AIChE

Perspective

DataNet: An Emerging Cyberinfrastructure for Sharing, Reusing and Preserving Digital Data for Scientific Discovery and Learning

Jae W. Lee
Dept. of Chemical Engineering, The City College of New York, New York, NY 10031

Jianting Zhang
Dept. of Computer Science, The City College of New York, New York, NY 10031

Ann S. Zimmerman
School of Information, University of Michigan, Ann Arbor, MI 48109

Angelo Lucia
Dept. of Chemical Engineering, University of Rhode Island, Kingston, RI 02881

DOI 10.1002/aic.12085
Published online October 1, 2009 in Wiley InterScience (www.interscience.wiley.com).

Keywords: DataNet, enabling cyberinfrastructure, national data repositories, scientific discovery and learning for the 21st century

Introduction

DataNet is an emerging, enabling cyberinfrastructure for open access, and long-term preservation of digital data funded by the National Science Foundation. DataNet will create unprecedented opportunities for advancing research and education in science and engineering, by creating an environment consisting of data repositories and new tools and services for sharing and reusing important digital data across disciplines. This will foster transformative, interdisciplinary scientific collaboration on a grand scale, permit scientists to build knowledge that will provide better understanding of and sustainable solutions to important scientific challenges, offer better public understanding of technological problems through openly accessible information in well articulated repositories, and provide a format for resolving the data-driven nature of scientific and technological problems of societal importance (i.e., grand challenge and other problems).

This perspective article presents an overview of the DataNet initiative and its vision, and outlines various opportunities that DataNet will afford science and engineering. Opportunities specifically available to chemical engineers include the possibility of (1) sharing data with scientists and engineers from other disciplines on technical challenges ranging from molecular modeling to carbon capture/storage to energy and biofuels production, and others, (2) initiating new collaborations on "big" science or grand challenges, (3) depositing and preserving data for research and education in secure and protected national repositories, (4) getting involved in various cyberinfrastructure activities, (5) receiving credit for sharing data by publishing that data in a peer-reviewed manner, (6) enhancing grantsmanship, and (7) more.

Many branches of science and engineering are becoming increasingly data driven. For example, the second generation of the panoramic survey telescope and rapid response system (Pan-Starrs), called PS4, is an effort for continuously scanning the entire sky for the purpose of scientific discovery and for monitoring near earth objects. For general information, see http://pan-starrs.ifa.hawaii.edu/public/. Pan-Starrs is expected to go online in 2012 and, when completed, will consist of four wide field imaging telescopes (three in addition to the one that already exists on Maui, HW called PS1). It will gather somewhere in the neighborhood of 30 terabytes ($1TB = 10^{12}$ bytes) of data per night. This is a large amount of data, equivalent to 30,000 one gigabyte (GB) memory sticks or several million pages in a digitized book. Computational chemistry is another discipline in which large amounts

Correspondence concerning this article should be addressed to A. Lucia at lucia@egr.uri.edu

of digital data can be generated. Simulations of large molecular clusters can take days, weeks, and even months to complete and can yield TB of computer output. However, researchers do not usually store results of simulations; they store input files and versions of computer programs (also digital data), and if needed regenerate the output, which is a dangerous practice that can waste time and effort and result in the loss of valuable information. Chemical engineering has also experienced an explosion in digital information (i.e., audio, video, textual material, graphics, etc.) in areas such as proteomics and genomics, *de novo* drug design, process modeling and simulation of oil and gas operations, and so on. The picture on the cover illustrates the data-driven nature of modeling and validating the interactions between the natural carbon cycle and anthropogenic activities. It is estimated that digital information worldwide now exceeds exabytes (10^{18}) of storage. The growing amount of data creates significant challenges for storage, access, preservation, and knowledge creation while cost and other factors make many important data sets both valuable to share and impossible to replace.

