

© 2005 by Aniruddha M. Bhagwat. All rights reserved.

PRELIMINARY CYBERINFRASTRUCTURE NEEDS ASSESSMENT AND
TECHNOLOGY REVIEW FOR CLEANER (COLLABORATIVE LARGE-SCALE
ENGINEERING ANALYSIS NETWORK FOR ENVIRONMENTAL RESEARCH)

BY

ANIRUDDHA M. BHAGWAT

B.E., University of Delhi, 2003

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Civil and Environmental Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2005

Urbana, Illinois

ABSTRACT

A new initiative called Collaborative Large Scale Engineering Analysis Network for Environmental Research (CLEANER) is being proposed by the National Science Foundation. CLEANER is envisioned to be a network of large-scale environmental field sites integrated by a collaborative infrastructure to enable groups of environmental engineering and science investigators to study landscapes stressed by human activities. An integral part of this collaborative environment would be cyberinfrastructure - a network of computational resources that will provide access to and integrate shared repositories for data, models, and tools with facilities for experimentation and computation. Since the cyberinfrastructure will meet the requirements of diverse researchers and educators across the environmental engineering and science field, a thorough assessment of these needs has to be done prior to the development of CLEANER.

This work is a preliminary study of cyberinfrastructure needs and provides inputs to the next phase of the requirements elicitation process. It also reviews various technologies that can be incorporated in the cyberinfrastructure and describes the development of a prototype CLEANER collaborative environment called the CyberCollaboratory.

To My Family

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my advisors Dr. Barbara Minsker and Dr. Wayland Eheart, for providing me the opportunity to work on a really interesting research project. I am grateful for their constant guidance, insight and support throughout my research.

I am also grateful to Luigi Marini and Yong Liu of the National Center for Supercomputing Applications (NCSA) for helping me understand the intricacies of portlet programming, webcrawling and other assorted computational and information technology.

I would like to acknowledge the generous inputs on requirements assessment procedures from Dr. Michael Twidale, of the Graduate School of Library & Information Sciences at the University of Illinois; Dr. Katherine Lawrence and Il-Hwan Kim, of the School of Information at the University of Michigan; and David Ribes of the University of California, San Diego.

I am also indebted to the following people who contributed to the CyberCollaboratory and scenario development: Barbara Minsker, Wayland Eheart, Michael Twidale, Cameron Jones, Xavier Llorà, Hua Xie, Gayathri Gopalakrishnan, Steve Downey, Andrew Wadsworth, Yong Liu, Timothy Wentling, Luigi Marini, Tom Prudhomme, Loretta Auvil, Lisa Gatzke, Ingbert Floyd, Jessica Lam, Nosh Contractor, Michael

Welge, and Paul Karpenko of the University of Illinois; and Tom Finholt and Katherine Lawrence of the University of Michigan.

This work was supported by National Science Foundation grant BES-0414259. Funding for the development of the CyberCollaboratory came from NSF grant SCI-0525308 and the Office of Naval Research grant N00014-04-1-0437.

It would have been impossible for me to have finished this project without the support of my family and friends. Additionally, I thank the members of the Minsker research group for giving me a sympathetic ear during the tough times.

TABLE OF CONTENTS

| | |
|--|------|
| LIST OF FIGURES | viii |
| 1. INTRODUCTION | 1 |
| 1.1 The Collaborative Research Paradigm..... | 1 |
| 1.2 A Brief History of Collaboratory Efforts..... | 4 |
| 1.3 The CLEANER Initiative | 5 |
| 1.4 Purpose of this Study | 8 |
| 2. BACKGROUND ON CYBERINFRASTRUCTURE NEEDS ASSESMENT..... | 10 |
| 2.1 Cyberinfrastructure Background..... | 10 |
| 2.2 Cyberinfrastructure for Environmental Research & Education..... | 13 |
| 2.3 Needs Assessment: A Precursor to Cyberinfrastructure Development | 15 |
| 2.4 The Needs Assessment Process | 16 |
| 2.5 Needs Assessment Exercises for Other Collaborative Endeavors | 22 |
| 3. NEEDS ASSESSMENT METHODOLOGY | 26 |
| 3.1 Preliminary Tasks | 26 |
| 3.2 Development of a Prototype CLEANER CyberCollaboratory | 27 |
| 3.3 Requirements Elicitation..... | 44 |
| 4. RESULTS | 47 |
| 4.1 Use-Case Scenarios..... | 47 |
| 4.2 General Questions and Responses | 52 |
| 4.3 Community Response to Collaboration Technologies | 66 |
| 5. CONCLUSIONS..... | 70 |
| BIBLIOGRAPHY..... | 72 |
| APPENDIX A – LIST OF PROJECT TEAM AND ADVISORY COMMITTEE MEMBERS | 79 |
| A.1 Project Team | 79 |
| A.2 Advisory Committee..... | 80 |
| APPENDIX B – DETAILED QUESTIONS FOR THE NEEDS GATHERING EXERCISE | 82 |
| B.1 Detailed Questions on Sharing Information | 82 |
| B.2 Detailed Questions on Data and Metadata..... | 86 |
| B.3 Detailed Questions on Modeling and Decision Support..... | 90 |
| APPENDIX C – LIST OF URLs USED FOR CLEANER LIBRARY WEBCRAWLING | 92 |
| C.1 Other Large-Scale Collaboratory Projects | 92 |
| C.2 Cyberinfrastructure information technologies | 93 |
| C.3 Research on science of collaboratories | 94 |
| C.4 Environmental Engineering & Hydrology..... | 95 |
| C.5 CLEANER Project Data | 96 |

LIST OF FIGURES

| | |
|---|----|
| Figure 1.1 – Venn diagrams of the scope of an individual researcher’s activities under the traditional research paradigm vs. the collaborative paradigm | 2 |
| Figure 1.2 – CLEANER Network and Examples of Stressed Environments | 7 |
| Figure 2.1 – Conceptualization of the Cyberinfrastructure | 11 |
| Figure 2.2 – Integrated Cyberinfrastructure Services to enable Collaborative environments..... | 12 |
| Figure 3.1 – CyberCollaboratory prototype for CLEANER..... | 28 |
| Figure 3.2 – Forums module in the CLEANER CyberCollaboratory | 30 |
| Figure 3.3 – Text chat module in the CyberCollaboratory | 31 |
| Figure 3.4 – Notebook module in the CyberCollaboratory | 32 |
| Figure 3.5 – Solution Center module in the CLEANER CyberCollaboratory | 33 |
| Figure 3.6 – Collaborative Editing module in the CyberCollaboratory..... | 34 |
| Figure 3.7 – Data section of the CyberCollaboratory | 36 |
| Figure 3.8 – Tracking simulated oil spill in the Data section | 37 |
| Figure 3.9 – Viewing and discussing oil spills in Data section | 38 |
| Figure 3.10 – Mockup of a workflow framework..... | 39 |
| Figure 3.11 – Community documents archive in the CyberCollaboratory library | 41 |
| Figure 3.12 – Web documents archive in the CyberCollaboratory library..... | 42 |
| Figure 3.13 – Search section of the CyberCollaboratory..... | 43 |
| Figure 4.1 – Creating a new account in the CLEANER CyberCollaboratory | 49 |
| Figure 4.2 – Survey results for Question 1 | 53 |
| Figure 4.3 – Survey results for Question 2 | 54 |
| Figure 4.4 – Survey results for Question 3 | 55 |
| Figure 4.5 – Survey results for Question 4 | 56 |
| Figure 4.6 – Survey results for Question 5 | 57 |
| Figure 4.7 – Survey results for Question 6 | 58 |
| Figure 4.8 – Survey results for Question 7 | 59 |
| Figure 4.9 – Survey results for Question 8 | 61 |
| Figure 4.10 – Survey results for Question 9 | 62 |
| Figure 4.11 – Survey results for Question 10 | 63 |
| Figure 4.12 – Survey results for Question 11 | 64 |
| Figure 4.13 – Survey results for Question 12 | 66 |

1. INTRODUCTION

Great discoveries and improvements invariably involve the cooperation of many minds.

~ Alexander Graham Bell

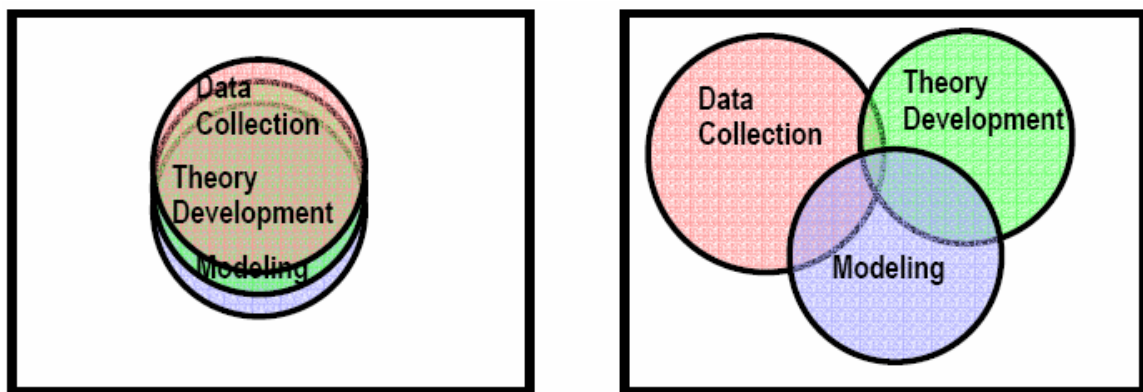
The scientific community has long recognized the value of collaborative investigation to address research problems. This recognition, combined with new information technology (IT), has led to a new collaborative paradigm of research. This section introduces the origins, fundamental ideas, and initial efforts to create this research model. It then presents an overview of the Collaborative Large Scale Engineering Analysis Network for Environmental Research (CLEANER) – a new initiative envisioning application of this collaborative paradigm to environmental engineering and science research. Finally, the purpose of this study is delineated.

1.1 The Collaborative Research Paradigm

Traditionally, most scientific research activities focus on individual projects with relatively narrow scopes and research objectives, which are conducted by a single investigator or a small team consisting of 2-3 members. However, it has been asserted that as the size, scope and complexity of research tasks grows, so does the need for collaboration among researchers (*Price*, 1963; *Zare*, 1997). Such collaboration may be necessitated by insufficient data, expertise, instrumentation or funds of individual researchers that could be alleviated by research partnerships among investigators having

complementary and supplementary resources. This need for collaboration lays the foundation of a new paradigm for research: the collaborative paradigm.

This concept is further illustrated in **Figure 1.1**. The rectangle represents all possible areas of research interest, and the circles depict the general tasks in a research activity – data collection, theoretical development and modeling – conducted by individual researchers. The traditional paradigm (**Figure 1.1a**) is characterized by a single researcher focusing on research tasks to fulfill his/her own project objectives. Collaboration, if any, is an informal process where a colleague asks the researcher to do some additional work “as a favor”. In contrast, under the collaborative paradigm (**Figure 1.1b**) in which a large number of individual investigators work in a partnership, each individual investigator’s research activities will not only serve his/her own objectives, but will also serve the purposes of others (*Eheart, 2004*).



1.1a: Traditional Paradigm

1.1b: Collaborative Paradigm

Figure 1.1 – Venn diagrams of the scope of an individual researcher’s activities under the traditional research paradigm (**Figure 1.1a**) vs. the collaborative paradigm (**Figure 1.1b**) (*Eheart, 2004*)

Customarily, collaborative research has depended heavily upon physical proximity. Studies have shown that quality and frequency of collaboration is inversely proportional to the distance between collaborators (*Allen, 1977; Kraut et. al., 1990; Katz, 1994*). To facilitate collaboration among geographically-disparate researchers, the conventional response has been residency, either temporary or permanent, at a common location. Despite its benefits, this solution has been found to cause a loss of productivity among researchers due to dislocation from a familiar environment (when a researcher must travel to a remote facility), as well as an inevitable exclusion and isolation from those located elsewhere (*Finholt, 2003*).

Another solution proposed by computer scientists and visionaries in the late 1980s was the concept of a virtual entity called the collaboratory: a combination of the words ‘collaborate’ and ‘laboratory’. As defined by *Wulf (1989)*, it is a “... a center without walls, in which researchers can perform their research without regard to physical location – interacting with colleagues, accessing instrumentation, sharing data and computational resources, and accessing information in digital libraries”. The terms co-laboratory, grid community/network, virtual science community, and e-science community are also used as synonyms for the collaboratory. In this model, the researchers remain at their native locations but interact with each other, access data, information and computational resources, and control instrumentation remotely through computer networks.

1.2 A Brief History of Collaboratory Efforts

Recognizing the potential of these collaboratories early on, the National Science Foundation (NSF) initiated programs in the early 1990s to support their design and coordination (*Rosenberg, 1991*). Researchers in physical oceanography, worm genomics and space physics were the first to implement collaboratory prototypes – SCIENCEnet, Worm Community System and Upper Atmospheric Research Collaboratory (UARC) respectively (*Hesse et. al., 1993; Schatz, 1991; Finholt and Olson, 1997*). Over time, with the advent of high-speed networking technologies, technological improvements in hardware, software and infrastructure, and the relatively modest costs associated with these technologies, conditions were optimal for the application of Internet-mediated collaboratories to various research disciplines (*Teasley et. al., 2001; Johnston et. al., 1997*). Indeed, numerous collaboratory initiatives have since been established across various disciplines – the Environmental Molecular Sciences Collaboratory (*Bair, 1999; Kouzes et. al., 1996*) sponsored by the Department of Energy; the National Institutes of Health (NIH)-supported Biological Collaborative Research Environment (BioCoRE) (*Bhandarkar et. al., 1999*) and the Great Lakes Regional Center for AIDS Research (GLRCFAR) (*Teasley et. al., 2001*); the Zebrafish Information Network (ZFIN) and the George E. Brown, Jr. Network for Earthquake Engineering Simulation (NEES) backed by the NSF, to name a few (*Finholt, 2003; Henline, 1998*). The NSF-funded Science of Collaboratories project has listed further detailed information about the many collaborative endeavors that have been undertaken over the past few decades on their website at <http://www.scienceofcollaboratories.org/Resources/colisting.php>. The

significant interest in the development of the scientific research and development collaboratories (excluding those which focus primarily on scientific education) can be gauged from the financial budgets of these projects, which ranged from \$447,000 to \$10,890,000 (*Finholt, 2001*).

More recently, NSF has proposed four large-scale environmental observatory initiatives that would be linked to form environmental collaboratories. These initiatives are in planning stages for obtaining Major Research Equipment and Facilities Construction (MREFC) funding. An MREFC is the NSF vehicle for developing large, shared-use research facilities. The proposed initiatives are the Consortium of Universities for the Advancement of Hydrologic Sciences Inc. (CUAHSI), the National Ecological Observatory Network (NEON), the Ocean Research Interactive Observatory Network (ORION) and the Collaborative Large Scale Engineering Analysis Network for Environmental Research (CLEANER) (*NSF, 2004*). Examples of other NSF MREFC programs include the Large Hadron Collider, the South Pole Station, EarthScope, the Atacama array of radio telescopes and NEES (*Reitherman, 2005*).

1.3 The CLEANER Initiative

To address the large-scale environmental problems facing the United States in the 21st century, fundamental knowledge is needed about the sources of contaminants, their links to different types and levels of human activities, their persistence, transport processes and degradation mechanisms and the risks they pose to the environment and humans (*NRC,*

2001). Because pollutants move between air, water, and land, there is a need to understand the interplay between these media and how efforts to control pollutants in one compartment affect environmental quality in other media. In addition, more effective ways to select among management strategies (*e.g.*, promoting the use of alternative materials versus developing enhanced waste treatment options) to address complicated environmental problems are required. To enable the engineering research and education communities to address these challenges of large-scale human-stressed complex environmental systems, the Directorate for Engineering proposed the CLEANER initiative (*NSF*, 2005).

Since 2001, a series of workshops and a national symposium (FAME: Frontiers in Assessment Methods for the Environment) have defined the general concept of CLEANER, which is envisioned to consist of groups of investigators studying landscapes stressed by human activities, supplemented by a national set-up of sensor-equipped interacting field sites (called environmental field facilities) and specialized support personnel and technology. These investigators are to be assisted by a network of computational resources that will provide access to and integrate shared repositories for data, models, and tools with shared facilities for experimentation and computation. This network, referred to as *cyberinfrastructure*, will also support data acquisition, analysis, searching, and information extraction; real-time simulation; and distributed collaboration. Furthermore, CLEANER is expected to enable more effective adaptive management approaches for human-stressed complex environmental systems based on enhanced observations, experimentation, modeling, engineering analysis, and design. It is also

expected to promote participation from the broad engineering and science community; and engage the academic community collaboratively in large-scale and complex real-world problems (Brezonik, 2005). **Figure 1.2** is a representation of the types of human-stressed complex environmental systems that could benefit from CLEANER.

A more detailed summary of previous CLEANER workshops has been created by *Gaber*, 2005.

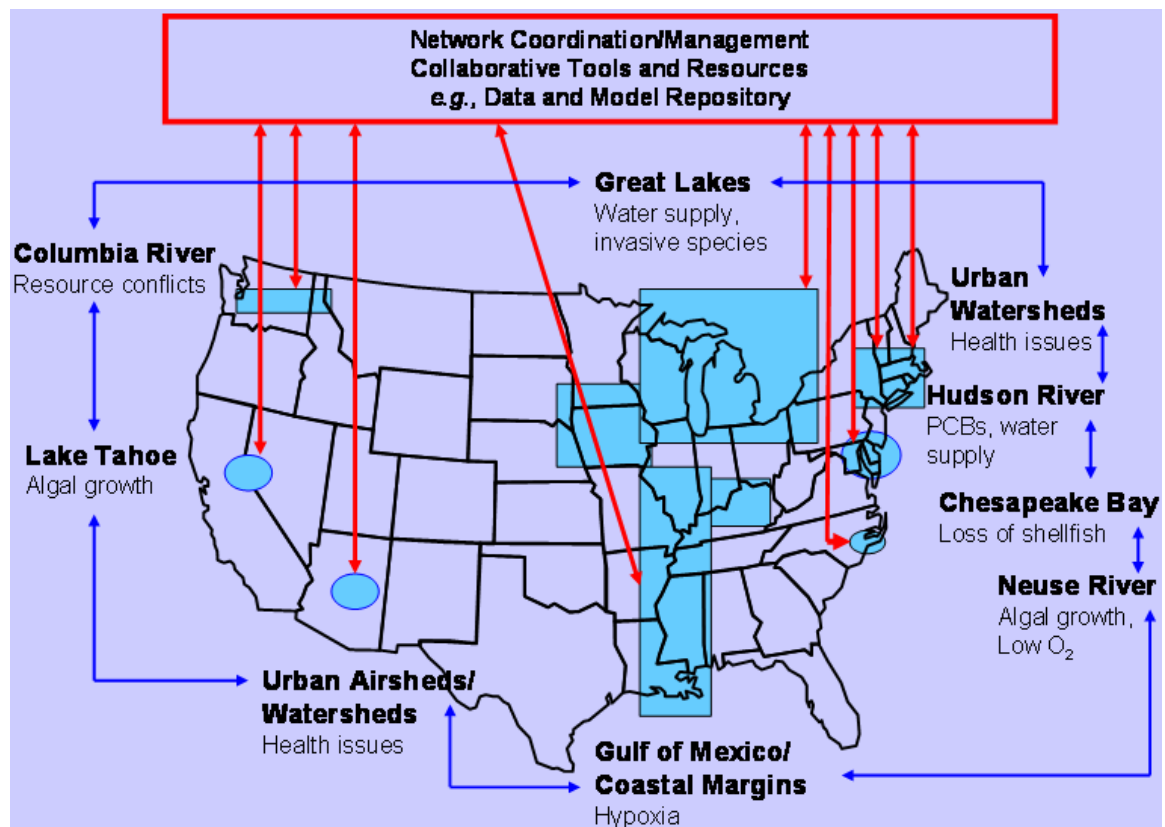


Figure 1.2 – CLEANER Network and Examples of Stressed Environments (Brezonik, 2005)

1.4 Purpose of this Study

In 2004, the NSF Directorate of Engineering awarded \$1 Million in planning grants to 12 projects focusing on cyberinfrastructure and management for the national CLEANER network and refining the nature of environmental field facilities (EFFs). Researchers at the University of Illinois Urbana-Champaign (UIUC) are participating in three of these planning grants, including one grant to identify needs for the cyberinfrastructure and develop a draft management plan for the Engineering Analysis Network (EAN) within CLEANER. This thesis is an outcome of that planning grant, with significant input from researchers involved in the other two UIUC planning grants. This thesis also focuses on identifying cyberinfrastructure requirements by learning from prior collaboratories in other fields and from gathering the opinions of researchers in environmental engineering and science.

