This is a preprint of an article published in *JASIST* (Stvilia, B., Hinnant, C., Wu, S., Worrall, A., Lee, D. J., Burnett, K., Burnett, G., Kazmer, M. M., & Marty, P. F. (2015). Research project tasks, data, and perceptions of data quality in a condensed matter physics community. *Journal of the Association for Information Science and Technology*, *66*(2), 246-263. http://onlinelibrary.wiley.com/doi/10.1002/asi.23177/abstract)

Research Project Tasks, Data, and Perceptions of Data Quality in a Condensed Matter Physics Community

Besiki Stvilia, Charles C. Hinnant, Shuheng Wu, Adam Worrall, Dong Joon Lee, Kathleen Burnett, Gary Burnett, Michelle M. Kazmer, and Paul F. Marty

College of Communication and Information, Florida State University, Tallahassee, FL, 32306-2100, United States

E-mail: {bstvilia, chinnant}@fsu.edu, {sw09f, apw06, dl10e}@my.fsu.edu, {kburnett, gburnett, mkazmer, marty}@fsu.edu

Abstract

To be effective and at the same time sustainable, a community data curation model needs to be aligned with the community's current data practices, including research project activities, data types, and perceptions of data quality. Based on a survey of members of the Condensed Matter Physics (CMP) community gathered around the National High Magnetic Field Laboratory, a large national laboratory, this paper defines a model of CMP research project tasks consisting of ten task constructs. In addition, the study develops a model of data quality perceptions by CMP scientists consisting of four data quality constructs. The paper also discusses relationships among the data quality perceptions, project roles, and demographic characteristics of CMP scientists. The findings of the study can inform the design of a CMP data curation model that is aligned and harmonized with the community's research work structure and data practices.

1. Introduction

Scientific communities have long established, culturally justified and sustainable models for curation and quality evaluation of scholarly publications. As science becomes increasingly data driven (Kell & Oliver, 2004), the need for building similar shared, sustainable community models for data curation and for integrating them with publication models is of significant interest and concern for funding agencies, scientific institutions, and research communities (Atkins et al., 2003).

To manage or curate data, one has to have knowledge of what constitutes data in a particular domain and for a particular process, as well as the characteristics of data, such as type, format, scale, ownership, quality, and provenance. Although Condensed Matter Physics (CMP) is a small team science (Stvilia et al., 2011), a typical research process in CMP is complex and distributed in time and space. A CMP research process may produce and use/reuse different kinds of data generated by different scientists with different specializations and from different labs and institutions around the world. Data types, metadata, and formats are ultimately defined by the activities that create and operate on the data and the tools used in those activities. To a CMP scientist designing and building instruments for an experiment, the computer aided

design (CAD) files of the instruments can be important data with high reuse potential. A scientist growing a new material may consider the chemical formula or an actual sample of the material as the most important data, while for another scientist measuring the properties of the same material, data could be readings of the instruments or sensors attached to that sample. The same research project may produce multiple papers published in different journals and pre-print archives. Without preserving the process metadata and support data (e.g., CAD files), replicating the research and linking and discovering related datasets and literature can be difficult and costly.

One of the main inhibitors of data curation and sharing could be concerns about data quality. Data owners may be concerned about the quality and potential misuse or misinterpretation of their data. Users, on the other hand, may not have sufficient resources or access to the processes that generated the data to evaluate their quality, and hence may mistrust and not use the data (Birnholtz & Bietz, 2003; Hinnant et al., 2012b). Data quality determines the quality of findings and decisions (Stvilia, Gasser, Twidale, & Smith, 2007). Long-term access to high quality data is essential if one is to make high quality decisions or to justify, validate, and evaluate existing decisions and results. Research data have their own lifecycles, where the same data or metadata can be used for different purposes and can have different levels of importance in different activities and at different times (Greenberg, 2001; Stvilia & Gasser, 2008a). Data, metadata, and data quality assurance needs to be analyzed at the process level to enable reproducible research (Mesirov, 2010; Peng, 2011).

There have been many information and data quality frameworks proposed in the literature (see Ge and Helfert, 2007, for a review), creating valuable knowledge for guiding data quality assurance practices and research. At the same time there is still a significant need for studying data curation, including data quality assurance work of different communities, to identify context specific structures of data quality problems, priorities of quality, and different sociotechnical aspects that may affect the community's data curation work. This would help design a data quality assurance model well-aligned with the community's data culture, needs, and priorities. A context-specific knowledge base of data quality assurance processes, tools, and skills needed can be used for effective infrastructure-support planning, training, and cross-contextual quality-based data selection and integration (Stvilia, Al-Faraj, & Yi, 2009). These knowledge bases—sociotechnical repertoires—can be used to assemble an effective cyber-infrastructure configuration to support research projects on demand (Foster, Jennings, & Kesselman, 2004). To be robust to changes, such as adding or dropping members or requirements, the repertoires need to be as complete as possible and aligned well with the real world community and team relationships and practices of work organization (Stvilia, Twidale, Smith, & Gasser, 2008).

This paper contributes towards the above objectives by examining project tasks, perceptions of and priorities for data quality, and data management practices of a CMP community gathered around the National High Magnetic Field Laboratory (NHMFL). Findings of the study can be used by the Lab and scientific funding agencies to provide the CMP community and other similar communities and labs with effective data curation infrastructure support.

2. Related Research

The most closely related work to this research is the survey of the data practices of a large interdisciplinary sample of scientists conducted by Tenopir and colleagues (2011). In addition to the questions related to research data collection, tools, storage and reuse, the researchers also investigated scientist's perception of different barriers that might hinder data sharing and/or reuse, as well as the relationships among the demographic characteristics of scientists and their perceptions and data practices. This study differs from their work by focusing on the data practices of a specific scientific community. In addition, this study defines the community specific models of research project tasks and perception of quality based on

the analysis of empirical data. Regardless of these differences, the work by Tenopir and colleagues provides an excellent context for discussing some of the findings of this study. Another closely related work to this research is the study of data practices of an interdisciplinary research center carried out by Borgman et al (2007). In particular, their study provides a good insight into the types of data and tools used by scientists from the areas of terrestrial ecology, marine biology, environmental contaminant transport, and seismology.

Several taxonomies and typologies of scientific activities have been proposed in the literature (e.g., Earth Observing System Data Panel, 1986; Qin, Ball, & Greenberg, 2012; Sufi & Mathews, 2004). Each phase or activity of a research project may produce useful and potentially reusable data. Data curation is the process of managing data for long term availability and reuse (Curry, Freitas, & O'Riáin, 2010; Lord & Macdonald, 2003). The data curation literature provides a number of models (e.g., the DCC Curation Lifecycle Model) of research data and related curation activities and processes (e.g., Burton & Treloar, 2009; Higgins, 2008). In addition, the data curation community has developed tools for identifying an organization's data assets; determining data curation tasks, architecture components, and risks; and devising appropriate policies and strategies for digital data archives and repositories. Characteristic examples of such tools include the Reference Model for an Open Archival Information System (OAIS), the Digital Repository Audit Method Based On Risk Assessment (DRAMBORA; DRAMBORA Consortium, 2008), the Data Audit Framework (Jones, Ross, & Ruusalepp, 2009), and the Trustworthy Repositories Audit and Certification (TRAC; Center for Research Libraries & Online Computer Library Center, 2007).

To promote reusability and sharing of data management infrastructure components, the library, preservation, and data management communities have started developing registries of data types and formats (e.g., PRONOM[1], Research Data Alliance[2]), metadata schemas (e.g., schema.org), and data curation templates (Data Curation Profiles[3]), as well as tools for validating file objects against file format specifications (e.g., JHOVE[4]).

Quality assurance is an essential part of data curation. Quality, in general, is defined as "fitness for use" (Juran, 1992). Consequently, data quality can be defined as the degree the data meets the needs and requirements of the activities in which it is used (Stvilia et al., 2007; Wang & Strong, 1996). Quality is dynamic and multidimensional, and the criticality of different quality problems is contextual. Data quality can be changed actively through direct modification of data objects, or indirectly through changes in the context of their interpretation and use (Stvilia et al., 2007; Stvilia & Gasser, 2008a). Hence, to aggregate or reuse data and data quality measurements from different contexts, there is a need for context-specific studies of data quality. A change in context may lead to not only changes in conceptual data quality measurement models, vocabularies, metrics, and measurement representations with regard to scale, precision, and formatting, but also changes in value structures, reference sources, and quality requirements (Stvilia et al., 2008). Data quality can be evaluated directly by examining intrinsic properties of data, or indirectly by evaluating the records of their provenance and use (Simmhan, Plale, & Gannon, 2005; Stvilia, 2006). The quality of data can be affected by the quality of any components of the data creation process and infrastructure, including the quality of data contributors and project teams, reference sources, and quality assurance tools (Stvilia et al., 2007). Several conceptual Information Quality (IQ) assessment models—both general and information type-specific—have been proposed in the IQ literature (e.g., Bruce & Hillman, 2004; Eppler, 2003; Fallis & Frické, 2002; Stvilia, 2007; Stvilia et al., 2007; Wang & Strong, 1996). For example, Stvilia et al. (2008) studied article creation and quality control processes in the English Wikipedia, while Sheppard and Terveen (2011) discussed data quality assurance in a citizen science community and the impact of data quality assurance work on science education. There is also a significant body of research on data quality issues in large-scale library digitization projects (e.g., Conway, 2010, 2011; Rieger, 2008). Arazy, Nov, Patterson, and Yeo (2011), Hinnant et al. (2012a), and Stvilia et al. (2011) discussed the relationships between project team composition, productivity, and information product quality. Nichols, Chan, Bainbridge, McKay, and Twidale (2008) and Stvilia (2008)

proposed architectures for information and data quality visualization and assessment tools. Open source tools have been developed to assess the validity of file objects (e.g., JHOVE) and enhance the quality of tabular data (e.g., OpenRefine[5]).