DataNet is the acronym given to the large National Science Foundation initiative (NSF 07-601) entitled Sustainable Digital Data Preservation and Access Network Partners that is funded through the Office of Cyberinfrastructure (OCI) and Computer Information and Science for Engineering (CISE). It is part of an overall national CI plan[1] that includes TeraGrid and other related programs for data tools and services. For a very clear understanding of how this national CI plan relates to the initiatives from the Office of Cyberinfrastructure we refer the reader to Seidel.[2] The primary purpose of the DataNet effort is to create an enabling cyberinfrastructure for open access and long-term preservation of important digital data for discovery, research, education, and training. The DataNet initiative is in direct response to several factors: (1) the fact that technological challenges in science and engineering are becoming increasingly data driven; (2) the increased demands that are being placed on scientists and engineers in many disciplines to manage and share their data, and (3) the need to preserve important digital data. NSF has plans to establish five DataNet partnerships at a cost of $100,000,000 across the US to serve all of the disciplines that it supports. Two DataNet awards were made in 2008 and are led by the University of New Mexico and Johns Hopkins University. The DataNet partnership led by the University of New Mexico is called DataONE, where ONE stands for Observation Network for Earth. The title of the DataNet partnership led by Johns Hopkins University is the Data Conservancy. Three more awards are anticipated in 2009-10.

The goal of this article is to give readers an overview of the emerging DataNet cyberinfrastructure. We describe what DataNet is, the opportunities it will offer, and how scientists and engineers can become active in DataNet. To do this we focus on the following: (1) the universal difficulties with data, (2) domain coverage and science drivers, (3) data repositories, (4) anticipated impacts, (5) computing platforms and architecture requirements, (6) long-term technological and economic sustainability, (7) full life cycle data management, (8) discovery, research and education, (9) user engagement, (10) outreach, (11) organizational structure and management, (12) the diversity of users and partners, (13) assessment and evaluation, and (14) a national functional data network.

Universal Difficulties with Data

We all share the same difficulties and frustrations associated with data — managing our data and sharing data with others. The difficulties with managing data often come from a lack of formal training. Many scientists and engineers do not have a working knowledge of basic data management concepts like metadata and onotologies. Thus, data sets within a given group of researchers can contain different formats, data types, and descriptions, making them heterogeneous. As the sizes of data sets grow, managing heterogeneous data sets can become a tremendous burden. Difficulties associated with sharing and re-using data are also universal.[3,4,5,6] Many important computer and information science problems such as searching and merging large data sets are unresolved. It can also be difficult to discover, share or reuse data because: (1) valuable data may have been discarded, (2) computer models may be unusable or difficult to use due to legacy issues, (3) incompatible formats can make data difficult or impossible to integrate, (4) data flow across domains may be impeded by incomplete, inaccurate, and/or poorly constructed metadata, and (5) many scientists are reluctant to share data due to a lack of reward, intellectual property issues, and documentation effort.

Domain Coverage and Science Drivers

It cannot be overemphasized that any scientist or engineer can participate in DataNet by electing to become a member of a particular DataNet partnership and by depositing and sharing data. While each of the eventual five DataNet partnerships will be built around a specific science driver of societal importance taken together all five partnerships are expected to cover all of the disciplines that NSF supports. For example, the science drivers for the DataONE and Data Conservancy partnerships are ecology informatics and the carbon cycle and urbanization, respectively. However, the data collections for DataONE, Data Conservancy, and other partnerships must contain data sets that go well beyond those dictated by the science drivers.

Data Repositories

Data in DataNet repositories will be freely available to anyone since they are paid for by US taxpayers. Table 1 gives an example of a digital data repository.

Table 1 shows the types of data sets that might be part of a repository, how the amount of each data type in the repository might increase over a five-year period, the associated scientific or engineering phenomena covered by the data, and some of the disciplines that might find particular types of data of value in research, education, and/or outreach. Many of the data sets in any given DataNet repository will be unique and fundamental data sets generated and deposited by DataNet partnership domain scientists or engineers. Other

