From Shared Databases to Communities of Practice: A Taxonomy of Collaboratories

Nathan Bos

Applied Physics Laboratory
Johns Hopkins

Ann Zimmerman

School of Information
University of Michigan

Judith Olson

School of Information
University of Michigan

Jude Yew

School of Information
University of Michigan

Jason Yerkie

Corporate Executive Board

Erik Dahl

MAYA Design, Inc.

Gary Olson

School of Information
University of Michigan

Promoting affiliation between scientists is relatively easy, but creating larger organizational structures is much more difficult, due to traditions of scientific independence, difficulties of sharing implicit knowledge, and formal organizational barriers. The Science of Collaboratories (SOC) project conducted a broad five-year review to take stock of the diverse ecosystem of projects that fit our definition of a collaboratory and to distill lessons learned in the process. This article describes one of the main products of that review, a seven-category taxonomy of collaboratory types. The types are: Distributed Research Centers, Shared Instruments, Community Data Systems, Open Community Contribution Systems, Virtual Communities of Practice, Virtual Learning Communities, and Community Infrastructure Projects. Each of the types is defined and illustrated with one example, and key technical and organizational issues are identified.

doi:10.1111/j.1083-6101.2007.00343.x

Introduction

Why are scientific collaborations so difficult to sustain? Inspired by the vision of Wulf (1989, 1993) and others, researchers over the last 25 years have made a number of large-scale attempts to build computer-supported scientific collaboration environments, often called collaboratories (National Research Council, 1993). Yet only a few of these efforts have succeeded in sustaining long-distance participation, solving larger-scale problems, and initiating breakthrough science.

Should we consider this surprising? Scientific progress is by nature uncertain, and long distance collaboration always faces many barriers (Olson & Olson, 2000). Still, the difficulties of sustaining large-scale collaboratories were unexpected to many scientists and funders, partially because modern studies of science have repeatedly emphasized the social nature of scientific communities. Pioneers in the social studies of science documented how the basic activities of scientists, such as deciding what counts as evidence, are fundamentally social undertakings (Collins, 1998; Latour & Woolgar, 1979). Kuhn (1963) showed how scientific peer groups determine what theories will be accepted as well as make more mundane judgments about what papers will be published and what grants will be funded. Crane (1972) first described the loosely-affiliated but highly interactive networks of scientists as "invisible colleges." Compared to other peer groups, scientific communities are often surprisingly egalitarian and broadly international. Newman's (2001) social network analyses of scientific communities in biomedicine, physics, and computer science showed that each of these fields formed a well-interconnected or "small world" network (Watts & Strogatz, 1998). Scientific users were early adopters and promoters of many of the technologies that long-distance collaboration now relies on, including email, ftp servers, and the World Wide Web.

Given this context, it was natural for visionaries to predict that scientists would lead the way in making boundaries of distance obsolete and would be the first take advantage of new technologies to assemble larger-scale efforts across distance. However, previous research failed to document some crucial barriers that make scientific collaboration more difficult than expected. There is a key distinction between informal, one-to-one collaborations, which have long been common between scientists, and more tightly coordinated, large-scale organizational structures, which are a less natural fit. In particular, our research has highlighted three types of barriers.

First, scientific knowledge is difficult to aggregate. While *information* has become very easy to transmit and store over great distances, *knowledge* is still difficult to transfer (Szulanski, 1992). Scientists generally work with ideas that are on the cutting edge of what is understood. This knowledge often requires specialized expertise, is difficult to represent, may be tacit, and changes rapidly. This kind of knowledge is the most difficult to manage over distances or disseminate over large groups. Scientists can often negotiate common understandings with similar experts in extended one-to-one interactions but may have great difficulty communicating what they know to larger distributed groups. Standard tools for knowledge management may presume

an ability to codify and disseminate knowledge that is not realistic in cutting-edge scientific enterprises.

Second, scientists work independently. Scientists generally enjoy a high degree of independence, both in their day-to-day work practices as well as in the larger directions of their work. Scientific researchers have greater freedom to pursue high risk/high reward ideas than do individuals in many other professions. Most practicing scientists would strongly resist controls that many corporate employees accept as normal, such as having their work hours, technology choices, and travel schedules dictated by others. The culture of independence benefits science in many ways, but it also makes it more difficult to aggregate scientists' labors. Scientific collaborations must work harder than other organizations to maintain open communication channels, adopt common toolsets, and keep groups focused on common goals.

The third barrier is the difficulty of cross-institutional work. Crossing boundaries between institutions is frequently a greater barrier than mere distance (Cummings & Kiesler, 2005). Even when all of the scientists are ready to proceed, collaboratorations can run into institutional-related problems, especially legal issues, that cannot be resolved (Stokols, Fuqua, Gress, Harvey, Phillips, Baezconde-Garbanati, et al., 2003; Stokols, Harvey, Gress, Fuqua, & Phillips, 2005). Universities often guard their intellectual property and funding in ways that hinder multi-site collaboration. Since the biggest science funding sources are federal government based, international or even inter-state collaboration is often hindered. In corporate settings, the largest international collaborations are often made possible by mergers, but there has been no such trend in university research, again due to the funding sources. Very few universities operate across state lines, much less national boundaries.

These barriers that are specific to scientists are compounded by the normal challenges of working across distance. Distance collaboration challenges coordination and trust building (Jarvenpaa & Leidner, 1999), fosters misunderstandings (Cramton, 2001) and inhibits communication of tacit knowledge (Lawson & Lorenz, 1999) and transactive knowledge, or knowledge of what colleagues know (Hollingshead, 1998).

