Intel, and Yahoo. The Open Cirrus testbed is a collection of federated datacenters for open-source systems and services research. As shown in Figure 1, the initial testbed is composed of six sites in North America, Europe, and Asia. Each site consists of a cluster with at least 1000 cores and associated storage. Authorized users can access any Open Cirrus site using the same login credential.

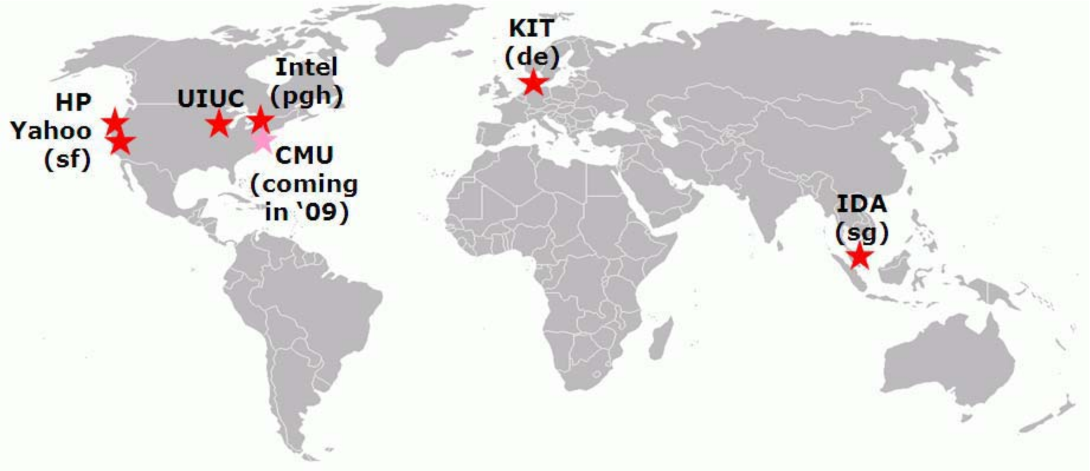
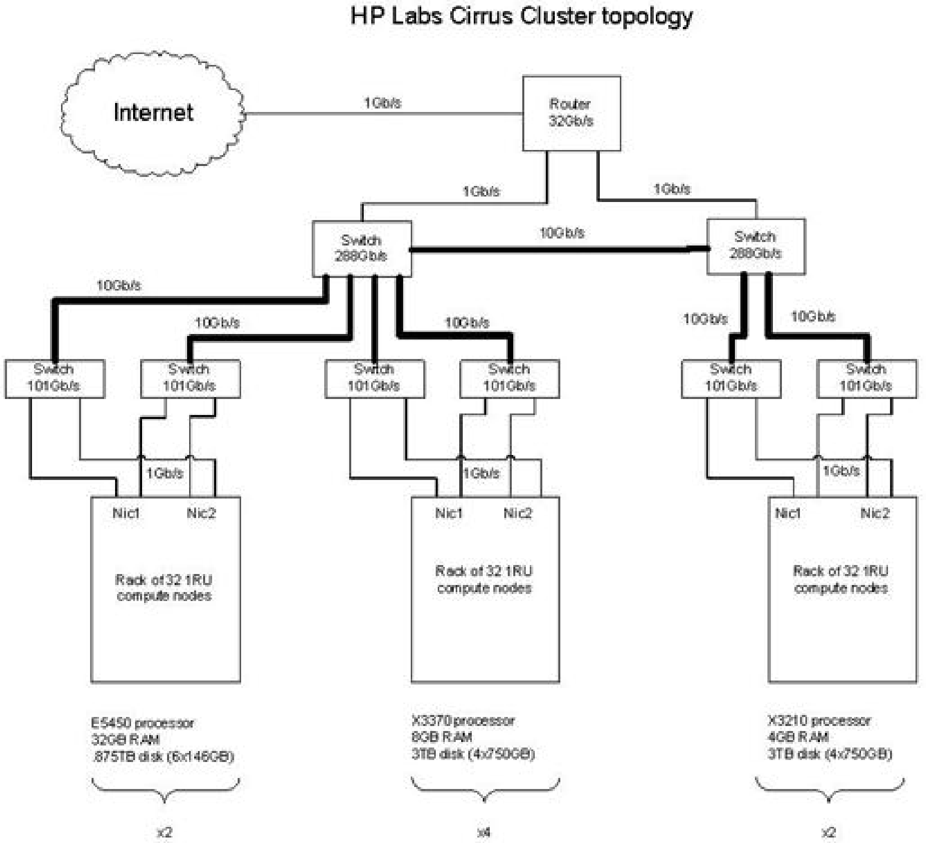


Figure 1. Open Cirrus testbed circa Q1 2009.

The Open Cirrus project focuses on research in all aspects of datacenter federation and datacenter management, new interactive services, new data-intensive applications, and the infrastructure for such services and applications, and will be actively soliciting such projects from the community.

