Designing Cyberinfrastructure to Support Science

Charlotte P. Lee
Donald Bren School of ICS
University of California, Irvine
cplee@ics.uci.edu

Matthew Bietz
mbietz@umich.edu

David Ribes
School of Information
University of Michigan
dribes@umich.edu

ABSTRACT

The aim of this workshop is to discuss current and future directions for CSCW research pertaining to cyberinfrastructure (CI). CSCW and CI intersect in their aims to support collaboration within heterogeneous groups and across physical distribution. Furthermore the development of CI–or large-scale informational resources–is itself a form of collaborative work worthy of CSCW research. We will address four themes in this workshop: designing for emerging groups; designing for the long-term; designing for interoperability; and the role of CSCW practitioners in CI.

INTRODUCTION

Recent years have seen the rise of new forms of large-scale distributed scientific enterprises supported primarily through advanced information infrastructures. These advanced infrastructures are called "cyberinfrastructure," although terms such as *grid computing*, *collaboratories*, and *eScience* are also commonly used. These projects have attracted significant investment from major funding agencies, substantial participation from domain scientists, and are of considerable interest to researchers whose practice focuses on issues in CSCW. Large-scale investments in cyberinfrastructure have largely taken the form of partnerships between domain scientists and information technologists, who jointly create a new form of infrastructure for conducting science. Cyberinfrastructure projects to date are largely developmental efforts. Cyberinfrastructure technologies are still emerging; they cannot be plucked off the shelf, but must be crafted and developed in situ.

The growth of large-scale Cyberinfrastructure projects reflect trends related to scientific collaboration including: 1) the rise of "big science" which is generally traced in particular to post-WWII efforts in physics and a movement towards large scale enterprises that physicists undertake; and 2) the increasing emphasis on interdisciplinarity which is related to the growth of big science but is also strongly associated with changes in funding for science and the sites and contexts of knowledge production. The gap between disciplines in modern day science is perceived as a natural place for new disciplines to evolve: "The real-world research problems that scientists address rarely arise within orderly disciplinary categories, and neither do their solutions. Thus, the information needed to solve complex research problems is distributed across disciplines and takes many different forms, physically and intellectually" [7]. Cyberinfrastructure development is thought of as requiring interdisciplinary collaboration, particularly between technologists and domain scientists (e.g. physicists, hydrologists, biologists, etc.), but CI is also often meant to stimulate a new discipline entirely. The interdisciplinarity and novelty of the science itself creates a challenge for designers of cyberinfrastructure, namely, figuring out who will be using the system and for what. As most CI projects remain at the stage of development, the first people using the systems tend to be the people who themselves are designing the system. Development projects typically include expert domain scientists who both advise and serve as alpha users [2,5].

Latour notes that science has two faces: established science and science in action [4]. Established science is what scholars teach to their students about what is known and is exemplified by what is written in undergraduate textbooks. Science in action however, is forward facing; it is research that is exploratory, ongoing, or novel. Cyberinfrastructure is an example of science in action. The most ambitious CI projects are attempting to create new scientific fields entirely and not only anticipate, but are actively cultivating new scientific communities and new practices [2].

In studying the development of infrastructure we must pay attention to the full range of participant's activities as they go about their work. Building infrastructure is much more than a computer science research venture, it is also a matter of reaching out to the institutions of science, organizing the work of heterogeneous experts and enacting stable and accessible technologies [9]. Institutionalizing is the work of generating sustainable goods and services linked to social or collective purposes with a connotation of permanence. Organizing work is concerned with the internal organizational arrangements that serve to motivate

Computer Supported Cooperative Work Conference Workshop, San Diego, 2008

participants and produce outcomes consistent with current developmental goals. Enacting technology is the work that is necessary to shift from experimental technologies to a functioning, stable infrastructure available for everyday use. At each of these 'scales' (institutionalizing, organizing work, and enacting technology) participants have concerns about i) aligning the end-goals of the diverse experts who participate in CI development; ii) designing tools that will support actual scientific research rather than utopian technologies that sit unused on shelves; and ii) motivating the sustained contributions of participants. We seek to address these cyberinfrastructure development challenges in this workshop.