This thesis is organized as follows. Chapter 2 gives background on the concepts used in our study and the needs assessment process used for other collaborative endeavors. Chapter 3 outlines the methodology used in our study and describes the pilot CLEANER CyberCollaboratory interface developed for requirements gathering. Chapter 4 gives results of a survey conducted via the CyberCollaboratory on CLEANER management and cyberinfrastructure. Finally, the conclusions of this exercise and review of cyberinfrastructure technology forms the closing chapter. Note that the requirements gathering process for CLEANER is still underway and will continue under the auspices of the newly-created CLEANER Project Office (<http://cleaner.ncsa.uiuc.edu>). Therefore

results presented in this thesis should serve only as initial input to the project office planning process, not as the final recommendations for CLEANER.

2. BACKGROUND ON CYBERINFRASTRUCTURE NEEDS ASSESMENT

This section discusses the emergence of the cyberinfrastructure concept as an integral part of the collaboratory, and outlines previous steps taken to implement this concept in the environmental engineering and science domain. It then explains why a needs assessment process is required for a cyberinfrastructure-enabled collaboratory environment, and the methods used in a generic needs assessment process. Finally, the needs assessment procedures adopted by other collaboratory exercises are explored.

2.1 Cyberinfrastructure Background

The term ‘cyberinfrastructure’ was coined by a NSF Blue Ribbon Advisory Panel to describe “the set of reliable, well-specified, and interoperable connections of electronic hardware and software that allows people to discover, learn, teach, collaborate, disseminate, access, and preserve knowledge in their domain” (*NCAR*, 2003). In the collaborative paradigm described earlier, cyberinfrastructure acts as the backbone of a collaboratory. Analogous to infrastructure (roads, power grids, telephone systems, bridges, rail lines, and similar public works) that is essential for an industrial economy to function, cyberinfrastructure is conceptualized to be indispensable to a knowledge economy (*Atkins et. al.*, 2003).

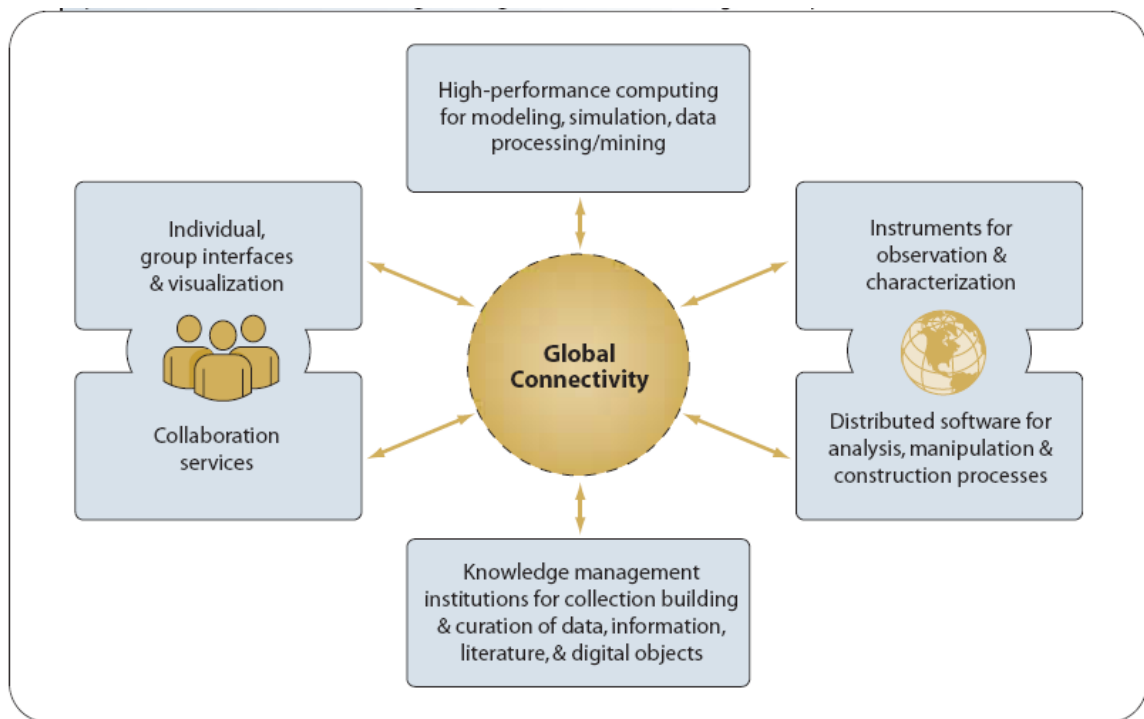


Figure 2.1 – Conceptualization of the Cyberinfrastructure (*NCAR*, 2003)

Figure 2.1 denotes the various components of the cyberinfrastructure: Networks that link wired and wireless communications devices; facilities for high-performance computing, shared data access, and generation of experimental results; instrumentation and sensor technologies for rapid estimation of mechanical, electrical, chemical, and biological responses; hardware and software technologies to support synthesis and fusion of data from various experimental and computational sources; software components that aid in domain-specific applications development, human-computer interaction, data replication and transfer, and remote authentication; services that allow collaboration among individuals (*NSF*, 2003).

One aspect of the conceptualized cyberinfrastructure being emphasized by the Blue-Ribbon Panel on Cyberinfrastructure is its interoperable design, *i.e.* its design should enable research communities to tailor application-specific collaborative environments around it. This concept is illustrated in **Figure 2.2**.

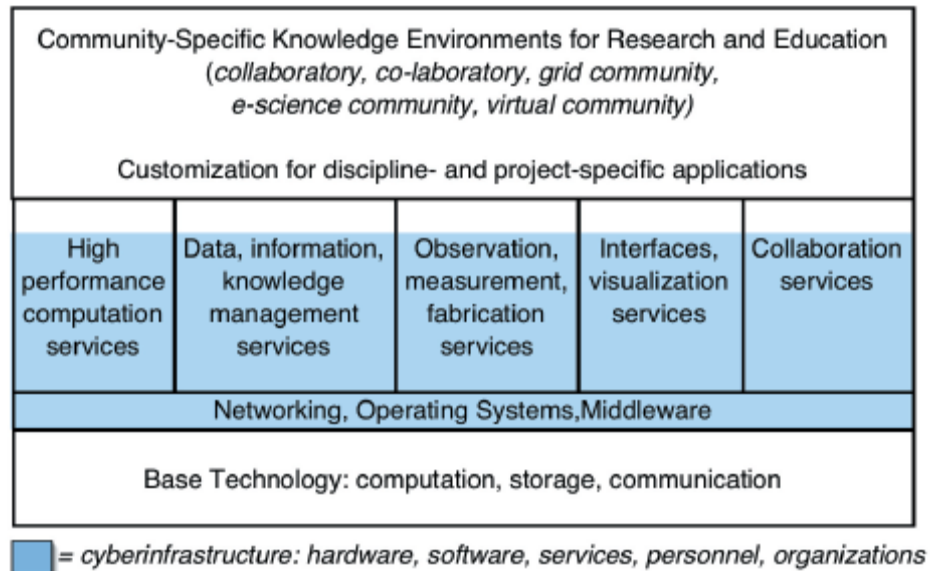


Figure 2.2 – Integrated Cyberinfrastructure Services to enable Collaborative environments (*NSF, 2003*)

With the advent of reliable high-speed networks, cheap and accessible computation, visualization and storage facilities, greater reliance on numerical computation and simulation for research, growing acceptance for Internet-mediated activities among the research community and previous strategic investments in base technologies, the Blue-Ribbon Advisory Panel feels that the time is right for major investment in cyberinfrastructure.

Cyberinfrastructure-enabled collaborative environments are poised to serve various disciplines such as atmospheric sciences, forestry, ocean sciences, computer science, medicine, astronomy, engineering, bioinformatics, physics, and social sciences, to name a few. It is expected that these collaboratories will revolutionize what individuals, teams and organizations can do, how they do it, and who participates (*Atkins et. al.*, 2003).

2.2 Cyberinfrastructure for Environmental Research & Education

There has been a significant movement towards addressing grand environmental issues by broadening the scope of environmental research. In its report, “Environmental Science and Engineering for the 21st Century” (2000), the National Science Board (NSB) recommended that NSF expand its efforts in environmental research and education by supporting critical needs in disciplinary and interdisciplinary environmental areas, advancing the current scope of environmental research using new tools and technologies, forming partnerships with other agencies, and enhancing environmental education and critical infrastructure. In a subsequent report, “Grand Challenges in Environmental Sciences” (2001), the National Research Council (NRC) identified priority areas for study by NSF and other agencies. An external Advisory Committee for Environmental Research & Education (AC-ERE) was formed to provide overall strategic guidance about environmental research and education areas specifically aligned with NSF’s mission (*Pfirman et. al.*, 2003).

In 2003, the AC-ERE published the report, “Complex Environmental Systems: Synthesis for Earth, Life, and Society in the 21st Century”, in which it outlined the recommended direction of cross-Foundation environmental research and education at NSF during 2003-2012. In a follow-up report titled “Complex Environmental Systems: Pathways to the future” (2005), the AC-ERE further highlighted timely pathways for NSF’s Environmental Research & Education. These reports identified cyberinfrastructure as a suite of critical enabling tools and research essential to the study of complex environmental systems (*Futrell et. al.*, 2003).

The Blue-Ribbon Advisory Panel on Cyberinfrastructure noted that the environmental science & engineering community is one that would benefit greatly among the various disciplines to which advanced cyberinfrastructure technology could be applied. There are a range of attributes of the environmental research domain that make it an ideal adopter of the cyberinfrastructure-enabled collaborative environments: many research activities are observationally oriented, highly collaborative and interdisciplinary; spatial and temporal scales of interest vary widely from the very large to the very small; there is a plethora of data formats and manipulation techniques and a marked range in the volumes of raw data generated; an increasing demand for quicker data analysis using sophisticated tools; and an increased focus on policy and management issues. Indeed, cyberinfrastructure has been identified as a *sine qua non* for successful next-generation research programs in this domain (*NCAR*, 2003).

2.3 Needs Assessment: A Precursor to Cyberinfrastructure Development

The Blue-Ribbon Advisory Panel on Cyberinfrastructure warned that there are significant risks and costs involved for the United States if prompt action with sufficient investment is not undertaken towards creating these cyberinfrastructure-enabled collaboratories. Noteworthy among the risks is lack of compatibility among various disciplinary research activities causing loss of interoperability; wasteful redundant system-building activities among science fields or between science fields and industry; and lack of synergy among the IT research, industry and domain science users resulting in under- or overestimating technological futures. Other perils of inaction include loss of observational data due to non-development of well-curated, long-term archives; lack of understanding of social/cultural barriers to new ways of doing research; inadequate support for educational activities; and a loss of leadership to other countries and a falloff of research and economic vigor.

Even though there is a need for timely action, it should be realized that this collaborative endeavor faces fundamental cultural, sociological and technological challenges and a dearth of examples to study. Part of the challenge lies in enabling deeper collaboration between computer scientists (who build the applications for the cyberinfrastructure) and domain researchers (who are the end-users of these applications). Incentives are needed for involvement for all involved in creation of these collaborative endeavors. The technological tools and services required for the collaboratory process are still maturing. Moreover, this development process requires a major investment of time and money -

e.g., The Blue-Panel Panel's recommendation to NSF is an annual investment of \$1 billion for cyberinfrastructure development and allied activities.

Given all these considerations, it is imperative that careful planning be done to ensure that the promises of the cyberinfrastructure revolution are not unfulfilled. Hence, an initial needs assessment is a necessary exercise for the CLEANER cyberinfrastructure.

2.4 The Needs Assessment Process

Needs assessment, or requirements elicitation, is an exercise that is routinely conducted for software engineering. Indeed, it is a part of the generic requirements engineering process which involves feasibility study, requirements elicitation and analysis, requirements specification and documentation, and requirements validation, often followed by requirements management to accommodate a change in requirements (*Sommerville, 2004*).

A feasibility study is designed to evaluate whether the proposed system is contributing to the objectives of the organization; if the system can be implemented using current technology and cost constraints; and whether the system can be integrated with other systems already present.

The next step, requirements elicitation and analysis, generally involves developers of the system working with the end-users to discover more about the application domain and the

services, hardware, and performance requirements of the proposed system. It is essential to involve as many *stakeholders* (individuals who would be have a direct or indirect influence on the system requirements), as possible in this step to broaden the results of the elicitation exercise.

This requirements elicitation and analysis process involves the following activities: domain understanding by the developers, requirements collection, classification, conflict resolution, requirements prioritization, and checking. This process has many hurdles due to the ambiguity of the requirements expressed by the stakeholders, conflict of requirements, and simultaneous change in requirements caused by the dynamic nature of the organizational environment (*Sommerville, 2004*).

Some of the techniques used in the needs gathering process are viewpoint-oriented elicitation, interviews, surveys, scenario-based design and ethnography. Apart from these conventional techniques, Human Based Genetic Algorithm (HBGA) is an emerging technique that is used in this project. Any elicitation process generally involves a combination of some or all of these techniques, and might also involve other methods. The sources of information in all these techniques are the system stakeholders and the specifications of similar systems. The fundamentals of some of the requirements elicitation techniques are explained in the following subsections.

2.4.1 Viewpoint-Oriented Analysis

Viewpoints are a way of structuring the requirements to represent the perspectives of different types of stakeholders. Stakeholders of a system may be classified under different viewpoints such as end-users, developers, operators, users of collaborating systems, *etc.* These stakeholders then elaborate their specific needs, which are kept in mind during the development of the system. This multi-perspective analysis is important as it recognizes the diversity of the requirements, and provides a framework for discovering conflicts (*Sommerville, 2004*).

2.4.2 Interviews and Surveys

Interviews are formal or informal communications between the system developers and the stakeholders where the stakeholders respond to a pre-defined set of questions (closed interviews) and are asked opinions on a range of issues which are not pre-defined (open interviews). Though interviewing is a good way of getting an overall understanding of the domain, it is not a complete way of understanding the domain requirements unless the developers are conversant with the domain terminology (*Sommerville, 2004*).

Surveys or questionnaires are another way of gathering opinions of a participating group of stakeholders, and like interviews, can be open-ended or close-ended. Surveys are administered in paper format or via the Internet, and a certain time frame is allotted for the stakeholders to provide responses. Surveys can be a beneficial component of the

requirements elicitation exercise as they let the participants prepare their responses at their leisure and allow more time to develop responses for open-ended questionnaires. The drawback of this method is that it requires a captive participating group to dedicate their time and attention, a requirement which can be difficult to meet. Online surveys are gaining in popularity, but they capture only the opinions of participants willing to go online or who have web access at the right time.

2.4.3 Scenario-based Design

Scenario-based design is “a family of techniques in which the *use* of a future system is concretely described at an early point in the development process” (*Rosson et. al.*, 2002). In the analysis of human-computer interactions, scenarios can be used to describe existing tasks using existing technology and future tasks using envisioned technology (*Carroll et. al.*, 1998). The scenarios created generally provide a description of the system at the starting situation; describe normal flow of events; anomalous behavior; information about other concurrent activities and the state description when the scenario finishes (*Sommerville*, 2004). Use-case scenarios are a subset of scenario-based design techniques in which scenarios are crafted as stories in which certain actors, called *personas*, encounter and manipulate various tools and objects in a setting. Scenario-based design is a popular technique because it lucidly exemplifies the system usage possibilities and concerns among various stakeholders. Also, scenarios are easy to write and modify, and hence easy to incorporate in a requirements elicitation exercise.

2.4.4 *Ethnography*

Ethnography, an observational technique that can be used to understand social and organizational requirements, involves the requirements engineer/systems developer immersing himself/herself in the working environment of the end-user. This technique has the advantage that the stakeholders do not have to explain or articulate their work, and therefore richer details about the system requirements may be gathered that are tacitly understood in the organization. But this technique has a drawback in that it cannot always identify new features required by the system, and hence is not a complete approach to elicitation. When combined with simultaneous system prototype development, this approach is called focused ethnography as it provides a direction to the ethnographic analysis (Sommerville, 2004).

2.4.5 *Human-Based Genetic Algorithms*

Human-Based Genetic Algorithms (HBGAs) are a new variant of the Genetic Algorithm (GA) technique. GAs are search procedures learned from Nature and are based on mechanics of natural selection and genetics like reproduction, crossover and mutation.

The basic premise of the GA technique is that if sets of genes (ideas) are representations of possible solutions to a certain optimization problem, then the initial sets of genes (ideas) are generally created by a random drawing, and the genetic operators (of reproduction, crossover and mutation) are then carried out on them. The objective is to

come up with a set of genes that constitutes the optimal value to a certain fitness function (evaluation) (*Goldberg*, 1989).

Interactive Genetic Algorithms (iGAs) are another variant of GAs generally used in domains that have a subjective judgment of fitness, such as evolving images, music, and aesthetic arts. As the fitness functions are not rigorously defined in these domains, humans are used to evaluate the fitness of ideas generated via the genetic operations (see *Goldberg*, 2003 for an application of iGA).

The concept behind HBGA is based on the variation of iGA where the initial stream of sets of genes (ideas) is also generated by humans and humans also act as critics (evaluators) of those ideas. In essence, humans perform all the genetic operations in these GAs (*Kosorukoff et. al.*, 2002).

HBGA could also be used as a requirements elicitation tool when coupled with the survey technique, to help elicit needs of the system by outlining some needs as initial ideas and letting the stakeholders define other needs based on variations and combinations of those needs. For a more detailed discussion on HBGA and its applications, please refer to *Kosorukoff*, 2005.

2.5 Needs Assessment Exercises for Other Collaborative Endeavors

According to the personnel at the Science of Collaboratories project, only a few collaborative endeavors have undergone extensive needs assessment studies. This section outlines the requirements elicitation procedures employed by three ongoing collaboratory projects – the George E. Brown, Jr. Network for Earthquake Engineering Simulation (NEES), the Consortium of Universities for Advancement of Hydrologic Sciences, Incorporated (CUAHSI) and the GEOscience Network (GEON).

2.5.1 *NEES*

NEES, another MREFC initiative, has built a collaboratory for conducting earthquake engineering research by earthquake, structural and tsunami researchers. Under a project management team consisting of domain scientists, computational experts and collaborative-environment developers, NEES conducted a scoping study in the year 2001, three years prior to the deployment of the NEESgrid collaborative environment (now called NEESit, see <http://it.nees.org/> for more details). During the community needs assessment phase of this scoping study, an external advisory committee was formed to augment community representation. Regular site visits and interviews were conducted at locations with NEES grants to study the nature of research and the potential and existing collaboration across sites. In addition, web-administered questionnaires with close-ended technical and functional inquiries were developed to gather the opinions of the NEES community on current practices and expectations from NEES. Finally, community

requirements workshops were organized with breakout sessions providing a forum for discussions on numerous earthquake engineering and IT issues (*Prudhomme et. al.*, 2001). All of these activities led to the development of a document detailing the identified core user requirements, called the “NEESgrid User Requirements” (see *Finholt et. al.*, 2002 and 2003 for more details).

