Two-Layer Structures in Scientific Collaboratories

Position Paper for Workshop on The Changing Dynamics of Scientific Collaboration at CSCW 2010

Andrea H. Tapia
College of Information Sciences and Technology
Penn State University
atapia@ist.psu.edu

Rosalie Ocker
College of Information Sciences and Technology
Penn State University
rocker@ist.psu.edu

Mary Beth Rosson
College of Information Sciences and Technology
Penn State University
mrosson@ist.psu.edu

Bridget Blodgett
College of Information Sciences and Technology
Penn State University
bward@ist.psu.edu

Timothy Ryan
Center for Quantitative Imaging
Anthropology Department
College of Information Sciences and Technology
Penn State University, tmr21@ist.psu.edu

ABSTRACT

We report preliminary results from a socio-technical analysis of scientific collaboration situated in physical anthropology research. We analyze the two-layer structure of the collaboration: one loosely coupled through shared access and use of scientific equipment, and one tightly coupled through shared creative development of research questions, data analysis and interpretation. We conclude with implications for both process and technology support.

Author Keywords

Scientific collaboration, collaboratories, virtual organizations

ACM Classification Keywords

H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

INTRODUCTION

Virtual organizations in science and engineering often take the form of *collaboratories* that encompass collaboration across geographic locations [1, 13]. As a socio-technical system, a collaboratory consists of technology (e.g. scientific instruments and software; communication and coordination software for remote interaction) and social-organizational practices (e.g., values, norms, and procedures). In the past decade, scientific collaboratories have emerged as an important context for socio-technical analysis and support (for summaries, see [4, 6, 7]).

In the research reported here, we describe a scientific collaboratory that has emerged around a rare piece of scientific equipment, a High Resolution Computed Tomography (HRCT) scanner. The collaborators include physical anthropology researchers around the world, and university staff who manage the scanner. While this collaboratory is an example of scientific collaboration structure "glued together" by a central critical resource [1], its goals are more complex than simply coordinating access to a tool.

The complexity of the HRCT collaboratory arises from its two-layer structure that is constantly changing shape. In one layer, it is a loosely-coupled organization with *persistent support* for managing the scanning tasks and resulting data; in the other it offers *project-specific support* that requires tightly-coupled communication and coordination for a cohort of distributed projects that emerge, coalesce, and follow their own trajectories. Within the project layer, projects develop their own organizational sub-structures and project-specific relationships to the scanner and results (e.g., type and timing of scanning; research protocols that must be followed and documented; meta-data and so on).

In this paper we report a preliminary analysis of the HRCT collaboratory, using case study data. We show how a two-layer analytic structure helps to identify the obstacles and possible enhancements to this complex virtual organization.

COLLABORATORY TYPOLOGIES

In the past decade, several researchers have offered classifications of scientific collaboratories [1,3,10]; however none of these has addressed the multi-layered structure present for most of these operations and the issues that arise from this structure. For example, Shrum, Genruth and Chompalov (2007) classified 53 scientific collaborations as bureaucratic, leaderless, non-specialized, or participatory [10]. Even though organizational structure, hierarchy, leadership, and formalization are key criteria in their categorization, these works do not discuss multi-layered structures.

Beginning in 2002 the "Science of Collaboratories" alliance (SOC) has been studying the nature and characteristics of scientific collaborations. One outcome is a categorization schema that offers distinctions among distributed research centers [2], but this too fails to consider the impact of multi-layer structures. Shared instrument, data sharing, and virtual community collaboratories emphasize loosely-coupled sharing and aggregation activities that can occur asynchronously across distance [2]. However, a distributed research center is a more complex structure because it aims to support co-creation of ideas, investigations, and research products across distant locations, including reliance at times on synchronous interactions [2].

STUDY CONTEXT: SCANNER AND HOMINID PROJECT

The HRCT is hosted by the Center for Quantitative Imaging, an NSF-funded research facility that is part of the anthropology research infrastructure at a major U.S. university. It is a hub for scientists working on advanced imaging

technologies. The HRCT presents challenges to researchers in need of the high-resolution images it can provide. Even with the increasing availability of powerful computers, datasets often overwhelm computational resources.

At the scanning layer (hence Scan-layer), there is a regular flow of requests for services. To respond, the staff must understand the project requirements of each request (size, timing, deadlines), specimen-related concerns (ownership, lending policy, special handling), and data needs (format, media, delivery, checking). Thus any one request can involve considerable negotiation, clarification and follow-up.

One project that relies centrally on scanning is the Hominid project – a primary source of observations analyzed here. Hominid integrates studies of primate morphology and paleontology with gene mapping in baboons and mice. Skulls of baboons, mice and fossilized hominids are scanned and studied across academic fields and institutions. This project consists of a team of senior and junior members located at three research sites in the U.S.