Table 1. An Example of a Data Repository

Data Set Description	Availability*		Diversity of Data Sets	
	Initial	5 yrs	Phenomena	Some Disciplines Served
Atmospheric Science				
NO_x, SO_2, CO_2 data	5 TB	25 TB	transport processes	meteorology, geo-science
full 4D sampling data	–	100 TB	transport in time	meteorology
Chemistry				
force field & structure	2 TB	10 TB	molecular simulation	chemistry, physics, biology
kinetic & equilibrium	300 GB	10 TB	reactions	chemistry, chemical eng
residue curves	50 MB	1 GB	separations	materials science geology
cluster trajectories	–	1 TB	nucleation & growth	chemistry, geochemistry
Environmental Sciences				
thermodynamic data	100 MB	10 GB	multiphase behavior	chemistry, engineering
paleo-climate data	500 GB	4 TB	climatology	climate science
habitat mapping data	11 TB	96 TB	ecology, pollution	environmental sciences
Geology				
mineral deposition data	20 MB	100 MB	nucleation & growth	geology, earth science
dissolution kinetics	10 MB	50 MB	mineral reactions	chemical eng, geochemistry
Modeling				
general/specific software	200 MB	500 MB	modeling physics	math, CS, operations res
fluid flow simulations	20 TB	50 TB	CFD, fluid dynamics	ocean science, engineering
water-rock simulations	100 MB	200 MB	multiphase flow	geology, oceanography
molecular simulations	2 TB	50 TB	phase transitions	physics, engineering
reservoir & ocean codes	50 MB	100 MB	multiphase flow	geophysics, engineering
ocean wave simulations	20 GB	50 GB	wave/wind forcing	civil eng, oceanography
Oceanography				
temperature, pH, salinity	1 TB	5 TB	transport processes	oceanography
micro-organism data	–	100 GB	bio-geochemistry	geochemistry, engineering
carbon flux data	100 GB	200 GB	mass transport	plant science, biology
ocean observatory data	100 GB	5 TB	heat/mass transfer	ocean, environmental sci
Polar Science				
ice core/tree ring data**	125 GB	200 GB	paleo-climate	polar & climate sciences
tectonic (GPS)***	2 GB	5 GB	plate movement	earth/ocean/polar sciences
Plant Science				
CO_2 uptake	100 GB	500 GB	carbon storage	ocean & environ sciences
carbon pathways	10 MB	15 MB	photosynthesis	plant science, cell biology
taxonomy data****	400 GB	2 TB	bio-fuels production	chemistry, engineering
Images				
molecular conformation	200 MB	1 TB	crystal structure	chemistry, material science
colloidal images	50 GB	300 GB	vesicle formation	biophysics, adv materials
process flow diagrams	100 MB	100 GB	systems behavior	process engineering
phase diagrams	500 MB	200 GB	phase equilibrium	chemistry, geology
confocal microscopy	200 GB	10 TB	morphology	chemistry, engineering
Textual				
theses, journal articles	500 MB	50 MB	research/education	all disciplines
progress reports	10 MB	100 MB	research	all disciplines
presentations, tutorials	50 GB	500 MB	education, outreach	all disciplines, public
Total	$\sim$ 43 TB	400 TB		

*Initial means indicated size is available at start of project; 5 yrs means projected size available in 5 yrs.
**National Geophysical Data Center (NGDC);
***Janus Database (International Ocean Drilling Program);
****Data from USDA[6].

data will be application specific in nature. Yet other publicly available data sets (e.g., plant data available through the USDA;[7] data from NASA Distributed Active Archive Centers (DAAC), and so on) will be made interoperable through data network tools. Fundamental data sets in a repository enable scientists and engineers to address a wide spectrum of important technological challenges. These fundamental data sets can be merged with application-specific and publicly available data sets from a wide variety of disciplines to address complex problems. For example, production of fuels from biomass (corn stalks, other lingo-cellulosic waste) is one of several proposed technologies for meeting large-scale future energy demands. To do this, plant composition data is needed to understand the characteristics of various renewable raw materials, elementary reaction kinetic data is needed to quantify various chemical reaction pathways to biofuels (Fischer-Tropsch, pyrolysis, fermentation, hydrolysis), and other property and model data (vapor pressures, equation of state models, etc.) are required to design separations to isolate fuels from byproducts. However, biofuel production requires rapid crop replenishment, impacts the nitrogen (N_2) cycle, and fuel production still produces CO_2 emissions. Moreover, it is well known that the carbon and nitrogen cycles are coupled[8] and quantifying this coupling requires N_2 cycle plant model and field data (fertilizer use, overcultivation, N_2/CO_2 uptake data, soluble nitrogen runoff into

oceans, etc.) to be merged with data for carbon storage; thus, challenges in energy from biofuels, management of the N_2 cycle, and carbon storage are all strongly interrelated.