Conceptualizations of research data quality and studies of scientists' perceptions of and priorities for data quality and data quality assurance skills are also found in the literature (e.g., De Roure, 2010; Gutmann, Schürer, Donakowski, & Beedham, 2004; Huang, Stvilia, Jorgensen, & Bass, 2012). Wu, Stvilia, and Lee (2012) reviewed the sources of data quality problems and knowledge organization approaches and tools used for data quality control in molecular biology. The perception of what constitutes quality and useful data, or when the data become useful, may vary within the same process or discipline, as well as across different processes within disciplines (Ball, 2010; Earth Observing System Data Panel, 1986). Furthermore, scientists may rely on different properties and cues of data to assess their relevance, value, and reusability (Bechhofer et al., 2013; Faniel & Jacobsen, 2010).

Although significant research has focused on the data practices and data curation activities of individual scientific projects and collaborations, to the best of our knowledge there has been no systematic study of the data practices in CMP, the largest interdisciplinary community within physics (National Research Council [U.S.], 2007). To provide effective data management support for community level data curation and reuse, better understanding and knowledge is needed of the CMP community's existing data practices and relationships, including the community's typical project tasks and perceptions of and priorities for data quality.

3. Research Design

This research examines data practices of the CMP community gathered around the NHMFL. The NHMFL (2012) 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 to its facilities for research involving magnetic fields, superconducting magnetometry, magnetic resonance imaging, and magnetic spectroscopy.

To study data practices at the community level one needs a theoretical framework, which can provide not only high-level conceptualizations of different data-intensive activities of the community, but also mechanisms for integrating, learning, and harmonizing conceptualizations of the community's data practices by different stakeholder groups. The theoretical framework used for this research consists of activity theory (Engeström, 1990, 2001; Kuutti, 1995; Leont'ev, 1978) and an information quality assessment framework and a value-based quality assessment model developed by one of the authors in previous research (Stvilia et al., 2007; Stvilia & Gasser, 2008b). The framework provides general conceptualizations of activity structure (i.e., goal oriented actions, tools, roles, rules, strategies, division of labor, etc.), its community and cultural context (i.e., language, norms, conventions, social networks, and relationships), and the structure of activity-specific data quality problems and related quality criteria. These conceptualizations were used to guide the development of semi-structured interview protocols, a survey instrument, and coding schemas for data analysis.

This paper reports on the data quality assurance part of a comprehensive survey of the data practices of the CMP community. To better understand the community's data practices, issues, and problems and develop the survey instrument, the authors first conducted 12 semi-structured interviews with representatives of different groups of the community, including sample material growers, experimentalists, theorists, visiting scientists, local scientists, administrators, senior scientists, junior scientists, postdoctoral researchers, and students. The authors used concepts and

relationships from activity theory, the information quality assessment framework (Stvilia et al., 2007), and the literature to develop questions for the interview protocol. The audio recordings of the interviews were transcribed and content analyzed.

The study then used interview findings to expand and refine the set of interview questions and develop a survey instrument. The survey instrument was pretested with nine participants from the CMP community for readability and validity. The finalized survey was distributed online to 672 scientists in the fall of 2012 using Qualtrics survey software. The scientists were invited using their email addresses, which were obtained from the NHMFL's database of researchers who conducted experiments using the Lab's facilities between 2008-2011. Only scientists who indicated CMP as their discipline were selected. The survey consisted of 7 sections and 89 questions. Although participants completed early sections of the survey at higher rates, 160 participants completed all the questions, resulting in an overall response rate of 24%. This paper reports results of the data quality evaluation section of the survey which was completed by 172 participants (26% response rate).

Before participating in an interview or completing an online survey, participants were given a consent form approved by the Human Subjects Committee of Florida State University. The form contained information about the project, including information about potential risks associated with participation in the data collection. Participants who completed an interview or a survey were emailed a $50 Amazon gift card.

4. Research Questions

The paper examines the following research questions:

1. What are the typical activities of a CMP research project?
2. What are the types of data these activities produce and/or use?
3. What are the tools that the CMP scientists use to manage data?
4. What are the project roles that the CMP scientists play?
5. What are the types and sources of data quality problems in CMP?
6. What are the perceptions of data quality in CMP?

5. Findings

5.1 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. CMP research projects consist of multiple activities and are usually performed by small teams of scientists, often with complementary skills and knowledge, playing different roles. There have been several general models and typologies of scientific activities and project tasks identified in the literature (e.g., Ball, 2010; Levitin & Redman, 1993). According to Bailey (1994), there are two ways of categorization schema construction: the classic, deductive approach and the

inductive, data driven approach. The later involves empirical data collection, clustering and then assigning conceptual labels to the data clusters. To identify a more specific structure of the CMP community's research work, this study used the empirical, bottom-up approach to identify the types of research project tasks.

In particular, the survey asked participants for some of the typical tasks that they perform in a research project. 171 participants answered this open-ended question. Responses were tokenized, terms normalized to lemma forms, and synonyms were merged using Java codes developed by one of the authors and the Stanford CoreNLP[6] Java natural language processing libraries. The analysis produced 294 unique terms ordered by frequency. The 34 most frequent terms were selected for the next phase of the analysis, an application of factor analysis to identify underlying semantic relationships among the terms and cluster related terms into a fewer number of factors or task constructs. Factor analysis and Principal Component Analysis (PCA) are frequently used dimension reduction techniques (Duda et al., 2000; Hair, Black, Babin, Anderson, & Tatham, 2005).

Starting with a model of 34 variables ensured 5 cases for each variable, which is the minimum number of cases required by factor analysis (Hair, Black, Babin, Anderson, & Tatham, 2005). In addition, variables with a Measure of Sampling Adequacy (MSA) lower than 0.5 were removed from the model one by one until the MSA of all of the variables was higher than 0.5. The resulting model included 27 variables with an overall MSA of 0.635 and the MSA of each variable higher than 0.5. The Bartlet test of sphericity was significant at the 0.0001 level. PCA was used to extract factors. Factors with Eigenvalues above 1 were selected for inclusion in the factor model. These 10 factors captured 65% of the total variance. The PCA factor matrix was rotated using the Varimax rotation algorithm with Kaiser normalization. Based on the total number of cases (171), factor loadings of 0.45 and above were identified as significant (see Table 1).

Table 1. Factor loadings for task terms.

Dimension	Component									
	1	2	3	4	5	6	7	8	9	10
data	0.25	0.10	***0.83***	-0.11	-0.08	0.05	0.08	0.19	0.08	-0.03
analyze	0.32	0.24	***0.71***	-0.16	-0.06	0.00	0.19	0.08	0.16	-0.04
experiment	0.23	0.08	-0.03	-0.20	-0.01	-0.17	0.00	***0.74***	0.32	-0.01
material	-0.08	0.07	-0.05	***0.82***	-0.03	-0.01	0.14	-0.03	-0.03	0.02
write	***0.79***	0.15	0.11	-0.11	0.25	-0.11	0.06	0.08	0.09	-0.01
paper	***0.76***	0.12	0.20	-0.06	0.15	-0.04	0.01	-0.04	0.14	-0.03
design	0.32	0.10	-0.02	-0.06	0.11	0.42	-0.13	0.36	-0.11	-0.25
characterize	-0.09	-0.05	-0.08	***0.70***	-0.04	0.02	-0.09	-0.09	-0.05	-0.05
synthesize	0.01	-0.04	-0.04	***0.79***	-0.04	-0.09	0.03	-0.03	-0.01	0.07
acquire	-0.04	-0.10	***0.79***	0.00	0.09	-0.06	-0.13	0.01	-0.11	-0.08
device	0.00	-0.01	0.05	-0.06	-0.04	***0.87***	0.01	-0.06	0.06	-0.01
fabricate	-0.07	0.23	-0.06	-0.02	0.02	***0.82***	-0.04	-0.01	-0.06	0.03
interpret	0.00	-0.06	0.02	-0.07	0.06	-0.01	-0.02	0.01	***0.85***	-0.03
discuss	0.25	-0.06	0.15	0.10	-0.04	0.01	***0.69***	0.13	0.15	0.01
literature	0.03	***0.66***	0.12	-0.05	-0.03	0.07	-0.22	-0.05	0.18	0.04
property	-0.08	-0.09	-0.12	0.08	0.09	-0.05	-0.02	0.07	-0.10	***0.84***
study	0.28	0.33	0.00	-0.11	-0.17	0.07	-0.14	-0.09	0.13	***0.51***
manage	-0.03	***0.62***	-0.04	0.01	0.08	0.03	0.20	0.07	-0.08	-0.05
idea	0.05	0.14	-0.05	-0.01	***0.78***	0.08	0.04	-0.06	0.14	-0.04
coordinate	0.02	0.03	0.06	-0.10	***0.71***	-0.06	0.39	-0.05	0.04	0.17
supervise	0.30	-0.12	0.01	-0.05	***0.68***	-0.04	-0.26	0.18	-0.21	-0.10
setup	-0.06	-0.09	0.28	-0.03	-0.01	0.07	0.02	***0.68***	-0.19	0.06
edit	***0.45***	0.43	-0.10	0.03	0.14	0.13	0.16	0.04	-0.18	-0.06
report	***0.70***	-0.19	0.11	-0.05	-0.15	0.07	0.11	0.10	-0.15	0.09
team	-0.02	0.11	-0.08	-0.02	0.13	-0.07	***0.69***	-0.10	-0.16	-0.10
simulate	0.05	***0.60***	0.06	-0.09	-0.13	0.05	-0.15	-0.12	-0.13	0.02
develop	0.07	***0.73***	0.04	0.12	0.21	0.08	0.24	0.08	0.05	0.04

Note. Extraction method: PCA; rotation method: Varimax with Kaiser normalization.

Based on the variable loadings, factor 1 can be interpreted as *scholarly communication activities*; factor 2 as *simulation*; factor 3 as *data collection and analysis*; factor 4 as *sample material synthesis*; factor 5 as *administration and coordination*; factor 6 as *device building*; factor 7 as *team discussion*; factor 8 as *experiment setup*; factor 9 as *interpretation*; and factor 10 as *research objective* (see Table 2).