During the DataSpace Project, MIT project personnel will be able to work with their HP Lab collaborators in IIML to design and test various DataSpace configurations using the HP Open Cirrus Testbed. The current HP Testbed architecture is depicted in Figure 2.



				Totals
Nodes/cores:	64/256	128/512	64/256	256/1024
RAM (GB):	2048	1024	256	3328
Storage (TB):	56	384	192	632
Spindles:	384	512	256	1152

Figure 2. HP Open Cirrus testbed circa Q1 2009.

DataNet Full Proposal –DataSpace Project
Appendix A4: Key Personnel

PIs and Co-PIs

Stuart Madnick, the Principal Investigator, is the John Norris Maguire Professor of Information Technology at the MIT Sloan School of Management and Professor of Engineering Systems at the MIT School of Engineering; co-Head, Total Data Quality Management (TDQM) Program; co-Head, MIT Productivity from Information Technology (PROFIT) Program. Professor Madnick has had a long-term research agenda to address ways to integrate information systems, giving organizations a more global view of their operations. He has led a project to develop new technologies for gathering and analyzing information from many different sources including conventional databases and the World Wide Web. Building on his existing base of research and publications, he will take primary responsibility for DataSpace research related to data semantics, data mediation and conversion, data quality, and data interoperability and integration. In addition, in his role as PI, he will take responsibility for the overall coordination and direction of the DataSpace research efforts.

The co-PIs for the DataSpace Project are:

Hal Abelson is the Class of 1922 Professor of Computer Science and Engineering in the MIT Department of Electrical Engineering and Computer Science, and the Computer Science and Artificial Intelligence Laboratory (CSAIL). Professor Abelson was a founding member of the CSAIL Decentralized Information Group to explore the technical, institutional, and public policy questions necessary to advance the development of global, decentralized information and will apply his expertise to addressing these aspects in the DataSpace environment.

Jerry Grochow is MIT's Vice President for Information Services and Technology. Dr Grochow is responsible for the management and operation of all information services at MIT, and prior to joining MIT had 30 years of experience in technology management for government, industry and nonprofit organizations. Dr Grochow's organization will be leading the deployment of the DataSpace platform and its initial operation at MIT, together with the MIT Libraries.

MacKenzie Smith is the Associate Director for Technology in the MIT Libraries. She has extensive experience developing and deploying open source data archiving software for open access and long-term preservation (e.g. the DSpace program) as well as research and development experience in federated, policy-based data management, Semantic Web models for data access and interoperability, and long-term preservation of new digital data. She will be responsible, in collaboration with Dr. Grochow, with the supervision of the DataSpace development and deployment team.

John Gabrieli is the Grover Hermann Professor in Health Sciences and Technology and Cognitive Neuroscience at the Department of Brain and Cognitive Sciences and Harvard-MIT Division of Health Sciences and Technology, an Associate Member of the McGovern Institute for Brain Research and Director of the MIT Martinos Imaging Center. Dr. Gabrieli will play a key role in providing research data from neuroscience research, and defining and coordinating the DataSpace efforts related to use of this data in the neuroscience field.

Ed DeLong is Professor in the Division of Biological Engineering and the Department of Civil and Environmental Engineering, co-Director of the Center for Microbial Oceanography: Research and Education (C-MORE), and a member of the National Academy of Sciences. He will play a key role in providing metagenomics and related data from his own research, and defining and coordinating the DataSpace efforts related to use of data in the biological oceanography field.

The DataSpace PIs collectively reflect the combination of scientific research, data management, and IT operations expertise that are required to achieve the DataSpace program goals.

Senior Personnel at MIT

Sir Timothy Berners-Lee is the inventor of the World Wide Web and the Director of the World Wide Web Consortium (W3C) and the World Wide Web Foundation, co-Director of the Web Science Research

Initiative (WSRI), and the 3COM Founders Professor of Engineering in the School of Engineering at MIT. Berners-Lee will research the architecture and standards for the DataSpace platform, and its support of Semantic Web standards for data and policy management.

David Karger is Professor in the MIT Electrical Engineering and Computer Science Department and the Computer Science and Artificial Intelligence Laboratory (CSAIL). His research is in the field of information retrieval and analysis of algorithms. He has also spent some time working at Akamai and consulting for Google and Vanu Inc. Professor Karger will work on technology to incentivize scientists to archive and share their research data, and to visualize that data with lightweight Semantic Web-based tools.

Thomas Malone is the Patrick J. McGovern Professor of Management at the MIT Sloan School of Management, the founding director of the MIT Center for Collective Intelligence (CCI), and one of the two founding co-directors of the MIT Initiative on "Inventing the Organizations of the 21st Century". His research focuses on how new organizations can be designed to take advantage of the possibilities provided by information technology. He will work on new models for public engagement with research, leveraging data collected by the DataSpace effort and informing the DataSpace architecture in the area of public outreach and engagement for collective decision making on a global scale.