The Science of Collaboratories Project

The Science of Collaboratories (SOC) was a five-year project funded by the National Science Foundation (NSF) to study large-scale academic research collaborations across many disciplines. The overall goals of the SOC project were to: 1) perform a comparative analysis of collaboratory projects, 2) develop theory about this new organizational form, and 3) offer practical advice to collaboratory participants and to funding agencies about how to design and construct successful collaboratories. Through our research, we identified many of the barriers, both organizational and technological, that made these projects difficult. On a more positive note, we also assembled a database with many success stories. The SOC database (http://www.scienceofcollaboratories.org) contains 75 summaries of collaboratories that

achieved some measure of success and analyses of the technology and other practices that enabled them. Additional information on this project and extended case studies will be published in a forthcoming book, *Science on the Internet* (Olson, Zimmerman, & Bos, 2007).

This article reports one of the main outputs of the SOC project, which is a seven-category taxonomy of collaboratories. This taxonomy has proven useful and robust for documenting the diversity of collaboratories that now exists, identifying associated strengths and key challenges, and framing a research agenda around these types.

Collaboratory Typologies

This is not the first typology of its kind, although it is unique in its scale and purpose. A great deal of previous work in computer-supported cooperative work (e.g., Grudin, 1994; DeSanctis & Gallupe, 1987) has classified technology as to how well it supported different task types and different configurations of local and distant workers. Bafoutsou and Mentzas (2002) reviewed this literature and mapped it onto the specific technology functionalities of modern groupware systems. This type of classification yields insights about what kinds of task/technology matches are most apt (e.g., text chat is a good choice for maintaining awareness but a poor choice for negotiation.) The SOC project conducted a similar technology inventory as part of its research, but this level of classification is not as useful for classifying large-scale projects because these projects perform many different task types using numerous tools over the course of their lives. Any single project will at different times engage in negotiation, decision-making, and brainstorming, and will make use of email, face-to-face meetings, and real-time communication tools. Low-level task/technology matching may be one factor in project success, but it is not a sufficient predictor of overall success.

A larger-scale classification scheme has been developed by Chompalov and Shrum (1999) based on data from Phase I of the American Institute of Physics (AIP) Study of Multi-Institutional Collaborations (AIP, 1992, 1995, 1999). This large-scale, three-phase study looked at a large number of collaborations in high-energy physics, space science, and geophysics. Chompalov and Shrum analyzed data from a subset of 23 of these projects and performed cluster analysis that made use of seven measured dimensions: project formation and composition, magnitude, interdependence, communication, bureaucracy, participation, and technological practice. Their analysis sought to find relationships between these dimensions and the outcome measures of trust, stress, perceived conflict, documentary process, and perceived success. Most of these categories had little relationship to success measures; nor did they correspond strongly to particular sub-disciplines. One of the researchers' findings was particularly intriguing: The technological dimension (whether the project designed and/or built its own equipment and whether their technology advanced the state of the art) corresponded to all five success measures. It is unclear from these data whether the technology measures actually caused better success or

corresponded in some other way; that is, led to a different sort of project. It is difficult to believe that every project should design its technology to work on the "bleeding edge" in order to ensure success (nor do Chompalov and Shrum make any such claim). It seems more likely that other features of these cutting-edge design projects, such as intrinsic interest, tangible products, or funding levels, contributed to their success.

By observing the value that could be obtained from "bottom-up" studies using large datasets of heterogeneous projects, our project learned a great deal from the groundbreaking AIP studies. The classification system we developed, however, differs fundamentally in purpose from that of Chompalov and Shrum. While they sought to explain success after the fact, our project sought to identify organizational patterns, somewhat similar to design patterns (after Alexander, Ishiwaka, & Silverstein, 1977), which could be used by funders and project managers in designing new collaborations. Rather than focusing on the technology or the emergent organizational features, the scheme is tightly focused on the goals of the projects. The result of this classification should be identification of key challenges and recommendation of practices, technology, and organizational structures that are appropriate for a stated set of goals.

Dataset and Sampling Methods

In spring of 2002 the Science of Collaboratories project started putting together a database of collaboratories that would be the most comprehensive analysis of such projects to date. The published part of the collaboratories database is viewable online from www.scienceofcollaboratories.org. The database currently contains 212 records of collaboratories. Of these, 150 have received a classification, and summaries have been published for 64. Nine broad disciplinary categories are represented using the National Science Foundation's field of study classifications.

Attendees of an SOC workshop together constructed and agreed to this definition of a collaboratory:

> A collaboratory is an organizational entity that spans distance, supports rich and recurring human interaction oriented to a common research area, and fosters contact between researchers who are both known and unknown to each other, and provides access to data sources, artifacts, and tools required to accomplish research tasks.

This definition is restricted to scientific endeavors, thus excluding many (albeit not all) corporate and government projects. Within the sciences, however, it is quite broad, covering many disciplines and many more organizational forms than did previous studies such as those of the AIP. For the purposes of data collection, the notion of distance was operationalized to include only collaborations that crossed some kind of organizational boundary (in this case following the AIP lead). For academic research this usually meant that nominees would have to be multi-university

or university/other partnerships; most that were merely cross-departmental or cross-campus were excluded. Few other restrictions were placed on entry, however, in order to be as inclusive as possible.

The breadth of this definition of a collaboratory complicated the choice of a sampling technique. There did not seem to be any way to create a truly representative sample, because the true boundaries of the population to be sampled were unknown. Some options were to choose to sample certain subsets of the population, such as all multi-site projects sponsored by NSF, all projects appearing in Google searches of the word "collaboratory," or all projects nominated by members of a certain professional organization. Each of these possibilities would inevitably exclude interesting areas of inquiry.

Doing so required a type of nonrandom sampling, namely purposive sampling. Patton (1990) provides a taxonomy of purposive sampling techniques. The technique used in this project is similar to what Patton calls s*tratified purposeful sampling*, which organizes observations to cover different "strata" or categories of the sample. The complication of this project was that the groups were themselves unknown at the beginning of the study. The technique chosen needed to be flexible enough to both classify and describe, so elements of *extreme and deviant case sampling,* which pays special attention to unusual or atypical cases, were incorporated.