Table 2. CMP task constructs.

Factor	Task construct	Terms
1	**Scholarly communication**	write, paper, edit, report
2	**Simulation**	literature, manage, develop, simulate
3	**Data collection and analysis**	data, analyze, acquire
4	**Sample material synthesis**	material, characterize, synthesize
5	**Administration and coordination**	idea, coordinate, supervise
6	**Device building**	fabricate, device
7	**Team discussion**	discuss, team
8	**Experiment setup**	setup, experiment
9	**Interpretation**	interpret
10	**Research objective**	study, property

5.2 Data

Data is one of the main tools and products of scientific activities. To manage or curate data, one has to have knowledge of what constitutes data in a particular domain and for a particular process, as well as the characteristics of the data, such as type, format, scale, ownership, quality, and provenance. The Office of Management and Budget (OMB, 1999, section 36(d)(2)(i)) defines scientific data as "the recorded factual material commonly accepted in the scientific community as necessary to validate research findings," emphasizing the importance of validation tasks in scientific research. However, the lifecycle of a scientific research project comprises multiple tasks. As the literature suggests (e.g., Buckland, 1997; Redman, 1998), scientists even within the same discipline may have different perceptions of what constitutes data depending on their specializations or the particular research tasks they perform. A scientist growing a new material may consider the chemical formula or an actual physical sample of the material as data, while 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, data could be obtained from simulations and/or analytical calculations. To a reader or a reviewer of a manuscript submitted to a scholarly journal, data can be graphs and analytical calculations included in the manuscript. Furthermore, one person's metadata can be another person's data (Redman, 1998). The process of generating a successful sample of a material documented in a sample grower's notebook can be metadata to the sample grower, but data to another scientist who wants to replicate the sample.

When asked in the survey what types of data they create or use, 91% of participants indicated that they produced or used raw data generated by instruments or simulation programs. The next most frequently selected data type was text documents (77%), followed by slides (65%) and laboratory notes (63%). Other frequently identified data types included spreadsheets, software codes, drawings, and statistical analysis data files (see Table 3). Participants specified other types of data that included software specific data analysis and visualization files, such as the Mathematica, MatLab, OriginPro, and LabView file types.

Table 3. Types of data created or used.

Data Type	Responses	%
Raw data generated by simulation programs, sensors or instruments	168	91%
Text documents (e.g. Word, PDF, LaTeX, TXT)	141	77%
Slides (e.g., PowerPoint)	120	65%
Laboratory notes	115	63%
Spreadsheets (e.g. Excel)	72	39%
Software codes	68	37%
Drawings (e.g., CAD)	57	31%
Statistical data files	34	18%
Website (e.g., project Website)	15	8%
Other	14	8%
Databases (e.g. Access, MySQL, Oracle)	9	5%

Data may not be very useful without "data about data": documentation or metadata. Metadata enables interpreting the data meaningfully, connecting to related data and knowledge, assessing their quality, and making them discoverable and reusable. When asked at which project phase documentation was created, participants most frequently selected the stages of data analysis, scholarly communication, and data collection (see Table 4). More than half of the participants indicated that they created documentation at the research design stage. Less than 30% mentioned creating documentation when preparing data for preservation. Only 10% indicated that they added documentation when depositing data in an institutional or subject data repository.

Table 4. When is documentation created?

Stage	Responses	%
Analyzing data	177	96%
Writing a paper	170	92%
Collecting data	163	89%
Presenting findings at a conference	151	82%
Publishing a paper in a peer-reviewed journal	142	77%
Publishing a paper in a pre-print archive (e.g., arXiv.org)	112	61%
Research design	95	52%
Preparing data for preservation	49	27%
Data management planning	40	22%
Depositing data in an institutional or subject data repository	18	10%
Other	3	2%
Never	0	0%

To enable and support effective and efficient metadata creation, it is important to know what tools the community uses to document data. When asked what tools they used to document and make data meaningful, 84% of the participants identified graphing or charting software, followed by presentation software such as PowerPoint (80%), paper-based lab notebooks (73%), and email (68%). Only 22% of the participants indicated that they used specialized electronic documentation software. In addition, only 4% of the participants indicated the use of a metadata schema to document data (see Table 5). Participants also mentioned using text "read me" files to describe the content of file folders.

Table 5. Tools used to create documentation.

Tool	Responses	%
Graphing or charting software (e.g. MATLAB, Excel)	155	84%
Presentation software (e.g. PowerPoint)	148	80%
Paper-based lab notebook	134	73%
Email	125	68%
Word processor (e.g., MS Word)	120	65%
Electronic lab notebook (e.g. OneNote, Labnotes)	40	22%
Other	16	9%
One or more metadata schema	7	4%
None	0	0%

A majority of the participants (60%) indicated that they did not use any specific standards or guidelines for documenting data. 22% of the participants indicated that they followed specific documentation guidelines; of this group, 61% of them named their team as requiring the use of guidelines, followed by the lab (46%) and the National Science Foundation (NSF; 24%). Only 34% of the participants indicated that their typical project had a data management plan. A majority of the participants felt that the metadata they created was of good (43%) or very good (36%) quality. Only 12% stated that the quality of their metadata was fair (11%) or poor (1%). 8% stated that their metadata was excellent.

In terms of reusing data, 47% of the participants indicated that they had used data generated outside of their projects. The most frequently mentioned type of reused data was text documents, followed by PowerPoint slides and raw data generated from sensors or simulation programs.

5.3 Participant Demographics and Project Roles

The overwhelming majority of the participants indicated that they would characterize their research as experimental (97%). Fewer participants indicated that they did computational (9%) or theoretical (9%) physics research. The majority of participants hold either academic or research positions (36% each). The next largest group was participants who hold both academic and research positions (14%). Only 1% of the participants indicated that they held an administrator position. Out of those who held academic positions, the largest group was graduate students (29%), followed by full professors (27%) and associate professors (22%). The largest group among those who held research positions was research scientists (43%), followed by student research assistants (26%), postdocs (19%), associate research scientists (8%), and assistant research scientists (4%).

On average participants worked on three collaborative projects per year and the average project team size was five. 48% indicated that they worked only on funded collaborative projects. 50% worked on both funded and unfunded collaborative projects each year. Only 2% worked on purely unfunded collaborative projects.

36% of the participants identified the Principal Investigator (PI) as their primary role in a typical funded research project. The second largest group was student research assistants (22%) followed by Co-PIs (16%) and postdoctoral researchers (15%). The Senior Investigator and Project Manager were the least frequently identified roles (7% and 2%). The role distribution for unfunded projects was more or less similar, with 33% of participants identifying themselves as project leads; however, no postdoctoral researchers or student research assistants identified a role for themselves on an unfunded project.

When asked who was ultimately responsible for managing research data in a typical funded project, 51% of the participants selected the PI. Other frequently selected choices included a student research assistant (14%) and that no one was responsible (10%; see Table 6). Participants used the other category (5%) to specify the following data management arrangements: "data generator", "everybody", and "PI has ultimate authority, but graduate students taking the data have immediate and primary responsibility".

For unfunded projects, in addition to the project lead (44%) and student research assistant (14%), participants also selected project researcher (17%) and postdoctoral researcher (12%). Similar to funded projects, 11% indicated that no one was responsible for managing the project's data. For funded projects only 1% indicated that IT staff of the department or data repository were responsible for managing project data, while for unfunded projects no participants indicated this was true.

Table 6. Who is ultimately responsible for managing data in a typical funded research project?

Answer	Response	%
Principal investigator (PI)	90	51%
Student research assistant	25	14%
No one (ultimate responsibility is not clearly defined)	17	10%
Postdoc	13	7%
Co-Principal investigator (Co-PI)	9	5%
Other	8	5%
Senior investigator	6	3%
Don't know	3	2%
Project manager	2	1%
IT staff in your department or institution	2	1%
Data repository or data archive	2	1%
Research technician	0	0%

5.4 Quality

5.4.1 Data Quality Problems

Quality is usually defined as "fitness for use" (Juran, 1992). Data quality can be defined as the degree the data meets the needs and requirements of the activities in which it is used (Stvilia et al., 2007; Wang & Strong, 1996;). The concept of data quality and dimensions of quality are usually perceived through experiencing data problem incidents. They also are learned through data quality training and apprenticeship into the community's data management culture (Stvilia et al., 2008). A data quality problem occurs when the existing data quality is lower than the level of quality needed on one or more data quality dimensions for a particular activity (Gertsbakh, 1977; Strong, Lee, & Wang, 1997).

Data quality problems may arise in any of the CMP project activities. These activities may include manufacturing material samples, designing an experiment, manufacturing instruments and parts for the experiment, measuring or simulating the characteristics of the sample under different treatments and conditions, interpreting the results of measurements, theorizing possible characteristics or relationships, and communicating findings to the community. However, not all data problems are data quality problems; some may be purely hardware related, such as insufficient data storage space. When asked about the data problems they had encountered, participants most frequently noted a lack of file naming conventions (52%), followed by difficulties in interpreting data due to poor or lost documentation (50%); a lack of version control (36%); and an inability to access data due to obsolescence, proprietary formats, expired software licenses, or other issues (35%; see Table 7).

Table 7. Data problems encountered.

Problem	**Responses**	**%**
Lack of file naming conventions	95	52%
Difficulty interpreting data due to poor or lost documentation	91	50%
Lack of version control	66	36%
Inability to access data due to obsolescence, proprietary formats, expired software license, and etc.	64	35%
Insufficient storage space	42	23%
None	25	14%
Problems establishing ownership of data	11	6%
Other	11	6%
Problems establishing provenance of data	9	5%

The most frequently reported sources of data quality problems were human errors (67%), impure sample materials (59%), imprecise instruments (51%), external environmental interferences in measurements (50%), errors in experiment design (44%), software errors (44%), and incomplete documentation (37%). Participants also reported changes in the context of data interpretation and purposeful reduction of data quality as problem sources, but at a much lower rate (see Table 8).