Michael Siegel is a Principal Research Scientist at the MIT Sloan School of Management and the MIT Productivity from Information Technology (PROFIT) Program. His research includes modeling of systems for assessment and improvement of operations, the use of information technology in health care systems, financial risk management and global financial systems, software benchmarking for financial risk management, applications of computation social science to analyzing state stability, digital business opportunities, ROI analysis for online financial applications, heterogeneous database systems, managing data semantics, query optimization, intelligent database systems, and learning in database systems. He will have two major roles in the DataSpace Project: he will (1) participate in research and development on semantic integration, and (2) direct much of the sustainability effort, providing both business development and outreach research and direction.

Daniel Weitzner is the Technology and Society Policy Director for the World Wide Web Consortium (W3C), a member of the Web Science Research Initiative (WSRI), and co-Director of the Decentralized Information Group in the MIT Computer Science and Artificial Intelligence Laboratory (CSAIL). He is a leader in the Internet public policy field, and researches architectural scalability for the Web, Web-based policy models, and infrastructure to build global consensus on technical designs and community operating conventions. Weitzner will contribute to the DataSpace research on scalable distributed architecture for data archives and Web-based policy management, as well as advising on virtual organizational models and best practices.

Senior Personnel at Partner Organizations

John Erickson is a principal scientist with HP Labs working in the area of enterprise informatics, and his research focuses on distributed, policy-driven data management. Dr. Erickson has been closely involved with research on the current DSpace platform, and is also working on personalization tools to motivate reposit of data into archives. Dr. Erickson will work on the distributed management architecture of the DataSpace platform.

Alon Halevy heads the Structured-Data Management Research Group at Google Inc. His group is focused on research related to integrating data from multiple sources and resolving schema heterogeneity. He has championed an effort called Open Information Integration (OpenII), in collaboration with other organizations, such as IBM and MITRE, to develop an open-source set of tools for information integration. He will help to facilitate and guide the effective use of OpenII in the DataSpace effort.

Geneva Henry is the Executive Director of the Digital Library Initiative at Rice University and will be responsible for managing Rice's participation in the MIT DataSpace project, working with the MIT DataSpace team to ensure that their scientific data needs are understood by the system architects. She will oversee Rice's implementation of DataSpace, coordinate with the rest of the DataSpace partners, and oversee project planning, implementation, budget and reporting of this project in coordination with MIT.

She will also work with faculty at Rice in the life sciences and energy/environment fields to gather requirements and capture current workflows for how these scientists do their research.

Meichun Hsu is a Lab Director with HP Labs where she researches large-scale data warehousing, data mining and scientific databases, data-intensive parallel computing, service-oriented integration, business process and workflow computing and transaction management. She will collaborate with the DataSpace Project on database architectures for scientific research data, highly scalable computation services that can be easily composed and applied to fresh or archived data sources whereby the results and the derivation processes can be automatically preserved in DataSpace, and advanced data visualization to allow scientists to navigate and drill down to data attributes of individual records.

Michele Kimpton is the Executive Director of the DSpace Foundation. Michele was formerly the Director of the Web Archive at the Internet Archive, and has expertise in scalable, long-term digital archiving systems, as well as developing new organizations in this domain. She will provide leadership for the education and outreach activities of DataSpace, as well as advising on potential governance and business models for the new virtual organization we create.

Joe Pato is a Distinguished Technologist with the Systems Security Lab at HP Labs. He has previously served as Chief Technology Officer for Hewlett-Packard's Internet Security Solutions Division. Currently resident at MIT as a Visiting Fellow with the Decentralized Information Group, Pato is researching the combination of trusted systems and information accountability to enhance collaboration. He will explore the inclusion of flexible security models for the DataSpace Project leading to effective sharing of scientific data as well as appropriate protection of provenance over the complete lifecycle of that data.

Terry Reese is the Gray Chair for Innovative Library Services at Oregon State University and will be responsible for managing OSU’s participation in the DataSpace project, working with the MIT DataSpace team and other partner institutions to communication their scientific data requirements to the development team, and to design the local data curation service for OSU researchers. He will oversee their planning, implementation, service development, assessment and business planning. He will also coordinate the federated research between MIT and OSU biological engineering researchers in the first phase of the project.

Stephen Todd is a Distinguished Engineer at EMC where has been responsible for architecting and delivering storage software products for many years. He has documented the John F. Kennedy Library's digital archiving solution and is an active participant in EMC's global Innovation Network. He will provide guidance and leadership on the DataSpace storage architecture and coordinate access and usage of EMC infrastructure resources for the testing and evaluation of DataSpace storage architectures.

Tyler Walters is the Associate Director for Technology and Resource Services at the Georgia Institute of Technology Library and Information Center. His expertise is in library technology, digital library programs, electronic resources management, metadata, and archives and records, and he is co-PI for the MetaArchive Cooperative, one of the digital preservation partnerships with the Library of Congress’ NDIIPP. Tyler will oversee Georgia Tech’s implementation of the DataSpace platform, their engagement with GT researchers, and develop a localized service and cost model to serve as an exemplar to other institutions.