2.5.2 CUAHSI

CUAHSI is an organization whose purpose is to develop infrastructure and services for the advancement of hydrologic science and education in the United States. With NSF support, CUAHSI is in the process of developing demonstrations and specifications for a Hydrologic Information System (HIS) to meet the needs of faculty, students, and researchers. The HIS needs assessment process included three main steps: preliminary information gathering, a pilot survey, and a web survey. The process began with collaborators to the HIS project team gathering preliminary information from their institutions. This preliminary information was presented at the HIS Symposium in March 2005. At the symposium, a pilot paper survey was conducted using feedback from the preliminary information gathering efforts. The results of both the information gathering and pilot survey were then used to develop the web survey that was conducted in May 2005 (*Maidment et. al.*, 2005).

In addition to these surveys, CUAHSI also built prototypes of tools, libraries and database designs to help demonstrate the utility of the HIS to the hydrologic science community (*Maidment et. al.*, 2005).

2.5.3 GEON

GEON is a five-year multi-disciplinary research initiative conducted jointly by earth scientists and computer scientists aimed at developing a cyberinfrastructure for the geosciences. GEON has been developed in response to geoscience researchers' needs to interlink and share multidisciplinary data sets and tools in order to better understand the complex dynamics of Earth systems (*Dogan et. al.*, 2004). The GEON cyberinfrastructure consists of a network of powerful computers (called the *grid*) having nodes all across the country, while the technical core team is based at the San Diego Supercomputer Center (SDSC).

While conducting conventional requirements elicitation exercises like scenario-based design, the primary needs assessment process of GEON was based on the concept that needs elicited while addressing real scientific questions will lead to the discovery of needs for any generalized activity in that domain (*Ribes*, 2005). In the beginning of the GEON project in 2002, two test-beds – the Rocky Mountains and the Mid-Atlantic region – were identified and the topics being studied in these test-beds were intra-continental deformation and *terrane* recognition and analysis, respectively (see *Keller*, 2003 for more details). Through a close and regular collaboration between IT experts and the

participating primary investigators (PIs) involved at these test-beds, cyberinfrastructure tools and technologies are being developed to meet the PI needs. It is expected that technologies developed for application to these real science questions and actual research sites would eventually be suitable for meeting the requirements of a broader scientific community.

3. NEEDS ASSESSMENT METHODOLOGY

This chapter presents the approach taken towards requirements elicitation for CLEANER. Section 3.1 outlines the tasks performed in this study. Section 3.2 provides an overview of the prototype collaboratory interface used during this exercise.

3.1 Preliminary Tasks

In order to form a community consensus on the needs of the national CLEANER network (EAN), a project team was formed in the summer of 2004 to spearhead the needs assessment effort. University of Illinois personnel in the project team included researchers at the National Center for Supercomputing Applications (NCSA), the Department of Civil and Environmental Engineering, and other researchers with interests in human-computer interactions, cognitive science & psychology, and communication & organization theory. University of Michigan personnel in the project team were from the Collaboratory for Research on Electronic Work (CREW). See Appendix A for a specific listing of the project team members.

An external advisory committee was also created to augment participation in this needs assessment process. It has been emphasized time and again that one of the major challenges in creating such cyberinfrastructure-enabled environments is guaranteeing smooth communication between domain scientists and computer and computational researchers (*NCAR*, 2003; *Atkins et. al.*, 2003; *NSF*, 2003). To ensure this, committee

members were chosen from both cyberinfrastructure and collaboratory experts and representative members of the engineering community (see Appendix A for a listing of the advisory committee members). The representatives from the engineering community were drawn primarily from the principal investigators of the other 12 CLEANER planning grants. Other members were chosen to represent domain diversity within environmental engineering, research approach diversity (*e.g.*, laboratory experiments, field experiments, modeling, *etc.*) and institutional diversity (*e.g.*, undergraduate, minority serving and graduate institutions).

3.2 Development of a Prototype CLEANER CyberCollaboratory

The NSF Workshop on Cyberinfrastructure for Engineering, Research & Education (2003) observed that “the long-term requirements for national cyberinfrastructure are in large part unknown *a priori*, so they cannot be estimated with complete accuracy without concomitant incremental efforts at construction of this electronic infrastructure”. Keeping this in mind, a prototype electronic collaboratory – named *CyberCollaboratory* – was begun under this project, which is accessible from a website on the Internet (<http://cleaner.ncsa.uiuc.edu>). The purpose of the CyberCollaboratory was to help the community visualize a potential collaboratory user interface through mockups and demonstrations.

The CyberCollaboratory prototype was built using portal technology from Liferay Portal Enterprise (see <http://www.liferay.com/web/guest/home> for more details). Liferay was

chosen because it is open-source, a mature and stable portal technology, and compliant with the latest standards for portlets (Java Specification Request 168 - JSR 168) and web services (WSRP - Web Services for Remote Portlets). Furthermore, its Java-based platform-independent approach meant easier deployment on the various operating systems being used. The module-based environment of a portal system makes the process of adding or removing various tools relatively simple, allowing ready integration of demonstrations created by different groups. **Figure 3.1** shows the current version of the CyberCollaboratory, which includes tools and demonstrations organized into six sections: Home, Collaboration, Data, Analysis, Library, and Search.

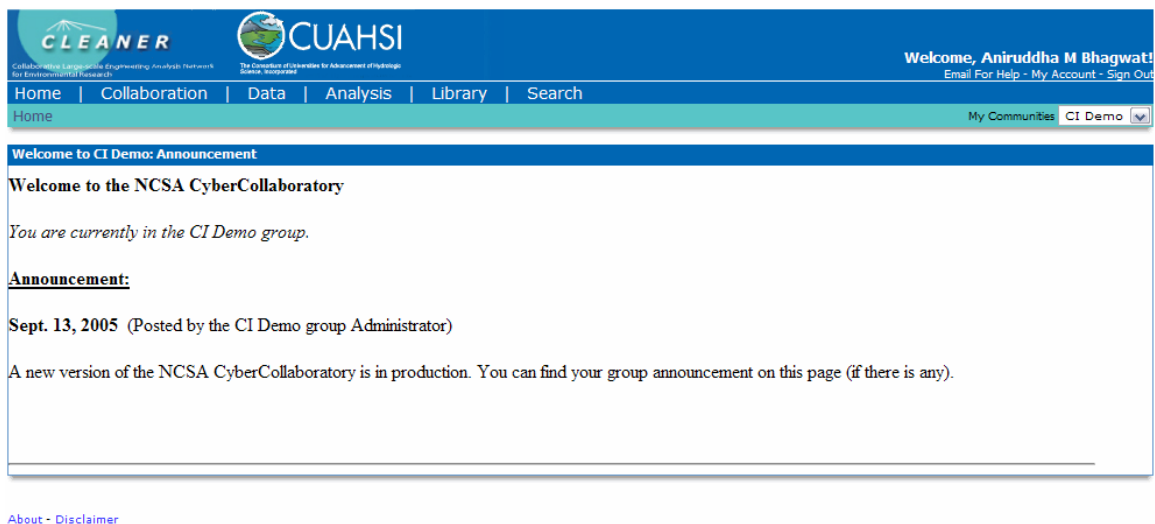


Figure 3.1 – CyberCollaboratory prototype for CLEANER

The specific sections and views in the CyberCollaboratory are easily customized for different groups, which are currently defined by the administrator. The drop down menu on the upper right side of **Figure 3.1** enables users to rapidly switch from one group to another. The view shown in **Figure 3.1** is specific to the cyberinfrastructure

demonstrations; working groups in the CLEANER Project Office only have access to functioning tools in the Collaboration, Library, and Search sections. This section documents the full set of technologies and demonstrations currently deployed in the “CI Demo” group within the CyberCollaboratory, organized by section below.

3.2.1 Home Section

This section of the CyberCollaboratory is a personal space for every individual user where members can find relevant announcements here from their group administrator. This section can be customized by advanced users, although this capability is currently disabled to avoid confusion among early adopters.

3.2.2 Collaboration Section

The Liferay portal came with basic communication modules like forums (or bulletin boards) and text chat used in other collaboratory environments, which were deployed in the CyberCollaboratory for community evaluation. Forums allow users to participate in asynchronous threaded discussions (*i.e.*, where users communicate at different times and messages are grouped by themes called threads). **Figure 3.2** shows the forums module in the CyberCollaboratory.

Text chat facilitates real-time communication between remote users. Users can create chat rooms, and others can join the chat and type text messages that are visible to all who have joined the chat room. The text chat module of the CyberCollaboratory can be seen in **Figure 3.3**.

CLEANER
Collaborative Learning and Engineering Analysis Network
Real Engineering, Real Research

CUAHSI
The Consortium of Universities for Advancement of Hydrologic Science, Incorporated

Welcome, Aniruddha M Bhagwat!
Email For Help - My Account - Sign Out

Home | Collaboration | Data | Analysis | Library | Search

Collaboration » Forums

My Communities CI Demo

Forums

The forum allows users to participate in asynchronous threaded discussions [i.e., where users communicate at different times and messages are grouped by themes ("threads")].

If this is your first time here, a short tutorial is available at
<http://cleaner.ncsa.uiuc.edu/tutorials/forums.htm> (opens in a new window)

Forums

[Add Forum](#) - [List Forums](#) - [Search](#)

| Forum | # of Threads | # of Posts (New/Total) | Last Post Date |
|----------------------------------|--------------|------------------------|------------------|
| Cyberinfrastructure Organization | 0 | 0 / 0 | Never |
| Joint workshop November 2005 | 0 | 0 / 0 | Never |
| CLEANER-CUAHSI Discussions | 5 | 11 / 12 | 8/26/05 1:22 PM |
| Gulf Hypoxia | 2 | 1 / 2 | 7/29/05 11:42 AM |

[About](#) - [Disclaimer](#)

Feedback Window

Feedback for Current Page

Please type in your comments in the following box:

☒ Post to public forum
☐ Report problem(s) to System

Administrator

Figure 3.2 – Forums module in the CLEANER CyberCollaboratory

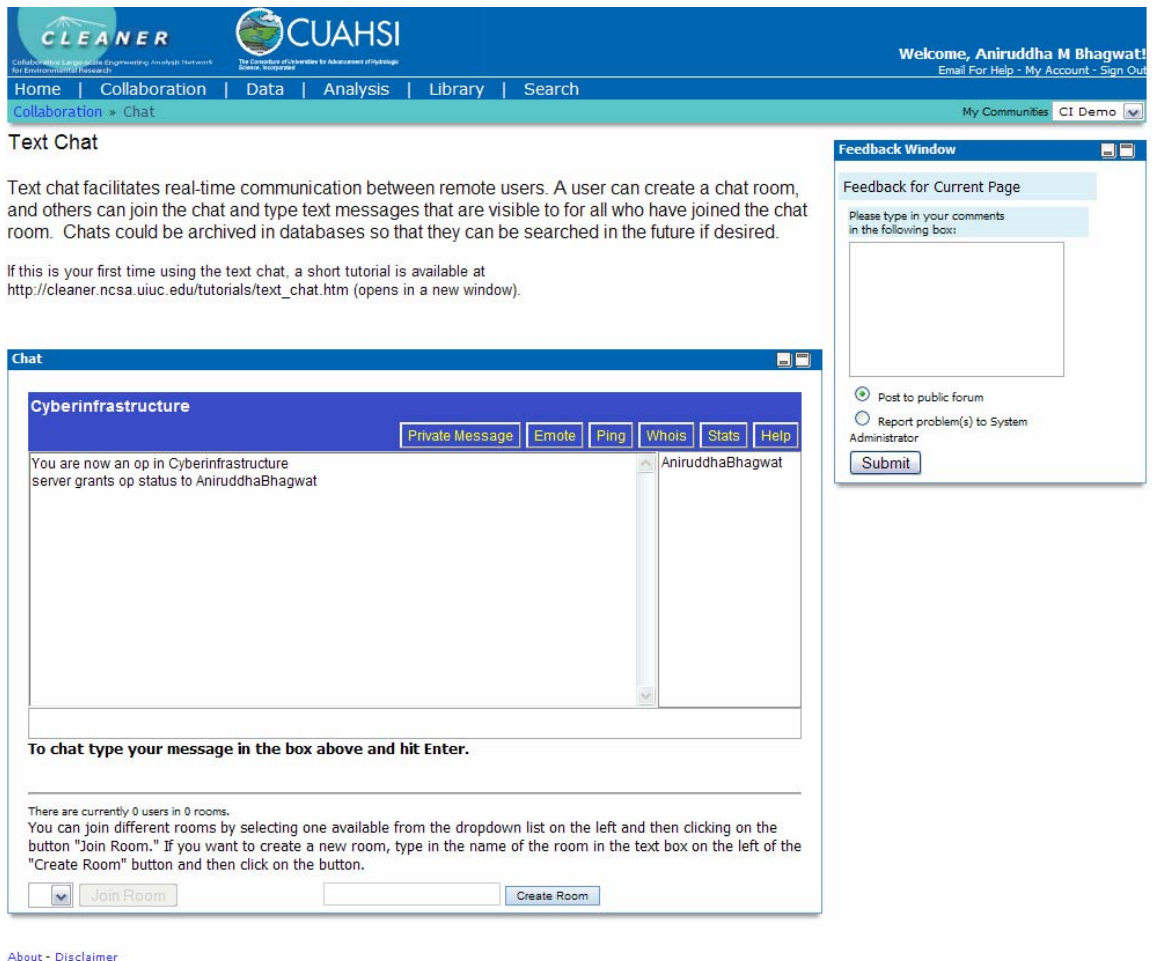


Figure 3.3 – Text chat module in the CyberCollaboratory

Another collaboration module that came with the Liferay portal was the Notebook. The notebook provides every user with a personal space for preparing notes, capturing data, and recording results. Other users can view the user's notebook (if they are given permission) and provide comments or ratings. Shared editing of notebooks is not currently enabled, but could be implemented in the future. **Figure 3.4** shows the notebook module in the CyberCollaboratory.



Figure 3.4 – Notebook module in the CyberCollaboratory

In addition to these generic tools, a new tool was added to the CyberCollaboratory that was adapted from an emerging system called Distributed Innovation and Scalable Collaboration in Uncertain Settings (DISCUS). DISCUS is a web-based technology under development in the Automated Learning Group (ALG) at the NCSA and the Illinois Genetic Algorithms Laboratory (IlligAL) (for details see *Goldberg et al.*, 2003). The HBGA module in DISCUS - called the Solution Center - can be used to gather free-form solutions to presented problems, vote for them, and encourage users to form new solutions based on existing solutions. This Solution Center module was adapted to

conduct the requirements gathering for CLEANER within the CyberCollaboratory, as seen in **Figure 3.5**.

CLEANER CUAHSI
 Collaboration | Data | Analysis | Library | Search
 Welcome, Aniruddha M Bhagwat! Email For Help - My Account - Sign Out

Solution Center

The Solution Center helps geographically-dispersed groups with reaching consensus on questions or problems, and is based on technology developed at the Illinois Genetic Algorithms Laboratory. Users can post questions/problems, propose solutions, and vote for promising solutions. All questions are grouped into categories. To post a question, first create a category for it (if that particular category does not exist already). Then add the question into that particular category.

If this is your first time using the solution center, a short tutorial is available at http://cleaner.ncsa.uiuc.edu/tutorials/solution_center.htm (opens in a new window)

| Category Name | # of Questions | Last Posted Date | Initiated By |
|---|----------------|-----------------------|---------------------|
| CLEANER Survey: General Questions | 13 | 2005-06-21 12:33:26.0 | Aniruddha M Bhagwat |
| CLEANER Survey: Detailed Questions on Sharing Information | 11 | 2005-04-22 04:54:27.0 | Aniruddha M Bhagwat |
| CLEANER Survey: Detailed Questions on Data and Metadata | 8 | 2005-04-21 20:20:21.0 | Aniruddha M Bhagwat |
| CLEANER Survey: Detailed Questions on Modeling and Decision Support | 3 | 2005-04-21 20:27:47.0 | Aniruddha M Bhagwat |

Feedback Window

Feedback for Current Page

Please type in your comments in the following box:

☒ Post to public forum
☐ Report problem(s) to System Administrator

Figure 3.5 – Solution Center module in the CLEANER CyberCollaboratory

To allow users to collaboratively create and edit documents, a wiki module (see <http://wiki.org/wiki.cgi?WhatIsWiki> for details on the wiki concept) was added to the collaboration section. Powered by the open-source MediaWiki portal, this collaborative editing tool was used to create documents containing instructions for providing input to the CLEANER planning project as well as to create use-case scenarios illustrating the possible usage of CLEANER cyberinfrastructure. This module can be seen in **Figure 3.6**. The Liferay portal system also contains an internal wiki that was used for posting descriptions of the demonstrations, but is not available for public use because the project team concurred that it was significantly more difficult to use than the MediaWiki system.

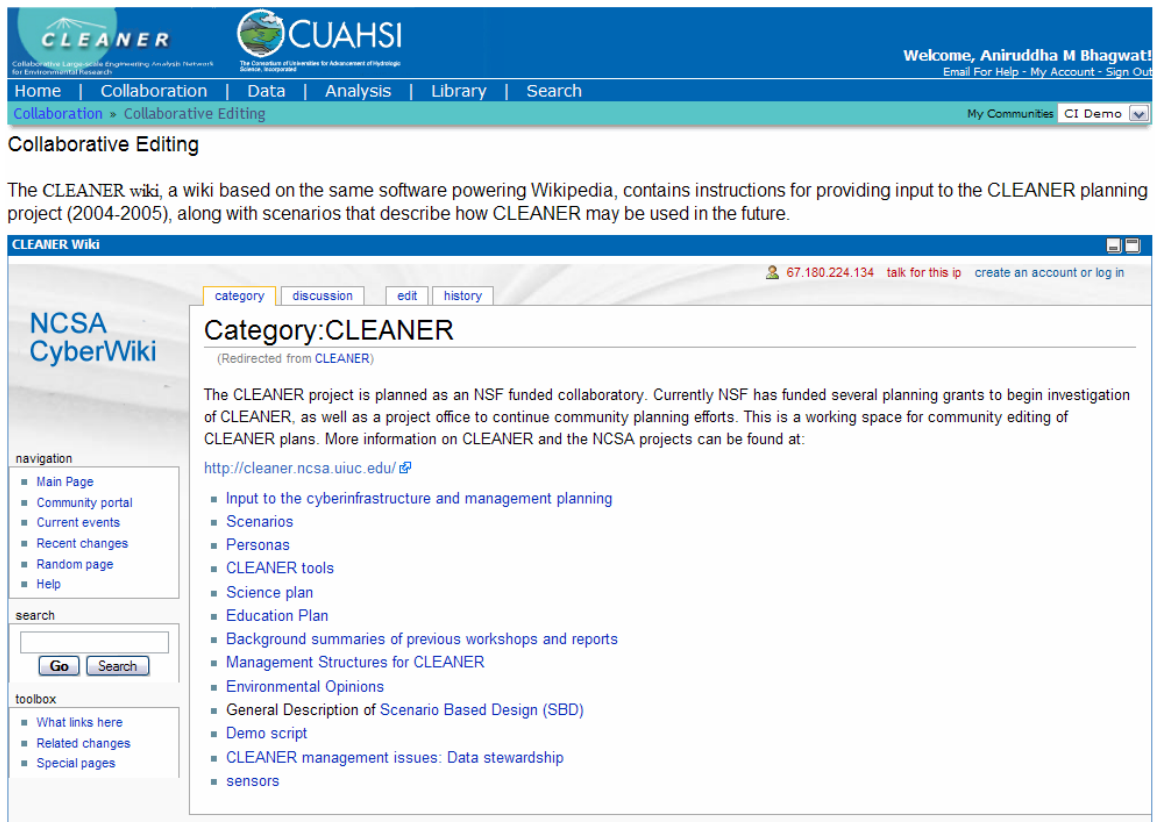




Figure 3.6 – Collaborative Editing module in the CyberCollaboratory

3.2.3 Data Section

The data section is envisioned as providing functionality for browsing data, managing meta-data (information about the data), searching data, data assimilation and syndication, and managing live data feeds from sensors. It is expected that this section would have links to each environmental field facility (EFF) in the CLEANER network, as well as capabilities for searching across the EFF datasets. Ideally, the data section should also allow users to identify and integrate data available from other sources (e.g., Federal agency data) for each EFF, creating a digital environmental observatory similar to the proposal of the CUAHSI HIS team for a digital hydrologic observatory (*Maidment et. al.*,

2005). Further, users would have the ability to initiate live collaboration with other researchers who are viewing the same data, via a text chat or audio link.