Table 8. Sources of data quality problems.

Source	Responses	%
Human error	115	67%
Impure sample material	102	59%
Imprecise instruments	88	51%
Interferences from external environment	86	50%
Error in experiment design	76	44%
Software error	73	42%
Incomplete documentation	63	37%
Changes in the context of data interpretation (e.g., over time data becomes obsolete as new knowledge or technology emerges; data is aggregated from multiple contexts)	33	19%
Purposeful reduction of data quality (e.g., including incomplete data in a publication to keep a competitive edge)	16	9%
Don't know	7	4%
Other	1	1%

5.4.2 Data Quality Perception

The structure and cost of data quality problems of the community determine the community's perceptions of and priorities for data quality—a community specific model for data quality (Huang et al., 2012; Stvilia et al., 2008). To help define the community's model for data quality, participants were asked to rate the importance of 14 quality dimensions on a 7-point Likert scale from extremely unimportant to extremely important. The authors used the information quality assessment framework (Stvilia et al., 2007), findings of the semi-structured interviews (Stvilia et al., 2013), and literature analysis to determine which 14 data quality dimensions to include in the survey (see Appendix, Table 16). 172 valid responses to the data quality perception question were obtained. The study used a factor analysis to determine the underlying structure of the community's perception of quality. The analysis treated each quality dimension as a variable. The MSA of each of the variables was higher than 0.8, with the Bartlet test of sphericity significant at the 0.0001 level.

The study used PCA to extract factors. A Scree plot suggested selecting the first four eigenvalues, which captured 69% of the total variance of the data. The component analysis factor matrix was rotated using the Varimax rotation algorithm with Kaiser normalization. Based on the total number of cases (172), factor loadings of 0.45 and above were identified as significant (see Table 9).

Table 9. Factor loadings for the data quality (DQ) criteria.

DQ Criteria	Component			
	1	2	3	4
Accessibility	0.24	0.16	0.17	***0.80***
Accuracy	***0.83***	0.15	0.10	0.17
Authority	0.18	***0.66***	0.20	-0.04
Completeness	***0.72***	0.15	0.16	0.31
Consistency	***0.55***	0.25	0.13	***0.46***
Currency	0.02	***0.70***	0.27	0.32
Precision	***0.61***	***0.45***	0.07	0.15
Informativeness	***0.45***	***0.73***	0.00	0.07
Relevance	0.41	***0.64***	0.06	0.28
Reliability	***0.80***	0.25	0.04	0.25
Simplicity	-0.04	0.25	***0.75***	0.23
Stability	0.35	0.06	***0.78***	0.12
Validity	***0.62***	0.18	***0.56***	-0.23
Verifiability	***0.75***	0.26	0.31	-0.07

Note. Extraction method: PCA; rotation method: Varimax with Kaiser normalization.

The first round of the factor analysis found that four variables (*consistency*, *precision*, *informativeness*, and *validity*) were loaded significantly on more than one factor. Since the *validity* variable had the highest cross-loading, it was deleted from the model and the loadings were recalculated. A Scree plot still suggested 4 factors, and the MSA of each of the variables still was higher than 0.8, with the Bartlet test of sphericity significant at the 0.0001 level. The first four eigenvalues captured 70% of the total variance of the data. In the reduced model, each variable was loaded significantly on only one factor (see Table 10). The mean importance ratings for each data quality dimension are shown in Table 11.

Table 10. Factor loadings for the DQ criteria (reduced model).

DQ Criteria	Component			
	1	2	3	4
Accessibility	0.25	0.09	0.20	***0.79***
Accuracy	***0.86***	0.15	0.11	0.09
Authority	0.22	***0.71***	0.25	-0.16
Completeness	***0.77***	0.14	0.22	0.16
Consistency	***0.57***	0.22	0.16	0.41
Currency	0.01	***0.65***	0.26	0.45
Precision	***0.61***	0.41	0.05	0.23
Informativeness	0.45	***0.71***	-0.01	0.17
Relevance	0.40	***0.59***	0.03	0.42
Reliability	***0.79***	0.21	0.00	0.32
Simplicity	0.03	0.23	***0.82***	0.11
Stability	0.36	0.04	***0.72***	0.18
Verifiability	***0.76***	0.27	0.26	-0.04

Note. Extraction method: PCA; rotation method: Varimax with Kaiser normalization.

Table 11. Mean importance ratings of the DQ dimensions.

Dimensions	Mean	Median	Std. Deviation
Accuracy	6.49	7	1.11
Reliability	6.27	7	1.29
Verifiability	6	6	1.39
Completeness	5.96	6	1.3
Consistency	5.95	6	1.35
Precision	5.73	6	1.41
Accessibility	5.62	6	1.5
Informativeness	5.43	6	1.46
Relevance	5.3	6	1.48
Stability	5.25	6	1.56
Authority	4.99	5	1.64
Currency	4.48	5	1.72
Simplicity	4.35	4	1.63

Based on the significant loadings, the four factors were labeled as *accuracy*, *informativeness*, *simplicity*, and *accessibility* (see Table 12). The authors evaluated the internal consistency of the factor constructs with Cronbach's alpha (excepting *accessibility*, which consisted of only one dimension). The alpha values of the *accuracy*, *relevance*, and *simplicity*

constructs were 0.89, 0.76, and 0.60, respectively. Although the alpha value of the *simplicity* construct was below the generality accepted lower limit of 0.70, it still can be considered as acceptable for exploratory research (Hair et al., 2005).

The variables that loaded significantly on each factor were then used to develop summated scales. Four scales were developed by averaging scores of the variables assigned to each factor. The *accuracy* data quality scale had the highest average importance score followed by the *accessibility* scale. The *simplicity* data quality scale had the lowest average perceived importance score. The scores of the summated scales were added to the rest of the data and used in examining the relationships among perception of data quality and other aspects of the community's scientific work and data management activities, including scientists' demographic characteristics.

Table 12. Mean importance scores of the data quality scales.

Data Quality Scale	Mean Rating
Accuracy (Accuracy, Completeness, Consistency, Precision, Reliability, Verifiability)	6.07
Accessibility (Accessibility)	5.62
Informativeness (Authority, Currency, Informativeness, Relevance)	5.05
Simplicity (Simplicity, Stability)	4.8

5.4.3 Perceptions of quality and research work context

To provide effective infrastructure support for the community's data quality assurance work, it is important to understand how the work is divided, what roles are played, what reference sources and tools scientists use to assess and/or enhance data quality, and what tools they use to communicate and collaborate on data quality assurance.

Only 9% of the participants indicated that they were familiar with data quality assessment criteria used by a specific academic or research community, or a funding agency; 23% were not sure, and 68% were not familiar. Participants who indicated familiarity with some data quality assessment models named not only traditional academic societies and funding agencies such as the American Physical Society and NSF, but also communities gathered around online databases such as the Inorganic Crystal Structure Database (ICSD), Cambridge Structural Database (CSD). Participants also referenced the quality criteria of scholarly journals and national and international laboratories, such as the NHMFL and the International Centre for Diffraction Data (ICDD).

Those who were familiar with some existing data quality assessment models indicated they used the quality criteria and models to develop their own laboratory standards for data quality and repeatability, review data for quality problems, and review manuscripts for journal publication.

Participants selected Origin analysis software as the most frequently used software for data quality evaluation (70%), followed by MatLab (40%). Participants also indicated the use of open source tools and locally written software applications.

A majority of the participants (59%) indicated that they cooperated with other scientists in controlling the quality of project data. They most often cooperated with people inside the project (98%). 29% indicated that they cooperated with

people outside the project to control project data quality. The most frequently used communication tool was email (99%), followed by phone (53%) and Skype (33%).

The Kruskal-Wallis test of dependence of the data quality scales on participant characteristics found several significant relationships. The analysis found significant difference of the data *simplicity* scale scores on whether participants used external data or not. Participants who had not used external data had a higher mean rank for the data *simplicity* construct than the scientists who did.

Likewise, the Kruskal-Wallis test found a significant dependence of the distribution of *simplicity* scale scores on the use of documentation guidelines and the project having a data management plan. Scientists who used specific guidelines when documenting data or whose typical research projects had data management plans had a higher mean rank for the *simplicity* scores than the scientists who did not (see Table 13). In addition, a binary logistic regression of the project having a data management plan into the data being archived after the project ends (model fit likelihood ratio: $\chi 2 = 11.86$; $p = 0.001$) showed the project not having a data management plan to be a significant predictor of data not being archived after the project ends (Wald = 9.69, $p = 0.002$).

The study found a statistically significant relationship between the *accuracy* scores and the type of collaborative project. Participants who worked on funded or both funded and unfunded projects had higher mean ranks for *accuracy* than participants who worked on unfunded projects only. In addition, a Kruskal-Wallis test found significant dependence of the *accessibility* scores on the scientist's primary role in funded projects. Student research assistants had a lower mean rank for the *accessibility* than the other roles. The study did not find significant relationships between data quality construct scores on the scientist's primary role in unfunded projects (see Table 13).

In addition to examining the relationships between the project related characteristics of scientists and quality perception, the study also looked at the relationships between data quality perceptions and the demographic characteristics of participants, including participants' methodological specialization and academic or research position in the organization. A Kruskal-Wallis test found significant dependence of the *accessibility* scale scores on research position (see Table 13). Assistant research scientists and postdoctoral research associates had higher mean ranks for *accessibility* than the other groups.

Table 13. Kruskal-Wallis test of dependence of the data quality scales on participant characteristics (significant relationships only).