John Wilbanks is the Executive Director of the Science Commons, and has a background in bioinformatics and Semantic Web applications in the life sciences, and the policy and legal issues surrounding data sharing. John will work with the DataSpace Project on policy contracts and technology to increase access to research literature and materials, and increasing the utility of online data in the context of the DataSpace platform.

Wei Lee Woon is Assistant Professor of Computer Science at the Masdar Institute of Science and Technology in Abu Dhabi. His research is in the area of text/document analysis, biomedical data mining, image processing and technology forecasting. In addition coordinating the testing and operation of the DataSpace platform at the Masdar Institute, he will collaborate with the DataSpace research team on data analysis and conversion problems.

DataNet Full Proposal – DataSpace Project
Appendix A5: DataSpace Project Participating Organizations and Sectors

Massachusetts Institute of Technology

MIT will provide the overall leadership to the project, identify important science challenges, perform the core research and development following the agenda defined in the project description, and will build a prototype institutional data archiving operation. The project will leverage not only explicit members of the project team, but the relevant diverse resources and research of the university towards the project goals. In particular,

- **Department of Brain and Cognitive Sciences (BCS)** stands at the nexus of neuroscience, biology and psychology combining these disciplines to study specific aspects of the brain and mind including: vision, movement systems, learning and memory, neural and cognitive development, language and reasoning. Researchers in BCS will provide expertise and challenge problems related to neuroscience to the DataSpace effort.
- **Department of Civil and Environmental Engineering (CEE)** seeks to understand natural systems, to foster the intelligent use of resources and to design sustainable infrastructure systems. **Department of Biological Engineering (BE)** has the mission of defining and establishing a new discipline fusing molecular life sciences with engineering to advance fundamental understanding of how biological systems operate and to develop effective biology-based technologies for applications across a wide spectrum of societal needs including breakthroughs in diagnosis, treatment, and prevention of disease, in design of novel materials, devices, and processes, and in enhancing environmental health. Researchers from the Department of Civil and Environmental Engineering and the Department of Biological Engineering will provide combined expertise and challenge problems related to Biological Oceanography to the DataSpace effort.
- **Sloan School of Management** will provide overall leadership for the DataSpace Project and specific expertise on several important DataSpace topics, such as data integration strategies, incentives for data sharing and re-use, data governance policies, and sustainable business models.
- **Libraries** will provide domain experts working directly with scientists to identify ontologies and other standards for data description and encoding; to perform ingest, description, and linking activities; and to define curation policies for the archives, and curate the data in centrally-managed archives. They will oversee project engineers developing DataSpace software.
- **Information Services & Technology** will provide expertise on the design, development and operation of enterprise cyberinfrastructure including software, hardware and storage architectures, network and security infrastructure, system support, database administration, and other core enterprise operations.
- **Computer Science/Artificial Intelligence Laboratory (CSAIL) and Engineering Systems Division (ESD)** will perform research and prototyping of new technologies required for the DataSpace infrastructure, including data visualizations and visualization frameworks, data interoperability standards and protocols, policy and curation frameworks, and innovative database designs.

Rice University conducts significant research in many areas in common with MIT, as well as using and contributing to the current DSpace digital archive platform. Library staff from Rice will inform the DataSpace Project's requirements, test and provide feedback on the DataSpace platform, and work with local scientists (e.g. high energy physicists) on additional data formats and requirements than those specified by MIT's local science domain experts. This will help to insure cross-disciplinary effectiveness of the DataSpace infrastructure across the range of science and other disciplines. Rice will also generate an institutional service and cost model for its DataSpace data archiving activities to inform the project's education and outreach activities.

Georgia Institute of Technology also has many scientific research areas in common with MIT, notably neuroimaging research, and is also an active participant in the current DSpace federation of institutions. The Georgia Tech Libraries will engage researchers in the fields of biology and biomedical engineering, initially, to participate in the data modeling for DataSpace and requirements for its use and management. They will test the DataSpace infrastructure and develop a local service and cost model for their campus that can be disseminated to other organizations as part of our education and outreach. Georgia Tech staff are experienced in digital preservation activities through, for example, their partnership with the Library of

Congress's National Digital Information Infrastructure and Preservation Program via the MetaArchive Cooperative, and they will advise DataSpace Project on digital preservation strategies.

Oregon State University conducts scientific research in other areas in common with MIT (notably biological oceanography) and was an early adopter and frequent contributor to the original DSpace repository system. OSU will, along with MIT, Rice, and Georgia Tech, form the core of institutional partners implementing and testing the DataSpace platform with a group of local research scientists to provide user requirements and feedback on the platform, and a model data curation service. OSU staff are currently gaining experience with scientific data curation and are active in national digital library efforts.

DSpace Foundation

In 2007 MIT and HP jointly created the DSpace Foundation, a non-profit organization to provide leadership to, and coordinate the activities of, the now approximately 500 organizations world-wide using and contributing to the DSpace software and related activities, to manage the long-term development strategy, and to provide a forum for outreach and education of the DSpace community around new initiatives such as DataSpace. For the DataSpace Project, the DSpace Foundation will perform education and outreach activities to the current DSpace community and beyond. Dedicated staff will be retained in years three through five of the Project to create awareness of the DataSpace work, promote and be the technical advocate for the platform, and coordinate training and education opportunities.