Anticipated Impacts

DataNet and DataNet partnerships are anticipated to provide scientists and engineers (and the nation) with new enabling cyberinfrastructure capabilities that will lead to new science, novel ways to conduct science, and sustainable solutions to many of the challenges facing the US and the world in the 21st century (e.g., clean and abundant energy, clean water, mitigation of greenhouse gases, etc., for an estimated world population of 10 billion by 2050).

Interdisciplinary collaboration is critical to achieving the DataNet vision. All DataNet partnerships will be driven by leading edge computer science (CS), information technology (IT), and science and engineering research that will produce significant advances in interoperability through new data mining techniques, new schema and metadata management methodologies, data curation tools and techniques, new ontological data structures, standards and best practices, more effective data visualization and analysis tools, and new considerably more cost-effective ways of archiving and preserving digital data. Each partnership will have the following anticipated impacts on discovery and learning: It will (1) be an exemplar vehicle for potentially transformative, interdisciplinary scientific collaboration on a grand scale, (2) permit scientists to build knowledge that will provide better understanding of important scientific challenges, (3) provide better public understanding of important technological problems through openly accessible information in a well articulated repository, and (4) provide a format for collaboration among all DataNet partnerships for resolving the data-driven nature of science[9,10,11] and technological problems of societal importance. DataNet partnerships will enable transformative scientific advancement and technological sustainability by directly addressing the data-driven nature of a wide array of challenges that require sharing and reuse of diverse data sets in a domain agnostic manner.

Full Data Life-Cycle Management Activities

Users need to deposit data, retrieve data, know that data will be migrated to new storage technologies as they emerge, and know that their data is secure, protected, and will be there when they need it. They do not necessarily care how these tasks are completed; they just want an effective system so they can use data. These challenges are encompassed by full life-cycle data management activities and DataNet partnership data life-cycle activities are expected to be user-centered, serve diverse disciplines, and facilitate data sharing and cross-discipline problem solving by developing novel solutions for (1) data deposition, acquisition and ingestion, (2) metadata/ontology management, (3) data security, (4) data integration and interoperability, and (5) data analysis and visualization. Since many of these topics may be unfamiliar to AIChE readers, we describe them briefly in the following sections.

Data deposition/acquisition/ingestion

This activity centers on developing tools and techniques to simplify and automate the deposition and ingestion of new data into a given repository with a focus on data quality, interoperability between new/existing data and metadata, enhancement of data gathering and evaluation, and usability of data for analysis and cross-discipline problem solving by a wide range of users.

Metadata management

Metadata is simply data about data. Metadata management tools must be domain agnostic, address a wide range of repository holdings, incorporate important parameters, adapt/integrate existing discipline-specific metadata management standards, and describe individual data sets, as well as relationships between different data sets.

Data security and protection

Data security tools must be flexible and designed with policy and technology considerations. Data access and privacy concepts must also be used to ensure distributed data are secure, protected against iterative discovery,[12] and built on existing federal standards of encryption, data access and management.

Data discovery, access and dissemination

When data are provided as a service, users are expected to query using specific vocabulary. However, individual and cross-discipline data from multiple producers require robust semantic correspondences between user and service description ontologies, which can be prescribed using existing frameworks (web services description languages,[13] and semantic web technologies[14,15]).

Standards-based interoperability and integration

To be effective, data must be integrated within and across disciplines in ways that are transparent to users, so that scientists and engineers can focus on science. This can be accomplished using platforms such as the data space support platform (DSSP)[16,17] to manage distributed data that might include basic attributes such as charts, video, wikis, animations and simulations, etc., or semantic maps between existing disciplinary taxonomies/ontologies.[18]

Evaluation, analysis and visualization

Scientists and engineers must be able to validate and verify results. Thus, analysis and annotation of very large data sets and scalable visualization of results in the form of graphs, cognitive maps, and other representations are important to users. Tools for evaluation, analysis, and visualization need to address data distribution, heterogeneous artifacts, spatial and temporal dimensions, varying provenance and sources, and other attributes.[19] It is also often useful to reuse visualizations and analyses through a repository of annotated reusable digital artifacts.