		Accuracy	Accessibility	Informativeness	Simplicity
Use of external data	Chi-square	0.19	0.20	0.63	***4.91***
	df	1	1	1	***1***
	Asymp. sig	0.67	0.66	0.43	***0.03***
Research Position	Chi-square	1.88	***12.90***	3.32	5.38
	df	4	***4***	4	4
	Asymp. sig	0.76	***0.01***	0.51	0.25
Primary role in funded projects	Chi-Square	9.00	*12.90*	7.68	3.20
	df	6	*6*	6	6
	Asymp. Sig.	0.17	*0.05*	0.26	0.78
Use of documentation guidelines	Chi-Square	1.01	4.56	1.33	***8.80***
	df	2	2	2	***2***
	Asymp. Sig.	0.60	0.10	0.51	***0.01***
Project to have a data management plan	Chi-Square	2.15	1.85	3.27	***9.81***
	df	2	2	2	***2***
	Asymp. Sig.	0.34	0.40	0.20	***0.01***
Worked on funded or unfunded projects	Chi-Square	***7.20***	1.68	5.28	5.79
	df	***2***	2	2	2
	Asymp. Sig.	***0.03***	0.43	0.07	0.06

6. Discussion

6.1 Activities

The analysis of survey responses identified 10 research project task constructs: *research objective*, *simulation*, *sample material synthesis*, *device building*, *experiment setup*, *data collection and analysis*, *interpretation*, *team discussion*, *scholarly communication*, and *administration and coordination* (see Table 2). From now on these ten project task constructs will be referred as the CMP project task model. Kerzner (2003) defined 6 general project phrases: *planning*, *designing*, *testing*, *validating*, *analyzing*, and *reporting*. Although most of the task constructs from the CMP project task model can be mapped to Kerzner's project phases, the CMP model is of finer granularity. For instance, *simulation*, *sample material synthesis*, *device building*, and *experiment setup* can be mapped into the *designing* stage of Kerzner's model. In addition to general project tasks, the CMP project task model includes tasks that are characteristic of a research project (e.g., *research objective* and *interpretation*).

Although CMP is a small team science performed by individuals and/or small teams, (Stvilia et al, 2011), the CMP project task model includes the *administration and coordination* task construct, which are not included in Kerzner's model. This shows that PIs and project leaders do a significant amount of administration work that needs to be taken into account when planning and designing an infrastructure for the project, including the infrastructure support needed for research data management.

Another model that could be informative to compare the CMP model to is the Joint Information Systems Committee's (JISC)[7] conceptualization of research stages. The JISC model attempts to define the relationships between research project and data lifecycles and has been frequently cited in the data curation literature (e.g., Tenopir et al, 2012). Even though the JISC model is research context specific, the CMP model is still more detailed. For example, the *ideas* stage from the JISC model can be mapped to at least three tasks from the CMP model: *research objective, administration and coordination,* and *simulation*. A project PI or lead researcher, in addition to pure administrative tasks, may facilitate the coordination of research ideas, and definition of research objectives. In addition, researchers analyze the literature to develop idea(s) for a research project. Likewise, the *simulate, experiment and observe* stage can be mapped to at least two CMP tasks (see Table 14). Since both the *administration and coordination,* and *team discussion* tasks of the CMP model reference the team aspects of research work, they can be mapped to the *partners* stage of the JISC model. Furthermore, both the *sample material synthesis* and *device building* tasks of the CMP model are research processes on their own, even though there is no direct semantic match between the terms of these constructs and the terms of the research stage names of the JISC model. On the other hand, the *share data* stage of the JISC model does not have a match in the CMP model. This difference could be explained by the objective of the JISC model to provide a conceptualization of research task and data relationships. Although some of the participants mentioned managing and sharing data (other than publications) in their responses to the survey question, the frequencies of the mentions were not sufficient for these tasks to enter into the factor analysis model of CMP project tasks.

Table 14. The comparison of the CMP project task model to the JISC model of research lifecycle.

		JISC Model of Research Lifecycle							
		Ideas	**Partners**	**Proposal Writing**	**Research Process**				**Publishing**
					Simulate, Experiment, Observe	**Manage Data**	**Analyze Data**	**Share Data**	
CMP Project Task Model	**Research Objective (study, property)**	**X**							
	Administration and Coordination (idea, coordinate, supervise)	**X**	**X**						
	Team Discussion (discuss, team)		**X**						
	Simulation (literature, manage, develop , simulate)	**X**			**X**	**X**			
	Sample Material Synthesis (material, characterize, synthesize)								
	Device Building (fabricate, device)								
	Experiment Setup (setup, experiment)				**X**				
	Data Collection and Analysis (data, analyze, acquire)						**X**		
	Interpretation (interpret)								
	Scholarly Communication (write, paper, edit, report)								**X**

One of the important components of research project infrastructure is a support for metadata management. Qin, Ball, & Greenberg (2012) conceptualized 10 user tasks involving scientific data: *discovery*, *identify*, *select*, *obtain*, *verify*, *analyze*, *manage*, *archive*, *publish*, and *cite*. Although the scope of the survey question was general and did not focus on data tasks, there is still an overlap between the CMP project task model and Qin et al's model. In particular, the CMP project task model includes the *analyze* and *publish* activities in Qin et al's model. There are differences as well. In addition to the project infrastructure development and administration activities (e.g., *administration and coordination*, *device building*), the CMP model includes the *interpretation* and *simulation* constructs, which are data intensive activities. This study found that project infrastructure building activities can produce data (e.g. CAD files of the instruments used in the project) that can be useful and reusable for future research projects. Furthermore, collecting and preserving project process data is essential for data quality evaluation and replication and validation of research results. Hence, it is important that research data repositories and project management systems collect and curate these types of data as well. Informed by the Functional Requirements for Bibliographic Records (FRBR[8]) task model, Qin et al's model has a more detailed structure for data discovery activity (i.e., *discovery*, *identify*, *select*, *obtain*). The data discovery verbs, with an exception of "acquire," were not mentioned often by participants and did not enter the CMP task model. This could be a result of CMP scientists reasoning about data management and data discovery at a more granular level than librarians or data curators. Similarly, participants did not emphasize the *archive* task in their responses when asked to list project tasks. At the same time, 65% of them stated that they typically archived data after their projects ended. This could be explained by the lack of formal archiving and more reliance on publications and data backups as informal data archives, even if they realized that these practices could not be a substitution for long-term data preservation.

> "[Project work] is generally published and the data remains accessible via back-up storage. But I do not consider this the same thing as archiving the data. As noted in previous responses, this can lead to difficulties when retrieving data many years later." (s68)

The lack of formal archiving could be caused by the less structured and more dynamic and iterative nature of scientific inquiry. Project objectives and goals may change or additional new research objectives may emerge in almost any phase of the research project in the scientist's mind. The term "sensemaking" is often associated with such unstructured processes (Gasser, Sanderson, & Zdonik, 2007; Weick, 1995). Sensemaking is an iterative way to achieve understanding and accumulate knowledge about a particular phenomenon or process (Boland & Tenkasi, 1995). As one of the survey participants noted he did not do data achieving because his projects did not have a clear end and remained "work in progress":

> "My research is small science, and continuously changing on a week to week basis. … [T]he projects never 'end' in a clear manner." (s5)

6.2 Data

Borgman, Wallis, and, Enyedy (2007) identified six types of data: *(a) raw data, (b) processed data, (c) verified data, (d) data certified using some standard, (e) models,* and *(f) software and algorithms*. In this study, participants considered digital publications—such as preprint and journal articles and presentation slides—as data, because they contained graphs with embedded tabular data sets. Indeed, these data types were the second and third most frequently selected data types by survey participants. Hence, the above typology of data from Borgman et al (2007) could be extended with at least three additional types of data: *text documents, presentations,* and *visualization* data, such as graphs.

One of the important data types scientists create and use is metadata, especially since one person's metadata is another person's data (Redman, 1992). To the scientist who synthesized a material sample, the chemical formula of the material is metadata. However, to another scientist who wants to generate the same material and/or test its properties under different physical conditions, the formula could be data. Bibliographic data could be metadata to the scientist who wants to find publications relevant to his or her research, while to another scientist who does scientometric or bibliometric analysis of a particular research lab or community, the same bibliographic records could serve as data (e.g., Hinnant et al., 2012a; Stvilia et al., 2011). As was expected, an overwhelming majority of participants indicated that they created metadata/documentation when analyzing and collecting data. It was surprising, however, that metadata also was created "after the fact" in the scholarly communication stage when writing a paper, or presenting a paper at a conference (see Table 4). This suggest that the value of raw data in CMP is explicated through a presence of related derived data, such as related presentations or papers using the data, and the scholarly impact of those publications (e.g., the number of citations received).

In addition, less than 27% of the participants indicated that they created or added metadata when depositing data for preservation, while 65% of them stated that they archived data after their typical project ends. In addition, more than 70% of the participants reported creating metadata when writing and publishing a paper in a peer-reviewed journal or when working on a conference presentation (see Table 4). These findings suggest that the cost of metadata creation could be a disincentive to scientists to document their data, unless the cost is balanced by some immediate benefit, such as the value of related publications or presentations. These also could indicate that most of them did not do formal archiving of primary data in a specialized data repository for community sharing, and instead did informal archiving on their personal computers. This conclusion is further strengthened by a majority of the participants indicating that they did not use any specific documentation guidelines, and only 34% indicating that their typical project had a data management plan.

The use of specialized tools for creating structured documentation seems to be rare among CMP scientists. Only 22% of the participants indicated that they used electronic lab notebooks, which was a much lower percentage than the use of email (68%), presentation software (80%), or graphing software (84%). For metadata creation, CMP scientists might prefer using the same tools that they use to generate, analyze, or present data, instead of using standalone specialized metadata tools.

60% of the participants stated that they did not use any specific metadata standard for documenting their data. 22% of the participants used specific documentation guidelines of their research team, lab or a funding agency. These are in a clear consensus with the findings of the earlier study by Tenopir et al (2011). They found that a majority (56%) of their respondents, who represented multiple disciplines, did not use any metadata standard and 22% followed their local lab documentation guidelines. These results show that the adoption of metadata standards in CMP remains a challenge that warrants more research of the barriers and facilitators to the adoption.