Science Commons

Science Commons develops free solutions for faster, more efficient Web-enabled scientific research by building a toolkit of policy contracts and technology. Science Commons targets areas where barriers to research are most common: access to research literature and materials, and increasing the utility of online data. Science Commons' mission is to accelerate the research cycle – the continuous production and reuse of knowledge that is at the heart of scientific method. Science Commons is a project of the Creative Commons whose copyright sharing licenses cover ninety million digital objects on the Web.

Unfunded Partners

HP Labs

Hewlett Packard Labs is a global research group that helps to shape HP strategy and invests in fundamental science and technology in areas of interest to HP. HP Labs has identified digital data archiving as a core technology for its customers in a variety of industries, and performs research on multiple aspects of the problem. Three groups at HP Labs – the Web Services and Systems Lab (that originally created the DSpace platform), the Intelligent Information Management Lab (IIML) and the Systems Security Lab – will collaborate with the DataSpace Project on federation architectures, policies for data curation, information security and accountability, data flow architectures for high performance scientific data analyses, and data visualization.

EMC

EMC Corporation is a major provider of information infrastructure systems, software and services. It is headquartered in Hopkinton, Massachusetts and its flagship product, the Symmetrix, is the foundation of storage networks in many large data centers. The EMC Innovation Network is EMC Corporation's global research program, a collaboration among researchers and technologists across the company, and university partners. The Innovation Network's mission is to explore the technologies that will shape future information infrastructures, the systems by which people and enterprises store, protect, optimize, and leverage information. The DataSpace project offers insights of potential interest to several of the company's product areas including archiving, enterprise content management, intelligent information management, security, and storage, as well as to new initiatives that deliver these capabilities as cloud computing services. EMC's participation will contribute a perspective on industry requirements and implementation considerations that will help direct the research to areas of high practical impact. The company has significant experience with the XAM specification for object management, and with digital curation more generally, which are directly related to the DataSpace project. The company also has academic research programs in information assurance and semantic search, which are also of relevance.

Google

Google Inc. is a public corporation focused on Internet search, e-mail, online mapping, office productivity, social networking, and other information services. The Google headquarters, the Googleplex, is located in Mountain View, California. Environmentalism, philanthropy and positive employee relations have been important tenets during the growth of Google and it has been identified multiple times as Fortune Magazine's #1 Best Place to Work. The Company describes its mission as: "… to organize the world's information and make it universally accessible and useful." The Google Structured-data Management Research Group is focused on research related to integrating data from multiple (structured and unstructured) sources and approaches to resolving schema heterogeneity. A main goal of the group is work (both on the research and product development) to build tools that simplify people's access to data, typically in complex data environments, which they refer to as *dataspaces*. A key effort, that is particularly relevant to this DataSpace proposal, is their role in the development of the Open Source Information Integration Suite (OpenII), an open-source set of tools for information integration.

Masdar Institute of Science and Technology

The Masdar Institute is a private, not-for-profit, independent, research-driven institute in Abu Dhabi developed with the support and cooperation of the Technology and Development Program (TDP) at the Massachusetts Institute of Technology (MIT). Masdar will collaborate on both the technology and data management research agenda as well as testing the DataSpace infrastructure as a member of the archives network. As a new organization in a rapidly developing region of the world, Masdar's participation in the project will demonstrate and validate the international effectiveness of DataSpace.

Additional Unfunded Partners - DataSpace Advisory Board (planned initial membership to be 10)

The DataSpace Project recognizes the importance of involving diverse sectors (e.g., commercial, international, etc.) in the problem of scalable sustainable solutions to long-term data management. The project's advisory board will bring a diverse range of expertise. The following people have already agreed to serve on the Advisory Board.

- Christine L. Borgman, Professor & Presidential Chair in Information Studies, Department of Information Studies, **Graduate School of Education and Information Science, University of California, Los Angeles;** *advice on skill requirements and training for data curation experts*
- Randy Buckner, Principal Investigator of the Cognitive Neuroscience Laboratory (CNL) and Professor of Psychology at **Harvard University**; *advice on integration and publishing of neuroscience data*
- Scott Doney, Senior Scientist in Marine Chemistry & Geochemistry at **Woods Hole Oceanographic Institution**; *advice on integration and publishing of oceanography data and its use in training researchers*
- Keith Jeffery, President, **European Research Consortium of Informatics and Mathematics (ERCIM)** and Director, Information Technology and International Strategy, **UK Rutherford Appleton Laboratory:** *advice on common long-term scientific data storage needs and approaches from research organizations across the European Union*;
- Liz Lyon, Director, **UKOLN** and Associate Director, **UK Digital Curation Centre** (DCC); *advice on data curation operations and long-term preservation of research data*
- Ed Roberts, David Sarnoff Professor of Management of Technology, MIT Sloan School of Management; MIT Technological Innovation & Entrepreneurship Program; and **MIT Entrepreneurship Center**:*advice on developing business model for long-term sustainability of DataSpace*;
- Pam Samuelson, Professor, **University of California at Berkeley School of Information and School of Law**; *advice on legal and policy issues related to research data publishing and reuse*
- Dan Schutzer, Director, **Financial Services Technical Consortium (FSTC):** *advice on best practices from the financial services industry as well as the potential for the financial services industry providing long-term support for DataSpace-type systems and technology*;
- Andrew Treloar, Director and Chief Architect, ARCHER Project, **Australian National Data Service (ANDS),** Monash University, Clayton, Australia; *advice on international efforts in scientific research data archiving standards and technologies*
- Wanda Orlikowski, Eaton-Peabody Professor of Communication Sciences at MIT, and **Professor of Information Technologies and Organization Studies at MIT Sloan School of Management**: *Expertise and advice on the development of the DataSpace Federation virtual community*.