A purposive sampling method called "landscape sampling" was devised to produce a sample as comprehensive as possible in type, but not in frequency. It is similar to what an ecologist would do in a new area, which would be to focus on finding and documenting every unique species, while putting off the job of assessing how prevalent each species is in a population. An ecologist in this kind of study focuses on novelty rather than representativeness; once a particular species is identified from a few instances, most other members of that species are disregarded unless they have unusual or exemplary features.

In searching out new cases, we cast the net very broadly, using convenience and snowballing techniques, along with other more deliberate strategies. Any type of project could be nominated by having an initial entry created in the database. Nominations were also solicited from the following sources: SOC project staff, SOC workshop attendees, three major funding sources, (the National Science Foundation, the National Institutes of Health, and the Department of Energy), program officers of each of those sources, and review articles in publications such as the annual database list published in *Nucleic Acids Research* (e.g., Baxevanis, 2002). Throughout the project the SOC website included a form for nominating projects that any visitor could fill out, and some nominations were received this way. Finally, a snowball technique was used, whereby project interviewees were asked to nominate other projects. These methods led to nomination of more than 200 projects, a richer and broader sample than could have been obtained otherwise.

Landscape samples must have criteria for inclusion/exclusion of cases that fit the definition. Resources were not available to investigate every project that fit the definition of a collaboratory. Instead energy was focused where the most learning

could happen and the most interesting sample could be obtained. The criteria for collaboratories that would be included were:

1. *Novelty*. The sampling technique was strongly biased toward finding examples of collaboratories that were different than what had been seen before. Projects were pursued that were novel in their use of technology, their organizational or governance structures, or the scientific discipline that they covered. The emergence of identifiable types (discussed below) greatly aided identification of novel cases.
2. *Success*. Projects that were particularly successful were also of special interest, regardless of whether they were novel. The success criterion had also been explored at a project workshop (SOC, 2001). Success usually manifested as either producing a strong body of scientific research or attracting and retaining a large number of participants, but there were other possible criteria as well, such as generativity.
3. *Prototypicality*. In some cases, collaboratories were included not because they were novel, but because they seemed prototypical of a certain type. (Identification of types aided this process.) This helped us correct and re-center the dataset when it turned out that the first one or two collaboratories of a certain type were atypical in some respects, just as the first member of a species to be identified may happen to be an outlier on some category.

Social vetting was also used to check and validate these decisions. Few collaboratory nominees were either included or excluded on the basis of one person's judgment. The process was for one investigator to do an initial summary of the project and report back to a subcommittee of three to five researchers who would make the decision whether to pursue the investigation further. This served to improve the decision process in the same way that multi-rater coding improves other qualitative rating methods.

Use of Landscape Samples

Landscape sampling is useful for expanding the horizons of a particular area of inquiry and producing a rough map of a new problem space. Landscape sampling is not useful for making some kinds of generalizations about a sample. For example, the collaboratories database could not be used to make claims about the average size of collaboratories or average success rate; for that, a representative sampling method would be needed. A landscape sample is useful for identifying characteristics, such as identifying key organizational issues and technology issues.

Multiple-Category Collaboratories

The process of categorizing collaboratories was a social one, as described above. A small group of experienced investigators examined the data and decided which classification best fit each project. Many projects were also given multiple classifications. One category was always chosen to be primary, but projects could have any number of secondary classifications. Often this was because a project had multiple

components. For example, the main work of the Alliance for Cellular Signalling is coordinated multi-site lab work, making it a clear-cut Distributed Research Center. However, this project was also managing the "Molecule pages" Community Data System on a related topic with different participants. Sometimes projects were given multiple classifications because they legitimately had multiple goals. For example, many of the projects list the training of new scientists as one of their goals, but in most cases this is not the primary goal. Therefore, many projects are assigned a secondary category of Virtual Learning Community. A few, upon further investigation, actually did prioritize training and dissemination ahead of new research; these were assigned the primary categorization of a Virtual Learning Community.

Seven Types of Collaboratories

Our seven-category classification system is presented below. For each classification, the following information is given:

1. Collaboratory type definition
2. An example collaboratory of this type
3. Key technology issues of this collaboratory type
4. Key organizational issues of this collaboratory type

Shared Instrument

Definition

This type of collaboratory's main function is to increase access to a scientific instrument. Shared Instrument collaboratories often provide remote access to expensive scientific instruments such as telescopes, which are often supplemented with videoconferencing, chat, electronic lab notebooks, or other communications tools.

Example

The Keck Observatory, atop the Mauna Kea summit in Hawaii, houses the twin Keck Telescopes, the world's largest optical and and infrared telescopes. Keck has been a leader in development of remote operations (Kibrick, Conrad, & Perala, 1998). Observing time on the Keck Telescope is shared between astronomers from Keck's four funders: the University of California system, the California Institute of Technology, NASA, and the University of Hawaii. Each institution is allocated time in proportion to its financial contribution. Because of the extreme altitude of the observatory, Keck's instruments have been made remotely accessible from Waimea, Hawaii, 32 km away. Remote observation employs a high-speed data link that connects observatories on Mauna Kea with Internet-2 and runs on UNIX. To prevent data loss, remote sites also have automated backup access via ISDN. Remote scientists have contact with technicians and scientists at the summit and at Waimea through H.323 Polycom video conferencing equipment. Future plans include online data archiving. Remote access facilities have also been

constructed at the University of California, Santa Cruz; the University of California, San Diego; and the California Institute of Technology. These remote facilities allow astronomers to do short observation runs (one night or less) without traveling to Hawaii, and allow late cancellations to be filled, increasing productivity.