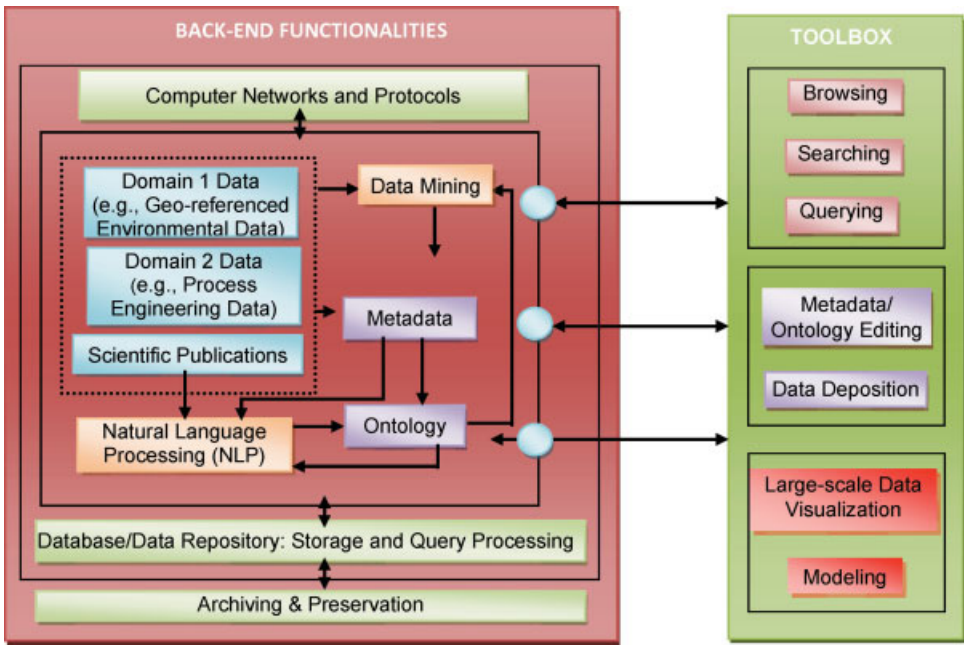

Figure 1. A view of a DataNet partnership architecture

Computing Network, Platform and Architecture Requirements

There are many computer platforms and architectures that can be designed for storing, searching, and retrieving large heterogeneous data sets, protecting data, visualizing data in scalable ways, etc. However, regardless of the design, there are three essential needs — high-performance computing (HPC), memory, and storage.

Network

All network activities (e.g., replication, data deposition, archiving and preservation, etc.) must be highly coordinated to be effective. Also the network architecture should provide a web-based front end portal that allows users to (1) gain access to those parts of the repository for which they have designated privileges, (2) interact with data in a repository through a flexible, robust user toolbox that provides secure user interfaces for depositing and querying data, and manages metadata and ontologies, and (3) processes schemas and queries.

Computing platforms

Viable computing platforms include the use of large "in-house" clusters, which have high maintenance and operations costs, or cloud computing which provides services on demand. A key attribute of any platform for handling DataNet tasks is a high performance, scalable file management system for accessing a common set of files from hundreds of computers,[20] capable of handling petabytes of data.

Architecture

Figure 1 is a schematic of a user-centered, standards-based, extensible, and seamless architecture for full life-cycle data management.

The User Toolbox. It makes complete sense to provide users with a toolbox with a web interface that supports data visualization, data mining, scientific workflow, tutorials, help, etc., and that is linked to data deposition and data query. Users should be provided with a basic set of capabilities to (1) search data by name and/or view relationships among data in graphical form, (2) query data using an *ad hoc* query language or graphical interfaces, (3) deposit data in an existing database using flexible interfaces, and (4) insert new schema using graphical tools for defining schema. The toolbox should also contain a rich set of tools that are transparent to users (e.g., schema matching should be transparent, but is critical for seamlessly connecting new data to an existing repository).

Long-Term Technological and Economic Sustainability

Will DataNet disappear after NSF funds are expended? No! Taxpayers and Congress have the right and the responsibility to demand success. Therefore, each DataNet partnership is charged with planning for long-term technological and economic sustainability with a time horizon of 50+ years.