WORKSHOP THEMES

For this workshop we have identified four key themes at the intersection of CSCW and CI:

- Designing for Emerging Groups
- Designing for the Long-Term
- Designing for Interoperability
- The Role of CSCW practitioners in CI

These themes are not discrete, but rather interlock. We will treat each in turn.

Designing for Emerging Groups

Understanding user requirements is particularly difficult when creating systems for future science. Scientists have a hard time articulating their needs when they don't necessarily know what could be possible [11]. Furthermore in trying to build a system for the long term, user needs, and even notions of the "the user" or "the community" are a moving target. We believe that CSCW researchers and practitioners are uniquely positioned to address these challenges. CSCW has long wrestled with the difficulty of both understanding the practices and activities of cooperative work and designing to support the same.

CSCW researchers recognize that the concept of "the user" is problematic [1]. Notions of "the user" too easily gloss over differences among people who differ in their experiences, abilities, roles, and interests. Furthermore, these characteristics can be context dependent: a user skilled in one context may be a novice in another. The *user* has been described as a "boundary label" to delineate between developers and non-developers [6]. Similar tensions arise in CI projects where the notions of *community* and *discipline* serve as a boundary labels between groups of actors.

In order to design for emerging groups we need to dig beneath the surface of such terms to uncover collaborative structures. Those of us who research cooperative work must be constantly mindful of the ways in which imposed dichotomies—e.g. users vs. developers, domain scientists vs. computer scientists—can obscure the complexities of how infrastructures are built. Creating productive interactions relies on the slipperiness of categories such as "users," and "scientific community" [2]. Over the long term, we must expect that categories and groups will emerge, decline, and shift [5].

This creates two challenges for CI designers. First, even with the theoretical issues surrounding the concept of "users," developers must still ask: for whom? How can we characterize a "scientific community" in a way that recognizes its amorphous and shifting nature while still providing structure for development efforts?

The second related challenge is to develop CI systems that can change and evolve over time. As science progresses, people, organizations, communities, and technologies will change. Collaboration tools are needed that interoperate while supporting work activities within larger and dynamic contexts [10].

Designing for the Long-Term

Infrastructures are meant to last not just for years, but for decades or longer. Data should be meaningfully preserved for as long as it remains useful, and past work should not be lost at each upgrade or update of software. Designing systems that can continue to function and remain relevant is a challenge. We do not yet know how to design long-term stable systems which can be tailored to the shifting needs of users over time.

We often think of computing resources as being in continuous flux, but increasingly it is clear we must design for system sustainability. Tools should be modular and backwards compatible to facilitate future upgrades; data should be carefully archived with its metadata and contextual information. This is a contemporary research question in CSCW, computer science and archival science. But 'the long term' is more than a technical feat, we must also consider the human work of maintenance or repair, and the institutional aspects of securing stable funding. Today, most funding for cyberinfrastructure comes through project grants lasting for just a few years, yet the work of science evolves over time. For example, it is only recently that ecologists have scaled up from field research and begun asking global questions. Infrastructures must become institutions, and their technologies must be flexible enough to allow for long-term change.

Designing for Interoperability

Many CI projects are conceived to support the aggregation and reuse of tools and data. Data may be aggregated to provide greater analytic power or to answer "big" questions that no single investigator or data set could alone. Data may also have broad applicability across disciplines. For example, historical weather data may be used to predict future weather, but could also help understand human migration patterns or explore the evolutionary development of microorganisms. Similarly, computational tools may have broad applicability across scientific domains e.g., visualization tools can serve to render many kinds of data.