Technology Issues

Shared Instrument collaboratories have often pushed the envelope of synchronous (real-time) communications and remote-access technology. Keck's recent innovation of allowing access to the Hawaii-based observatory from California is pushing the envelope of what has been done in this area. Other interesting technology problems that often arise are those involved with managing very large instrument output datasets and providing security around data. One product of the EMSL collaboratory (Myers, Chappell, & Elder, 2003) was a high-end electronic notebook that improved on paper notebooks by saving instrument output automatically, allowing access from many locations, and providing the level of security needed for lab notebooks.

Organizational Issues

Shared Instrument collaboratories must solve the problem of allocating access, which becomes trickier when instruments are oversubscribed (i.e., there is more demand than time available.) Collaboratories typically solve this by appointing committees to award time based on merit. A less well-handled problem is providing technical support. Local technicians are often critical to using the instruments effectively; remote participants may not have the social relationships and contextual knowledge to work with them effectively.

Community Data Systems

Definition

A Community Data System is an information resource that is created, maintained, or improved by a geographically-distributed community. The information resources are semi-public and of wide interest; a small team of people with an online filespace of team documents would not be considered a Community Data System. Model organism projects in biology are prototypical Community Data Systems.

Example

The Protein Databank (PDB) is the single worldwide repository for the processing and distribution of 3-D structure data of large molecules of proteins and nucleic acids (Berman, Bourne, & Westbrook, 2004). PDB was founded in 1971 and was a pioneer in Community Data Systems. As of October 2003, the PDB archive contains approximately 23,000 released structures, and the website receives over 160,000 hits per day. Government funding and many journals have adopted guidelines set up by the International Union of Crystallography (IUC) for the deposition and release of structures into the PDB prior to publication. IUC was additionally instrumental in establishing

the macromolecular Crystallographic Information File (mmCIF), now a standard for data representation.

Technology Issues

Community Data Systems are often on the forefront of data standardization efforts. Large shared datasets can neither be constructed nor used until their user communities commit to formats for both storing and searching data. PDB's role in creating the mmCIF standard is very typical; there are many other examples of standards and protocols that have emerged in conjunction with Community Data Systems.

A second area of advanced technology that often seems to co-evolve with community datasets is modeling and visualization techniques. Modelers find opportunities among these large public datasets to both develop new techniques and make contact with potential users. The Visible Human project, for example, has unexpectedly become a touchstone for new developments in 3-D anatomical visualization because of the dataset and user base it provides (Ackerman, 2002).

Organizational Issues

Community Data Systems can be viewed as Public Goods projects that may find themselves in a social dilemma related to motivating contribution (Connolly, Thorn, & Heminger, 1992). In addition to figuring out how to motivate contributors, these projects also must develop large-scale decision-making methods. Decisions about data formats and new developments for such community resources must take into account the views of many different stakeholders from many different locations.

Open Community Contribution System

Definition

An Open Community Contribution System is an open project that aggregates efforts of many geographically separate individuals toward a common research problem. It differs from a Community Data System in that contributions come in the form of work rather than data. It differs from a Distributed Research Center in that its participant base is more open, often including any member of the general public who wants to contribute.

Example

The Open Mind project is an online system for collecting "common sense" judgments from volunteer participants ("netizens") via its website (Stork, 1999). Participants contribute by making simple common sense judgments and submitting answers via a Web form. Participation is open, and contributors are encouraged to return to the site often. The aggregated data are made available to artificial intelligence projects requiring such data. Two currently active projects are on handwriting recognition and common sense knowledge. The site is hosted by Ricoh Innovations, and individual projects are designed and run by academic project teams. Current project teams are from MIT, Stanford, and Johns Hopkins.

Inspiration for this system came when David Stork, founder of the project, reviewed many different pattern recognition systems and came to the conclusion that rapid advances in this field could take place if very large datasets were available. These datasets would generally be too large for hired project staff to construct, but they might be assembled with help from many online volunteers.

The Open Mind initiative only collects and aggregates data; it does not develop products (although Ricoh Innovations does.) Data from the project are made freely available to both commercial and noncommercial users.

Technology Issues

The main technology challenge for these collaboratories is to create a system that operates across platforms and is easy to learn and use. Users must be able to do productive work in the system very quickly without much advanced training. Administrators of such collaboratories do well to utilize the tools of user-centered design early and often. These projects also must address the challenge of standardized data formatting, without expecting contributors to learn complex entry methods.

Organizational Issues

Open systems must address the problem of maintaining quality control among a large and distributed body of contributors. Some projects rely on sheer mass of data to render mistakes or inconsistencies harmless. NASA's Clickworkers project, for example, found that by averaging together the crater-identification work of several community volunteers, they could create a dataset as high in quality as would be produced by a smaller number of trained workers. Wikipedia uses community vetting in a different way. Mistakes in the data are usually caught by repetitive viewing and vetting by knowledgeable readers. Intentional biases, editorializing, or vandalizing of the data are also generally caught and corrected quickly. Some volunteer editors take on the responsibility of being notified automatically when certain controversial entries, such as the entry on "Abortion," are edited (Viegas, Wattenber, & Dave, 2004). As with Community Data Systems, Open Community Contribution Systems must also address the challenge of reaching and motivating contributors.

Virtual Community of Practice

Definition

This collaboratory is a network of individuals who share a research area and communicate about it online. Virtual Communities may share news of professional interest, advice, techniques, or pointers to other resources online. Virtual Communities of Practice are different from Distributed Research Centers in that they are not focused on actually undertaking joint projects. The term "community of practice" is taken from Wegner and Lave (1998).