An overwhelming majority of the participants felt that metadata they created served its purpose and was of good, very good, or excellent quality. This finding echoes those of data quality studies in other domains, where data and metadata creators felt satisfied with good enough data quality which met their local needs. However, the literature has shown that meeting local needs may not be sufficient for making data globally sharable and interoperable (Shreeves et al, 2005; Stvilia, Gasser, Twidale, Shreeves, & Cole, 2004).

When asked about data management roles in a typical funded research project, 51% of the participants stated that the project PI was responsible for managing project data. This number was 44% for a typical unfunded project. For funded projects only 1% indicated that someone outside their project team, such as IT staff, was responsible for managing project data; no participants indicated this was true for unfunded projects. These results again strengthen the conclusion that in

CMP data curation is mostly informal and not centralized. Data is owned at the individual and project team levels instead of at the community level, and most of the times PIs, project leads, or students who generated the data are considered the owners and curators of the data. As one of the participants stated:

> "…data is not archived in part because the generator of data is seen as steward of data throughout their career." (s89)

6.3 Quality

The most frequent categories of data problems were metadata problems: lack of file naming conventions and the difficulty of interpreting data due to poor or lost documentation (see Table 7). The most frequently reported sources of data quality problems were related to experiment process quality, such as human errors, impure sample materials, imprecise instruments, interference from external environment, or errors in experiment design (see Table 8).

Data quality can be changed actively through direct modification of data objects. CMP scientists may clean data to enhance its quality. In some cases, however, these attempts to remove background "noise" or amplify a "good" signal may inadvertently introduce errors if they are not done correctly:

> "Once you have the quantities you think you are measuring, there are additional contributions to these quantities, some of which you are not interested in. And so how do you subtract that? So often… well, some are worried about what these subtractions are." (p3)

Scientists may also deliberately degrade the quality of their data or metadata to keep their "know how" and specialized knowledge secret from competitors,

> "You maybe spend one year or two years to make this sample without the notebook. But as long as you know its chemical formula, if you are a good scientist, you will figure out how to make it. So many labs … don't want to mention the name [chemical formula] anywhere. They treat their sample's name as a code." (p6)

Only 9% of the survey participants selected the purposeful degradation of data quality as a source of data quality problems, though (see Table 8). This might distinguish the CMP community from open online communities of knowledge creation such as Wikipedia, where purposeful degradation of quality is a more frequently occurring problem (Stvilia et al., 2008).

Data quality can be changed not only through direct manipulation or modification of the data, but also indirectly, with changes in the context of its interpretation and use in time and space. This could be caused by changes in culture, community composition, the set of activities using the data, or in the state of knowledge and technology (Stvilia & Gasser, 2008a). 33% of the survey participants identified the context change as a source of data quality problems. Over time, as the research methods and techniques change, data quality can be evaluated differently.

> "There's an understanding of the limitations of the technique of the time, but, anyone should go and look at that data, with that eye, say, 'there's some limitations, technique of that time, but this data would still be valid within these limitations.' And anybody can go, given those limitations of time, and analyze that data again." (p4)

Data quality also is influenced by the technology used to generate or collect the data. 51% of the participants selected imprecise instruments as a source of data quality problems (see Table 8). Alternatively, improvements in the research technology may result in a higher level of data quality and novel contributions to the science. One of the interviewees discussed how a new technology enabled his team to obtain high quality measurement data, which eventually led to a successful paper in an important journal of the field:

> "This paper was done, because the quality of the data, the signal to noise was such, because of those pair amplifiers, they allowed us to get this paper accepted to *Physical Review Letters* in like two, three weeks. Just because one detail of the electronics." (p4)

Also, as time goes and the literature and knowledge of a particular area evolves, not only can the quality of the same data can be evaluated differently, but it also can be interpreted differently.

> "The field evolves, the reputation of that data evolves as people discuss, as theorists get involved, more data is acquired. There's a deeper understanding of what you saw really means, and we, you start talking about it in a different way." (p7)

Different scientific communities may have different scholarly communication cultures, and may have different expectations and norms for the quality (e.g., completeness) of the research process description in a publication. One of the interview participants discussed the differences in and expectations for the completeness of the description of a sample material generation process when publishing in physics versus chemistry journals:

> "[In a physics journal] you just can say I made a single crystal, and I measured that. That is it. But, for a chemistry journal, you need to list your X-ray data or how you make it in a very detailed way." (p8)

To identify the community's understanding of data quality, the study asked survey participants to rate 17 quality dimensions by their importance. The factor analysis produced four data quality constructs: *accuracy*, *accessibility*, *informativeness*, and *simplicity*. The literature includes studies of data quality in different communities. When examining consumers' perception of online health information quality, Stvilia et al. (2009) identified five information quality constructs: *accuracy*, *completeness*, *authority*, *usefulness*, and *accessibility*. In a different study, Huang et al. (2012) investigated the understanding of and priorities for data quality by genomics scientists and data curators. They defined a data quality model consisting of five data quality constructs. These three models share at least two constructs, *accuracy* and *accessibility*, even though the construct compositions differ slightly across the models (see Table 15). It is important to note that the *accuracy* constructs are rated the highest in all three models, followed by the *accessibility* constructs.

The differences among the three quality perception models could be attributed to the differences in the communities studied, the types of data these models were intended for, as well as the number of dimensions included in the starting models before factor analysis was applied. Huang et al (2012) started with 19 quality dimensions. Stvilia et al (2009) used 21 quality dimensions as a starting model. This study used 14 dimensions in its starting model of quality perception. The starting sets of quality dimensions were obtained both from the literature (e.g., Stvilia et al, 2007; Wang & Strong, 1996) and pre-survey interviews of representatives of the studied communities. Hence, these starting models did reflect some of the differences in community and context specific vocabularies, and priorities for data and information quality dimensions. For example, the precision of measurements, and consistency in naming and representing data files were very important for experimentalists in the CMP community, while the other two groups assigned lower priorities to those quality dimensions, particularly the consumers of online health information (see Table 15). This could be explained by health information consumers not conducting original research and/or collecting and managing raw research data, but rather using derived information products such as articles, blogs, factsheets, and Q&A pages.

Table 15. Comparison of data quality models. Constructs in italics are common across all three models.

CMP DQ Model		**Online Consumer Health IQ Model (Stvilia et al., 2009)**		**Genomics DQ Model (Huang et al., 2012)**	
DQ Scales	**Mean Rating**	**IQ Scales**	**Mean Rating**	**DQ Scales**	**Mean Rating**
Accuracy (Accuracy, Completeness, Consistency, Precision, Reliability, Verifiability)	6.07	***Accuracy*** (Accuracy, Credibility, Reliability)	4.41	***Accuracy*** (Accuracy, Unbiased, Believability, Traceability)	6.01
Accessibility (Accessibility)	5.62	**Completeness** (Completeness, Clarity)	4.17	***Accessibility*** (Accessibility, Believability, Appropriate amount of information)	5.8
Informativeness (Authority, Currency, Informativeness, Relevance)	5.05	**Authority** (Authority)	3.8	**Usefulness** (Interpretability, Understandability, Ease of manipulation, Consistent representation, Value added)	5.52
Simplicity (Simplicity, Stability)	4.8	**Usefulness** (Ease of use, Objectivity, Utility)	3.75	***Relevance*** (Relevant, Concise representation, Up-to-date, Reputation, Value)	5.08
		Accessibility (Accessibility, Cohesiveness, Consistency, Volatility)	3.57	**Security** (Security, Traceability)	4.56

The survey data showed that CMP scientists may assign different importance to different quality criteria. On average, CMP scientists rated highest the dimensions loaded on the accuracy construct such as accuracy, reliability and verifiability (see Table 11). Data used in CMP ranges from sample materials and CAD files for designing and manufacturing parts for an experiment, to simulation data and computer codes. Quality assurance methods used in CMP can vary from those used in manufacturing—such as reducing signal variation through controlling noise introduced by the equipment and environment—to those used in metadata and software quality control. However, in contrast to manufacturing, outcomes of scientific work are uncertain. Often, physical properties of a material are not known and there are no "gold standards" or reference sources to evaluate the quality of experimental data against other than the data obtained from previous experiments. Hence, reliability or reproducibility is one of the most important quality criteria for scientific data:

> "The most important thing is reproducibility. You get a particular material, you're looking for particular effects, whatever ... You measure it, and find a particular behavior in a particular crystal. You switch the crystal, and you get another one. And, you get a very similar if not identical data set, and so, and you get a third one, you still get the same. And, perhaps you get a fourth one, and go and use another measurement system, another cryostat, and other electronics, in base they're still the same, same data set. Then someone in Japan and China does the same and gets consistently the same results." (p4)

Alternatively the reputation and consequently the value of data may decline if other researchers cannot replicate the data or a flaw is discovered when reusing the data

> "If at the end of the day one guy just says, 'OK, I cannot repeat it [data],' it means either you are doing something wrong or you are cheating." (p6)

The value of data and its quality can be a function of the activity success or failure, the amount of data use, the cost of activity, the amount of novel or unexpected information, the amount of payoff received by the agent, or a combination of many or all of the above (Machlup, 1983; Marschak, 1971; Radner, 1986; Stvilia & Gasser, 2008b). The purpose of scientific research is to acquire new knowledge and advance the state of art of a particular discipline. In CMP, data and its quality gain in importance when the researcher perceives that it may contain new information that could lead to a novel contribution to the literature:

> "So, let's say [a] good sample gives you good data, but I still divide them … [into] useful data or not useful data. Yes. Sometimes I make a sample. I measure something. Then I think, OK, that's not so exciting physics. So I cannot publish a high quality paper. So then I just slow down or just put them aside. ...Let's say [the] sample determines [whether] the data is good or not. Physics determines [whether] the data is excellent or not." (p6)

A formal quality assessment model can be developed collaboratively by a community. It can also be developed and enforced in a top-down manner by governments, businesses, or organizations. In addition, data quality assessment models can be offered by third-party quality rating and certification entities (Stvilia et al., 2009). In this study, only 9% of the participants were familiar with some formal data quality assessment criteria or model. They stated familiarity with data quality criteria and models offered by government agencies such as the NSF, scholarly societies, databases, and labs. Future research directly related to this one could compare the CMP data quality model to the models and principles used by those entities. This would help further refine and contextualize the CMP data quality model and develop a best practices guide for data quality assurance, which would be harmonized with the data management needs of project teams, the community as a whole, and their funding partners.