A demonstration of this type of collaborative data viewing was implemented in the Data section (see **Figure 3.7**). The demonstration illustrates how users in remote locations can easily discuss oil spill and surface current information generated at Texas A&M University in real time. By clicking on “Track Simulated Oil Spill,” selecting a date for the spill, and clicking on the map of Corpus Christi Bay in Texas to create the release of oil and begin tracking the trajectory of a spill from that location, one can watch the spill being transported based upon near-real time or historical surface current data generated from high frequency (HF) radar, as seen in **Figure 3.8**. All simulations are stored in a back-end database and can be shared and discussed with other users under the “View and Discuss Created Oil Spills” link as shown in **Figure 3.9**. Each simulation is listed in the table in **Figure 3.9**, including information on the user who created it, the date of the oil spill, the time of the simulation, and the number of users currently discussing the results. In the future, users could also be able to leave comments on previous results for other users, as well as access audio or videoconferencing technology to share results. Other features in the CyberCollaboratory could also be made available in the future for collaborative viewing (*e.g.*, shared documents or analyses).

Collaborative Large-scale Engineering Analysis Network for Environmental Research

The Consortium of Universities for Advancement of Hydrologic Science, Inc.

Welcome, Aniruddha M Bhagwat!

[Email For Help](#) - [My Account](#) - [Sign Out](#)

[Home](#) | [Collaboration](#) | [Data](#) | [Analysis](#) | [Library](#) | [Search](#)

[Data](#) > CCBay Data Sharing Demonstration

[My Communities](#)
[CI Demo](#)

Data

The data portion of the system will provide functionality for browsing data, managing meta-data (information about the data), searching data, data assimilation and syndication, and managing live data feeds from sensors. This page would have links to each Environmental Field Facility (EFF) in the CLEANER network, as well as capabilities for searching across the EFF datasets.

One key feature planned for this page will be the ability to initiate live collaboration with other researchers who are viewing the same data, via a text chat or audio link. Below is a demonstration of how this would work using live and historical sensor data from the Shoreline Environmental Research Facility.

Another planned feature would be the ability to annotate data or data summaries (e.g., the visualizations shown in the demonstration), with the annotations stored with the data for others to view (i.e., part of the metadata).

CCBay Data Sharing Demonstration

[Track Simulated Oil Spill](#) | [View and Discuss Created Oil Spills](#)

Collaborative Data Viewing Demo: Oil Spills

Please select one of the options in the menu above.

Track Simulated Oil Spill allows simulating an oil spill in Corpus Christi Bay based on high frequency radar data accumulated over the course of several years.

Once a simulated oil spill has been created, multiple users can access that simulation from the second option above, **View and Discuss Created Oil Spills**, and discuss the results through a text chat room specific to that simulation.

Feedback Window

Feedback for Current Page

Please type in your comments in the following box:

☒ Post to public forum
☐ Report problem(s) to System Administrator

Figure 3.7 – Data section of the CyberCollaboratory

[Track Simulated Oil Spill](#) | [View and Discuss Created Oil Spills](#)

Track Simulated Oil Spill

To simulate on oil spill based on sensor's data please follow the following simple steps:

(1) Select a date and a time of the day.

Date: 2005 January 01

Hour of the Day: 01

(2) Select the number of hours the simulation will cover.

Number of Hours: 12 (1 hour per frame)

(3) Select a point on the map where the oil spill would take place by clicking on the map. Clicking on the map will automatically submit the information.

Please be patient as the animation requires some time to load. If the animation includes no oil spill, then the data for that period is not available. In that case, please try a different date.

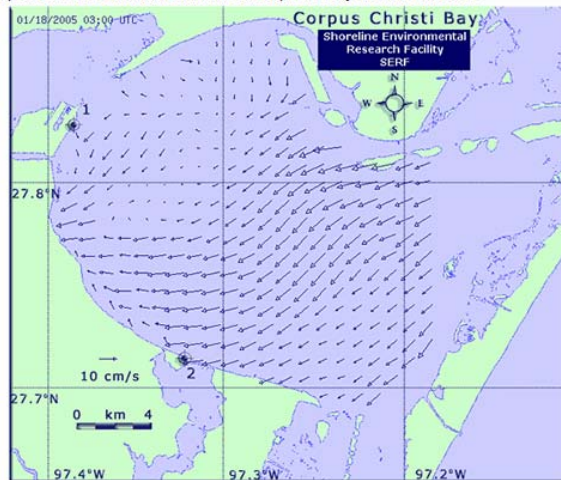




Figure 3.8 – Tracking simulated oil spill in the Data section

Collaborative Long-term Engineering Analysis Network for Environmental Research

The Consortium of Universities for Advancement of Hydrologic Science Research

Welcome, Aniruddha M Bhagwat!

[Email For Help](#) - [My Account](#) - [Sign Out](#)

[Home](#) | [Collaboration](#) | [Data](#) | [Analysis](#) | [Library](#) | [Search](#)

[Data](#) » CCBay Data Sharing Demonstration

[My Communities](#)
[Ct Demo](#)

CCBay Data Sharing Demonstration

[Track Simulated Oil Spill](#) | [View and Discuss Created Oil Spills](#)

View and Discuss Created Oil Spills

Select the item from the list that you would like to view and discuss. Most columns can be sorted for easier browsing.

16 items found, displaying all items.

| Spill Id | System | User | Date Selected | Time Created | Users | View and Discuss |
|----------|--------|-----------|--------------------|--------------------|-------|----------------------------------|
| 207 | CCBY | dtrujillo | September 20, 2005 | September 21, 2005 | 0 | View and Discuss |
| 205 | CCBY | dtrujillo | September 19, 2005 | September 20, 2005 | 0 | View and Discuss |
| 204 | CCBY | dtrujillo | September 19, 2005 | September 20, 2005 | 0 | View and Discuss |
| 203 | CCBY | dtrujillo | September 19, 2005 | September 20, 2005 | 0 | View and Discuss |
| 202 | CCBY | dtrujillo | September 19, 2005 | September 20, 2005 | 0 | View and Discuss |
| 201 | CCBY | dtrujillo | September 20, 2005 | September 20, 2005 | 0 | View and Discuss |
| 200 | CCBY | dtrujillo | September 19, 2005 | September 19, 2005 | 0 | View and Discuss |
| 199 | CCBY | dtrujillo | September 15, 2005 | September 15, 2005 | 0 | View and Discuss |
| 197 | CCBY | dtrujillo | September 15, 2005 | September 15, 2005 | 0 | View and Discuss |
| 196 | CCBY | dtrujillo | September 15, 2005 | September 15, 2005 | 0 | View and Discuss |
| 195 | CCBY | lmarini | September 14, 2005 | September 14, 2005 | 0 | View and Discuss |
| 194 | CCBY | lmarini | September 14, 2005 | September 14, 2005 | 0 | View and Discuss |
| 193 | CCBY | dtrujillo | September 12, 2005 | September 14, 2005 | 0 | View and Discuss |
| 192 | CCBY | dtrujillo | September 13, 2005 | September 14, 2005 | 0 | View and Discuss |
| 191 | CCBY | dtrujillo | September 14, 2005 | September 14, 2005 | 0 | View and Discuss |

Figure 3.9 – Viewing and discussing oil spills in Data section

3.2.4 Analysis Section

The Analysis section contains a few mockups of workflow technologies that could be used for data analysis and visualization. Workflow tools have been traditionally used in business management for streamlining of operational procedures. Lately, they have been gaining interest in the research community. Technologies such as D2K, D2K-SL (see <http://alg.ncsa.uiuc.edu>) and Kepler (see <http://kepler.informatics.org>) allow sequential linking of different analytical modules to perform a set of tasks. In the workflow example

shown in **Figure 3.10**, two modules deploy web services to access data from a national database and another module transforms the data to a common spatial grid. Three modules assimilate the data into simulation models, which could run in parallel across the national computational grid (*e.g.*, Teragrid). The last module would generate forecast visualizations. Each workflow could be saved and shared among users, creating a need for a workflow library. A “meta-workflow” tool that enables existing workflows created in different tools to be combined as a sequence of web services is under development and will be incorporated into future versions of the CyberCollaboratory.

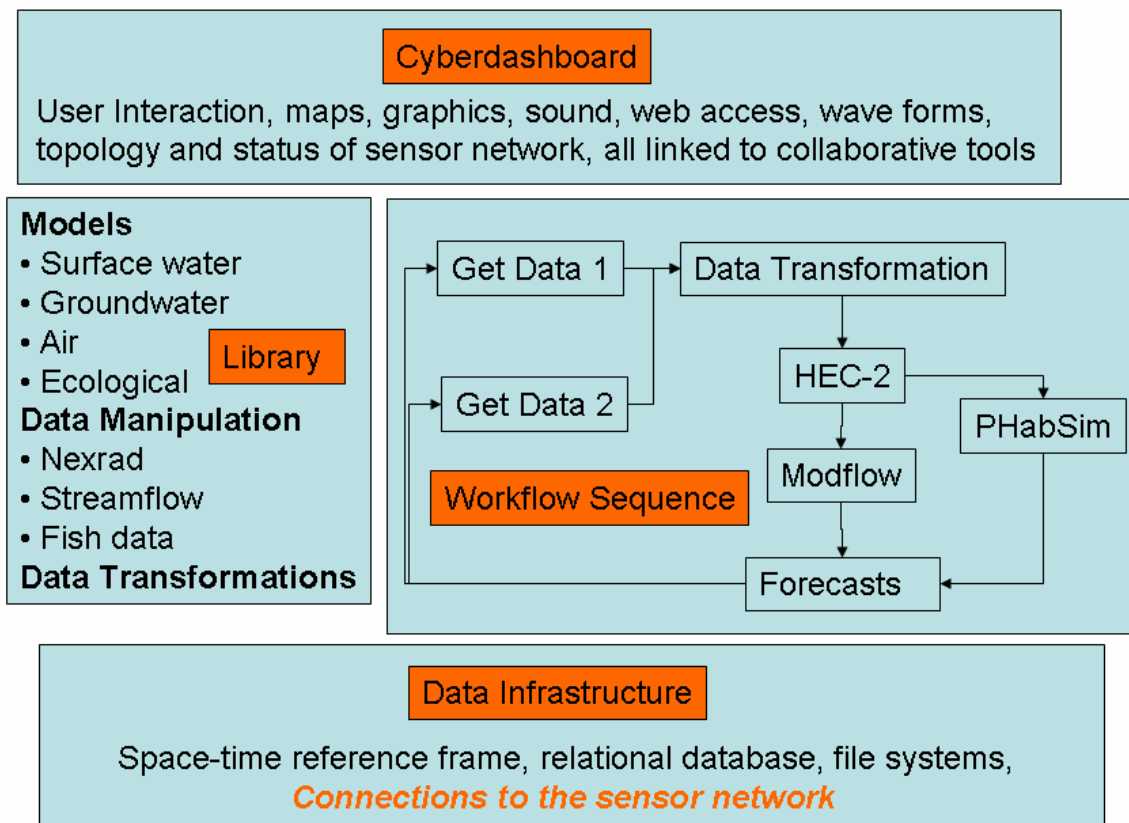


Figure 3.10 – Mockup of a workflow framework (*from Workshop on Sensor Array Cyberinfrastructure and Informatics: Identifying Generic Solutions for Environmental Observatories*, 2005)

3.2.5 Library Section

The library section of the CyberCollaboratory contains a community documents archive for documents uploaded by users and a web documents archive on topics relevant to the environmental engineering and science domain. The community documents archive (shown in **Figure 3.11**) came from the original Liferay portal system. This archive provides a space for the CyberCollaboratory users to upload documents in file cabinets or folders, which can be accessed by the entire community. Users can create and edit these folders. There is also a section called “Image Gallery” in which community images can be stored in folders. The web documents archive (shown in **Figure 3.12**) was built using open-source webcrawling technology provided by Heritrix (see <http://crawler.archive.org/> for more details). Webcrawling is the process of scanning websites on the Internet based on specified criteria, and archiving their content in a database. This content can be in various data formats – *e.g.*, text, HTML, spreadsheet files, documents in ‘pdf’ or ‘doc’ formats, *etc.* The objective is to create a database of websites and files that can be searched for a particular query to yield results relevant to a particular domain. In essence, webcrawling leads to the creation of a search engine for researchers of a particular domain. The criteria for the webcrawl can be the presence of certain keywords in the website’s content, or webcrawling can be done beginning with a seed set of Universal Resource Locators (URLs) and crawling on to other links present in the seed URLs. Dynamic web crawlers can update the database at scheduled intervals to ensure that searches are always based on up-to-date information.

The library of the CLEANER CyberCollaboratory was built to demonstrate the utility of these archives. The Java-based Lucene indexing technology was used to create databases from the archives created by Heritrix. Based on certain seed URLs of interest to the environmental engineering domain (listed in Appendix C of this thesis), a webcrawl was done in March 2005 to create the static library database currently deployed in the CLEANER CyberCollaboratory. In the future, this webcrawling is expected to be dynamic.

CLEANER CUAHSI
 Collaboration Large-scale Engineering Analysis Network for Environmental Research The Consortium of Universities for Advancement of Hydrologic Science Research

Welcome, Aniruddha M Bhagwat!
 Email For Help - My Account - Sign Out

Home | Collaboration | Data | Analysis | Library | Search
 Library » Community Documents My Communities CI Demo

Community Documents

This page allows you to access and create community documents.

"Document Library" lists all the repositories (like file cabinets) of community documents available for viewing and editing. Within each repository, documents can be organized into different folders. You can create new repositories or folders and upload documents to share.

"Image Gallery" lists all archived community images, also organized into folders.

"Recent Documents" lists any documents that you have recently accessed.

Document Library

[Add Repository](#) - [View Repositories](#) - [Search](#)

| Repository | # of Documents | Last Post Date |
|--|----------------|------------------|
| CLEANER VideoConference | 8 | 7/12/05 3:57 PM |
| CLEANER Project Office Documents | 5 | 9/21/05 12:57 PM |
| CLEANER Quarterly Update | 1 | 9/23/05 11:08 AM |

Image Gallery

[Add Folder](#) - [View Gallery](#)

| Folder Name | # of Sub Folders | # of Images |
|-----------------------------------|------------------|-------------|
| CLEANER Logo | 0 | 0 |
| CLEANER Mockups | 0 | 1 |
| D2K SL | 0 | 16 |
| Kepler | 0 | 1 |
| Newsletter images | 1 | 3 |
| Site Images | 0 | 1 |

Recent Documents

[View Repositories](#) - [Search](#)

[CLEANER_Videoconf_Connections.xls](#)

Feedback Window

Feedback for Current Page

Please type in your comments in the following box:

☒ Post to public forum
☐ Report problem(s) to System Administrator

Figure 3.11 – Community documents archive in the CyberCollaboratory library

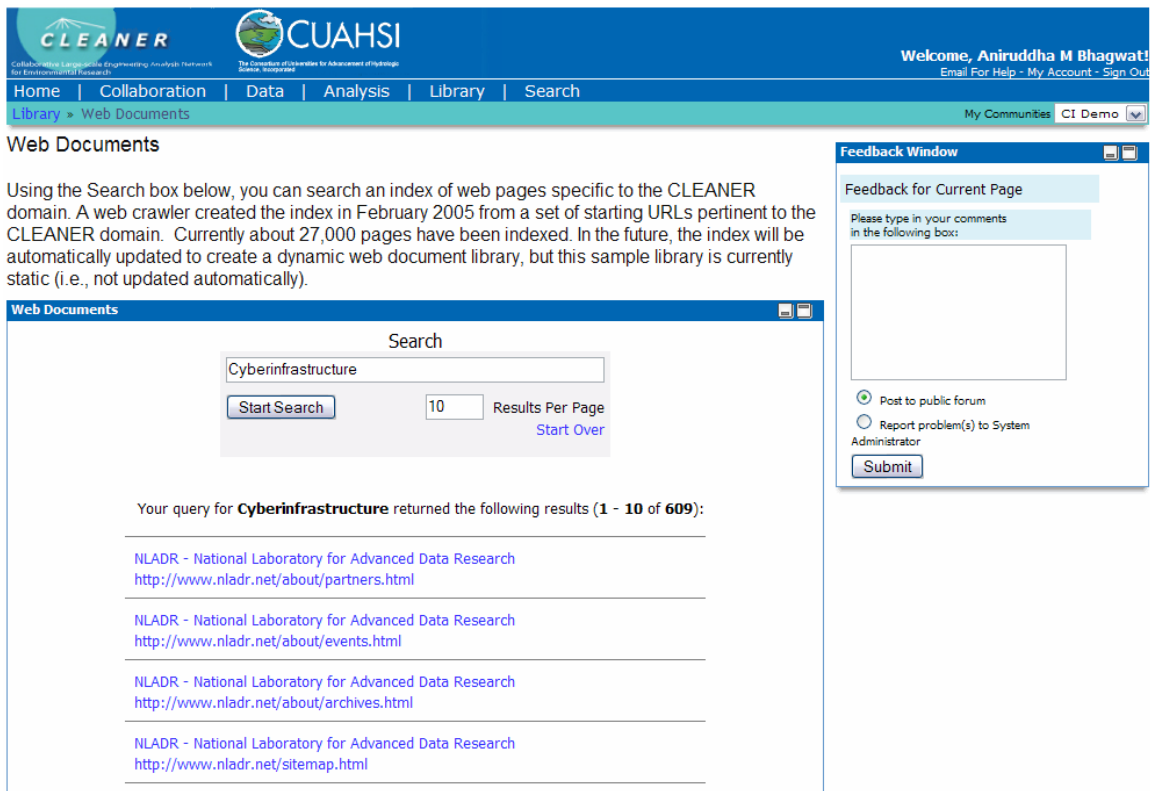


Figure 3.12 – Web documents archive in the CyberCollaboratory library

3.2.6 Search Section

The Search section uses the basic search facility available in the Liferay portal for searching the entire CLEANER CyberCollaboratory or several external web searchers (e.g., Google and AskJeeves). A mockup of an advanced search facility was added to demonstrate how this facility could be augmented to specify where the user wanted to search and what kind of results the user wanted – data, tools, people or documents. Moreover, inclusion of an online learning system to learn users' interests would enable a referral system that could make recommendations to users. For example, a user

performing a search on a particular topic could be informed of other users who were interested in the same or related topic, and the documents they accessed. The mockup for these advanced search capabilities can be seen in **Figure 3.13**.

The screenshot displays the CLEANER CUAHSI website interface. At the top, there is a navigation bar with links: Home, Collaboration, Data, Analysis, Library, and Search. A user is logged in as Aniruddha M Bhagwat. Below the navigation bar, there is a search section with a text input field and a 'Go' button. To the right of the search section is a 'Feedback Window' with a text area for comments and a 'Submit' button. Below the search section is an 'Advanced Search' section with two columns: 'Areas' and 'Referrals'. The 'Areas' column has checkboxes for 'My Space', 'Collaboration', 'Data', 'Analysis', and 'Library'. The 'Referrals' column has checkboxes for 'People', 'Data', 'Topics', 'Documents', 'Tools', and 'Projects'. A 'Start Search' button is located between the two columns.

Figure 3.13 – Search section of the CyberCollaboratory

By spring 2005, the first release of the CyberCollaboratory was online and undergoing preliminary testing by developers and domain researchers at NCSA, the University of Illinois, and the University of Michigan. After various iterations and revisions, it was opened to the advisory committee and other community researchers in summer 2005. In September 2005, a second release was created to incorporate Liferay portal upgrades and

additional demonstrations. This release is now being used by the CLEANER Project Office for additional requirements gathering.

