Data Curation in Scientific Teams: An Exploratory Study of Condensed Matter Physics at a National Science Lab

Charles C. Hinnant
College of Communication and Information, Florida State University, PO Box 3062100, Tallahassee, FL 32306-2100
chinnant@fsu.edu

Besiki Stvilia
College of Communication and Information, Florida State University, PO Box 3062100, Tallahassee, FL 32306-2100
bstvilia@fsu.edu

Shuheng Wu
College of Communication and Information, Florida State University, PO Box 3062100, Tallahassee, FL 32306-2100
sw09f@my.fsu.edu

Adam Worrall
College of Communication and Information, Florida State University, PO Box 3062100, Tallahassee, FL 32306-2100
apw06@my.fsu.edu

Kathleen Burnett
College of Communication and Information, Florida State University, PO Box 3062100, Tallahassee, FL 32306-2100
kburnett@fsu.edu

Gary Burnett
College of Communication and Information, Florida State University, PO Box 3062100, Tallahassee, FL 32306-2100
gburnett@fsu.edu

Michelle M. Kazmer
College of Communication and Information, Florida State University, PO Box 3062100, Tallahassee, FL 32306-2100
mkazmer@fsu.edu

Paul F. Marty
College of Communication and Information, Florida State University, PO Box 3062100, Tallahassee, FL 32306-2100
marty@fsu.edu

ABSTRACT
The advent of big science has brought a dramatic increase in the amount of data generated as part of scientific investigation. The ability to capture and prepare such data for reuse has brought about an increased interest in data curation practices within scientific fields and venues such as national laboratories. This study employs semi-structured interviews with key scientists at the National High Magnetic Field Laboratory to explore data management, curation, and sharing practices within a condensed matter physics community. Findings indicate that condensed matter physics is a highly varied field. The field's work practices and reward structures may impede the development and implementation of highly formalized curation policies focused on sharing data within the broader community. This study is an extension of a larger mixed-methods study to examine the life-cycles of virtual teams and will serve as a foundation for a larger survey of the lab's user community.

Categories and Subject Descriptors
H.3.0 **[Information Storage and Retrieval]**: General

General Terms
Management, Documentation, Experimentation, Human Factors.

iConference 2012, February 7–10, 2012, Toronto, Ontario, Canada.
ACM 978-1-4503-0782-6/12/02

Keywords
Information management, data curation, scientific collaboration, information quality, information sharing

1. INTRODUCTION
Research processes are increasingly data driven, and there is a growing need to share, reuse, and aggregate data before, during and after discrete experiments are conducted. Access to original data and the record of its provenance is also necessary to replicate and validate the findings of scientific experiments, leading to an increased need for preserving and maintaining scientific data in an actionable, "usable" state for ongoing research, education, reporting, certification and verification. The curation of scientific data is therefore a topic of increased interest and concern [2, 7, 13, 14].

This study employs the theory of information worlds, which seeks to describe the intertwined processes of information exchange and social interaction in a wide variety of social and professional settings [8], to explore the data curation and sharing practices of scientific teams at the National High Magnetic Field Laboratory (NHMFL) in Tallahassee, FL. This poster presents the initial results from a qualitative pilot study of key researchers within the lab's Condensed Matter Physics (CMP) research community.

2. BACKGROUND
The NHMFL is the largest and most powerful magnet laboratory in the world. Over 900 scientists a year use its magnets to run a variety of experiments, applying diverse knowledge in physics, chemistry, biology, engineering, and other related fields [1]. The

NHMFL provides a unique environment with which to examine the technical and social factors that influence the data curation practices of a scientific subfield such as CMP.

Data curation is generally defined as "the activity of managing and promoting the use of data from its point of creation, to ensure it is fit for contemporary purpose, and available for discovery and re-use" [12]. Within any scientific field, the socio-technical nature of the scientific work itself, coupled with the norms of the specific collaborative team and the broader professional norms of the community, all influence data curation practices. However, scientific work processes, including activities, data, instruments, and culture are often hidden to outsiders and obtain an identity through final products, such as publications [11]. Different scientific fields also have different collaboration and reward structures [10] that may influence data curation practices. Scientists may not have economic incentives to share and document data for others to use. Even if data is non-competitive (e.g., already published), additional costs may be associated with sharing data that may not bring tangible benefits to the scientist or scientific team. These costs may include time spent on generating additional documentation and metadata, as well as the time spent on, and the cost of, infrastructure needed for distributing data, which can be significant as the scale of data, and the demand for it, grow [3, 4, 9]. Other types of motivations for data curation and, ultimately, sharing include expectations of co-authorship, reciprocity or an expectation of reciprocity, and sharing expensive instrumentation [3, 4, 5, 16].

