# Studying the Data Practices of a Scientific Community

Besiki Stvilia
bstvilia@fsu.edu

Charles C. Hinnant
chinnant@fsu.edu

Shuheng Wu
sw09f@my.fsu.edu

Adam Worrall
apw06@my.fsu.edu

Dong Joon Lee
dl10e@my.fsu.edu

Kathleen Burnett
kburnett@fsu.edu

Gary Burnett
gburnett@fsu.edu

Michelle M. Kazmer
mkazmer@fsu.edu

Paul F. Marty
marty@fsu.edu

College of Communication and Information, Florida State University, PO Box 3062100, Tallahassee, FL 32306

## ABSTRACT
To be effective and at the same time sustainable, a community data curation model has to be aligned with the community's current work organization: practices and activities; divisions of labor; data and collaborative relationships; and the community's value structure, norms, and conventions for data, quality assessment, and data sharing. This poster discusses a framework for developing a community data curation model, using a case of the scientific community gathered around the National High Magnetic Field Laboratory, a large national lab. The poster also reports findings of preliminary research based on semi-structured interviews with a sample of the main stakeholder groups of the community.

## Categories and Subject Descriptors
H.1.0 [**Models and Principles**]: General

## General Terms
Documentation, Management, Measurement.

## Keywords
Data curation, data management, condensed matter physics, data quality, activity theory.

## 1. INTRODUCTION
Scientific communities have long established, culturally justified and sustainable models of sharing scholarly publications. As science becomes increasingly data driven, the need for building similar shared, sustainable, community models for data curation and for integrating them with publication models becomes of greater interest and concern for funding agencies, scientific institutions and communities [1].

Data curation is "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" [2]. Many general models of research data and related curation activities, processes, architecture components, and risks have been proposed in the literature (e.g., DCC Curation Lifecycle Model, OAIS, DRAMBORA). However, data curation work is context specific. General models and tools are valuable knowledge sources to plan data curation activities, but they do not define context specific models or incorporate perceptions and value structures for data, metadata, and data quality problems. These general models do not address the social aspects of scientific work that may enable or alternatively hinder data curation activities. Furthermore, these models and toolkits provide little guidance and insight into data practices at the team level and how those practices interact with the data curation norms, policies, and infrastructure at the organization and community levels.


*JCDL'13,* July 22–26, 2013, Indianapolis, Indiana, USA.
ACM 978-1-4503-2077-1/13/07.

## 2. FRAMEWORK
To study data practices at the community level one needs a theoretical framework, which can not only provide high-level conceptualizations of different data intensive activities of the community, but also mechanisms for integration, learning, and harmonization of the community's data practice conceptualizations by different stakeholder groups. Activity Theory [3] and an information quality assessment framework and value based quality assessment model developed by one of the authors in earlier research [4,5] were selected to guide this research. According to Activity Theory, context can be viewed as an interplay between general cultural and community structures (language, norms, conventions, social networks, and relationships) and the structure of a particular activity or an activity system (goal oriented actions, tools, roles, rules, strategies, etc.). This activity theoretic framework helps conceptualize the typified activities of creation and use of particular data types: their structures, values, the types of quality problems that each data type may be prone to, and the criticality of these problems to the activity's success or failure. This conceptualization of the activity system can be used to guide an empirical analysis of data objects and other documents (e.g. lab notebooks), development of interview protocols and survey questionnaires, and iterative participatory design of metadata and knowledge organization tools and templates.

The framework was developed to study data work of a condensed matter physics (CMP) community gathered around the National High Magnetic Field Laboratory (NHMFL). The NHMFL is a unique interdisciplinary scientific center, one of the largest of its kind, collaboratively operated by Florida State University, the University of Florida, and Los Alamos National Laboratory. It provides scientists with free access for research involving magnetic fields, superconducting magnetometry, magnetic resonance imaging, and magnetic spectroscopy. In preliminary research, the activity theoretic framework was used to bootstrap the interview protocol and coding schema for semi-structured interviews. Twelve interviews were conducted. The interviewees

represented different stakeholder groups and roles at the NHMFL, including sample material growers, experimentalists, theorists, visiting scientists, local scientists, administrators, senior scientists, junior scientists, postdoctoral researchers, and students. The preliminary research was limited to exploring the following fundamental questions: *(1) What are the data-related activities present in the community? (2) What is data quality in the community? (3) What is the value of data quality and how it can be evaluated?*

# 3. FINDINGS OF PRELIMINARY RESEARCH

## 3.1 Data, Activities

CMP scientists study the properties, including the structure and state dynamics, of condensed matter. At the NHMFL, scientists measure and interpret the effects and dynamics of interaction of different stimuli, such as magnetic fields, on matter. Research processes at the NHMFL consist of multiple activities and are usually performed by small teams of scientists, often with complementary skills and knowledge, who play different roles. In addition to general project activities such as experimentation, computation, and recordkeeping, scientists also spend significant time in planning and preparing for a research project (e.g., writing research proposals, building instruments) and building and maintaining their social networks in the community.