At the project-specific layer (hence Project-layer), the Hominid activities are tightly coupled among research team members. After a project vision is initiated, the team plans and implements a set of inter-related activities that leverage expertise and resources at each site. In this case, the data and expertise relevant to baboons and mice are located at different sites; the HRCT equipment and expertise is at the Center. Coordination involves prioritization, HRCT scheduling, and transport of specimens and resulting datasets; co-creation and interpretation of the data analysis; and shared development and publication of scientific papers.

DATA COLLECTION

Our case study relies on two forms of data – interviews and textual documentation from project records. We conducted 13 interviews of the key Hominid stakeholders, including the PIs from the three research sites, postdoctoral fellows, and graduate students. Interviews lasted 30-75 minutes; each was audio and/or videotaped and transcribed. The texts were original project documents pertaining to the Hominid NSF proposal, the intellectual property agreements, emailed correspondence and HRCT documentation. The transcribed interviews and textual data records were analyzed using *analytic induction*, a mixture of deductive and inductive approaches [5].

FLEXIBILITY AT THE SCAN-LAYER

The Scan-layer is loosely coupled. Loosely coupled organizations are known for a lack of coordination, and an absence of regulations that promote organizational flexibility, adaptation and sensitivity to the environment [12]. The Scan-layer is persistent, through time, while projects come and go. The Scan-layer operates somewhat independently of its parent university. The changes it experiences have minimal impact on the university, so it can respond and adapt to project needs with a sensitivity and flexibility that it could not have if its activities were more tightly controlled by the university. This loose coupling also enables researchers and staff at this level to interact in parallel with multiple research projects. Importantly though, this loosely coupled operation has promoted the development of procedures and policies that are idiosyncratic, ad hoc, and just-in-time to fulfill its operating needs. It often operates without formal agreements or standard operating policies and procedures.

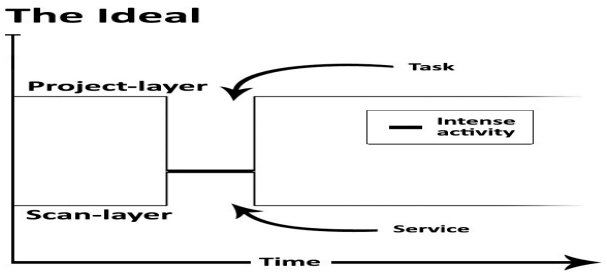

Figure 1: Idealized interaction between the two layers.

Figure 1 depicts an idealized view of the two-layer collaboratory: the Scan-layer supports one project at a time; its services exactly match the needs of the project; there is no overlap of project and Scan-layer staff. However, the virtual organization enabled by technology has led to more complex collaborative structures, the focus of the next section.

FINDINGS IN BRIEF

Our case study documents a range of problems for scientific collaboration emerging from the HRCT research context. In this brief paper we focus on the problems that arise through an interplay between the Scan-layer and the Project-layer.

Creating and Sharing Data Across Layers

The virtual character of the collaboratory has promoted many issues related to recording, storage and sharing of research data. For example, at the Scan-layer, very large files of raw scanner data are created, and file size is a significant remote collaboration barrier (for transfer, for local storage). Currently these files are stored in specialized storage arrays at the Scan-layer site. These files are protected by the university's stringent firewalls that are critical for intrusion protection, but that also reduce ease of access for external collaborators, even those who may "own" the data being produced. These large files are usually shared with Project-layer members via FTP or a custom DVD sent by post.

We have also observed problems for storage (and subsequent sharing) of data that is *not* scanned output. Despite an electronic format for scans and their metadata, the *process* data for scans (e.g., timing, resolution) is often stored in paper notebooks or even staff members' memories. This happens even though the staff recognize the limitations of paper records for search, reliability, access, and sharing.

At the Project-layer, researchers create "landmark data" (spatial coordinates from the raw data) for each scan. Once a scan is retrieved from the HRCT database, it is converted it to a viewable image and landmarked. The images and the

coordinates of the landmarks are stored in a separate site-hosted database that is proprietary to the institution and inaccessible to external collaborators. This local database is created and maintained by an IT support person who supports the entire Hominid project. None of the distributed PIs or other team members can access this local database. Thus the support person spends considerable time finding ways to share landmark information across institutions. Often the simplest solution is again to send a DVD by postal mail.

Electronic format issues also create problems across the two layers. Different data may be stored in different formats at different locations, or on different devices. The databases used to store data use formats that are proprietary to each PI institution; they are not linked in any systematic way. The scans are stored in one place; images built from the scans are stored in another place; landmark coordinates are stored in yet another place; the metadata that integrate these data via subject identifiers are stored in yet another location. In general, project team members do not know where these different datasets are stored or how to access them. Thus data creation, storage and management can be time consuming, increasing the costs of collaboration for all parties.