Similar to earlier studies of data quality perception in different scientific communities (e.g., Huang et al., 2012), this study found that, depending on different roles played in research projects and differences in overall data practices, there could be differences in the perceptions of data quality by scientists. Scientists who had not used external data had higher data *simplicity* scale scores on average than scientists who had. This might indicate that scientists' concerns about outside data complexity and interpretability could serve as barriers to data reuse. Alternatively, it might indicate that the scientists assigned a lower priority to the simplicity or ease of use of data because they had not had an experience of interpreting outside data.

Scientists who used documentation guidelines or had data management plans for their projects had higher *simplicity* scale scores on average than scientists who did not. In addition, the study found a positive relationship between a project having a data management plan and data being archived after the project ends. This could be a result of those scientists having higher data management literacy and being more conscious of data quality issues than scientists who were not introduced to or did not follow more formal data management practices. This also could be caused by scientists complying with data management requirements set by funding agencies, such as the NSF. A separate investigation of the effects of government data policies on the data practices of scientists and their perception of data quality could shed more light on these relationships.

7. Conclusions

The paper examined research project tasks and the perceptions of data quality in a CMP community. The study identified ten constructs of research project tasks and four data quality constructs. The community placed the highest importance on the intrinsic quality dimensions (i.e. *accuracy*). CMP scientists generate and use different data ranging from CAD files and physical material samples to scholarly publications and metadata. A CMP research process includes at least ten activities with multiple data inputs and outputs. The different activities of the process may have different primary data. The value and importance of primary and intermediary datasets are defined by the value and impact of final data products, such as publications. The study found that scientists performed value-adding actions, such as adding metadata to related datasets when presenting or publishing research findings. Future research may integrate the CMP project task model and data quality model developed by this study with data task models from the literature as well as with typologies of data, project roles, and quality assurance actions to develop a comprehensive data management ontology for CMP.

The study developed a typology of project tasks as well as identified the types of data generated and used by scientists, and the tools used to manage data. Future research related to the current study will examine the relationships among the types of projects tasks, data and tools. In addition, the study found that the quality priorities of the CMP scientists who worked on funded projects were different from those who had unfunded projects. Similarly, there were differences in quality priorities on some of the quality dimensions by scientists who played different roles. Future related research could further investigate the causes for these differences by collecting additional qualitative data.

The findings of this study can inform the design of data management policies, best practice guides, and infrastructure tools, such as a data management ontology aligned and harmonized with the data practices and priorities of the CMP community and other related communities.

Acknowledgements

The authors thank two anonymous reviewers for their helpful comments and suggestions on an earlier version of this manuscript. This research was supported in part by the Florida State University Office of Research and by the NSF under Grant OCI-0942855. The article reflects the findings and conclusions of the authors, and do not necessarily reflect the views of Florida State University, the NSF, or the NHMFL.

8. References

Arazy, O., Nov, O., Patterson, R., & Yeo, L. (2011). Information quality in Wikipedia: The effects of group composition and task conflict. *Journal of Management Information Systems*, *27*(4), 71-98.

Atkins, D. E., Droegemeier, K. K., Feldman, S. I., Garcia-Molina, 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.

Bailey, K. (1994). *Methods of social research* (4th ed.). New York, NY: The Free Press.

Ball, A. (2010). *Review of the state of the art of the digital curation of research data* (ERIM Project Report No. erim1rep091103abl2). Bath, UK: University of Bath. Retrieved from http://opus.bath.ac.uk/19022/

Bechhofer, S., Buchan, I., Roure, D., Missier, P., Ainsworth, J., Bhagat, J., & Goble, G. (2013). Why linked data is not enough for scientists. *Future Generation Computer Systems, 29.* doi:10.1109/j.future.2011.08.004

Birnholtz, J. & Bietz, M. (2003). Data at work: supporting sharing in science and engineering. In *Proceedings of the 2003 international ACM SIGGROUP conference on Supporting group work* (GROUP '03). ACM, New York, NY, USA, 339-348.

Borgman, C. L., Wallis, J. C., & 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.

Bruce, T., & Hillman, D. (2004). The continuum of metadata quality: defining, expressing, exploiting. In D. Hillman & E. Westbrooks (Eds.), *Metadata in Practice* (pp. 238–256). Chicago, IL: ALA Editions.

Buckland, M. (1997). What is a "document"? *Journal of the American Society for Information Science, 48*, 804-809.

Burton, A., & Treloar, A. (2009). Designing for discovery and re-use: The 'ANDS Data Sharing Verbs' approach to service decomposition. *International Journal of Digital Curation, 4*(3), 44-56.

Center for Research Libraries, & Online Computer Library Center. (2007). *Trustworthy Repositories Audit and Certification (TRAC): Criteria and checklist*. Chicago, IL: Center for Research Libraries. Retrieved from http://www.crl.edu/PDF/trac.pdf

Conway, P. (2010). Preservation in the age of Google: Digitization, digital preservation, and dilemmas. *Library Quarterly, 80*, 61-79. doi:10.1086/648463

Conway, P. (2011). Archival quality and long-term preservation: a research framework for validating the usefulness of digital surrogates. *Archival Science, 11*, 293-309.

Curry, E., Freitas, A., & O'Riáin, S. (2010). The role of community-driven data curation for enterprises. In D. Wood (Ed.), *Linking enterprise data* (pp. 25-47). New York, NY: Springer. doi:10.1007/978-1-4419-7665-9_2

Data Curation Center (2004-2013). Data Asset Framework Online Tool. Retrieved March 8, 2012 from http://www.data-audit.eu/tool2/.

De Roure. (2010, November 27). Replacing the paper: The twelve Rs of the e-Research. Retrieved from http://blogs.nature.com/eresearch/2010/11/replacing_the_paper_the_twelve_rs_of_the_e-research_record.html

de Solla Price, D. J. (1963). *Little Science, big Science*. New York, NY: Columbia University Press.

DRAMBORA Consortium. (2008). *DRAMBORA interactive: Digital Repository Audit Method Based on Risk Assessment*. Retrieved from http://www.repositoryaudit.eu/

Duda, R., Hart, P., & Stork, D. (2000). *Pattern classification* (2nd ed). Wiley.

Earth Observing System Data Panel. (1986). *Earth Observing System: Data and information system* (Technical Memorandum No. 87777). Washington, DC: National Aeronautics and Space Administration. Retrieved from http://hdl.handle.net/2060/19860021622

Ellingson, R. & Heben, M. (2011). Molecular and condensed matter physics laboratory [PowerPoint slides]. Retrieved from http://tinyurl.com/8sxuc6y

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

Engeström, Y. (2001). Expansive learning at work: Toward an activity theoretical reconceptualization. *Journal of Education and Work, 14,* 133-156. doi:10.1080/13639080020028747

Eppler, M. (2003). *Managing information quality: Increasing the value of information in knowledge-intensive products and processes.* Berlin, Germany: Springer-Verlag.

Fallis, D., & Frické, M. (2002). Indicators of accuracy of consumer health information on the Internet: A study of indicators relating to information for managing fever in children in the home. *Journal of the American Medical Informatics Association, 9*, 73–79.

Faniel, I. M., & Jacobsen, T. E. (2010). Reusing scientific data: How earthquake engineering researchers assess the reusability of colleagues' data. *Computer Supported Cooperative Work (CSCW), 19,* 355-375. doi:10.1007/s10606-010-9117-8

Foster, I., Jennings, N., & Kesselman, C. (2004). Brain meets brawn: Why grid and agents need each other. *Proceedings of the Third International Joint Conference on Autonomous Agents and Multi-Agent Systems, 1*, 8-15. doi:10.1109/AAMAS.2004.78

Gasser, L., Sanderson, A., & Zdonik, S. (2007). *NSF IIS-GENI workshop report: First edition.* Washington, DC: National Science Foundation. Retrieved May 22, 2008, from https://apps.lis.uiuc.edu/wiki/download/attachments/10304/iis-geni-report-first-edition.pdf

Ge, M., & Helfert, M. (2007). *A review of information quality research—develop a research agenda*. Paper presented at the 12th International Conference on Information Quality, Cambridge, MA.

Gertsbakh, I. (1977). Models of preventive maintenance. In H. Theil (Ed.), *Studies in Mathematical and Managerial Economics*. Amsterdam, Netherlands: North-Holland Publishing Company.

Greenberg, J. (2001). Quantitative categorical analysis of metadata elements in image applicable metadata schemas. *Journal of the American Society for Information Science and Technology, 52,* 917–924.

Gutmann, M., Schürer, K., Donakowski, D., & Beedham, H. (2004). The selection, appraisal, and retention of social science data. *Data Science Journal, 3*, 209–221.

Hair, J., Black, B., Babin, B., Anderson, R., & Tatham, R. (2005). *Multivariate data analysis*. Upper Saddle River, NJ: Prentice-Hall.

Higgins, S. (2008). The DCC curation lifecycle model. *International Journal of Digital Curation, 3*(1), 134-140. Retrieved from http://ijdc.net/index.php/ijdc/article/view/69

Hinnant, C., Stvilia, B., Wu, S., Worrall, A., Burnett, G., Burnett, K., Kazmer, M., & Marty, P. (2012a). Author team diversity and the impact of scientific publications: Evidence from physics research at a national science lab. *Library & Information Science Research, 34,* 249-257.

Hinnant, C., Stvilia, B., Wu, S., Worrall, A., Burnett, K., Burnett, G., Kazmer, M. M., & Marty, P. F. (2012b). Data curation in scientific teams: An exploratory study of condensed matter physics at a national science lab. In J.-E. Mai (Chair), *Proceedings of iConference 2012* (pp. 498-500). New York, NY: ACM.