Technological sustainability

Technological sustainability can be accomplished by creating an enabling cyberinfrastructure with exceptional reliability, adaptability and scalability, and by providing an environment that allows users to focus on core

competencies, participate in unique and fertile collaborations, and leverage infrastructure for innovation.

1. The computing platform must be able to adapt rapidly to changes in system and user needs, access technological developments in grid computing and management services, and provide hardware reliability and sustainability by managing networking and virtual cyberinfrastructure services using extensible back-end software.

2. Network reliability and data availability require caching and replication of data across geographically separate sites and archiving data to secure tape archives to provide data reliability and data preservation.

3. A user-centered architecture with a web toolbox to deliver products together with open-source, nonproprietary software tools, and the use of evolving metadata/ontology standards promotes long-term software sustainability.

4. Sustainable practices such as fostering synergy between users and network developers, reaching consensus among stakeholders with regard to partnership mission, network design, implementation, tools, etc., through multiple methods of user engagement, and educating stakeholders on the importance of standards and best practices all contribute to the long-term use and usability of DataNet partnership products and services.

Economic sustainability

Economic sustainability is accomplished by developing realistic revenue streams for value-added DataNet partnership products and services. Specifically,

1. The choice of computing platform can contribute to economic sustainability by bringing with it economies of scale, data security/retention, and by providing resources at low cost.

2. Institutional commitments (e.g., return/reduced overhead) contribute valuable resources to DataNet partnerships during early stages of development. Return/reduced overhead on future successful grants and other institutional commitments contribute to long-term economic sustainability.

3. DataNet cyber-infrastructure will provide valuable resources (facilities and services) to the community and beyond. Costs for using facilities and services in research should be built into proposals and grant winners should be expected to cover costs associated with using those resources.

4. DataNet partnerships should provide membership levels and services (data base organization, schema/metadata creation, training courses) at affordable prices. While data and CS/IT tools will be free, fees for services should follow some type of software business model (e.g., open source Red Hat Linux).

5. DataNet partnerships anticipate considerable domain science IP will result from member activities and contribute to long-term economic sustainability.

6. Commercialization of domain science IP and products is also a potential revenue stream.

Discovery, Research, and Education

Data lifecycle management activities help define open research topics and opportunities for DataNet partnerships to function at the frontiers of research and education in CS/IT and in science and engineering by (1) involving undergraduate and graduate students in leading edge research, (2) defining both traditional and nontraditional degree programs with interdisciplinary research projects at the crossroads of science/engineering and CS/IT, (3) industrial internships, and (4) cross-partnership exchange visits. These are just some of the activities that are required for partnerships to meet their DataNet responsibilities of facilitating and serving as objects of research.

User Engagement

It is each DataNet partnership's responsibility to engage users and build their user base, which can be effectively accomplished through various tried and true techniques.[21,22,23,24,25]

Promoting formal collaborations

DataNet partnerships must partner with a wide array of scientists and engineers to establish formal collaborations that will ensure that the partnerships are driven by relevant (grand challenge) problems of interest to scientists and engineers and that the repositories are populated with the data needed to address those problems.

Workshops and short courses

Workshops and short courses should use realistic science examples, educate and grow the user community, address barriers to sharing data, and provide feedback on usefulness, ease of use, changing user needs, and to help meet DataNet goals. Workshops and short courses should be offered at relevant conferences and venues.

Working groups

Working groups provide longer interactions around topics relevant to DataNet goals and objectives. Working groups must share responsibility for disseminating their findings and have the potential to provide rich opportunities for joint collaboration.

Special issue publications

Special issue publications can provide a means to give credit to data producers and users in peer reviewed outlets, to disseminate findings and results from activities, and to help ameliorate barriers to sharing data.

User outreach

User outreach activities are needed to grow the number of users and contributors to DataNet and can be accomplished through (1) presentations, demonstrations, handouts, and other materials to publicize DataNet, (2) real-life scenarios based on the use of DataNet in research, education, and policy, (3) focus groups and stakeholder forums at professional meetings to elicit information from researchers about their work and data needs, and (4) sharing information about DataNet methods and results that can be used by others.