A tension arises, however, in that while infrastructures tend to emerge over long time periods and seek to create interoperability, they are also the products of local, short-term, planned decisions. A key challenge in infrastructure development is to make something broadly useful while allowing for local optimization. Systems that are narrowly tailored to the needs of a specific scientific community may not be useful to other scientists.

This problem should be approached from multiple directions. For example, standards, metadata schema, and ontologies must be developed to allow for effective data sharing and aggregation. Curation and archival work are required to ensure data integrity over time. As cyberinfrastructure grows and involve a wider variety of participants from different sectors or countries, legal and cultural issues become increasingly important.

Interoperability also requires attention to the underlying technical systems. Scientific databases, middleware, high-speed data networks, and other connecting components are key parts of CI development. But developers again face the challenge of designing for broad, amorphous, and shifting user communities.

Role of CSCW in CI Projects

How do we as CSCW practitioners and researchers design tools and systems to contribute to the development of infrastructure? A key contribution of CSCW to CI is that our community can act as "boundary workers" who bridge the divide between system building and social analysis [3].

Practitioners and researchers in the field of CSCW have been involved with CI projects for several years. As the number of CSCW researchers and practitioners in this area continue to grow, reflecting on our own work practices, interests, and experiences can help us to define the state of CI research in CSCW [8]. In order to prompt discussion we will discuss the following topics related to CI development:

- Our own professional interests
- Roles and responsibilities of CSCW practitioners
- Available methods for study and participation
- Comparative analysis of individual projects

Reflexive discussions will serve as a way to identify commonalities, potential collaborators, and to learn from the challenges faced and overcome by others. By understanding the state of CSCW CI research, we lay the groundwork for thinking strategically about future directions.

FORMAT AND OUTCOMES

The workshop will be one full day divided into segments. During the first segment, participants will each be given 5 minutes to introduce themselves and briefly sketch their previous work, current interests, and research or development methods used as they pertain to studies of cyberinfrastructure/eScience. During the second segment, participants will discuss themes sketched above. The workshop organizers will cull additional discussion points from the submitted positions papers for group discussion. During the third segment the group will be divided into 2 to 3 groups according to their particular interests. The groups will be asked to identify points of commonality, divergence, and common obstacles. During the fourth segment, each group will present up to 3 agenda items for discussion.

The workshop will provide a unique opportunity for sharing problems and results amongst a growing research community. The workshop will help to identify diversity within the research area, stimulate future collaborations, and contribute to the evolution of CSCW involvement in cyberinfrastructure and e-Science.

PARTICIPANTS

The workshop will be limited to 20 participants (not including the organizers). In addition to the workshop listing on the CSCW web site, the workshop organizers will announce the workshop to appropriate discussion lists and invite authors who have published on cyberinfrastructure at CSCW or related venues to submit position papers.

Potential participants will be asked to submit position statements of two to four pages that address one or more of the themes described above. Organizers will review the position statements to select participants. Notices of acceptance will be sent by email.

ORGANIZERS

Charlotte P. Lee is a Research Scientist in the Donald Bren School of Information and Computer Sciences at the University of California, Irvine. She has a B.A. in Sociology from the University of California, Berkeley, an M.A. in Sociology from San Jose State University and a Ph.D in Information Studies from the University of California, Los Angeles. Her professional experience is in system administration, database administration, interaction design, and project management. Dr. Lee's research is in the fields of Social Informatics, Computer Supported Cooperative Work (CSCW), Design Studies, and Science and Technology Studies. Her work focuses on empirically describing and theorizing the informational practices, artifacts, and collaborative structures of communities of practice working towards a shared goal: collaborative design. Her paper entitled the "Human Infrastructure of Cyberinfrastructure" was nominated for the Best Paper Award at the ACM's 2006 Conference on Computer Supported Cooperative Work. Dr. Lee is the principle investigator of an NSF-funded project studying collaboration in the development of cyberinfrastructure using the Community Cyberinfrastructure for Advanced Marine Microbial Ecology Research and Analysis (CAMERA) as a case study.