The literature and our preliminary work suggests that, depending on their specialization and research task, scientists even within the same discipline may have different perceptions of what constitutes data. A scientist growing a new material may consider the chemical formula or an actual physical sample of the material as data. For another scientist measuring the properties of the same material, data could be readings of the instruments or sensors attached to that sample. To a theorist, additional types of data could be obtained from simulations and/or analytical calculations. Finally, to a reader or reviewer of a manuscript submitted to a scholarly journal, data could be the graphs and analytical calculations included in the manuscript. This study did not differentiate among the different types of data (e.g., physical samples, measurement data, metadata, publications) and referred all of them as data. Figure 1 presents an initial, preliminary model of the data curation activities of the CMP community associated with the NHMFL.

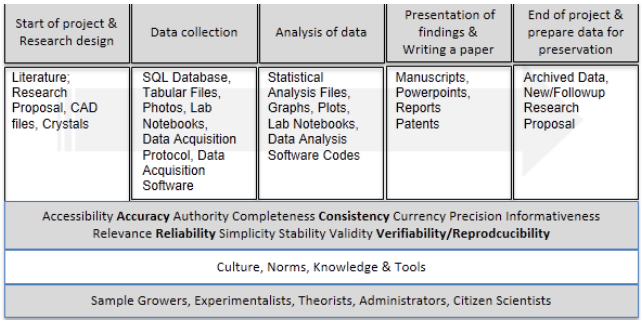


**Figure 1. Model of data curation activities of the CMP community**

## 3.2 Quality

Quality is usually defined as "fitness for use." Data quality problems may arise in any of the activities constituting the complex process of CMP research and scholarly communication. These activities may include manufacturing material samples, designing an experiment, manufacturing instruments and parts for the experiment, measuring and/or simulating the characteristics of the sample under different treatments and conditions, interpreting the results of the measurements, theorizing possible characteristics or relationships, and communicating findings to the community. Perception of quality may also change indirectly due to changes in the community's culture, knowledge and research technology. The interview data showed that scientists particularly cared about the *reproducibility*, *accuracy*, and *consistency* of data (see Fig. 1).

## 3.3 Value

The study found instances of all five categories of data quality value change from [4] (see Table 2).

**Table 2. Measures of data quality value**

| Measures | Explanation |
|---|---|
| A function of the activity success or failure | Success or failure of an activity |
| A function of the cost and rework | Cost of equipment use and the time spent by scientists to generate and/or massage data |
| A function of the amount of use | Amount of data use, or the use of derived data products such as publications |
| A function of the activity cost | Change in the cost of an activity in which data is used |
| A combination of the above factors | Some combination of the above factors |

# 4. CONCLUSIONS

This poster reported on a conceptual framework for studying the data work of a large scientific community. Future research will collect additional data and develop a more detailed community level data curation model through participatory design involving all major stakeholder groups within and having impact on the CMP community.

# 5. REFERENCES


[1] Atkins, D. E., Droegemeier, K. K., Feldman, S. I., Garcia-Monlina, H., Klein, M. L., Messerchmitt, D. G., Messina, P., & Wright, M. H. (2003). *Revolutionizing science and engineering through cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure.* Arlington, VA: NSF.

[2] Lord, P., & 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*. Bristol, UK: The JISC.

[3] Engeström, Y. (1990). Learning, working and imagining: Twelve studies in activity theory. Helsinki: Orienta-Konsultit Oy.

[4] Stvilia, B., & Gasser, L. (2008). Value based metadata quality assessment. *Library & Information Science Research, 30*, 67-74.

[5] Stvilia, B., Gasser, L., Twidale, M. B., & Smith, L. C. (2007). A framework for information quality assessment. *Journal of the American Society for Information Science and Technology, 58*, 1720-1733.