DataNet Full Proposal –DataSpace Project
Appendix A6: Results from Prior Research

(max 4 pages: Due to the length limitation only selected examples of prior research are included)

DSpace and Data Management Archives

DSpace™, an existing digital archiving and preservation system of the MIT Libraries was begun by the MIT Libraries and HP Labs in April of 2000 with funding from Hewlett Packard under the HP/MIT Alliance [SRW*04, STW05]. Over the seven years since its official launch, DSpace has developed into a groundbreaking digital archiving system that captures, stores, indexes, preserves, and redistributes digital research material of many types on the Web. At MIT, DSpace has been used to create an digital archive capable of capturing, managing, preserving and disseminating the intellectual output of MIT's research faculty in digital formats [TS*03a,TS*03b].

Although developed jointly as a research project of the MIT Libraries and HP Labs, DSpace is freely available to research institutions worldwide as open source software that can be adapted and extended to meet local needs. DSpace was designed for ease of use, with a Web-based user interface that can be customized. Since its official release as an open source software platform in November of 2002, DSpace has been adopted for use by approximately five hundred research organizations worldwide, most notably by research university libraries but also includes cultural heritage organizations, government agencies, non-profits, and private enterprise.

Over the past five years the DSpace community has evolved a governance model and a software development model that is truly diverse and distributed – an effective virtual organization. In 2007 MIT and HP jointly founded the DSpace Foundation, an independent 501c3 non-profit corporation, to lead, support and coordinate the activities of the global DSpace community [BTS04]. Because such a collaborative community has been built around DSpace with MIT at its epicenter, the research that we conduct has an immediate impact on a broad, worldwide audience. Furthermore, the DSpace community leverages the good work of other projects to the extent possible; and invests its own efforts in areas where new research is required, and where results benefit both DSpace and also the broader digital archives community.

Integrating Data Management with Data Grids (2004-2005) [Smith]

With funding from the National Archives through the NSF NPACI program as part of the NARA Collection on Persistent Archives project, the MIT Libraries collaborated with the University of California San Diego Libraries and the San Diego Supercomputer Center (SDSC) to explore integrating the open source DSpace digital archive software with data grids, and specifically with the Storage Resource Broker (SRB) system developed at SDSC to manage and provide long-term access to and preservation of digital material. SRB is also available to the academic research community as an open source software system and is used by a large number of organizations who handle scientific research data, but it had not yet become mainstream in more traditional library and archives settings. By showing how DSpace and SRB could exchange digital collections and be integrated to function seamlessly together, we provided the academic research community with a richer system and a more scalable solution to the problem of digital storage and management. By exploring options for federation among the many institutions using DSpace, we demonstrated a means of providing widely distributed replication of valuable content to help protect against another risk of digital preservation – the single point of failure. This integration work was subsequently adopted as part of the production digital archiving strategy for a number of research universities using DSpace and SRB.

Developing Scalable Data Management Infrastructure in a Data Grid-Enabled Digital Library System (the PLEDGE Project, (2005-2007) [Smith]

With continued funding from the National Archives through the NSF NPACI program, the MIT Libraries again collaborated with the University of California, San Diego Libraries and the San Diego Supercomputer Center (SDSC) on methods for data management policy definition and exchange across distributed grid-based systems, primarily DSpace and iRODS (a successor to SRB developed by SDSC in 2007) [MS07]. The project developed an analysis of the NARA/RLG Trusted Digital Repository (TDR) checklist, now known as TRAC (Trusted Repository Audit

Checklist), to produce a set of rules and metadata that could be used to automate enforcement and verification of archiving policies [SM07]. A policy expression language was selected and a set of TRAC policies encoded as a management policy test bed. Finally a mechanism was developed to exchange policies between DSpace and iRODS to demonstrate automated enforcement of local data management policies across heterogeneous, distributed digital archive environments.