Table 2. Sample Questions for Evaluation of DataNet Goals

Typical DataNet Goals	Representative Evaluation Questions
Enable new scientific discovery and understanding	•Are new interdisciplinary collaborations formed as a result of DataNet partnership activities? •How many publications, conference papers, and presentations are based upon DataNet data? •What impact do DatNet activities, collections, and tools have on the practices of scientists and engineers?
Ensure that tools and systems meet user needs	•How well do DataNet tools and interfaces correspond with usability metrics that assess ease of use, effectiveness, efficiency, and user satisfaction with tools and interfaces? •Do DataNet software and systems meet subjective & objective measures for speed, scale, consistency, robustness, and reliability? •Do DataNet tools and services reduce the amount of time required to create integrated sets of heterogeneous data?
Educate and train data users and contributors	•Does participation in education and training impact DataNet use? •How satisfied are participants with DataNet education and training activities? How likely are they to attend future events? •Are milestones related to the number of workshops and/or training sessions being met?
Increase environmental literacy	•Are learning objectives relative to environmental literacy in K-12 and the broader public being met? •Are DataNet data and tools being employed effectively in teaching and outreach?

Data advisory board

Data Advisory Boards drawn from domain experts, archivists, data librarians, digital preservation specialists, policy makers, representatives from other DataNets and international efforts are essential for providing advice, assessing plans, helping to set priorities for repository collections and tools, and promoting DataNet to their respective communities.

General Outreach

DataNet partnerships are expected to provide systems, tools and resources that (1) foster the use of digital data for educational and training purposes at all levels (K-12, undergraduate, graduate, professional, and public), and (2) enhance capabilities for integrating research and education at all levels. Many capabilities and resources of DataNet should be freely available for the purpose of enhancing education and training at all levels and made readily available to students, educators, and the general public.

Organizational Structure and Management

DataNet partnerships are virtual organizations that must be well organized and built on a shared vision, frequent communication, flexible response to changing needs, and shared governance.

Diversity of Partners and Users

DataNet partnerships are expected to foster membership from academia, commercial, not-for-profit, government, and international organizations. Similarly, the user base should be broad ranging and cover a wide variety of disciplines.

Assessment and Evaluation

Unbiased annual assessment and evaluation of goals, activities, and outcomes are essential for the success of any DataNet partnership. These assessments must necessarily be conducted by an external evaluator in collaboration with DataNet leadership personnel. Table 2 gives a set of generic sample questions for evaluating DataNet goals.

Functional Data Network

DataNet partnerships should not function in a vacuum. Coordination between DataNet partnerships and other organizations actively involved in a wide variety of cyberinfrastructure and digital preservation activities (e.g., MIT's DSpace efforts (http://dspace.mit.edu/), TeraGrid, the Library of Congress National Digital Information Infrastructure and Preservation Program (NDIIPP), etc.) is essential to the formation of a functional data network. Two basic ways of establishing a national functional data network are by:

1. Encouraging common workshops and working group meetings among all DataNet partnerships, interacting with other open access/preservation organizations, and by inviting other cyberinfrastructure organizations to attend DataNet-wide workshops and short courses.
2. Working together to share best practices and evaluation approaches and measures.

An Open Invitation

Chemical engineers are engaged in a wide variety of data-driven scientific investigations and activities. Therefore, we strongly encourage everyone in the chemical engineering community to get involved with DataNet as it evolves, in whatever capacity they feel comfortable with (i.e., as a user, a member, a member of a Data Advisory Board, etc.). It will offer unprecedented opportunities for new scientific

collaborations, new ways of conducting science, it will enhance education and outreach capabilities, and it will provide unique opportunities for better grantsmanship and entrepreneurship.