Huang, H., Stvilia, B., Jörgensen, C., & Bass, H. (2012). Prioritization of data quality dimensions and skills requirements in genome annotation curation. *Journal of the American Society for Information Science and Technology*, *63*, 195–207.

Juran, J. (1992). *Juran on quality by design*. New York, NY: The Free Press.

Kell, D. & Oliver, S. (2004). Here is the evidence, now what is the hypothesis? The complementary roles of inductive and hypothesis-driven science in the post-genomic era. *Bioessays, 26*, 99–105.

Kuutti, K. (1995). Activity theory as a potential framework for human-computer interaction research. In B. Nardi (Ed.), *Context and consciousness: Activity Theory and human computer interaction* (pp. 17-44). Cambridge, MA: MIT Press.

Leont'ev, A. N. (1978). *Activity, consciousness, and personality*. Englewood Cliffs, NJ: Prentice-Hall.

Levitin, A. V., & Redman, T. C. (1993). A model of the data (life) cycles with application to quality. *Information and Software Technology, 35*, 217-223.

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 Committee for the Support of Research. Retrieved from http://www.jisc.ac.uk/media/documents/programmes/preservation/esciencereportfinal.pdf

Machlup, F. (1983). Semantic quirks in studies of information. In F. Machlup & U. Mansfield (Eds.), *The study of information: Interdisciplinary messages* (pp. 641–671). New York: Wiley.

Marschak, J. (1971). Economics of information systems. *Journal of the American Statistical Association, 66*(333), 192–219.

Mesirov, J. (2010). Accessible reproducible research. *Science, 327*(5964), 415-416.

National High Magnetic Field Laboratory. (2012). *Mag Lab by the numbers*. Retrieved from http://www.magnet.fsu.edu/mediacenter/factsheets/numbers.html

National Research Council (U.S.). (2007). *Condensed-matter and materials physics: The science of the world around us.* Washington, DC: The National Academies Press.

Nichols, D. M., Chan, C. H., Bainbridge, D., McKay, D., & Twidale, M. B. (2008). A lightweight metadata quality tool. In R. Larsen (Chair), *Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 385-388). New York, NY: ACM. doi:10.1145/1378889.1378957

Office of Management and Budget. (1999). *Uniform administrative requirements for grants and agreements with institutions of higher education, hospitals, and other non-profit organizations* (OMB Circular 110). Retrieved from http://www.whitehouse.gov/omb/circulars_a110#36

Peng, R. D. (2011). Reproducible research in computational science. *Science, 334*(6060), 1226-1227.

Qin, J., Ball, A., & Greenberg, J. (2012). Functional and architectural requirements for metadata: supporting discovery and management of scientific data. *Proceedings of the International Conference on Dublin Core and Metadata*

Applications 2012. Kuching, Malaysia: University of Bath. Retrieved from http://dcevents.dublincore.org/IntConf/dc-2012/paper/view/107

Radner, R. (1986). Normative theory of individual decision: An introduction. In C. Mcguire, & R. Radner (Eds.), *Decision and Organization* (2nd ed., pp. 1–18). Minneapolis: University of Minnesota Press.

Redman, T. (1992). *Data quality: Management and technology*. New York, NY: Bantam Books.

Redman, T. (1998). The impact of poor data quality on the typical enterprise. *Communications of the ACM, 41*(2), 79-82.

Relevance, (2012). In *Wordnet*, Retrieved from http://wordnetweb.princeton.edu

Rieger, O. (2008). *Preservation in the age of large-scale digitization: A white paper*. Washington, DC: Council on Library and Information Resources.

Sheppard, S. A., & Terveen, L. (2011). Quality is a verb: The operationalization of data quality in a citizen science community. In F. Ortega (Chair), *Proceedings of the 7th International Symposium on Wikis and Open Collaboration* (pp. 29-38). New York, NY: ACM.

Shreeves, S. L., Cole, T. W., Knutson, E. M., Stvilia, B., Palmer, C. L., & Twidale, M. B. (2005). Is 'quality' metadata 'shareable' metadata? The implications of local metadata practices for federated collections. In H. Thompson (Ed.), *Proceedings of the Twelfth National Conference of the Association of College and Research Libraries* (pp. 223-237). Chicago, IL: Association of College and Research Libraries.

Simmhan, Y., Plale, B., & Gannon, D. (2005). A survey of data provenance in e-science. *ACM SIGMOD Record, 34*(3), 31-36.

Strong, D., Lee, Y., & Wang, R. (1997). Data quality in context. *Communications of the ACM, 40*, 103-110.

Stvilia, B. (2006). *Measuring information quality*. (Doctoral dissertation, University of Illinois at Urbana - Champaign). Retrieved from http://wwwlib.umi.com/dissertations/fullcit/3223727Stvilia, B. (2007). A model for ontology quality evaluation. *First Monday, 12*(12-3). Retrieved from http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/2043/1905

Stvilia, B. (2008). A workbench for information quality evaluation. In *Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital libraries* (pp. 469-469). New York, NY: ACM.

Stvilia, B., & Gasser, L. (2008a). An activity theoretic model for information quality change. *First Monday*, *13*(4). Retrieved from http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/2126/1951

Stvilia, B., & Gasser, L. (2008b). Value based metadata quality assessment. *Library & Information Science Research, 30*, 67-74. doi:10.1016/j.lisr.2007.06.006

Stvilia, B., Al-Faraj, A., & Yi, Y. (2009). Issues of cross-contextual information quality evaluation—The case of Arabic, English, and Korean Wikipedias. *Library & Information Science Research*, *31*, 232-239.

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. doi:10.1002/asi.20652

Stvilia, B., Gasser, L., Twidale, M. B., Shreeves, S. L., & Cole, T. W. (2004). Metadata quality for Federated Collections. In S. Chengulur-Smith, L. Raschid, J. Long, & C. Seko (Eds.), *Proceedings of the International Conference on Information Quality - ICIQ 2004* (pp. 111-125). Cambridge, MA: MITIQ.

Stvilia, B., Hinnant, C. C., Schindler, K., Worrall, A., Burnett, G., Burnett, K., Kazmer, M. M., & Marty, P. F. (2011). Composition of scientific teams and publication productivity at a national science lab. *Journal of the American Society for Information Science and Technology, 62*, 270-283.

Stvilia, B., Hinnant, C., Wu, S., Worrall, A., Lee, D. J., Burnett, K., Burnett, G., Kazmer, M. M., & Marty, P. F. (2013). Studying the data practices of a scientific community. In *Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries* (JCDL '13). ACM, New York, NY, USA, 425-426.

Stvilia, B., Twidale, M., Smith, L. C., & Gasser, L. (2008). Information quality work organization in Wikipedia. *Journal of the American Society for Information Science and Technology, 59*, 983–1001.

Sufi, S., & Mathews, B. (2004). *CCLRC Scientific Metadata Model: Version 2* (CCLRC Technical Report No. DL-TR-2004-001). Chilton, UK: Council for the Central Laboratory of the Research Councils (CCLRC). Retrieved from http://epubs.cclrc.ac.uk/work-details?w=30324

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A., Wu, L., et al. (2011) Data Sharing by scientists: Practices and Perceptions. *PLoS ONE 6*(6): e21101. doi:10.1371/journal.pone.0021101Wang, R., & Strong, D. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems, 12*(4), 5-33.

Wu, S., Stvilia, B., & Lee, D. J. (2012). Authority control for scientific data: The case of molecular biology. *Journal of Library Metadata, 12*, 61-82.

Appendix

Table 16. Definitions of DQ dimensions.

Dimension	Definition
Accessibility	Ease of locating and obtaining research data object relative to a particular activity (Stvilia et al, 2007).
Accuracy	The degree to which the data correctly represents an object, process, or phenomenon in the context of a particular activity or culture (Stvilia et al, 2007); the degree to which the measured value(s) fall at or very close to the true value (Ellingson & Heben, 2011).
Authority	The degree of reputation of data in a given community (Stvilia et al, 2007).
Completeness	The extent to which data is complete according to some general or contextual reference source (Stvilia et al, 2007).
Consistency	The extent to which similar attributes or elements of data are consistently represented using the same structure, format, and precision (Stvilia et al, 2007).
Currency	The age of data (Stvilia et al, 2007).
Informativeness	The amount of information contained in data (Stvilia et al, 2007).
Precision	The granularity of the model or content values of data according to some general or contextual reference sources (Stvilia et al, 2007); the degree to which repeated measurements fall reliably at or very near the same value (which may or may not be the correct value) (Ellingson & Heben, 2011).
Relevance	The extent to which data is related to the matter at hand (Relevance, 2012; Stvilia et al, 2007).
Reliability	The degree of confidence in data in the context of a particular activity.
Simplicity	The extent of cognitive complexity/simplicity of data measured by some index or indices (Stvilia et al, 2007).
Stability	The amount of time data remains valid in the context of a particular activity (Stvilia et al, 2007).
Validity	The extent to which data is valid according to some stable reference source, such as a dictionary or set of domain constraints and norms (Stvilia et al, 2007).

Verifiability	The extent to which the correctness of data is verifiable or provable in the context of a particular activity (Stvilia et al, 2007).

1 PRONOM. http://www.nationalarchives.gov.uk/PRONOM/Default.aspx
2 Research Data Alliance. https://rd-alliance.org/
3 Data Curation Profiles. http://www4.lib.purdue.edu/dcp/about
4 JHOVE. http://jhove.sourceforge.net/
5 OpenRefine. https://github.com/OpenRefine/OpenRefine
6 Stanford CoreNLP. http://nlp.stanford.edu/software/corenlp.shtml
7 Joint Information Systems Committee (JISC). Stages of the research and data lifecycle. http://www.jisc.ac.uk/whatwedo/campaigns/res3/jischelp.aspx
8 Functional Requirements for Bibliographic Records (FRBR). http://archive.ifla.org/VII/s13/frbr/frbr_2008.pdf