Example

Ocean.US is an electronic meeting place for researchers studying oceans, with a focus on U.S. coastal waters (Hesse, Sproull, Kiesler, & Walsh, 1993). The project runs an

active set of bulletin boards/email listservs used to exchange professional information (e.g., job openings), along with some political and scientific issues. Ocean.US also provides online workspace for specific projects and develops online support for workshops and distance education in this field. The project began in 1979 as ScienceNet, providing subscription-based electronic discussions and other services before email and Web services were widely available. ScienceNet was shut down in the mid-1990s when the technology became ubiquitous and the project could no longer be supported with paid subscriptions. It was re-implemented as a set of web-based services, and renamed Ocean.US. The service is owned and run by a for-profit company, Omnet.

Technology Issues

As with Open Community Contributions Systems, the main technology issue is usability. Successful Communities of Practice tend to make good use of Internet-standard technologies such as listserv, bulletin boards, and accessible web technology. A key technology decision for these projects is whether to emphasize asynchronous technologies such as bulletin boards, or invest time and energy into synchronous events such as online symposia.

Organizational Issues

Communities of Practice, like other for-profit e-communities, must work hard to maintain energy and participation rates with a shifting set of participants. Faced with stiff competition for online attention, many Community of Practice websites are moving away from all-volunteer efforts toward professional or for-profit management.

Virtual Learning Community

Definition

This type of project's main goal is to increase the knowledge of participants but not necessarily to conduct original research. This is usually formal education, i.e., provided by a degree-granting institution, but can also be in-service training or professional development.

Example

The Ecological Circuitry Collaboratory (ECC) is an effort to "close the circuit" between empiricists and theoreticians in the ecological sciences and to create a group of quantitatively strong, young researchers. The collaboratory is comprised of a set of seven investigators and their students. It is funded by the NSF Ecosystem Studies and Ecology programs. Participant researchers study the relationship between system structure (i.e., biodiversity) and the function of that system, and they also do work in terrestrial and aquatic habitats including forests, streams, estuaries, and grasslands.

The goal of the project is to educate young ecologists to combine empirical research methods with quantitative modeling, as well as to show that ecological modeling is a valuable resource in an ecologist's toolkit. Toward this end, students

and investigators meet regularly for short courses and exchange of ideas and information. The collaboratory also includes a postdoctoral researcher who leads the team in integration and synthesis activities, coordinates distributed activities, and supports faculty mentors.

Technology Issues

In multi-institutional educational projects there is often a large disparity in technology infrastructure, especially when well-equipped American universities collaborate with K-12 institutions or non-western universities. Educational projects can make use of specialized e-learning software, but there are often tradeoffs involved. In currently available software, one often has to choose between software primarily designed for one-to-many broadcasts (e.g., lectures) and those designed to support small groups working in parallel. Many software packages are designed only for Windows-based systems, despite the continued prevalence of Macintoshes and the growing popularity of Linux in educational settings.

Organizational Issues

Compared to other collaboratory types, the organizational issues related to Virtual Learning Communities are relatively easy to address. Key challenges are aligning educational goals and aligning assessments so that learners from multiple sites are having their needs met. Projects such as the VANTH biomedical engineering collaboratory (Brophy, 2003) have spent a great deal of up-front time negotiating goals, and project staff have spent much time and energy developing cross-site assessments with good success, demonstrating viability. Despite this, only a very few Virtual Learning Communities were found and added to the database, suggesting that they are not very common.

Distributed Research Center

Definition

This collaboratory functions like a university research center but at a distance. It is an attempt to aggregate scientific talent, effort, and resources beyond the level of individual researchers. These centers are unified by a topic area of interest and joint projects in that area. Most of the communication is human-to-human.

Example

Inflammation and the Host Response to Injury is a large-scale collaborative program that aims to uncover the biological reasons why patients can have dramatically different outcomes after suffering similar traumatic injuries. This research aims to explain the molecular underpinnings that lead to organ injury and organ failure, while also helping to clarify how burn and trauma patients recover from injury. Inflammation and the Host Response to Injury consists of an interdisciplinary network of investigators from U.S. academic research centers. Participating institutions

include hospitals that participate in clinical research studies, academic medical centers that perform analytical studies on blood and tissue samples, and informatics and statistics centers that develop databases and analyze data.

The program is organized into seven core groups. Each of the core groups is composed of a core director, participating investigators, and other experts. Core personnel are accomplished and highly successful basic scientists working in the areas of research relevant to the focus of each individual core. In addition to researchers who are experts in identifying and quantifying molecular events that occur after injury, the program includes experts who have not traditionally been involved in injury research but have been integrated into the program to expand the multi-disciplinary character of the team. These experts include biologists who are leaders in genome-wide expression analysis, engineers who do genome-wide computational analysis, and bioinformatics experts who construct and analyze complex relational databases. Program scientists are mutually supported by core resources that provide the expertise, technology, and comprehensive, consensus-based databases that define the success of this program.

Technology Issues

Distributed research centers encounter all of the technology issues of other collaboratory types, including standardization of data and providing long-distance technical support. Distributed Research Centers also should pay attention to technologies for workplace awareness, which try to approximate the convenience of face-to-face collaboration. Awareness technologies such as Instant Messaging and more exotic variants (Gutwin & Greenberg, 2004) allow distant collaborators to know when others are interruptible, in order to engage in the quick consultations and informal chat that are the glue of co-located interaction.

Organizational Issues

As the most organizationally ambitious project type, these collaboratories experience all previously mentioned issues with a few additional concerns. They must gain and maintain participation among diverse contributors, work to standardize protocols over distance, facilitate distributed decision-making, and provide long-distance administrative support. Distributed research centers also must settle questions of cross-institutional intellectual property (IP). Universities have gotten more proactive about protecting in-house IP, and getting them to agree to multi-site sharing agreements necessary for open collaboration often proves challenging. Both the Alliance for Cellular Signaling and the Center for Innovative Learning Technologies spent much up-front time negotiating IP policies with partner institutions.