3.3 Requirements Elicitation

To form a preliminary idea about the requirements of the cyberinfrastructure, a literature study was done on the needs assessment procedures of other large-scale collaborative projects. Also, discussions with domain and computational experts at UIUC and NCSA were held. This led to the creation of a 28-page document that identified the various issues that needed to be discussed in detail. Issues ranged from technological, such as the provision of communication and data manipulation tools, to functional issues such as successfully engaging the community. Moreover, a glimpse of the prototype CyberCollaboratory under development was also provided in this document. This document was discussed with the advisory committee in fall 2004, and their inputs were noted on the various issues.

These issues were then refined and classified into categories based on their interest to focus-groups within the community, and presented as questions that would prompt responses about the needs of the community:

- General Questions for all members (13 questions) - listed in the results section of this thesis.
- Detailed Questions on Sharing Information (11 questions) – listed in Appendix B of this thesis.

- Detailed Questions on Data and Metadata (8 questions) – listed in Appendix B of this thesis.
- Detailed Questions on Modeling and Decision Support (3 questions) – listed in Appendix B of this thesis.

As shown in **Figure 3.5**, these questions were then entered in the Solution Center module of the CyberCollaboratory and were posed to specific community members in a conversational style. These community members were from the project team, advisory committee, and members of the environmental engineering and science research community. When appropriate, prior responses, which arose from earlier discussions, were provided to these questions. Members were then asked to vote on the various responses, and to form new responses if they desired. Multiple voting was allowed to accommodate for a member's agreement with multiple responses, as all the responses were not exclusive options. After sufficient time, these responses were recorded, and are presented in the results section of this thesis. The HBGA-based Solution Center thus acted as an interview/survey mechanism for the requirements elicitation exercise. A total of 26 users provided responses to the 13 questions in the General Questions category.

In the current exercise, only the General questions geared towards the entire community were posed to the community members, and these are presented in the results section of this thesis. The other questions involving details on information sharing, data and metadata, and modeling and decision support are of interest to specific experts and are therefore reserved for more detailed discussion in the next phase of CLEANER planning.

These questions and some suggested responses can be found in Appendix B of this document.

In addition, to implement the scenario-based family of techniques and give instructions for completing the survey, a separate website was set up at <http://colab.ncsa.uiuc.edu/CyberWiki/index.php/CLEANER> (now embedded in the CLEANER CyberCollaboratory within the “Collaboration” section, see **Figure 3.6**). This website acted as an information portal about CLEANER as well as a platform where use-case scenarios for the collaboratory were jointly developed by the project team. These scenarios allow a viewpoint-oriented analysis of the needs of the various types of community members. A supplementary objective of this exercise was to familiarize the community members with and encourage their participation on the wiki technology.

4. RESULTS

This chapter presents the results generated from this preliminary needs assessment exercise. Section 4.1 presents the collaboratively developed use-case scenarios, annotated with screenshots of the CyberCollaboratory. Section 4.2 shows the results of the HBGA-based survey module. Finally, section 4.3 provides the response of the participating community members to current and prospective collaboration technologies being considered for the CyberCollaboratory.

4.1 Use-Case Scenarios

Various use-case scenarios were collaboratively developed by the project team. Initially, a few hypothetical personas modeled on prospective users of the CLEANER cyberinfrastructure were created, and then their envisaged activities on the collaboratory were developed into scenarios. These and other jointly-developed scenarios, including scenarios delineating the use of the CyberCollaboratory to conduct an online group meeting and to address the issues of nutrient pollution in the Gulf of Mexico, can be found on the CLEANER wiki, accessible through <http://cleaner.ncsa.uiuc.edu/cybercollab>.

4.1.1 Scenario 1: Getting Started

Jasmine Smith joined the University of Smallsville last year as an Assistant Professor. *Jose Gonzalez* is a Graduate Assistant. He works for Prof. Smith and is in the second year of his PhD program in environmental engineering. While Jose was doing a web search related to one of the tasks which Prof. Smith had assigned him, the CLEANER web site was one of the top 10 hits. After reading a bit about the collaboratory, he emailed the URL to Prof. Smith.

After checking her email later that afternoon, Prof. Smith read through the public CLEANER website. Immediately she became excited about the possibilities that CLEANER offered, and she submitted a request to join CLEANER via an online form. An administrator reviewed her application, verified that she is a legitimate environmental researcher, and created an account in the CLEANER CyberCollaboratory for her and the members of her lab, as shown in Figure 4.1. Being a busy professor, she asks Jose to learn how to use the software, so that later he can show her what she needs to know. Jose goes to the CLEANER web site, where he finds a log-in to the CLEANER CyberCollaboratory. He logs in and enters the system.

The screenshot shows the 'My Account' section of the CLEANER CUAHSI website. At the top, there is a blue header with the CLEANER logo (a stylized mountain and water) and the CUAHSI logo (a globe with a mountain). Below the logos, the text 'Welcome!' is displayed. The main content area is titled 'My Account' and contains a 'Registration Information' form. The form includes fields for First Name, Middle Name, Last Name, User ID, Email Address, Password, and Re-type Password. There is also a Birthday section with dropdown menus for month (January), day (1), and year (1984), and a Sex dropdown menu (Male). A Word Verification section is present, showing a distorted image of the word 'clonors' and a text input field for the user to enter the word. A 'Create Account' button is located at the bottom of the form. Below the form, there is a link for 'About - Disclaimer'.

Registration Information

First Name

Middle Name

Last Name

User ID

Email Address

Password

Re-type Password

Birthday

January 1 1984

Sex

Male

Word Verification

Enter the word shown below.

This helps ICSCA prevent automated registrations.

clonors

Create Account

[About - Disclaimer](#)

Figure 4.1 – Creating a new account in the CLEANER CyberCollaboratory

4.1.2 Scenario 2: Finding Information

Jose goes to the search section of the collaboratory, seen in **Figure 3.13**, to see what he can find out about nutrient transport in the Mississippi River and hypoxia in the Gulf of Mexico. Using these tools, Jose identifies:

- 2 CLEANER field facilities in the Mississippi River and Gulf of Mexico areas with major projects on nutrient transport and hypoxia
 - For each project, the names of the major investigators and graduate students and links to their public online resources.
- A Discussion Forum on nutrient transport and hypoxia
- 10 recent publications on this topic
- Archives from a recent on-line CLEANER conference on the Mississippi River and Gulf of Mexico nutrient issues.

Jose saves this information to his electronic notebook seen in **Figure 3.4**, and with the click of a button, gives his advisor access to the information and sends her an e-mail with the link to get to the notebook. Jose subscribes to the nutrient fate discussion forum so he can get e-mail updates on postings.

Prof. Smith receives Jose's e-mail and follows the link to the electronic notebook. There she reads about Prof. Jones' research at a CLEANER field site on the Mississippi River. *Terrence Jones* is an environmental engineering professor who specializes in nutrient fate and transport and best management practices and performs field-scale nutrient experiments at one of the CLEANER field sites.

Accessing the social networking tool, Prof. Smith identifies other researchers who have cited Prof. Jones' work and links to their papers and online resources. From this information, she realizes that Prof. Jones appears to be a leader in field-based

implementation of nutrient sensors, but doesn't appear to have performed any modeling with his data. She begins an e-mail correspondence with Prof. Jones who ultimately invites Prof. Smith to participate in an upcoming proposal for a large-scale field experiment coupled with watershed-scale modeling in the Mississippi River and Gulf of Mexico.

4.1.3 Scenario 3: Data Analysis

After acquiring some data from Prof. Jones, Jose saves the data in his CyberCollaboratory workspace and enters the workflow area for analysis. A message shows saved workflows that other researchers accessing this type of data have used. Jose pulls up a workflow that will integrate 3 types of data to a common format, assimilate the data into a community model for the Mississippi Region and Gulf of Mexico, and visualize the results. He uses the workflow to predict the effects of several proposed agricultural best management practices (BMPs) in his local watershed on nutrient levels in Corpus Christi bay on the Gulf of Mexico.

4.1.4 Scenario 4: Writing a Proposal

Prof. Smith and Prof. Jones form an online group to discuss their proposal. They brainstorm ideas on the CyberCollaboratory using their own discussion forum (see **Figure 3.2**), regular audio conferences and text chats (see **Figure 3.3**) and solution center

(see **Figure 3.5**). Their interactions are archived in the online system for their future reference. They collaboratively edit their proposal documents, which are stored in a shared online archive.

4.2 General Questions and Responses

The results of the Solution Center-enabled interview/survey mechanism used for the needs assessment exercise are outlined in the following pages. Responses are listed by selection frequency, with the most frequently selected response listed first.

1) **Grand challenge research questions that CLEANER will address.**

A major part of the CLEANER planning is to define the grand challenge research questions that the CLEANER infrastructure would allow researchers to address. Some of the topics below came from previous CLEANER workshops. Please select your top three choices, or add your own.

- a) How do trends in population, land-use and industrial and urban development affect environmental quality?
- b) How can engineered systems be designed to ensure environmental sustainability?
- c) How do we explore effective options to prevent/mitigate adverse environmental effects at different spatial and temporal scales?
- d) How do we predict and control the outcomes of potential mitigation strategies?
- e) What are the impacts of engineered systems on the environment?
- f) What are the emerging chemicals, effects, and pathogens that threaten future water quality and public health?
- g) How do we analyze and solve the problem of non-point source runoff from urban and agricultural land of nutrients and sediments, which is degrading the nation's water quality and aquatic life?
- h) How do we monitor, model, and forecast water quality of the nation's rivers, estuaries, and coastal waters?
- i) How can food and fiber production be made sustainable?
- j) What are the contributions of air pollution and atmospheric deposition on coastal water quality?
- k) How do we monitor, model, and forecast harmful algal blooms and hypoxia?

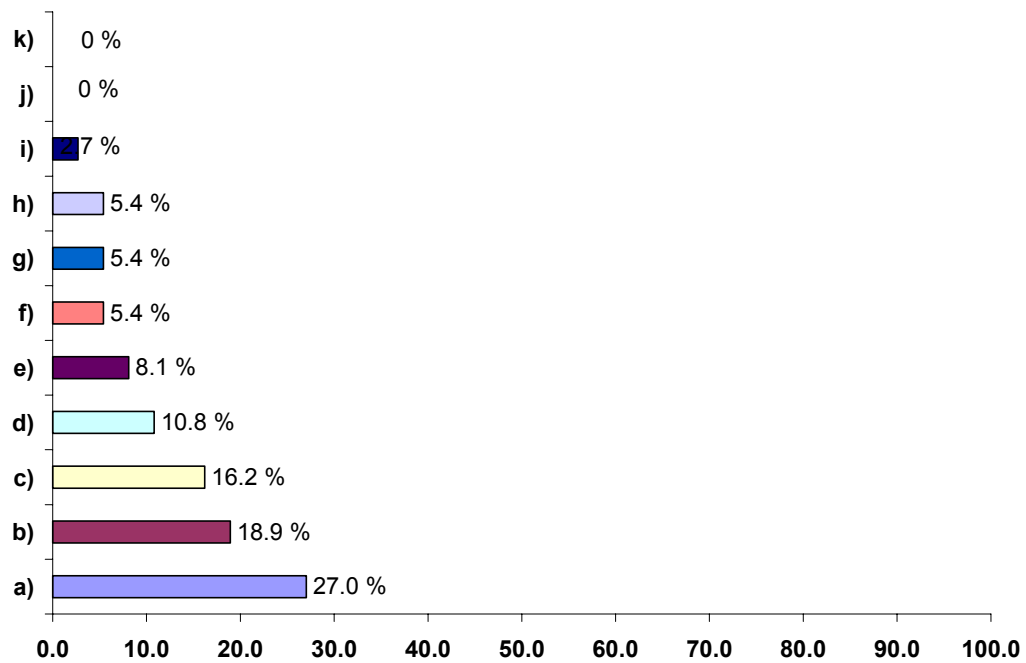


Figure 4.2 – Survey results for Question 1, Number of votes: 37

The results of the survey indicate which grand challenges in the environmental research are considered most relevant by the participating group.

2) How much do you know about CLEANER?

If you haven't heard about CLEANER, please review the highlights before proceeding with the rest of the questions.

Responses:

- a) A lot
- b) Some
- c) Very little
- d) Never heard of it before I got this request

The survey results shown in **Figure 4.3** indicate that there is a fair amount of awareness among these particular respondents about the CLEANER initiative, but this may be due to the involvement of most of the participants in other CLEANER grants.

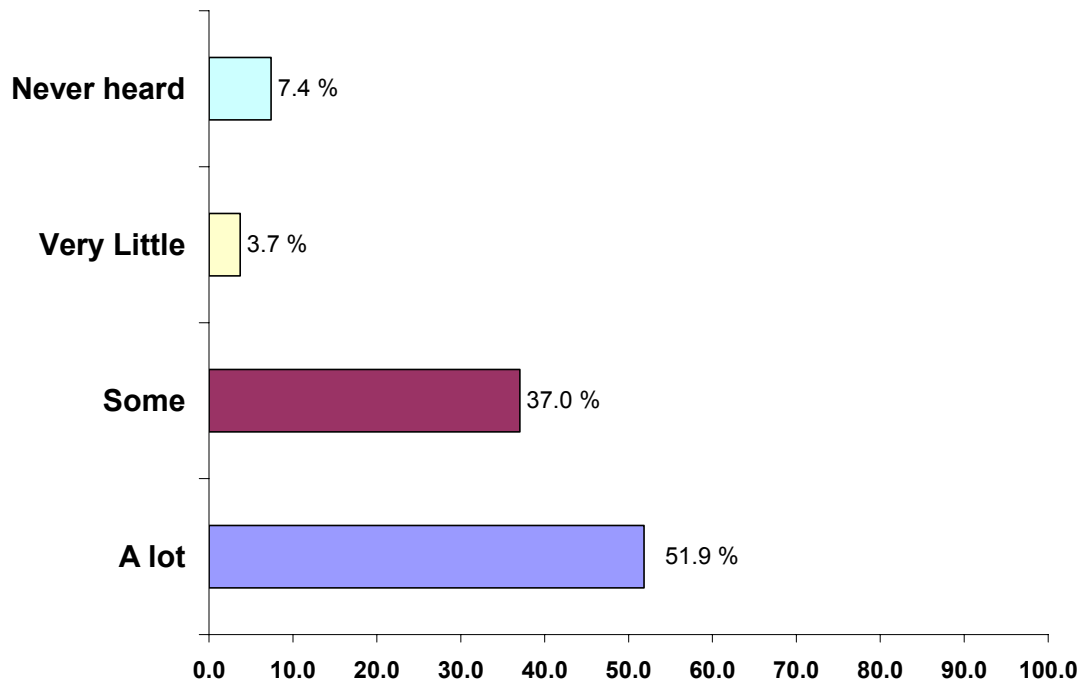


Figure 4.3 – Survey results for Question 2, Number of votes: 27

3) Should an infrastructure like CLEANER be created for environmental research and education?

Given what you know about CLEANER, do you think such an infrastructure should be created for environmental research and education?

Responses:

- a) Definitely
- b) Need more information
- c) Maybe
- d) No

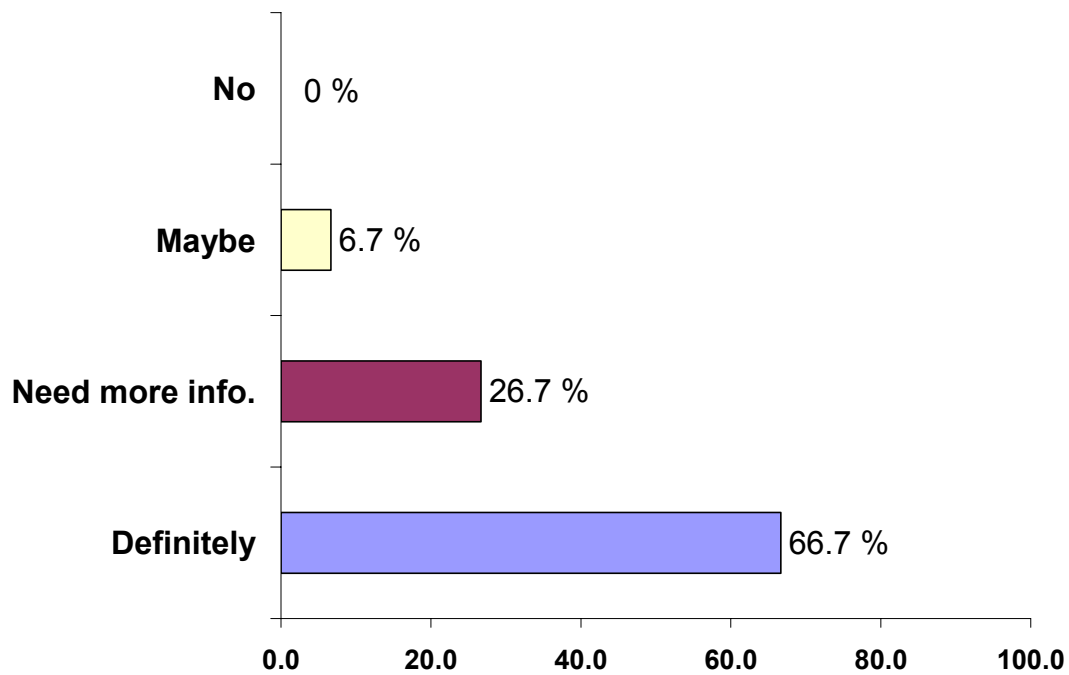


Figure 4.4 – Survey results for Question 3, Number of votes: 15

The survey results indicate that these particular respondents support the creation of CLEANER, but this may again be due to the involvement of the participants in other CLEANER grants.

4) How do you think the existence of CLEANER would affect your research?

Responses:

- a) It might help my research
- b) It would help my research substantially
- c) It wouldn't affect my research
- d) It would hurt my research

It can be seen in **Figure 4.5** that many participants agreed that the CLEANER initiative would benefit them in their work to a varying degree.

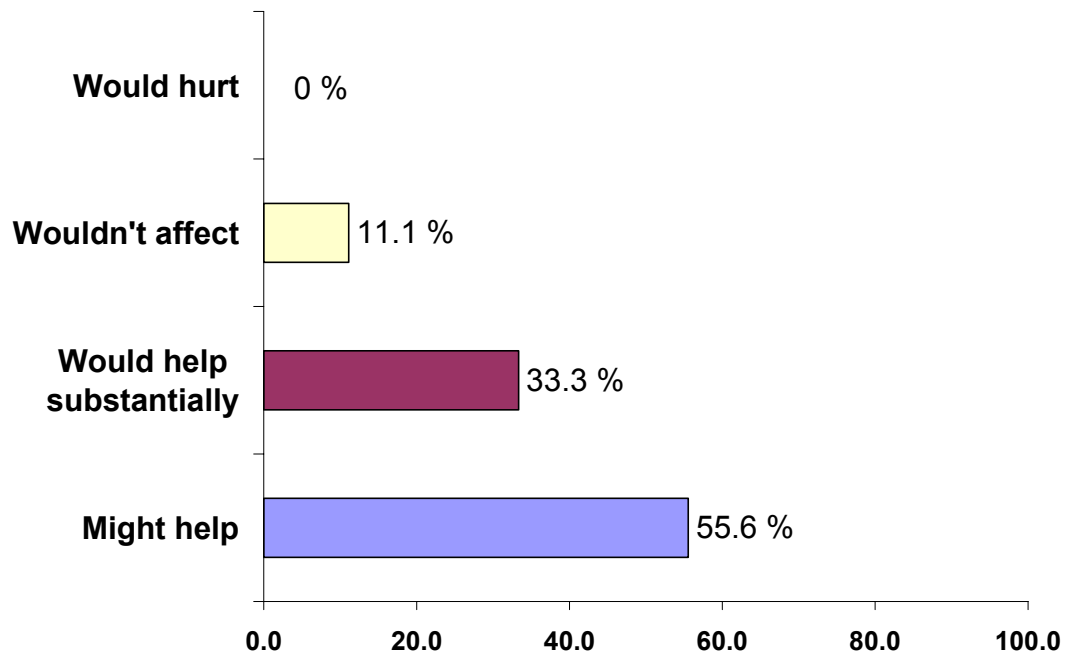


Figure 4.5 – Survey results for Question 4, Number of votes: 9

5) What types of incentives might persuade you to get involved in CLEANER?