Semantic Web and Policy Management

The architecture of the Semantic Web (or Data Web) builds on the existing Web architecture to add new layers to support structured data (i.e. RDF), data ontologies (i.e. OWL), rules and logic, proof, and trust. These new layers will be necessary to achieve the degree of complexity required to manage large-scale scientific research data on the Web. The new data layer standards (RDF and OWL) and the query language for them (SPARQL) are standard specifications of the W3C. The higher level specifications for rules expression, logic/proof reasoning, and trust are still in development through a series of research projects being conducted at MIT and elsewhere.

Transparent Accountable Datamining Initiative (the TAMI Project, (2005-2008) [Abelson, Berners-Lee, Weitzner]

The TAMI Project is creating technical, legal, and policy foundations for transparency and accountability in large-scale aggregation and inferencing across heterogeneous information systems [BLH*06, WAB*08]. The incorporation of transparency and accountability into decentralized systems such as the Web is critical to help society manage the privacy risks arising from the explosive progress in communications, storage, and search technology. The expansion of government use of large-scale data mining for law enforcement and national security provides a compelling motivation for this work. While other investigations of the impact of data mining on privacy focus on limiting access to data as a means of protecting privacy, a variety of social, political, and technical factors are making it increasingly difficult to limit collection of and access to personal information. The TAMI Project is addressing the risks to privacy protection and the reliability of conclusions drawn from increasing ease of data aggregation from multiple sources by creating methods and technologies for adding increased transparency and accountability of the inferencing and aggregation process itself. The project is developing precise rule languages that are able to express policy constraints and reasoning engines that can describe their results.

Creating the Policy-Aware Web: Discretionary, Rules-based Access for the World Wide Web (NSF ITR 04-012, 2005-2006) [Berners-Lee, Weitzner]

Development of a rules-based policy management system that can be deployed in the open and distributed milieu of the World Wide Web, and define the necessary features of such a system to create a “policy aware” infrastructure for the Web. The project demonstrated the integration of a Semantic Web rules language with a theorem prover designed for the Web to enable HTTP to provide a scalable mechanism for the exchange of rules, and eventually proofs, for access control on the Web.

Data Interoperability and Integration

Identification and Resolution of Semantic Conflicts Using Metadata (1991-1993) [Madnick, Siegel]

This research, funded by NSF, developed a theory and basis for much of the data semantics work that is now being performed under the Semantic Web. The research identified that there are efforts to establish high-performance nationwide “information highways” connecting information systems, but these efforts largely ignored the semantic “interchanges.”. Because the meaning of data acquired from a source environment is usually different from that needed or expected in the receiver’s environment, the effectiveness of such an information highway is directly proportional to its capability for context interchange, i.e., the representation, exchange, comparison and transformation of context knowledge. Manual techniques for context interchange do not scale-up to allow the integration of hundreds of sources with evolving context knowledge bases. This research constituted the basis for a theory of context interchange and for the development of effective, dynamic and scalable information systems that provides semantic interoperability. This research accomplished these goals by providing a formalization for context interchange including theories for semantic values.

This NSF effort also addressed the design of a context representation language and context mediator in this environment including the development of a more flexible language (a metadata query language as an extension to SQL [SSR94]), and a conflict identification mechanism based on logic and abstraction. Some of the results from this funding can be seen in the publications that have resulted related to heterogeneous database integration using metadata.

Context Interchange: Using Knowledge about Data to Integrate Disparate Sources and Context Interchange: Using Knowledge about Data to Integrate Disparate Sources -- Prototype Option [the COntext INterchange, aka COIN, project, funded by DARPA, 1992-1998 [Madnick, Siegel]

This project represented a large-scale research and proof-of-concept effort based on the earlier NSF-funded Identification and Resolution of Semantic Conflicts Using Metadata project described above. Some of the principal results of this effort included the formal definition of "context" as well as the development of a prototype context representation and reasoning mediation system, called COIN. In addition, methods for context mediation that include context knowledge comparison and selection of conversion methods resulting in the ability to manipulate context among disparate systems, identify semantic conflicts, explicate semantic differences, and where possible, automatically resolve conflicts and, context system architectures for facilitating semantic interoperability in multiple systems (e.g. client/server, multi-database systems). In addition, a set of prototypes were developed to demonstrate concepts in context mediation, web wrapping, and multi-data source query processing and optimization. For example, one of the earliest results was a demonstration of context mediation technology in a client-server system [DGH*95]. In this prototype we developed a context representation and context mediation based on LOOM. Subsequent versions of COIN used extensions to F-Logic. The COIN context mediator provided middleware services that examined the context specification at the source and at the receiver, identified the semantic conflicts and developed a query execution plan that included the resolution of those conflicts. Some of these results are described in papers such as [BGL*97, GBMS99, GM90, LM96, LCM*99, LMS96b, Mad95a, Mad95b, Mad96, Mad99, MSG91].