3. METHOD

This study is an extension of a larger mixed methods study of the collaboration of scientific teams at the NHMFL. The original study employed direct observations, semi-structured interviews, citation analysis, and social network analysis [6, 15]. This study used in-depth semi-structured interviews of key informants to better explore the data curation practices of collaboration teams within the CMP community at the NHMFL. Five key scientists at the NHMFL were identified to be interviewed by our project staff regarding the nature of scientific collaborations, as well as data management practices. Three experimentalists and two theorists were interviewed for this portion of the project. Participants also played different roles within the lab's broader scientific community. Two participants were full-time staff scientists, while two were academic faculty affiliated with the NHMFL. Finally, an external scientist was interviewed in order to more fully understand the practices of external lab users. All interviews included questions regarding the lifecycle of scientific collaborations, scientific work practices, perceptions of data ownership, as well as the rules, community norms and policies regarding data collection, data preparation, data analysis, data archiving, and data sharing. The interviews also included questions regarding the criteria used by scientists within the community to assess data quality. The interviews were recorded, transcribed, and coded within NVIVO 9.0 in order to facilitate the identification of major themes.

4. FINDINGS

Interviews took place with key informants between June 2011 and August 2011 at the NHMFL facilities in Tallahassee, FL. Interview times ranged from 40 to 153 minutes, with the average duration 80 minutes. All subjects indicated that research collaborations could be facilitated in a variety of ways, such as prior research relationships, interest in examining a specific sample, extended relationships through graduate students or post-doctoral fellows, and a scientist's reputation within the scientific community. Since the NHMFL is a facility oriented toward external users, research staff are often assigned to assist with the experiments being carried out by external users using the lab's magnets.

The experimentalists interviewed indicated that scientific practices associated with CMP research directly impacted perceptions of data and data curation. The CMP experiments conducted with the lab's magnets generate data by using sensors that measure the effects of stimuli (i.e. magnet field) on the sample being studied, as well as other variables of interest. A great value is placed on both the skill associated with developing more accurate sensors, as well as the analytical techniques employed to reduce "noise" in the data that may inhibit seeing small experimental effects in the sample. One subject tied the ability to reduce noise to the perceived reputations of scientists: "*So pushing the envelope frequently means you are improving the techniques to see things that other people have not been able to see so far. … But then, what you are evaluating is… which tricks did they do? And frequently they…don't want to give the entire recipe, they know they have an edge, they want to keep an edge for a little while."*

The uniqueness of the experimental work, especially in the way experiments are implemented and the way the resulting data is analyzed, highlights the importance of the contextual data associated with the experimental protocol. Several subjects indicated that scientists maintained their own "notebook" that contained contextual information regarding the actual experimental protocol. When asked if this highly contextual information was ever shared with other scientists such as journal referees, one subject stated, *"No, no, documentation, no, no. The notebook, whatever, would not make sense…Look, it's a notebook…and you're… getting the raw data …nobody's going to do that."*

All subjects indicated that the ownership of scientific data generated at the NHMFL rested firmly with the principal investigator of the user team generating the data. This was not only viewed as a norm within the CMP community, but is also reinforced by strict policies at the lab regarding the protection of data and information related to users' experiments conducted at its facilities. *"But the data, how it's shared, is, is quite tricky… Depends a lot on the collaborations. One thing you're going to learn is that physicists are quite paranoid about not being scooped."*

Several subjects indicated that the CMP community was highly diverse and, as a result, different from other fields within physics. They also highlighted how the work orientations of CMP may impact the ability to develop more formalized curation policies, rules and labs at NHMFL. One subject commented on how the nature of the work would impact the ability to implement curation policies by saying, "*In physics, especially condensed matter physics, I think it is still dirtier than many other fields. …you do not have a standard instrumentation that gives sort of standard output finals or something, you know. So I think that people do things very different ways …."* The same subject also highlighted that CMP is still often characterized by relatively small teams or even a single-investigator.

5. DISCUSSION AND CONCLUSIONS

The qualitative focus of this study, coupled with our ongoing research at the NHMFL, allows for a detailed and nuanced understanding of the data curation and sharing norms at the

NHMFL and the broader CMP community. This study also lays the foundation for a more comprehensive survey of data curation practices among the NHMFL users and staff community. Coupled with our ongoing work into the lifecycles of virtual scientific teams, this work highlights the variance in team norms and the highly contextualized nature of the scientific work environment of the CMP community at NHMFL.

A key finding for data curation is that data practices appear to be highly specific to individual teams and, therefore, may vary widely across teams. This variety in data management and curation practices may weaken the potential for data sharing, reuse and repurposing. For many scientists, the perceived cost of data management and curation may exceed their perceived value of sharing, reusing and repurposing the data, even as they acknowledge the value of such practices writ large within the greater scientific community.