Matthew Bietz recently received his Ph.D. from the School of Information at the University of Michigan. His research focuses on the interpersonal and relational dimensions of scientific collaboration. His dissertation was an experimental study of the delivery, interpretation, and use of interpersonal critical feedback in electronically mediated communication environments. He has also been involved with several e-Science projects, including the Waterford Project, the International AIDS Research Project and the Science of Collaboratories Project. He is currently working with Dr. Lee studying the CAMERA cyberinfrastructure. He received Bachelors degrees in English and Cello Performance from Lawrence University, a MA in Historical Musicology from SUNY Stony Brook, and a MS in Information Science from the University of Michigan.

David Ribes is a post-doctoral research investigator at University of Michigan, School of Information and a member of the Science Studies Program. His research interests include the organization of large-scale infrastructure development, the practice of science policy and the consequences of informational transformations in scientific research. His training is in Sociology with a specialization in Science and Technology Studies (STS). His primary methods are ethnographic. He has worked with multiple CI and observatory projects, including GEON, LEAD, WATERS and LTER. David completed his Ph.D. at the University of California San Diego (UCSD) in Sociology and Science & Technology Studies (STS). His dissertation was a three year ethnographic study of the development of the GEON CI. His Masters is from McGill University and his Bachelors is from York University.

REFERENCES

1. Bannon, L. J. From human factors to human actors: The role of psychology and human-computer interaction studies in system design In *Design at Work: Cooperative Design of Computer Systems*. J. Greenbaum & M. Kyng, Eds. L. Erlbaum,Hillsdale, NJ, 1991, 25-44.
2. Bietz, M. J. (2008). *Interactivity and Electronic Communication: An Experimental Study of Mediated Feedback.* Unpublished Dissertation, University of Michigan, Ann Arbor.
3. Edwards, P. N., Jackson, S. J., Bowker, G. C., & Knobel, C. P. (2007). *Understanding infrastructure: Dynamics, tensions, and design*. http://www.si.umich.edu/~pne/PDF/ui.pdf.
4. Latour, B. *Science in Action*. Harvard University Press, Cambridge, MA, 1987.
5. Lee, C. P., Dourish, P., & Mark, G. The human infrastructure of cyberinfrastructure In *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work*. ACM,New York, 2006, 483 - 492.
6. Mackay, H., Carne, C., Beynon-Davies, P., & Tudhope, D. Reconfiguring the user: Using Rapid Application Development. *Social Studies of Science, 30*, 5 (2000), 737-757.
7. Palmer, C. L. *Work at the Boundaries of Science: Information and the Interdisciplinary Research Process*. Kluwer, Boston, 2001.
8. Ribes, D., & Baker, K. S. Modes of Social Science Engagement in Community Infrastructure Design In *Proceedings of Third International Conference on Communities and Technology*. Steinfield & B. T. Pentland & M. Ackerman & N. Contractor, Eds. Springer,London, 2007, 107-130.
9. Ribes, D., & Finholt, T. A. Tensions across the scales: Planning infrastructure for the long-term In *Proceedings of the 2007 International ACM Conference on Supporting Group Work*. ACM,New York, 2007, 229-238.
10. Voss, A., Proctor, R., Poschen, M., Rodden, T., Olson, G. M., Slack, R., Hartswood, M., Jirotka, M., Carusi, A., & Budweb, S. (2007). *Realising and supporting collaboration in e-research [Workshop Proposal]*. ECSCW2007. http://www.e-researchcommunity.org/docs/ECSCW2007WorkshopProposal.pdf.
11. Zimmerman, A., & Finholt, T. A. Growing an infrastructure: The role of gateway organizations in cultivating new communities of users In *Proceedings of the 2007 Conference on Supporting Group Work*. ACM,New York, 2007, 239-248.