Context Interchange – On-going Research and Experimentation, 1998-2008 [Madnick, Siegel]

After the development of the initial COIN prototype, additional improved versions, referred to as eCOIN (including automatic composition of conversion programs and addressing temporal semantics) as well as experiments demonstrating the application of COIN in diverse applications ranging from financial services to counter-terrorism have been conducted through funding from organizations such as Banco Santander Central Hispano (BSCH), DARPA, Fleet Bank, Merrill-Lynch, Malaysia University of Science and Technology (MUST)-Motorola, MITRE, PricewaterhouseCoopers, Singapore Defense Science and Technology Agency (DSTA), Singapore-MIT Alliance (SMA), Suruga Bank, and TRW. This research has also addressed and analyzed legal regulations (current and proposed) that impact data reuse and repurposing. Some of the results of these research efforts can be seen in papers, such as [FGM07, FM00, FMG02a, FMG02b, FMG04, FMM05, FMS00, FMY*05, ZMS01, ZMS04a, ZMS04b, ZMS04c, ZM06, ZM08].

Semantic Interoperability of Metadata and Information in unlike Environments (the SIMILE Project, funded by HP and the Mellon Foundation, 2003-2008) [Smith, Karger]

SIMILE is a joint project of the MIT Libraries, MIT CSAIL, and the W3C. The project is investigating ways to enhance interoperability among digital assets, schemas, vocabularies, ontologies, metadata, and services. A key challenge is that the collections which must interoperate are often distributed across individual, community, and institutional stores. The project provides end-user services by drawing upon these distributed resources. SIMILE has provided tools to leverage and extend data in DSpace and other digital repositories, enhancing their support for arbitrary schemas and metadata, primarily though the application of RDF and semantic web techniques. The project also developed a digital data dissemination architecture based on Web standards that provides a mechanism to add useful "views" to a particular digital artifact (i.e. asset, schema, or metadata instance), and bind those views to consuming services.

SIMILE has focused on well-defined, real-world use cases in the library and higher education domains.

Distributed Systems and Scalable Database Architectures

Infrastructure for Resilient Internet Systems (NSF Cooperative Agreement No ANI-0225660. IRIS ITR) [Karger]

This project proposed a novel decentralized infrastructure, based on distributed hash tables (DHTs) that enabled a new generation of large-scale distributed applications. The key technology on which it is built, DHTs, are robust in the face of failures, attacks and unexpectedly high loads. They are scalable, achieving large system sizes without incurring undue overhead. They are self-configuring, automatically incorporating new nodes without manual intervention or oversight. They simplify distributed programming by providing a clean and flexible interface. And, finally, they provide a shared infrastructure simultaneously usable by many applications.

The approach advocated was a radical departure from both the centralized client-server and the application-specific overlay models of distributed applications. This new approach not only changes the way large-scale distributed systems are built, but could potentially have far reaching societal implications as well. The main challenge in building any distributed system lies in dealing with security, robustness, management, and scaling issues; today each new system must solve these problems for itself, requiring significant hardware investment and sophisticated software design. The shared distributed infrastructure relieves individual applications of these burdens, thereby greatly reducing the barriers to entry for large-scale distributed services.

III-COR: Data Homesteading: Tools to let Scientific Users Harvest, Husband, and Share Structured Information. (NSF Proposal Number: 0712793. 09/01/07) [Karger]

The only scalable way to let scientists integrate data from multiple, arbitrary sources is to give them the tools that they need to do it themselves, without requiring support from skilled application and web developers. This project aims to turn data integration into an activity that can be performed by individual scientists on an ad hoc basis, collecting, manipulating, and publishing precisely the information that they want to work with. This project provides tools for three aspects of the data integration problem: to let non-programmers extract the information they need from websites with disparate or missing scheme; to let them visualize and manipulate it in task specific ways, and to let them publish their data back to the community in re-purposable form.

Data and Information Quality

MIT Total Data Quality Management (TDQM) project (1988-present) [Madnick, Wang] and MIT Information Quality (MITIQ) program (2002-present) [Madnick, Wang]

The TDQM project (based in the MIT Sloan School of Management) and the MITIQ program (based in the MIT School of Engineering) are both multi-sponsor consortia addressing a broad array of issues related to data and information quality. Sponsors have included organizations such as Acxion, Bull, Dun & Bradstreet, First Logic, Lockheed Martin, S C Johnson, etc. The TDQM project, started in 1988, is generally recognized as the first major research effort addressing data quality issues. Research conducted ranged from early work on "data source tracking" and the development of the polygen relational algebra [MW89b, WMH89, MW90a, MW90b, MW90c, KMS95] to the identification of quality metrics for data [WMK93 and many other papers by MITers and affiliated researchers], to the definition of various data quality calculations, such as using provenance information [PM07, PM08].

In addition to the research activities at MIT, TDQM and MITIQ have been instrumental in establishing an international "movement" on data and information quality research through diverse activities, such as the establishment of the annual International Conference on Information Quality (ICIQ) in 1996 and the Information Quality Industry Symposium (IQIS) in 2007, the publication of a series of books on Data Quality, and assisting in the establishment of the first Masters Degree program in Information Quality at the University of Arkansas at Little Rock (UALR) in 2006 and the ACM Journal on Data and Information Quality in 2008.