Responses:

- a) The opportunity to perform collaborative interdisciplinary research at a scale that was not previously possible.
- b) Access to resources, ideas, and other people's data.
- c) Funding for researchers willing to form collaborative teams.
- d) The opportunity to learn what's going on more quickly.
- e) Other reward mechanisms to provide recognition for active collaborators.

Among the different incentives that might encourage further participation in the CLEANER collaboration, the ones that were considered foremost by the participating group were new opportunities to do interdisciplinary research at a larger scale; greater access to data, resources and ideas of other researchers; and enhanced funding opportunities for collaborative teams, as seen in **Figure 4.6**.

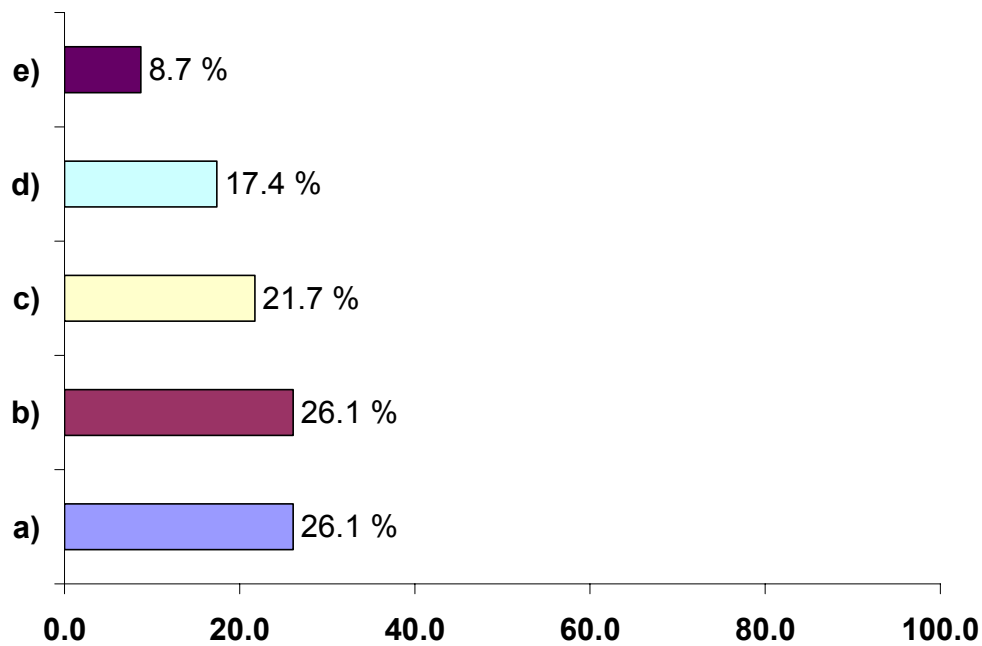


Figure 4.6 – Survey results for Question 5, Number of votes: 23

6) CLEANER is envisioned as a network of environmental field facilities (EFFs) that would be managed by a central project office. What roles do you think the project office should play?

Responses:

- a) Providing specialized expertise to support researchers (*e.g.*, on sensors, data, and information technologies to analyze the data).
- b) Developing guidelines or standards for data collection and storage to ensure compatibility.
- c) Providing seed funding for research projects.
- d) Developing guidelines or standards for models to ensure compatibility.
- e) Ensuring that data collected by the EFFs are well coordinated and meet broader community needs.
- f) General oversight of the EFFs.
- g) Gathering community input on resource needs.
- h) Providing an information clearinghouse on topics relevant to CLEANER.
- i) Sponsoring proposal competitions to meet resource needs (pre-screening for NSF).

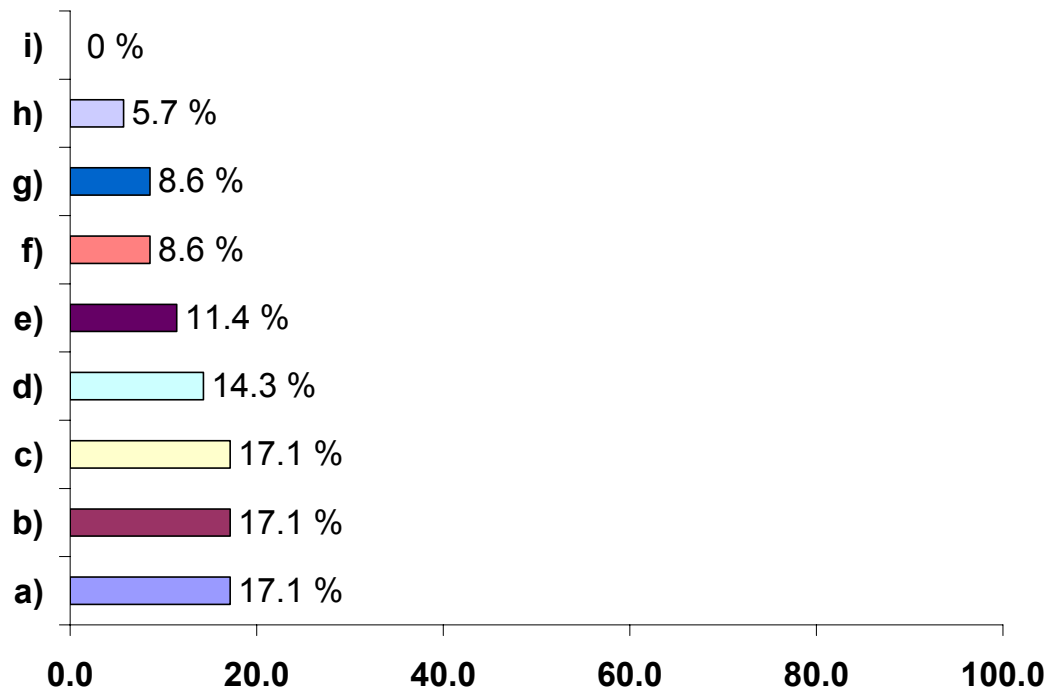


Figure 4.7 – Survey results for Question 6, Number of votes: 35

Respondents indicated that they wanted the Project Office to play a technical support and coordination role through activities such as providing specialized expertise and seed funding to researchers and developing guidelines or standards for data and modeling compatibility.

7) If CLEANER environmental field facilities (EFFs) were in place and generating data and you could access these resources from the web, what types of activities would you be most interested in doing that you cannot easily accomplish today?

Responses:

- a) Finding out what other researchers are doing.
- b) Finding field data.
- c) Communicating with other researchers.
- d) Finding documents and reports.
- e) Running models with field data.

- f) Analyzing field data.
- g) Remotely controlling field experiments.
- h) Integrating field data.

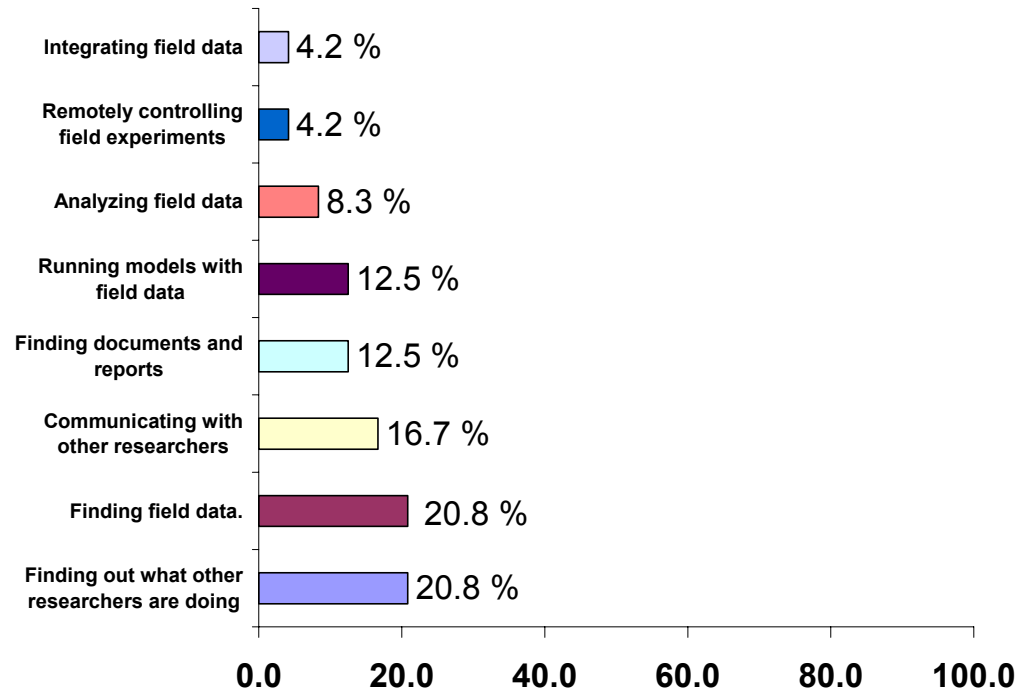


Figure 4.8 – Survey results for Question 7, Number of votes: 24

Participants indicated that the presence of the environmental field facilities (EFFs) could enhance their research by giving them a larger platform to know and interact with other researchers, as well as providing them with data for their research. Interestingly, not much preference was indicated for remote-controlled experiments by the participants, which was a big draw for the NEESgrid collaboratory.

- 8) Which of these types of tools described or implemented in this CyberCollaboratory would be helpful (either to your research or future CLEANER research)? CLEANER researchers will be supported by cyberinfrastructure, which will consist of computational infrastructure (computers, networks, and software) that the community needs. This CyberCollaboratory is an example of software infrastructure that could potentially serve CLEANER. Which of the following types of tools described or implemented in this CyberCollaboratory would be helpful (either to your research or future CLEANER research)?

Responses:

- a) Modeling tools that let CLEANER users use simulation models on the CyberCollaboratory.
- b) Group management tools that let you form smaller groups working on common projects and let you define separate access to resources of this group.
- c) Visualization tools that provide a graphical representation of data and research analyses.
- d) Data Management tools that let you upload, annotate and define access privileges for datasets.
- e) Collaborative Editing tool for co-authoring documents.
- f) Web library that is a fully-searchable repository of webpages collected on subjects of specific interest to CLEANER researchers.
- g) Communication tools that provide services like email, instant messaging, audio-video conferencing *etc.*
- h) Workflow tools for analysis, modeling, and visualization.
- i) Solution Center that lets you pose questions and gather votes on opinions.
- j) Social networking tools that let you find other people and their research interests, collaborations, data and information being accessed *etc.*
- k) Document library that lets users upload and share documents with other users.

As seen in **Figure 4.9**, among the various technologies demonstrated in the CyberCollaboratory, the participants preferred modeling and visualization tools, as well as effective group management. Newer concepts such as collaborative editing and workflow analysis tools did not find much favor, but this could be because of limited awareness of their utility to environmental research and education.

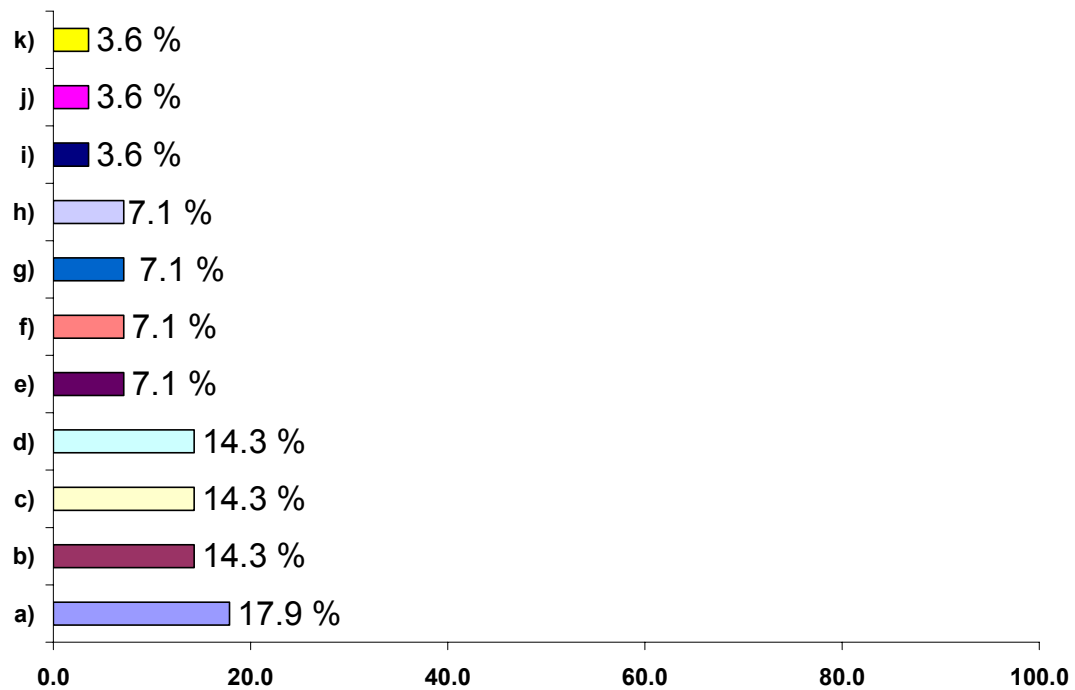


Figure 4.9 – Survey results for Question 8, Number of votes: 28

- 9) What type of management structure should be used for CLEANER?**
 Several alternate CLEANER management structures are proposed for discussion. The information, directives and funding flows in the proposed management structures can be found at http://colab.ncsa.uiuc.edu/CyberWiki/index.php/Management_Structures_for_CLEANER. What type of management structure should be used for CLEANER?

Responses:

- a) Management Structure II: Collaboratory with Coordinator.
- b) Management Structure I: “Traditional” Collaboratory.
- c) Management Structure III: Central Coordinating Authority and Coordinated Information Services Laboratory.

As seen in **Figure 4.10**, among the few management structures proposed, most of the participants were reluctant to commit to a certain organization structure, and hence the project office should delve deeper on this issue.

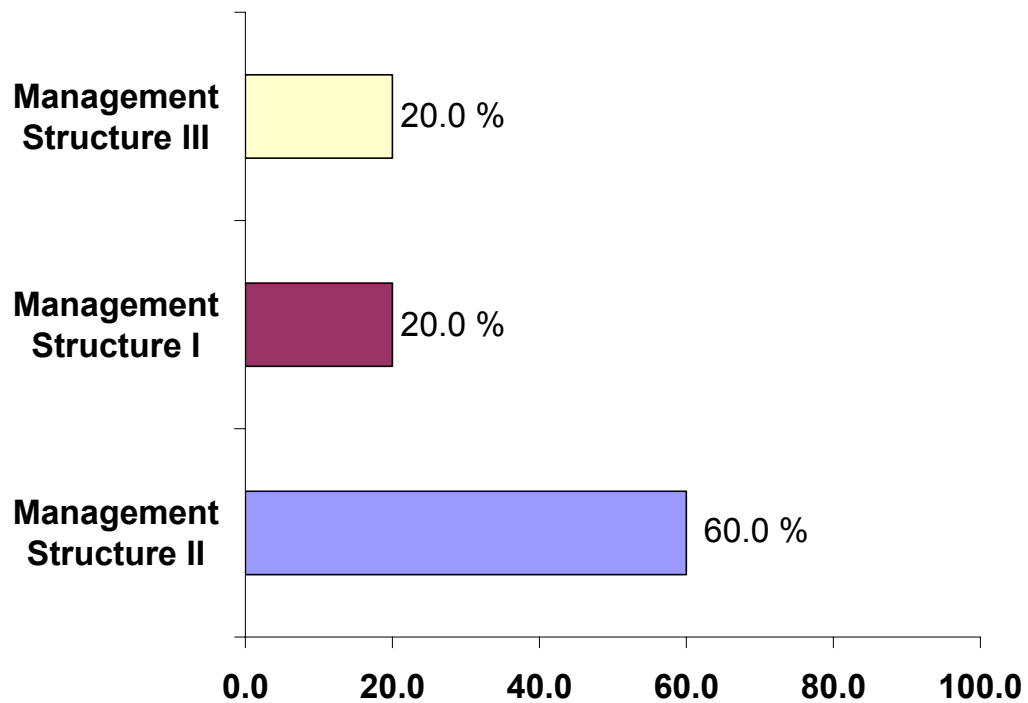


Figure 4.10 – Survey results for Question 9, Number of votes: 5

10) Would you have privacy concerns if your activities were monitored and stored in a database for future reference by other members? Some types of cyberinfrastructure features (*e.g.*, getting notifications of recently-posted information that may be of interest) would require monitoring users' activities. Would you have privacy concerns if your activities were monitored and stored in a database for future reference by other members?

Responses:

- a) No, not if I knew that was the policy ahead of time.
- b) Yes, therefore members should be able to define what they would like to be shared about them (including completely opting out of sharing).
- c) No, the benefits of free exchange of information outweigh the privacy concerns.

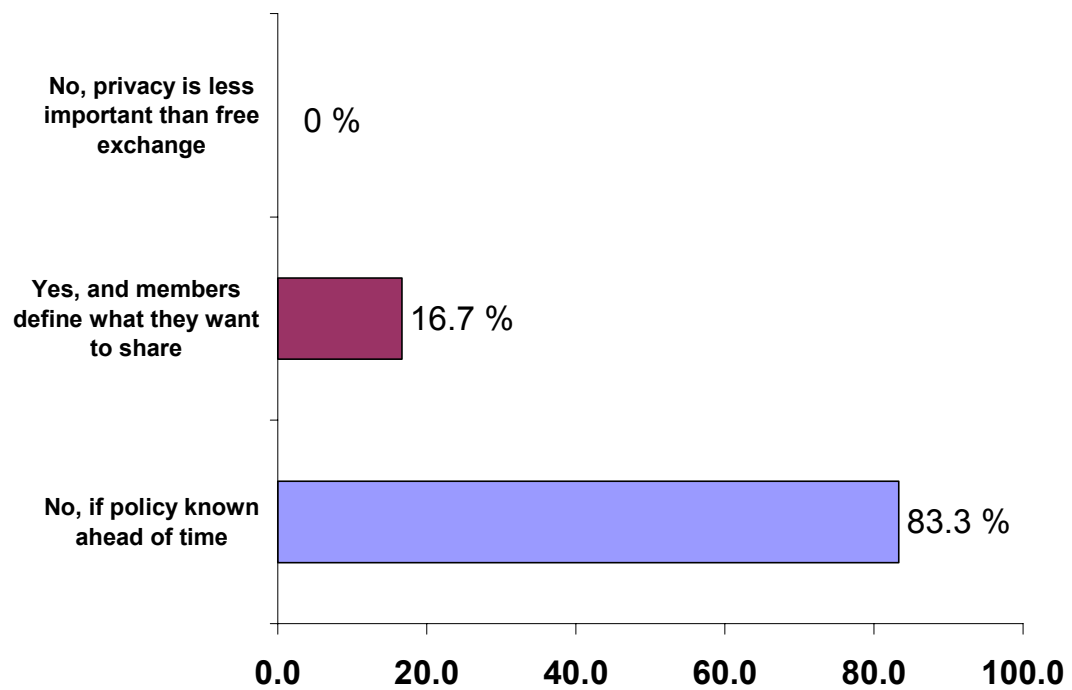


Figure 4.11 – Survey results for Question 10, Number of votes: 6

Another important issue is privacy concerns due to the nature of the collaboratory, in which users could know more about activities of other users. Among those who responded, there was an unequivocal acceptance of one's activities being monitored as long as the participants were cognizant of this monitoring ahead of time.

11) What should be the time interval between data generation and making it public on CLEANER?

Responses:

- a) Different classes of data are set - some immediately transferred to common CLEANER repositories, others involve various time lags.
- b) Specific time frame before data made public to CLEANER to allow for quality control.
- c) All data should be made public immediately.
- d) Data collector decides the scope of data sharing - how much and when.