Distributed Research Centers also must think about the career issues of younger participants. What does it mean for young scholars to be lower authors on one or two very large, potentially important papers, rather than first authors on a set of smaller works? Is it a good career decision for them to get involved in projects where they will spend considerable amounts of their time in managerial tasks and meetings

rather than individual data analysis and writing? These are very real tradeoffs that should be addressed explicitly for junior researchers and graduate students involved in distributed research centers.

Community Infrastructure Project

Definition

Community Infrastructure Projects seek to develop infrastructure to further work in a particular domain. By infrastructure we mean common resources that facilitate science, such as software tools, standardized protocols, new types of scientific instruments, and educational methods. Community Infrastructure Projects are often interdisciplinary, bringing together domain scientists from multiple specialties, private sector contractors, funding officers, and computer scientists.

Example

The GriPhyN (Grid Physics Network) is a team of experimental physicists and information technology (IT) researchers planning to implement the first Petabyte-scale computational environments for data-intensive science. GriPhyN will deploy computational environments called Petascale Virtual Data Grids (PVDGs) to meet the data-intensive computational needs of the diverse community of international scientists involved in the related research. The term "Petascale" in the name emphasizes the massive CPU resources (Petaflops) and the enormous datasets (Petabytes) that must be harnessed, while "virtual" refers to the many required data products that may not be physically stored but exist only as specifications for how they may be derived from other data.

GriPhyN was funded through the National Science Foundation as a large Information Technology Research (ITR) project. The group is focused on the creation of a number of tools for managing "virtual data." This approach to dealing with data acknowledges that all data except for "raw" data need exist only as a specification for how they can be derived. Strategies for reproducing or regenerating data on the grid are key areas of research for the virtual data community. The key deliverable of the GriPhyN project is the Chimera Virtual Data System, a software package for managing virtual data.

The collaboratory team is composed of seven IT research groups and members of four NSF-funded frontier physics experiments: LIGO, the Sloan Digital Sky Survey, and the CMS and ATLAS experiments at the Large Hadron Collider at CERN. GriPhyN will oversee the development of a set of production Data Grids, which will allow scientists to extract small signals from enormous backgrounds via computationally demanding analyses of datasets that will grow from the 100 Terabyte to the 100 Petabyte scale over the next decade. The computing and storage resources required will be distributed for both technical and strategic reasons and across national centers, regional centers, university computing centers, and individual desktops.

Technology Issues

As with other collaboratories, Infrastructure Projects often necessitate development of new field standards for data and data collection protocols. Current Infrastructure Projects like GriPhyN are also tackling the problem of managing very large datasets. Associated issues also arise in data provenance, which is keeping track of the editing and transformations that have occurred on datasets.

Organizational Issues

A critical issue for interdisciplinary projects is negotiation of goals among disciplinary partners. Whose research agenda will be paramount? In partnerships between disciplinary experts and computer scientists there is often conflict between pursuing the most technologically advanced solutions (which are of research interest to the computer scientists) and more immediately practical solutions (Weedman, 1998).

Infrastructure Projects sometimes must decide between having academic managers and private sector management. The AIP Phase III study (AIP, 1999) compared these and found tradeoffs; private sector managers were better at finishing projects on time and on budget, while academic managers were better at accommodating idiosyncratic needs of researchers.

A third common issue is how work on Infrastructure Projects should fit into the careers of younger scientists who participate in them. Should building infrastructure "count" as a contribution to the discipline in the same way as other publishable works? If not, should junior faculty and younger scholars avoid working on such projects?

Conclusions

Sample Limitation

Despite precautions taken, the SOC database has some limitations that could not be corrected during the time frame of the SOC project. One area of missing projects is military-funded collaborations. Although the military has a strong interest in long-distance collaboration, there was not sufficient information gathered to be able to enter any of them into the database. Informants were difficult to find, and those located could not provide the information requested. This may have been affected by the timing of the project: The years after the 9/11 terrorist attacks were marked by strong concerns about security and strict control of information about military projects and procedures.

Another known area of missing data is international projects. Attention was focused primarily on U.S. projects and concentrated on U.S. funders as informants. This was partly due to limitations of language (data collection relied on phone interviews) and was partly a practical decision regarding allocation of resources. However, European Union projects, particularly Framework 7 projects that mandate assembly of broad international teams, would be excellent candidates for future study.

Key Dimensions: Resources and Activities

Other categorization schemes have used a-priori dimensions based on technology, scientific disciplines, or consideration of theoretical issues. This system was intended to be a more "bottom-up" exercise, working from a large dataset and letting the relevant categories emerge with time and understanding. Having done this, it is useful now to go back and examine the categories again to ask what dimensions tend to differentiate the projects.

The two-dimensional classification shown in Table 1 seems to capture many of the important distinctions. Each collaboratory type is placed in one cell, based on its dominant type of resource and activity. The first dimension, along the x-axis, differentiates based on the type of resource to be shared. In the case of Shared Instrument and Community Infrastructure collaboratories, the resources are scientific tools or instruments, such as telescopes or laboratory equipment. Other categories are information and knowledge. Sharing of each of these types of resource requires different technologies, practices, and organizational structures. The second dimension, along the y-axis, is the type of activity to be performed. This distinction corresponds to the distinction often made in organizational studies between loosely-coupled and tightly-coupled work.

In general, the collaborations become more difficult to manage and sustain from the top left of this table to the bottom right. It is generally more difficult to share knowledge than data or tools, and it is generally more difficult to co-create than to aggregate.

This dimensional classification offers some insights. Over time, the field of collaboratories has been observed to move from the top left to the bottom right. The AIP studies and early collaboratory writings (1992, 1995, 1999; National Research Council, 1993) focused largely on tool sharing, with some of the greatest recent successes moving into data sharing. Some individual collaboratory efforts have also been observed to move along these dimensions in both directions. Recognizing that more effort is needed more in one direction than in the other may help manage and plan these projects.