This research provided rich qualitative data regarding the curation practices of collaborative teams at the NHMFL. Furthermore, the information provided by this pilot study assists in the development and implementation of a survey of the NHMFL staff and users regarding data curation and sharing practices.

6. ACKNOWLEDGMENTS

The authors would like to express their gratitude to Larry Dennis, Dean of the College of Communication and Information, for suggesting the NHMFL as a research site and facilitating access to the scientists and staff. We would also like to express our appreciation to the scientists who participated in the study. This research was supported in part by the National Science Foundation (NSF) under Grant OCI-0942855. The article reflects the findings, and conclusions of the authors, and do not necessarily reflect the views of the NSF or the NHMFL.

7. REFERENCES

[1] About the National High Magnetic Field Lab. 2011. http://www.magnet.fsu.edu/about/overview.html.

[2] Atkins, D. E. et al. 2003. *Revolutionizing science and engineering through cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure*. National Science Foundation, Arlington, VA. http://www.nsf.gov/od/oci/reports/atkins.pdf.

[3] Borgman, C. L., Wallis, J. C. and Enyedy, N. 2007. Little science confronts the data deluge: Habitat ecology, embedded sensor networks, and digital libraries. *International Journal on Digital Libraries*. 7, 17-30. DOI= http://dx.doi.org/10.1007/s00799-007-0022-9.

[4] Cragin, M. H., Palmer, C. L., Carlson, J. R. and Witt, M. 2010. Data sharing, small science, and institutional repositories. *Philosophical Transactions of the Royal Society A*. 368, 1926 (Sep. 2010), 4023-4038. DOI= http://dx.doi.org/10.1098/rsta.2010.0165.

[5] David, P. A. 2004. *Towards a cyberinfrastructure for enhanced scientific collaboration: Providing its 'soft' foundations may be the hardest part* (SIEPR Discussion Paper No. 04-01). Stanford Institute for Economic Policy Research, Stanford, CA. http://129.3.20.41/eps/le/papers/0502/0502002.pdf.

[6] Hinnant, C.C. et al. 2011. *Author team diversity and the impact of scientific publications: Evidence from physics research at a national science lab.* Manuscript submitted for publication.

[7] Howe, D. et al. 2008. Big data: The future of biocuration. *Nature*. 455, 7209 (Sep. 2008), 47-50. DOI= http://dx.doi.org/10.1038/455047a.

[8] Jaeger, P.T. and Burnett, G. 2010. *Information worlds: Behavior, technology, and social context in the age of the Internet*. Routledge, New York, NY.

[9] Karasti, H., Baker, K.S. and Halkola, E. 2006. Enriching the notion of data curation in e-science: Data managing and information infrastructuring in the Long Term Ecological Research (LTER) Network. *Computer Supported Cooperative Work*. 15, 4, 321-358. DOI= http://dx.doi.org/ 10.1007/s10606-006-9023-2.

[10] Knorr-Cetina, K. 1999. *Epistemic cultures: How the sciences make knowledge*. Harvard University Press, Cambridge, MA.

[11] Latour, B. and Woolgar, S. 1979. *Laboratory life: The social construction of scientific facts*. Sage, Beverly Hills, CA.

[12] Lord, P. and Macdonald, A. 2003. *e-Science Curation Report: Data curation for e-Science in the UK: An audit to establish requirements for future curation and provision*. JISC. http://www.jisc.ac.uk/media/documents/programmes/preservation/e-sciencereportfinal.pdf.

[13] National Science Board. 2005. *Long-lived digital data collections: Enabling research and education in the 21st century* (NSB Report No. 05-40). National Science Foundation, Arlington, VA. http://www.nsf.gov/pubs/2005/nsb0540/nsb0540.pdf.

[14] National Science Foundation. 2007. *Cyberinfrastructure vision for 21st century discovery* (NSF Report No. 07-28). National Science Foundation, Arlington, VA. http://www.nsf.gov/pubs/2007/nsf0728/nsf0728.pdf.

[15] Stvilia, B. et al. 2011. Composition of scientific teams and publication productivity at a national science lab. *Journal of the American Society for Information Science and Technology*. 62, 2 (Feb. 2011), 270-283. DOI= http://dx.doi.org/10.1002/asi.21464.

[16] Zimmerman, A. 2007. Not by metadata alone: The use of diverse forms of knowledge to locate data for reuse. *International Journal on Digital Libraries*. 7, 1-2 (May. 2007), 5-16. DOI= http://dx.doi.org/10.1007/s00799-007-0015-8.