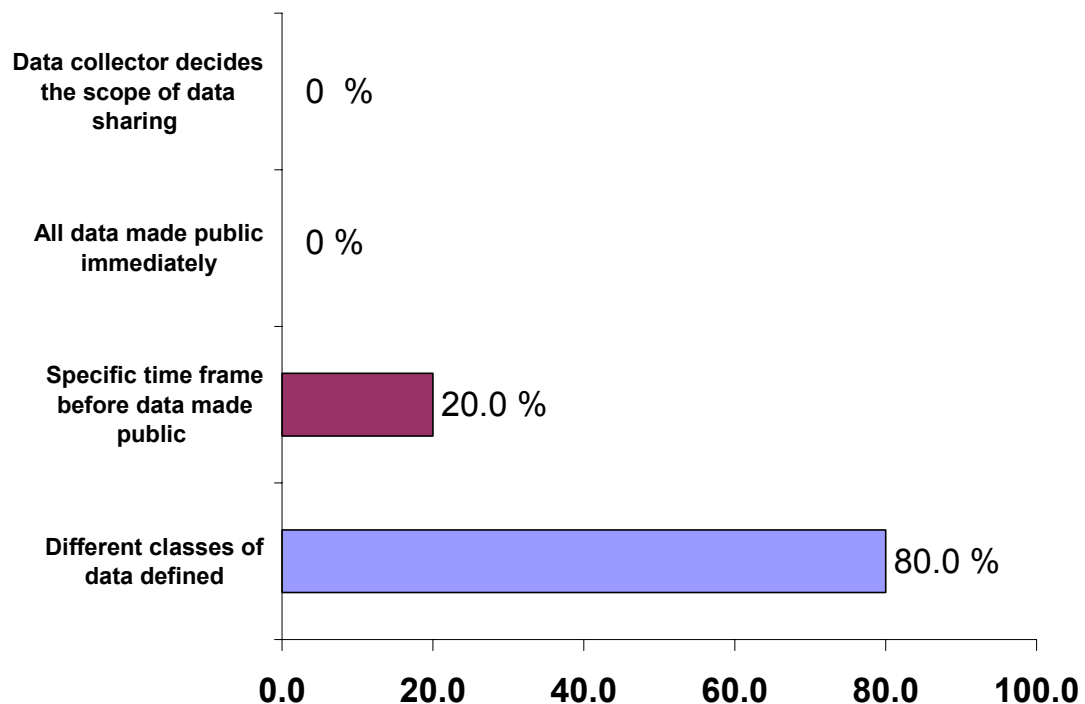


Figure 4.12 – Survey results for Question 11, Number of votes: 5

As can be seen from the results, the participants displayed a partiality towards a classification system of data on CLEANER that would allow different time periods between its generation and making it available to the public. This would allow the data gatherers to take exclusive advantage of some of their efforts in case they wished to do so.

12) How should CLEANER handle data ownership and intellectual property rights? Various users of the CLEANER cyberinfrastructure will be generating and sharing data relevant to their individual research. This brings into focus the issue of data ownership and intellectual property rights. Does the data/ knowledge shared by a particular user, when available to other users via CLEANER, become the property of the entire CLEANER network, a particular team of CLEANER, or does the original collector of the data still have absolute control over its use? If any commercial products are created from CLEANER technologies, who owns the rights to them?

Responses:

- a) All data and information on CLEANER should be available to the entire community immediately.
- b) The original user (owner of data/information) should define the scope and timing of usage of the data/ information.
- c) Inside a particular group, the members can come to a consensus on timing and amount of data/information sharing.
- d) There should be a hierarchy of access. First the group that collects the data (or is in charge of the specific project the data is generated for) followed by other collaborators within the Field Facility followed by other field facilities and finally open to the public. Each step in access should be after a specific amount of time from data collection. The group in charge of the data can open the data earlier and if the data is not the project of a specific group it should be open immediately. Commercial products should be left to the inventing institution.
- e) In groups involving a hierarchy (e.g., a research lab), a particular user administers these issues.
- f) Commercial products should have standard disclosures documenting contributors.

This question addresses the question of intellectual property rights and data ownership in a collaborative environment. As seen in **Figure 4.13**, no single answer could be the majority preference, and participant preferences were equally distributed towards public, individual, group and staggered group and public ownership of the data generated on CLEANER.

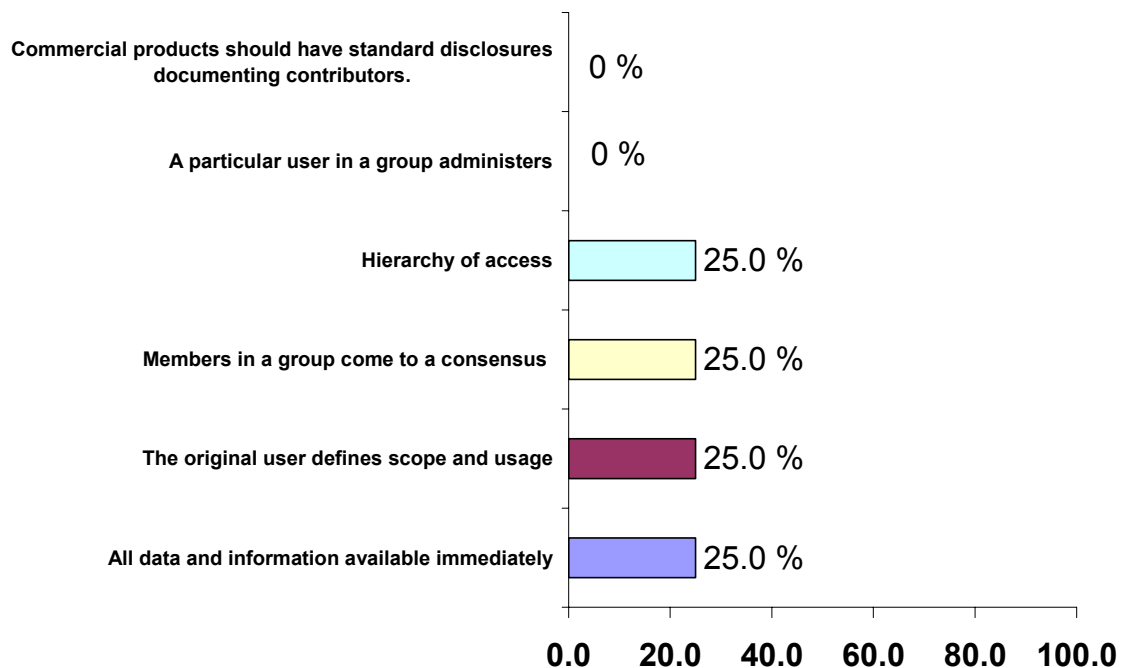


Figure 4.13 – Survey results for Question 12, Number of votes: 8

Various other issues were identified that were more detailed and of specific interest to focus-groups among the community. These issues are passed on to the CLEANER Project Office for detailed consideration in the next planning phase, and can be found in Appendix B of this document.

4.3 Community Response to Collaboration Technologies

This section records the response of the participating community members to the various technologies, both current and prospective, implemented in the Collaboration section of the CyberCollaboratory.

Forums (bulletin boards) have been on the Internet for a long time, and provide a structured approach to asynchronous communication among forum members. This technology is very mature, reliable and easily deployable in ‘free’ open-source formats (for *e.g.*, see phBB Forums at <http://www.phpbb.com/>). Communication through these forums does not require intricate knowledge of HTML scripting, and therefore these are gaining popularity on the Internet. It was observed that though the computational scientists and developers were comfortable with this tool, many environmental science and engineering researchers were relatively unaware of its presence or had not used it for research purposes. Moreover, the forums implemented in the CyberCollaboratory were not heavily used by project team members, who tended to rely more on e-mail for asynchronous communication.

Email has been steadily gaining popularity as a communication tool since its mainstream advent in the mid-nineties. It was noted that most of the researchers use email in their daily research activities efficiently, but have misgivings against irrelevant or junk mail coming into their inboxes and difficulties with organizing e-mail archives. Forums could be combined with email technology such that email notifications could be sent to subscribing users when a new message is posted in a forum of interest and responses could be sent via e-mail, enabling greater acceptance of the forum technology. A new forum that will be tightly integrated with e-mail is now being implemented in the CyberCollaboratory for community evaluation.

Text Chat is widely used in the Internet community, but it was observed that the domain scientists preferred to use it only for short communications or notifications. Researchers expressed frustrations with the Liferay chat rooms, which do not provide information on other users who are currently online or capabilities for inviting other users to join a chat room. Some users had difficulties with computers that were missing Java plug-ins and could not operate the chat room. Alternative text chat tools should be evaluated for inclusion in the CyberCollaboratory in the future.

Audio-video conferencing was not explicitly deployed on the prototype CyberCollaboratory due to limitations on the number of simultaneous participants allowed by open-source or demonstration software. Commercial software (such as Macromedia[®] Breeze) can be provided for larger numbers of users, but is available at a steep price and was therefore not considered for this demonstration. Audio-videoconferencing was conducted among the advisory committee members via the Access Grid and Polycom[®] technology, but it required the use of specialized hardware and software. Furthermore, it was observed that due to the limited view-ability of a few members on the screen, and due to the consciousness of being viewed by others, researchers involved in the project preferred regular telephonic audio-conferencing over audio-videoconferencing. It is possible that Voice over Internet Protocol (VoIP) technologies could be implemented in the CyberCollaboratory to replace telephonic audio-conferencing and integrate coupled document and analysis sharing with audio communication (similar to the shared data demonstration described previously).

The wiki module was heavily used by the project team for creation of scenarios and providing information about CLEANER, and is now being used by the CLEANER Project Office teams. Though this approach was met with enthusiasm during this project, some training is needed to help users to get comfortable with it and discover its utility for collaborative research. While the wiki user interface is relatively simple, formatting is significantly more difficult than in desktop word processing packages.

5. CONCLUSIONS

This study provides a preliminary assessment of the general needs of members of the environmental engineering and science community for the CLEANER cyberinfrastructure. This assessment used a prototype CyberCollaboratory to give the participating community a glimpse of a potential CLEANER cyberinfrastructure.

Consensus is needed on the other features/tools/technologies required for the CLEANER cyberinfrastructure for it to become a comprehensive medium facilitating collaboration. Many of these features are needs-driven, but some are capability driven; *i.e.*, new needs emerge from new capabilities. Also, there is an acceptance threshold for new technology in any field, and to overcome that, it is necessary to make the community aware of the utility of new technology. This study introduces various new technologies that could be incorporated in the CLEANER cyberinfrastructure, and records their response towards each of them. Awareness of the utility of these new features will help the community gauge their requirements in a new light. However, it must be kept in mind that technology should be developed based on the needs of the domain, so as to avoid the pitfall of adapting the community needs around technology.

With respect to the survey results, it can be concluded that the respondents are cognizant of the need for a national network that would enable research collaborations beyond current means to tackle complex environmental issues. Opportunities to interact with more researchers and discover their research interests, obtain large datasets, run

simulations and get better visualizations of results are incentives drawing the respondents to this collaboration initiative. Furthermore, feedback from users of the CyberCollaboratory indicated that users are interested in technologies that can be easily learned to allow them to enhance their current and potential research activities.

It is recommended that the CLEANER Project Office obtain community consensus on the issues identified here for data and information sharing, modeling, management and decision support, security, and hardware and networking issues. To ensure that needs of a diverse span of the community are understood and included in CLEANER development, measures must be taken to spread awareness of cyberinfrastructure technologies. Measures could include training workshops where participant responses are taken face-to-face, usability studies that carefully measure user responses to the technology, and continuation of the methods used in this exercise, albeit to a larger audience. Through iterative improvements of the system in response to community input, this process will eventually lead to a better and more comprehensive needs assessment for a successful CLEANER environment.

BIBLIOGRAPHY

AC-ERE, Complex Environmental Systems: Pathways to the Future. 2005, NSF: Arlington, VA. p. 12.

Allen, T.J., *Managing the Flow of Technology*. 1977, Cambridge, MA: MIT Press.

Atkins, D., and the NSF Blue Ribbon Advisory Panel on Cyberinfrastructure, Revolutionizing Science and Engineering through Cyberinfrastructure. 2003, NSF Blue Ribbon Advisory Panel on Cyberinfrastructure: Arlington, VA. p. 84.

Bair, R.A., Collaboratories: Building Electronic Scientific Communities, in *Impact of Advances in Computing and Communications Technologies on Chemical Science and Technology: Report of a Workshop*, T.H. Dunning, Editor. 1999, National Academy Press: Washington, D.C. p. 125-140.

Bhandarkar, M., Budescu, G., Humphrey, W.F., Izaguirre, J.A., Izrailev, S., Kale, L.V., Losztin, D., Molnar, F., Phillips, J.C. and Schulten, K. BioCoRE: A collaboratory for structural biology. in *Proceedings of the SCS International Conference on Web-Based Modeling and Simulation*. 1999. San Francisco, CA.

Brezonik, P.L., CLEANER Project Office Information Meeting. 2005: Arlington, VA.

Carroll, J.M., Rosson, M.B., Chin Jr., G. and Koenemann, J., Requirements Development in Scenario-Based Design. *IEEE Transactions On Software Engineering*, 1998. 24(12): p. 1156-1170.

Dogan, S., and Baru, C. GEON: The GEON Grid Software Architecture. in *Twenty-fourth Annual ESRI International User Conference*. 2004. San Diego, CA.

Eheart, W.J., *Personal communication regarding traditional vs. collaboratory paradigms*. 2004.

Finholt, T.A., and Olson, G. M., From Laboratories to Collaborators: A New Organizational Form for Scientific Collaboration. *Psychological Science*, 1997. 8(1): p. 28.

Finholt, T.A., Collaboratories, in Annual Review of Information Science and Technology, B. Cronin, Editor. 2001, *American Society for Information Science and Technology*: Washington, D.C. p. 73-108.

Finholt, T.A., Wierba, E.E., Birnholtz, J.P., and Hofer, E., NEESgrid user requirements document, Version 2.0. 2002, NEES.

Finholt, T.A., Collaboratories as a new form of scientific organization, in *Econ. Innov. New Techn.* 2003. p. 5-25.

Finholt, T.A., Horn, D., Birnholtz, J.P., Bae, S.J., and Motwani, D., NEESgrid User Requirements Document, Version 3.0. 2003, NEES.

Futrell, J., and the AC-ERE, Environmental Cyberinfrastructure (ECI): Tools for the Study of Complex Environmental Systems. 2003, NSF AC-ERE: Arlington, VA. p. 4.

Gaber, N., *Background summaries of previous CLEANER workshops and reports*, http://colab.ncsa.uiuc.edu/CyberWiki/index.php/Background_summaries_of_previous_workshops_and_reports, last accessed on September 25, 2005.

Goldberg, D.E., *Genetic Algorithms in Search, Optimization and Machine Learning*. 1989: Addison-Wesley Publishing Company.

Goldberg, D.E., Welge, M., & Llorà, X., DISCUS: Distributed Innovation and Scalable Collaboration in Uncertain Settings. 2003, Illinois Genetic Algorithms Lab., University of Illinois at Urbana-Champaign.: Urbana, IL. p. 17.

Henline, P., 8 Collaboratory Summaries, in *ACM Interactions*. 1998. p. 66-72.

Hesse, B.W., Sproull, S., Keisler, S.B. and Walsh, J.P., Returns to science: computer networks in oceanography, in *Communications of the ACM*. 1993. p. 90-101.

Johnston, W.E., Greiman, W., Hoo, G., Lee, J., Tierney, B., Tull, C. and Olson, D., High-Speed Distributed Data Handling for On-Line Instrumentation Systems. 1997, Lawrence Berkeley National Laboratory: Berkeley, CA.

Katz, J.S., Geographical Proximity and Scientific Collaboration. *Scientometrics*, 1994. 31(1): p. 31-43.

Keller, G.R., GEON (GEOscience Network) - A First Step In Creating Cyberinfrastructure For The Geosciences, in *Electronic Seismologist*. 2003.

Kosorukoff, A., and Goldberg, D. E. Evolutionary computation as a form of organization. in *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO)*. 2002.

Kosorukoff, A., *Human Based Genetic Algorithm*. <http://www.geocities.com/alex+kosorukoff/hbga/hbga.html>, last accessed on 28 May, 2005.

Kouzes, R.T., Myers, J.D. and Wulf, W.A., Collaboratories: Doing science on the Internet. *IEEE Computer*, 1996. 29(8): p. 40-46.

Kraut, R.E., Egidio, C. and Galegher, J., Patterns of contact and communication in scientific research collaboration, in *Intellectual Teamwork: Social and Technological Foundations of Cooperative Work*, R.E. Kraut, Egidio, C. and Galegher, J., Editor. 1990, Lawrence Erlbaum Associates: Hillsdale, NJ. p. 149-171.

Maidment, D.R., et. al., *Hydrologic Information System Status Report, Version 1*, D.R. Maidment, Editor. 2005, CUAHSI. p. 224.

NCAR, Cyberinfrastructure for Environmental Research and Education: Report from a workshop held at the National Center for Atmospheric Research October 30 - November 1, 2002, N.C.W.S. Committee, Editor. 2003, NCAR/ NSF: Boulder, CO. p. 19.

NRC, *Envisioning the Agenda for Water Resources Research in the Twenty-First Century*. 2001, NAS Press.

NSF, Cyberinfrastructure for Engineering Research & Education. 2003, NSF Directorate of Engineering: Arlington, VA. p. 16.

NSF, *Cyberinfrastructure Poised to Revolutionize Environmental Sciences and Other Disciplines*, Press Release 04-014, February 13, 2004.

NSF. CLEANER: Project Office to Coordinate Network Activities. 2005.

Pfirman, S., and the AC-ERE, Complex Environmental Systems: Synthesis for Earth, Life, and Society in the 21st Century, S. Pfirman, Editor. 2003, NSF: Arlington, VA. p. 68.

Price, D.J.d.S., *Little Science, Big Science*. 1963: Columbia University Press, New York.

Prudhomme, T., Kesselman, C., Finholt, F., Foster, I., Parsons, D., Abrams, D., Bardet, J-P., Pennington, R., Towns, J., Butler, R., Futrelle, J., Zaluzec, N., and Hardin, J., NEESgrid: A Distributed Virtual Laboratory for Advanced Earthquake Experimentation and Simulation. Scoping Study. 2001, NEES.

Reitherman, R., A Short History and Overview of NEES. 2005.

Ribes, D., *Personal communication about needs-assessment for GEON*. 2005.

Ribes, D., Baker, K.S., Millerand, F., and Bowker, G.C. Comparative Interoperability Project: Configurations of Community, Technology, Organization. in *Joint Conference on Digital Libraries*. 2005.

Rosenberg, L.C., Update on National Science Foundation Funding of the "Collaboratory". *Communications of the ACM*, 1991. 34(12): p. 83.

Rosson, M.B., and Carroll, J. M., Scenario-Based Design, in *The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications*, J.a.S. Jacko, A., Editor. 2002, Lawrence Erlbaum Associates. p. 1032-1050.

Schatz, B.R., Building an Electronic Community System. *Journal of Management Information Systems*, 1991. 8(3): p. 87.

Sommerville, I., Software Engineering. 7th ed. 2004.

Teasley, S., Wolinsky, S., Scientific Collaborations at a Distance, in *Science*. 2001. p. 2254-2255.

Workshop on Sensor Array Cyberinfrastructure and Informatics: Identifying Generic Solutions for Environmental Observatories. 2005. Center for Embedded Network Sensing, UCLA, CA.

Wulf, W.A. The national collaboratory - a white paper. in *Towards a National Collaboratory: Report of an Invitational Workshop at the Rockefeller University*. 1989.

Zare, R.N., Knowledge and distributed intelligence, in *Science*. 1997. p. 1047.