Table 1 Collaboratory types by resource and activity

	Tools (instruments)	Information (data)	Knowledge (new findings)
Aggregating across distance (loose coupling, often asynchronously)	*Shared Instrument*	*Community Data System*	*Virtual Learning Community, Virtual Community of Practice*
Co-creating across distance (requires tighter coupling, often synchronously)	*Infrastructure*	*Open Community Contribution System*	*Distributed Research Center*

These dimensions also help to differentiate some of the types from each other. The distinction between a Community Data System and an Open Community Contribution System was murky even to the research team, but understanding the distinction between aggregating and co-creating helped guide classifications and provide insight into the most difficult aspects of these projects.

Use of Collaboratory Typology

The SOC Collaboratory taxonomy has proven useful in guiding both research and assessment within the SOC project. A question that arose early on in the project was, "What technology should be recommended for collaboratories?" However, the nature of the projects that were being generalized across was so diverse as to make the question specious. The technology needs of a Shared Instrument Collaboratory are very different from those of a Virtual Community of Practice, for example. Identification of types enables more focused practitioner advice to be provided. Understanding these types has also framed research questions, such as helping to narrow the scope of our study of contributor motivation, and helping understand how collaboratories change in purpose as they evolve over time. Our future plans include continuing to develop this understanding of types. In the near future, we will focus on identifying best practices for different types. Expansion of types also seems inevitable. Finally, differentiation of sub-types within the classification system is another potentially rich area for exploration.

References

Ackerman, M. (2002, June 14). Personal communication.

AIP Study of Multi-Institutional Collaborations. (1992). *Phase I: High-Energy Physics*. New York: American Institute of Physics. Retrieved August 4, 2006 from http://www.aip.org/history/pubslst.htm#collabs

AIP Study of Multi-Institutional Collaborations. (1995). *Phase II: Space Science and Geophysics*. College Park, MD: American Institute of Physics. Retrieved August 4, 2006 from http://www.aip.org/history/pubslst.htm#collabs

AIP Study of Multi-Institutional Collaborations. (1999). *Phase III: Ground-Based Astronomy, Materials Science, Heavy-Ion and Nuclear Physics, Medical Physics, and Computer-Mediated Collaborations*. College Park, MD: American Institute of Physics. Retrieved August 4, 2006 from http://www.aip.org/history/pubslst.htm#collabs

Alexander, C., Ishikawa, S., & Silverstein, M. (1977). *A Pattern Language*. New York: Oxford University Press.

Bafoutsou, G., & Mentzas, G. (2002). Review and functional classification of collaborative systems. *International Journal of Information Management*, **22**, 281–305.

Baxevanis, A. D. (2002). The molecular biology database collection: 2002 update. *Nucleic Acids Research*, **30**(1), 1–12.

Brophy, S. P. (2003). Constructing shareable learning materials in bioengineering education. *IEEE Engineering in Medicine and Biology Magazine*, **22**(4), 39–46.

Berman, H. M., Bourne, P. E., & Westbrook, J. (2004). The Protein Data Bank: A case study in management of community data. *Current Proteomics*, **1**, 49–57.
Chompalov, I., & Shrum, W. (1999). Institutional collaboration in science: A typology of technological practice. *Science, Technology, & Human Values*, **24**(3), 338–372.
Collins, H. M. (1998). The meaning of data: Open and closed evidential cultures in the search for gravitational waves. *American Journal of Sociology*, **104**(2), 293–338.
Cramton, C. (2001). The mutual knowledge problem and its consequences for dispersed collaboration. *Organization Science*, **12**, 346–371.
Crane, D. (1972). *Invisible Colleges*. Chicago: University of Chicago Press.
Connolly, T., Thorn, B. K., & Heminger, A. (1992). *Social Dilemmas: Theoretical Issues and Research Findings*. Oxford, England: Pergamon.
Cummings, J. N., & Kiesler, S. (2005). Collaborative research across disciplinary and organizational boundaries. *Social Studies of Science*, **35**(5), 703–722.
DeSanctis, G., & Gallupe, R. B. (1987). A foundation for the study of group decision support systems. *Management Science*, **23**(5), 589–609.
Grudin, J. (1994). Computer-supported cooperative work: History and focus. *IEEE Computer*, **27**(5), 19–26.
Gutwin, C., & Greenberg, S. (2004). The importance of awareness for team cognition in distributed collaboration. In E. Salas & S. M. Fiore (Eds.), *Team Cognition: Understanding the Factors that Drive Process and Performance*. (pp. 177–201). Washington: APA Press. Retrieved May 4, 2005 from http://grouplab.cpsc.ucalgary.ca/papers/
Hesse, B. W., Sproull, L. S., Kiesler, S. B., & Walsh, J. P. (1993). Returns to science: Computer networks in oceanography. *Communications of the ACM*, **36**(8), 90–101.
Hollingshead, A. B. (1998). Retrieval processes in transactive memory systems. *Journal of Personality and Social Psychology*, **74**(3), 659–671.
Jarvenpaa, S., & Leidner, D. (1999). Communication and trust in global virtual teams. *Organization Science*, **10**, 791–815.
Kibrick, R., Conrad A., & Perala, A. (1998). Through the far looking glass: Collaborative remote observing with the W.M. Keck Observatory. *Interactions*, **5**(3), 32–39.
Kuhn, T. S. (1963). *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.
Latour, B., & Woolgar, S. (1979). *Laboratory Life: The Social Construction of Scientific Facts*. Beverly Hills, CA: Sage Publications.
Lawson, C., & Lorenz, E. (1999). Collective learning, tacit knowledge, and regional innovative capacity. *Regional Studies*, **33**(4), 305–317.
Myers, J. D., Chappell, A. R., & Elder, M. (2003). Re-integrating the research record. Computing in Science & Engineering, May/June, 44–50.
National Research Council (U.S.). (1993). *National Collaboratories: Applying Information Technology for Scientific Research*. Washington, D.C: National Academy Press.
National Science Foundation. (2004). *Science and Engineering Degrees: 1966-2001*. Division of Science Resources Statistics, NSF 04-311, Project Officers, Susan T. Hill and Jean M. Johnson. Arlington, VA. Retrieved June 12, 2005 from http://www.nsf.gov/statistics/nsf04311/htmstart.htm
Newman, M. (2001). The structure of scientific collaboration networks. *Proceedings of the National Academies of Sciences*, **98**(2), 404–409.
Olson, G. M., Zimmerman, A., & Bos, N. D. (forthcoming 2007). *Science on the Internet*. Cambridge, MA: MIT Press.