Literature Cited

1. National Science Foundation, Cyber-infrastructure Council. *Cyber-infrastructure Vision for 21st Century Discovery*. Arlington, VA: National Science Foundation; 2007.
2. http://www.nsf.gov/sbe/secure/advcom1108/Presentations/04.Seidel_SBE_Advisory.pdf.
3. Blue Ribbon Task Force on Sustainable Digital Preservation and Access. Sustaining the Digital Investment: Issues and Challenges of Economically Sustainable Digital Preservation; 2008. http://brtf.sdsc.edu/biblio/BRTF_Interim_Report.pdf.
4. Interagency Working Group on Digital Data. *Harnessing the Power of Digital Data for Science and Society*: Report of the Interagency Working Group on Digital Data to the Committee on Science of the National Science & Technology Council; 2009.http://www.nitrd.gov/About/Harnessing_Power_Web.pdf.
5. National Research Council. *Bits of Power: Issues in Global Access to Scientific Data*. Washington, DC: National Academy Press; 1997.
6. National Science Board. *Long-lived Digital Data Collections: Enabling Research and Education in the 21st century*. Arlington, VA: National Science Foundation; 2005.http://www.nsf.gov/pubs/2005/nsb0540/nsb0540.pdf.
7. http://plants.usda.gov/index.html.
8. Finzi AC, Schlesinger WH. Soil N-cycling in a pine forest exposed to five years of elevated CO_2. *Ecosystems*. 2003;6(5):444–456.
9. Emmott S. *Towards 2020 Science*. Redmond, WA: Microsoft Research; 2006.
10. Hey T, Trefethen AE. *The Data Deluge: An e-Science Perspective*. In: Berman F, Fox GC, Hey T, eds. *Grid Computing: Making the Global Information Infrastructure a Reality*.Chichester, UK: John Wiley & Sons, Inc; 2003:859–906.
11. National Science Foundation, Cyber-infrastructure Council. *Cyber-infrastructure Vision for 21st Century Discovery*. Arlington, VA: National Science Foundation; 2007.
12. Miklau G, Suciu D. A formal analysis of information disclosure in data exchange. *Proc SIGMOD 2004*. 2004:575–586.
13. WebService Modeling Ontology, http://www.wsmo.org/.
14. Akkiraju R, Farrell J, Miller J, Nagarajan M, Schmidt M, Sheth A, Verma K. Web Service Semantics - WSDL-S. In: *A Joint UGA-IBM Technical Note*. version 1.0; 2005.
15. http://www.w3.org/Submission/OWL-S/.
16. Franklin MJ, Halevy AY, Maier D. From databases to dataspaces: a new abstraction for information management. *SIGMOD Record*. 2005;34(4):27–33.
17. Salles MAV, Dittrich J-P Karakashian SK, Girard OR, Blunschi L. iTrails: Pay-As-You-Go Information Integration in Dataspaces. In: *VLDB 2007*. 2007:663–674.
18. Zhu H, Madnick SA. Lightweight Ontology Approach to Scalable Interoperability. In: *Proceedings of VLDB Workshop on Ontologies-based Techniques for Databases and Information Systems (ODBIS 2006)*; 2007:45–54.
19. Liu Z, Wu B, George R. Fuzzy Clustering for Knowledge Discovery in Oceanographic Data. Proceedings of the IEEE Conference on Granular Computing, Atlanta, GA, May 10-12, 2006:651–654.
20. Fadden, S. An Introduction to GPFS Version 3.2.1 IBM white paper (2008). http://www-03.ibm.com/systems/clusters/software/whitepapers/gpfs_intro.html.
21. Rogers EM. *Diffusion of Innovations*. 5th ed. New York, NY: Free Press; 2003.
22. Zimmerman A, Finholt TA. Growing an Infrastructure: The Role of Gateway Organizations in Cultivating New Communities of Users. In: *Proceedings of the International ACM SIGGROUP Conference on Supporting Group Work (GROUP '07)*. New York, NY: ACM Press; 2007:239–248.
23. Reichman OJ. NCEAS: promoting creative collaboration. *PLoS Biology*. 2004; 2(3):e72.http://dx.doi.org/10.1371/journal.pbio.0020072.
24. Senkowsky S. The ascent of NESCent. *Biosci*. 2007; 57(2):106–111.
25. Olson JS, Hofer EC, Bos N, Zimmerman A, Olson GM, Cooney D, Faniel I, eds. *A Theory of Remote Scientific Collaboration*. In: *Scientific Collaboration on the Internet*. Cambridge, MA: MIT Press; 2008:73–97.