APPENDIX A – LIST OF PROJECT TEAM AND ADVISORY COMMITTEE MEMBERS

(In alphabetic order)

A.1 Project Team

Jay Alameda – University of Illinois, Urbana Champaign

Loretta Auvil – University of Illinois, Urbana Champaign

Aniruddha Bhagwat – University of Illinois, Urbana Champaign

Nosh Contractor – University of Illinois, Urbana Champaign

Steve Downey – University of Illinois, Urbana Champaign

Wayland Eheart – University of Illinois, Urbana Champaign

Tom Finholt – University of Michigan

Ingbert Floyd – University of Illinois, Urbana Champaign

Lisa Gatzke – University of Illinois, Urbana Champaign

Gayathri Gopalakrishnan – University of Illinois, Urbana Champaign

Cameron Jones – University of Illinois, Urbana Champaign

Paul Karpenko – University of Illinois, Urbana Champaign

Jessica Lam – University of Illinois, Urbana Champaign

Katherine Lawrence – University of Michigan

Yong Liu – University of Illinois, Urbana Champaign

Xavier Llorà – University of Illinois, Urbana Champaign
Luigi Marini – University of Illinois, Urbana Champaign
Barbara Minsker – University of Illinois, Urbana Champaign
Tom Prudhomme – University of Illinois, Urbana Champaign
Michael Twidale – University of Illinois, Urbana Champaign
Von Welch – University of Illinois, Urbana Champaign
Andrew Wadsworth – University of Illinois, Urbana Champaign
Michael Welge – University of Illinois, Urbana Champaign
Timothy Wentling – University of Illinois, Urbana Champaign
Hua Xie – University of Illinois, Urbana Champaign

A.2 Advisory Committee

Carl Adams – University of Minnesota
Chaitan Baru – San Diego Supercomputer Center
James Bonner – Texas A&M University
Pat Brezonik – National Science Foundation
Randall Butler – University of Illinois Urbana-Champaign
Nick Clesceri – Rensselaer Polytechnic Institute
Jeff Glass – Duke University
Tom Harmon – University of California, Merced
Kim Jones – Howard University
Orie Loucks – Miami University of Ohio

Gary Olson – University of Michigan

Michael Piasecki – Drexel University

Jerry Schnoor – University of Iowa

APPENDIX B – DETAILED QUESTIONS FOR THE NEEDS GATHERING EXERCISE

This section lists the more detailed issues that came out during the discussions within the project team and the external advisory committee. Similar to the general questions posed earlier, these issues are posed as questions in a conversational style, and some preliminary responses are provided. Because of their interest to specific focus groups, these were not posed to the participating community, but are passed on to the project office for further investigation. They are divided into three categories – questions related to sharing information, questions related to data and metadata and finally, questions related to modeling and decision support.

B.1 Detailed Questions on Sharing Information

- 1) What information would you want to find on the CLEANER cyberinfrastructure about other researchers?** What information would you want to find on the CLEANER cyberinfrastructure about other researchers?

Responses:

- a) Current research interests and publications.
- b) Who has been communicating with whom?
- c) Who has been accessing particular data or information?
- d) Who has been using which models and tools?

- 2) How do you want to be able to communicate with other CLEANER researchers?** How do you want to be able to communicate with other CLEANER researchers?

Responses:

- a) Email
- b) Audio-Video conference
- c) Text chat
- d) Audio conference
- e) Discussion forum
- f) Sharing notes and other files

- 3) Collaboration among CLEANER members would likely require formation of smaller groups of members working on common projects. What type of access privilege capabilities are needed to allow members to form groups?**
 A decision has to be made on an administration model for these groups to enable smooth functioning of a collaborative environment. What type of access privilege capabilities are needed to allow members to form groups?

Responses:

- a) One group access privilege model could be: Moderators are full-members who create a new team and can define access privileges of the other team members; Team-members have full read-write access, but they cannot define access-privileges of other members and Affiliates have only read access privileges.

- 4) Various types of data and information would be available on the CLEANER website. What types of access privileges do you want for the entire CLEANER web site?**

Responses:

- a) General public would have full access to all resources.
- b) General public would have limited access to some resources (*e.g.*, controlling EFF experiments, collaborative editing of community resources), and would need to apply for full access.
- c) All users should have to apply for access to any information.

- 5) How would you prefer the cyberinfrastructure to suggest to you about other members, discussions, data, research proposals and resources of your interest?** Once members start using the cyberinfrastructure, the database of their activities, potentially enhanced with tools that learn patterns in the activity data, could be used to make suggestions to you about other members, discussions, data sets, research proposals, articles of interest *etc.* based on research areas and resources of interest to the member. How would you prefer that this be done?

Responses:

- a) This feature would not be useful to me.
- b) Automatic system learning by monitoring the usage patterns of members.
- c) Members manually supplying keywords of interest on a regular basis.
- d) I would like to be notified when a particular type of data is collected by anyone, but especially when they are collected by someone I specify.
- e) I'd prefer to get general news postings on new materials that I can search for my interests.

6) How should the results of the research done on CLEANER be disseminated to the broader public (schools, government agencies, etc.)?

The cyberinfrastructure should provide a method for the distribution of research results to the community within and outside of CLEANER. Educational Modules could be provided in the cyberinfrastructure to enable easier knowledge dissemination. How should the results of the research done on CLEANER be disseminated to the broader public (schools, government agencies, etc.)?

Responses:

- a) Newsletters
- b) Video lectures and slides
- c) Online encyclopedias
- d) Online quizzes

7) What type of online library would you want to have in the CLEANER cyberinfrastructure?

The Library hosted on the CLEANER cyberinfrastructure could be an invaluable resource for researchers as it would host documents gleaned from the web on topics relevant to CLEANER researchers as well as documents uploaded or produced by CLEANER members. It may also be used to archive communications taking place on the cyberinfrastructure (text chats, discussions, audio-video conferences *etc.*) for future reference. What type of online library would you want to have in the CLEANER cyberinfrastructure?

Responses:

- a) Archive of all public or group documents.
- b) Archives of all public or group communications
- c) Archive of documents and other resources gleaned from the web on topics relevant to CLEANER researchers.
- d) "Metalibrary" that contains only links and references, which would reduce required storage space.

8) If a web library is included in the cyberinfrastructure, what should be the criteria for crawling web documents and archiving them in the library?

Responses:

- a) Automated system learning should monitor the keywords being searched for by members and add them to the keywords used for web-crawling.
- b) Members should suggest keywords and their acceptability decided upon by carrying out a voting procedure.
- c) Use recursive crawling to automatically search for keywords based on their frequency in certain seed websites.
- d) Web crawlers should crawl on certain seed websites and all links from that website.
- e) The form of archive should depend on a relevance score. If the document is very relevant to CLEANER topics, keep a copy; less relevant, keep an abstract and a reference or link to the full text; less relevant still, just keep the link or reference

9) What criteria should be used for removal of obsolete content from the library? The burgeoning size of the library and data archives could create a strain on the CLEANER storage, indexing and retrieval resources. Also, dated knowledge and data may need to be removed to maintain accuracy of the archives. Hence, a system for periodic revision and inspection of the archives may be needed. What criteria should be used for removal of obsolete content from the library?

Responses:

- a) All documents and data should be permanently stored.
- b) Member-uploaded documents, data and archived communications should be periodically referred to them for obsolescence checks.
- c) Web documents should be updated automatically by the system and removed when they are no longer linked on the original web site.
- d) Automatic removal of certain kinds of documents and archived communications after a certain time period
- e) Activity-based removal with notification (like the real libraries currently do). After no activity accessing a source for a period of time, it is marked for deletion and users are notified and given the opportunity to “pardon” it. If no one steps forward to pardon it after a deadline, it is deleted.

10) What types of searches of the CLEANER cyberinfrastructure would be useful to you?

Responses:

- a) Searching for specific data
- b) Searching for specific models
- c) Searching for specific member groups
- d) Searching for specific members
- e) Searching for a member's social network (who they've been communicating with and about what topics; what activities they and their collaborators have been doing).
- f) Searching for funding opportunities
- g) Searching for archived communications
- h) Searching the results of all my previous searches

11) How should the CLEANER community decide what research areas come under the auspices of CLEANER? As environmental engineering research becomes more and more diverse and overlaps with various other allied branches of engineering and science, the limits of CLEANER may need to be defined. How should the CLEANER community decide what research areas come under the auspices of CLEANER?

Responses:

- a) CLEANER management agrees upon list of topics of interest to CLEANER researchers, and members can petition for additional topics to be added.
- b) CLEANER members are free to decide.
- c) Community consensus required for new or unconventional research areas to be included.

B.2 Detailed Questions on Data and Metadata

1) What should be the policy for storage of data and metadata on the CLEANER cyberinfrastructure? What should be the policy for storage of data and metadata (information about the data; *e.g.*, how it was collected) on the CLEANER cyberinfrastructure? Central storage ensures that the data are always available when needed, regardless of changes at the data collector's site, but would require significant resources for the central network.

Responses:

- a) The entire metadata and data should be stored in the central network repositories.
- b) Only metadata should be stored on the CLEANER system and the data should reside locally on member's own servers and only uploaded when requested.
- c) Only frequently requested data should be uploaded to the central network repositories.
- d) The system could provide options for data generators to choose among the above options.

2) As the data expected to be shared on CLEANER span different domains of environmental engineering, what should be the standard for the content of this metadata? Effective knowledge management requires that acquired data should be annotated with metadata. Metadata can also be created for community models linked to the cyberinfrastructure, allowing data to be easily assimilated into models. But as the data expected to be shared on CLEANER span different domains of environmental engineering, what should be the standard for the content of this metadata? Also, how compatible should it be with other metadata standards?

Responses:

- a) Metadata based on domain of environmental engineering.
- b) Metadata based on models to which the data would be applied.
- c) Metadata based on the specific measurement technique used.
- d) Development of a new and customized metadata standard
- e) User-specified metadata

3) What types of data analysis tools should be included in the cyberinfrastructure?

Responses:

- a) Statistical analysis tools
- b) Tools for automatically detecting anomalies in sensor data
- c) Tools for creating statistical and other data models (e.g., neural networks)
- d) Tools for assimilating data into numerical or analytical models

- 4) **What advice should the cyberinfrastructure provide when a user works on data?**

Responses:

- a) Advice on the quality of data (collector of data, procedure involved *etc.*)
- b) Comparison of the dataset with respect to similar datasets
- c) What analysis (data mining, modeling *etc.*) and visualizations have been done to a particular dataset by whom?

- 5) **What functionalities should the cyberinfrastructure (CI) provide for data fusion and validation?** Fusion and validation tools would provide integration of multiple data sources. For example, the stream flow data from multiple gauging stations of a river give a more complete picture of the hydrological conditions of the river than the data from a single station. Another example would be comparing parameters measured by different researchers and by different techniques to identify the accuracy or uncertainties of those quantities. While these tools are useful, data fusion and validation techniques vary from problem to problem and therefore it may be very difficult to provide a universal tool for the many types of environmental data that are expected to be available on CLEANER. What functionalities should the cyberinfrastructure (CI) provide for data fusion and validation?

Responses:

- a) CI provides various tools to the users and leaves it to their discretion to choose/modify the tools.
- b) CI fuses the data according to the wishes of the original collectors of data.
- c) CI provides no tools for data fusion, but a disclaimer is attached to the dataset.
- d) CI fuses the data according to its own best judgment.
- e) Experts at a central CLEANER information technology lab work with investigators to fuse their data in the most effective way with existing and new data.

- 6) **How long would you be willing to wait to receive archived data or information from the CLEANER web site after you request it?** When the CLEANER network is in place, it is expected that very large quantities of data will be generated and stored in the archives. If all that data were available for immediate access, the community would have to make major investments in large-scale, rapid data archiving systems, potentially at the expense of other infrastructure. With this in mind, how long would you be willing to wait to receive archived data or information from the CLEANER web site after you request it?

Responses:

- a) All data should be available within seconds.
- b) I would be willing to wait a few minutes for some data (*e.g.*, large satellite datasets or old data).
- c) I would be willing to wait a few hours for some data.
- d) I would be willing to wait a few days for some data.

7) What mechanisms are needed for remote control of the instrumentation at the environmental field facilities (EFFs) to allow members of the CLEANER community to conduct experiments remotely?

Responses:

- a) Any remote users should be able to control experiments at the EFFs; A peer review system is needed to approve experiments at the EFFs.
- b) Only experiments that enhance the observatory's mission (which may be only partly supported by CLEANER) should be allowed to be conducted remotely.
- c) No remote experimentation control is needed.

8) Should the data being collected by other agencies be re-archived by CLEANER or should it simply be linked to CLEANER? Many other agencies involved in data collection are collecting large volumes of environmental data. A specific example of this case is the data archived by USGS - it is comprehensive, well-documented, widely used and easily retrievable. Should the data being collected by these agencies be re-archived by CLEANER or should it simply be linked to CLEANER? Linking the data would require significantly less archive space, but would require ongoing communication between the agencies and CLEANER to address any changes in the data sources that require revisions to the links.

Responses:

- a) All external data should be archived locally.
- b) All external data should be linked to archives maintained by the relevant external agency.
- c) Data sharing agreements should be put in place with key external agencies to provide links and maintenance of their external data, and other external data should be archived locally.

B.3 Detailed Questions on Modeling and Decision Support

1) What functionalities should the cyberinfrastructure provide for modeling?

Responses:

- a) Allow users to post their simulation models for public download.
- b) Require model contributors to create metadata for models to provide easier links between data and modeling systems.
- c) Develop mechanism for peer review of community models, based on verification at multiple sites.
- d) Provide optimization tools for automated model and parameter selection.
- e) Allow models to be run interactively on the CLEANER cyberinfrastructure.
- f) Allow visualizations of model results to be made interactively on the CLEANER cyberinfrastructure.
- g) Allow models, and their inputs and outputs, to be viewed and run collaboratively by multiple users.
- h) Provide advice on what models might be used on particular datasets, or what datasets have been previously used for specific models and by whom.
- i) Provide seamless integration with high-end computing, so that computationally-intensive models can easily be run on the cyberinfrastructure.

2) Should there be a group of modeling experts, say at a central CLEANER information technology lab, which provides support to CLEANER users?

Responses:

- a) No, this is not needed.
- b) Yes, for providing integration support across environmental field facilities (EFFs), *e.g.* for large-scale models.
- c) Yes, for modeling tasks those are needed by many or all environmental field facilities (EFFs), but not for specialized modeling would serve only one or two EFFs. For those jobs, EFFs should have their own modeling groups.

3) What types of decision support should be provided within the CLEANER cyberinfrastructure?

Responses:

- a) General purpose optimization tools.

- b) Decision and fault tree tools.
- c) Visualization tools.
- d) Collaborative decision making tools (*e.g.*, see the solution center in this CyberCollaboratory).

APPENDIX C – LIST OF URLS USED FOR CLEANER LIBRARY WEBCRAWLING

This appendix lists the URLs used as seed for webcrawl conducted to build the Web Documents section of the CyberCollaboratory Library module. It is divided into sub-sections to reflect the category of the URLs.

C.1 Other Large-Scale Collaboratory Projects

- NSF <http://www.nsf.gov/>
- EPA <http://www.epa.gov/>
- CLEANER website at Berkeley <http://cleaner.ce.berkeley.edu/>
- CUAHSI <http://www.cuahsi.org/>
- NEES <http://www.neesgrid.org/> , <http://it.nees.org/>
- National Center for Ecological Analysis and Synthesis (NCEAS)
<http://www.nceas.ucsb.edu/>
- SPARC, the Space Physics and Aeronomy Research Collaboratory
<http://www.windows.ucar.edu/sparc/>
- GEON <http://www.geongrid.org/>
- CMCS <http://cmcs.ca.sandia.gov/>
- NCEAS <http://www.nceas.ucsb.edu/>
- NEON <http://www.nsf.gov/bio/neon/>

- US National Virtual Observatory <http://www.us-vo.org/>
- The National Science Digital Library (NSDL) <http://nsdl.org/>
- The Grid Physics Network <http://www.griphyn.org/>
- Earth System Grid <https://www.earthsystemgrid.org/>
- Digital Library for Earth Systems <http://www.dlese.org/>
- US Department of Energy <http://www.doeccollaboratory.org/>
- Japanese Earth Simulator Center <http://www.es.jamstec.go.jp/esc/eng/>
- Advanced Knowledge Technologies <http://www.aktors.org/>
- Telescope collaboratory (including Hubble)
- The Biomedical Informatics Research Network (BIRN) <http://www.nbirn.net/>
- National Institute of General Medical Sciences <http://www.nigms.nih.gov/>

C.2 Cyberinfrastructure information technologies

- Access GRID <http://www.accessgrid.org/>
- GLOBUS Alliance <http://www.globus.org/>
- TERAGRID <http://www.teragrid.org/>
- OGCE <http://www.collab-ogce.org/>
- NLADR <http://www.nladr.net/>
- SDSC <http://www.sdsc.edu/>
- NCSA <http://www.ncsa.uiuc.edu/>
- ALG <http://alg.ncsa.uiuc.edu/>
- KLSG <http://learning.ncsa.uiuc.edu/>

- IlliGAL <http://www-illigal.ge.uiuc.edu/>
- PACI <http://www.paci.org/>
- NSF Middleware Initiative <http://www.nsf-middleware.org/>
- DOE Scientific Discovery Through Advanced Computing
<http://www.osti.gov/scidac/>
- The CONDOR project <http://www.cs.wisc.edu/condor/>

C.3 Research on science of collaboratories

- Collaboratories project at the University of Michigan
<http://www.scienceofcollaboratories.org/>
- IEEE Computer society <http://www.computer.org/>
- International Society for Learning Sciences <http://www.isls.org/>
- International Journal of Computer-Supported Collaborative Learning
<http://ijcscl.org/>
- ACM Digital Library <http://portal.acm.org/>
- Access GRID <http://www.accessgrid.org/>
- GLOBUS Alliance <http://www.globus.org/>
- ASCE <http://www.asce.org/>
- The Directorate for Computer and Information Science and Engineering (NSF)
<http://www.cise.nsf.gov/>
- U.K. Research Councils E-science Program <http://www.rcuk.ac.uk/escience/>
- The National Academies Press <http://www.nap.edu/>

- Society for Industrial and Applied Mathematics (SIAM) <http://www.siam.org/>
- Lawrence Livermore National Laboratory <http://www.llnl.gov/>
- Los Alamos National Laboratory <http://www.lanl.gov/>

C.4 Environmental Engineering & Hydrology

C.4.1 Government/UN Bodies

- US Dept of Agriculture <http://www.usda.gov/>
- Natural Resources Defense Council <http://www.nrdc.org/>
- UN Environmental Programme <http://www.unep.org/>
- Food and agriculture organization (FAO) <http://www.fao.org/>
- NOAA <http://www.noaa.gov/>
- National Climatic Data Center <http://www.ncdc.noaa.gov/>
- US Geological Survey <http://www.usgs.gov/>
- US Long term Ecological Research network <http://lternet.edu/>
- The National Center for Atmospheric Research <http://www.ncar.ucar.edu/>

C.4.2 Other Organizations:

- American Water Works Association <http://www.awwa.org/>
- UNEP-Freshwater <http://freshwater.unep.net/>
- American Water Resource Association <http://www.awra.org/>
- Water Environment Federation <http://www.wef.org/>

- ASCE <http://www.asce.org/>
- American Geophysical union (AGU) <http://www.agu.org/>
- International Water Association (IWA) <http://www.iwahq.org.uk/>
- International Water Resources Association <http://www.iwra.siu.edu/>
- Water Environment Resource Foundation <http://www.werf.org/>
- International Association for Hydrological Sciences (IAHS)
<http://www.cig.enscm.fr/~iahs/>
- International Association of Hydraulics Engineering and Research (IAHR)
<http://www.iahr.net/>

C.5 CLEANER Project Data

- CLEANER website at Berkeley <http://cleaner.ce.berkeley.edu/>
- CLEANER website at NACSE <http://cleaner.nacse.org/workshops/>
- CLEANER workshop at RPI <http://www.rpi.edu/dept/research/>
- Other CLEANER workshops <http://nsf.gov/>, <http://eng.nsf.gov/>