Olson, G. M., & Olson, J. S. (2000). Distance matters. *Human Computer Interaction*, **15**, 139–179.
Patton, M. Q. (1990). *Qualitative Evaluation and Research Methods* (2nd ed.). Newbury Park, CA: Sage Publications.
Science of Collaboratories (SOC) research group. (2001). *Social underpinnings workshop report*. Retrieved January 26, 2006 from http://www.scienceofcollaboratories.org/Workshops/WorkshopJune42001/index.php
Stokols, D., Fuqua, J., Gress, J., Harvey, R., Phillips, K., Baezconde-Garbanati, L, et al. (2003). Evaluating transdisciplinary science. *Nicotine and Tobacco Research*, **5**(Suppl. 1), S21–39.
Stokols, D., Harvey, R., Gress, J., Fuqua, J., & Phillips, K. (2005). In vivo studies of transdisciplinary scientific collaboration: Lessons learned and implications for active living research. *American Journal of Preventive Medicine*, **28**(Suppl. 2), 202–213.
Stork, D. G. (1999). Character and document research in the Open Mind Initiative. *Proceedings of the Fifth International Conference on Document Analysis and Recognition*. Washington, D.C.: IEEE Computer Press, 1–12.
Szulanski, G. (1992). *Sticky Knowledge: Barriers to Knowing in the Firm*. London: Sage Publications.
Viégas, F. B., Wattenberg, M., & Dave, K. (2004). Studying cooperation and conflict between authors with history flow visualizations. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 575–582). New York: ACM Press.
Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of "small world" networks. *Nature*, **393**, 440–442.
Weedman, J. (1998). The structure of incentive: Design and client roles in application-oriented research. *Science, Technology, & Human Values*, **23**(3), 315–345.
Wegner, E., & Lave, J. (1998). *Communities of Practice: Learning, Meaning, and Identity*. New York: Cambridge University Press.
Wulf, W. A. (1989). The national collaboratory—A white paper. In J. Lederberg & K. Uncaphar (Eds.), *Towards a National Collaboratory: Report of an Invitational Workshop at the Rockefeller University, March 17-18, 1989* (Appendix A). Washington, D.C.: National Science Foundation, Directorate for Computer and Information Science Engineering.
Wulf, W. A. (1993). The collaboratory opportunity. *Science*, **261** (5123), 854–855.

About the Authors

Nathan Bos is a Senior Staff Researcher in Cognitive Engineering at the Johns Hopkins University Applied Physics Laboratory. His research is in computer-supported cooperative work and long-distance scientific collaborations. He has published recently on partially-distributed collaborations and use of multi-player simulations for learning.
Address: 11100 Johns Hopkins Rd., Laurel, MD 20723 USA

Ann Zimmerman is a research fellow in the School of Information at the University of Michigan. Her research interests include the design, use, and impact of cyberinfrastructure; the sharing and reuse of scientific data; and the effects of large-scale collaborations on science policy and research management.
Address: 1075 Beal Ave., Ann Arbor, MI 48109-2112 USA

Judith Olson is Professor in the School of Information, Ross School of Business and the Psychology Department, and Associate Dean for Academic Affairs at the School of Information, all at the University of Michigan. Her research interests are in the area of distance work, doing fieldwork, laboratory experiments, and agent based modeling, in science, engineering, non-profits, and corporations. She is trying to uncover the pitfalls of distance work and design either new social practices or technologies to overcome them.
Address: 550 E. University, Ann Arbor, MI 48104 USA

Jude Yew is a doctoral student in the School of Information at the University of Michigan. His research interests include cognition and learning through the use of technology. In particular he is interested in how social software can aid in the formation of group knowledge.
Address: School of Information North, 1075 Beal Ave., Ann Arbor, MI 48109-2112 USA

Jason Yerkie is a strategic research consultant with the Sales Executive Council at the Corporate Executive Board. His research interests are the organizational and financial performance impacts of business process re-engineering initiatives and decision support systems in Global 1000 companies.
Address: 2000 Pennsylvania Ave. NW, Suite 6000, Washington, D.C. 20006 USA

Erik Dahl is an interaction designer and anthropologist in the Human Sciences Group at MAYA Design, Inc. His research interests are information-centric computing architectures and interface design; the social impacts of new media and pervasive computing on work practices; semiotic meaning creation within fields of technological distantiation; emergent innovation within complex systems; and the practical application of social and information theory to product design.
Address: SouthSide Works, Building 2, Suite 300, 2730 Sidney St., Pittsburgh, PA 15203 USA

Gary M. Olson is Paul M. Fitts Professor of Human-Computer Interaction at the School of Information at the University of Michigan. His research investigates the socio-technical factors involved in geographically-distributed science and engineering.
Address: School of Information, University of Michigan, 1075 Beal Ave., Ann Arbor, MI 48109-2112 USA