Establishing Ownership and Control Across Layers

Issues of data, process, task and object ownership were another general consequence of the two-layer structure. The Project-layer is producing large amounts of data across three institutions. However the ownership, use and sharing of these data was never addressed in formal project agreements. As a result the PIs and the project staff have a sense of unease; they are wrestling with conflicting ideas about the mandate to share, and are confused as to what and how data can or should be shared with others.

This ownership tension has affected activities at both layers. While the HRCT is integral to the process of creating research data, this dependency rarely plays a role in debates and concerns over ownership. Instead a variety of stakeholders who have a role in the HRCT-related projects hold claim to ownership of the specimens and the data produced from them. For example, when a project creates scans of sculls owned by a museum or institute, the owning organization often believes that any data produced from their specimens are owned by them. At the same time, the university hosting the HTCT claims ownership of scan data produced via its equipment. Finally, the project researchers who request the scans and produce the landmark data also make claims to ownership of the data and findings.

At both the Scan-layer and the Project-layer, the scientists receive regular requests for access to scanned data. Data ownership is not called into question when specimen, scanner, data and institution are all one. However, virtual collaboratory research is multi-site and multi-institution, and this places ownership at the Scan-layer of many debates. The uncertain ownership of data leads to an uncertain procedure for granting permission for outside scholars to use the data. Management of data sharing with external researchers has emerged as a key issue for the project team.

A NEW VIEW OF THE TWO-LAYER ORGANIZATION

The needs and opportunities of virtual collaboratories have created new roles for the Scan-layer and its staff. The Scan-layer takes on a new role as a persistent research partner in a larger scientific effort. Not only does the Scan-layer create scanned data for its users, but also it often must store and manage parts of the scientific data in ways that are specific to individual projects; it must address issues of access, sharing and other protocols. These new requirements are particularly salient when the project is a multi-year, multi-institution endeavor [3,9]. Although the Scan-layer was designed for flexibility and responsiveness, when its collaborative engagements are remote and extended, more formalized structures may be required for success [8, 11].

The Hominid project involves collaboration across three large institutions, as well as several individual researchers at other locations. The scientists must interact and depend on each other on a daily or weekly basis to make progress on their analyses and dissemination goals. To support its underlying data collection, storage, access and ownership concerns, the project may need a more formalized structure for its operations, procedures and protocols. Interestingly, Hominid is turning to the more loosely coupled Scan-layer for guidance in formulating its research procedures and protocols. This raises an opportunity for Scan-layer collaborators to develop processes that work not only for Hominid but also for other projects now and in the future.

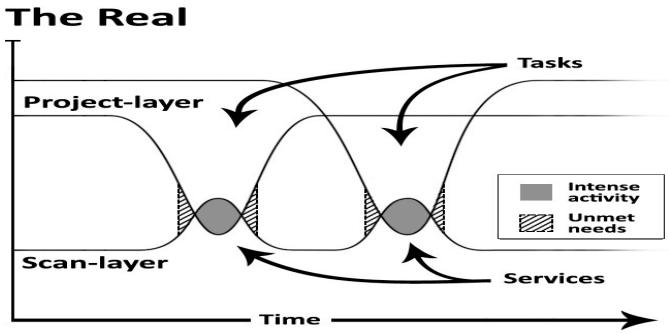

Figure 2: The actual interaction between the two layers.

Figure 2 is a more realistic view of the interaction we have observed between the Scan-layer and the Project-layer, including our view of how it might generalize to other HRCT-related projects (we are collecting data from other projects but currently have only analyzed Hominid). Each project is independent until it requires Scan-layer services. At that time it links with the Scan-layer to obtain HRCT services. However, also at that time each project begins to draw the Scan-layer into its activities, enmeshing its tasks and needs with the Scan-layer (e.g., data creation, storage, security, retrieval, access). Neither the Scan-layer nor the projects can function without the "glue-like" people and technology used to integrate the two layers. In the present

virtual organization, scientific staff at the Scan-layer serve as PIs on Hominid, thus taking a more substantive and active in role research activities. They have evolved from a role of technician to scientist, and expectations among all stakeholders must change to match.

We expect that interactions across the layers will change over time as a function of project needs. The Hominid project may intensively scan hundreds of specimens but then do no more scanning for several months. In the past the Scan-layer's interaction with Hominid would end with the completion of the last scan. However, the virtual organization enables the two-layer interactions (and corresponding dependencies) to become more constant. Once the scanning is done, a repository for the scan data is constructed, and data access must be managed. Because of technical expertise at the Scan-layer, a project may ask the for help with computational analysis of images and measurements of the raw data collected, prolonging the interaction and higher activity levels. When the Scan-layer provides scanning services to several projects at once, implicit resource dependencies may be created among projects with entirely different scientific goals.

It is this persistent "tangling" of the Project-layer and the Scan-layer that is enabled by the virtual collaboration in this setting. An important general consequence is the requirement for Scan-layer personnel to find and provide appropriate tools and procedures for handling multiple remote collaborations in ways that are smooth and reliable but also flexible enough to meet project-specific needs.

SOCIO-TECHNICAL DESIGN IMPLICATIONS

Our case study has pointed to three high-level concerns: physical management of multi-faceted scan data; data ownership; and dynamic cross-layer dependencies. One implication is to support specification (in advance) of the types of data that will be created through HRCT, the access and security each requires, and where such storage will be supported. A front-end tool that provides a transparent view of the project data thus created and stored would help remote participants to track where and the different data needed by their research activities can be found and accessed.

Ownership can also be addressed through an up-front planning process, analogous to the human subjects planning that takes place in social science research. While it is certainly possible that some data or results may be emergent and require ad hoc arrangements, it seems likely that a relatively simple online form can be used to formalize routine agreements concerning original specimens, raw scan data, secondary results (e.g., images with landmarks), meta-data from the scans and so on. The Scan-layer can play a guiding role in this, perhaps by requiring such agreements as part of any participation in a new research project.

With respect to the time-based changes in cross-layer dependencies, a more general awareness tool may be useful. For example, the Scan-layer can provide an abstract view of concurrent projects and their requests (i.e., without revealing confidential data such as personnel or research goals). Simply knowing how many and what sorts of other projects are underway might ease the interaction across the layers.

CONCLUSION

In this paper we have reported preliminary results from a socio-technical analysis of scientific collaboration. Through our analysis of interview and document data, we analyzed a two-layer structure of the collaboration, one that is loosely coupled at the Scan-layer and tightly coupled at the Project-layer. We have shown how virtual collaboratories of this sort face cross-layer challenges that raise requirements for process and technology support to increase the transparency and coordination within the complex and dynamic structure.

ACKNOWLEDGMENTS

This work is partially supported by a grant from the US National Science Foundation (0838400).

REFERENCES

1. Bos, N., et al., *From Shared Databases to Communities of Practice: A Taxonomy of Collaboratories*, in *Scientific Collaboration on the Internet*, G. Olson, A. Zimmerman, and N. Bos, Editors. 2008, The MIT Press: Cambridge, Massachusetts. p. 53-72.
2. Bos, N., et al., *From shared databases to communities of practice: A taxonomy of collaboratories.* Journal of Computer-Mediated Communication, 2007. **12**(2), article 16.
3. Chompalov, I., J. Genuth, and W. Shrum, *The organization of scientific collaborations.* Research Policy, 2002. **31**(5).
4. Cummings, J. and S. Kiesler, *Collaborative Research Across Disciplinary and Organizational Boundaries* Social Studies of Science, 2005. **35**(5): p. 703-722.
5. Epstein, L. and A.D. Martin, *Coding Variables*, in *Encyclopedia of Social Measurement*, K. Kempf-Leonard, Editor. 2004, Academic Press.
6. Finholt, T.A., *Collaboratories as a new form of scientific organization.* Economics of Innovation and New Technology, 2003. **23**(1).
7. Kouzes, R.T., *Electronic collaboration in environmental and physical sciences research.*, in *Electronic Collaboration in Science*, M.F.H. S. H. Koslow, Editor. 2000, Lawrence Erlbaum: Mahwah, NJ:. p. 89-112.
8. Maglaughlin, K.L. and D.H. Sonnenwald, *Factors that impact interdisciplinary scientific research collaboration: Focus on the natural sciences in academia*. 2005, University College of Bor√•s. Swedish School of Library and Information Science.
9. Olson, G., *The next generation of science collaboratories*, in *International Symposium on Collaborative Technologies and Systems*. 2009: Baltimore, MD, USA. p. xv-xvi.
10. Shrum, W., J. Genuth, and I. Chompalov, *Structures of Scientific Collaboration*. 2007, Cambridge, Massachusetts: The MIT Press.
11. Sonnenwald, D.H., *Scientific Collaboration.* Annual Review of Information Sciences and Technology, 2007(41): p. 643-681.
12. Weick, K.E., *Educational organizations as loosely coupled systems.* Administrative Science Quarterly, 1976. **21**(1-19).
13. Wulf, W., *The National Collaboratory*, in *Towards a National Collaboratory*. 1989, Rockefeller University: NY.