# Data Curation Perspectives and Practices of Researchers at Kent State University's Liquid Crystal Institute: A Case Study

A thesis submitted to the College of Communication and Information of Kent State University in partial fulfillment of the requirements for the degree of

Master of Library and Information Science

by

Shadi Shakeri

December, 2013

Thesis written by

Shadi Shakeri

B.A., University of Tehran, 2004

M.L.I.S., Kent State University, 2013

Approved by

_____

Karen F. Gracy, Ph.D., Advisor

_____

Tomas A. Lipinski, J.D., L.L.M., Ph.D., Director, School of Library and Information Science

_____

Stanley T. Wearden, Ph.D., Dean, College of Communication and Information

## Table of Contents

Page

TABLE OF CONTENTS ..... iii

LIST OF FIGURES ..... v

LIST OF TABLES ..... vi

ACKNOWLEDGMENTS ..... vii

CHAPTER

I. BACKGROUND ..... 1
- Problem Statement ..... 3
- Background of the Liquid Crystal Institute ..... 4

II. LITERATURE REVIEW ..... 6
- Research Objectives of the Study ..... 6
- Data Curation and Understanding of Data Practices ..... 6
- NSF Data Management Plans (DMPs) ..... 7
- Data Management for Research Data Produced in University Settings ..... 7
- Data Curation Needs of Small-Science Disciplines ..... 11
- Data Sharing Standards and Reuse ..... 12
- Cyberinfrastructure ..... 13
- Data Curation Case Studies ..... 14

III. METHODOLOGY ..... 18
- Overview of the Study's Methods ..... 19
- Data Collection Methods ..... 20
- Recruitment of Participants ..... 21
- Data Collection Activities ..... 23
- Maintaining Confidentiality of Participants ..... 24
- Data Analysis ..... 25
  - Development of data model for data curation research ..... 25
  - Definitions of the codes in Table 3 ..... 29

IV. RESEARCH FINDINGS ..... 35
- Data Management Perspectives and Practices ..... 35
- LCI Researchers' Perspectives on Data Characteristics ..... 35
- Data Management Perspectives and Practices of LCI Researchers ..... 36
  - Data sharing ..... 36
    - Data sharing in practice ..... 41
    - Intellectual property rights ..... 42
    - Data licensing ..... 44
  - Use of controlled vocabularies ..... 45

Data documentation and metadata....46
Note-taking perspectives and practices....48
Data storage and preservation....49
Data deposit practices....50
Data retrieval practices....51
Access management practices (security)....52
Data preservation perspectives and practices....53
Data selection and appraisal....56
Data Management Plans (DMPs)....58
Data Analysis Findings....59
Relationships and interactions among phenomena and concepts represented by codes....61
The Data World of Liquid Crystal Scientists....65
Development of the data model for liquid crystal scientists....65
Examples of interactions in the world of liquid crystal scientists....66
Research Questions and Findings....68

V. DISCUSSION....73
Research Findings and Recommendations....73
Merits of the Study....76
Limitations of the Study....77
Number of participants and degree of participation in the study....77
Scope of data collected....78
Techniques used to reduced bias....78
Data validation....78
Future Studies....79
Validation of the data world of liquid crystal scientists....79
Data curation issues and limitations....79
Campus-wide research data services....80
Data sharing and reuse....80
Conclusion....80

APPENDICES
A. Questionnaire Instrument....84
B. Interview Instrument....86
C. Participant Profiles....89
D. Glossary of Cyberinfrastructure and Digital Curation Terms....92
E. List of Acronyms Used....96

REFERENCES....97

## List of Figures


Figure Page

1. The interaction of the codes within the cluster of Researchers' Perspectives on Data Characteristics……62

2. Interactions between the codes within the cluster of Researchers' Perspectives in Data Management……63

3. Interactions between the codes within the cluster of Data Management Practices……64

4. Graphical representation of the clusters and the interactions among them within the data world of liquid crystal scientists……66

5. Interaction between perspectives on data quantifiability and data sharing, resulting in data sharing practice ……67

6. Interaction between perspectives on data replicability and reusability, affecting data sharing practices……68

## List of Tables


Table Page

1. Summary of Researchers' Participations in Interviews and Questionnaire ...23

2. Proposed Initial Code List for Analysis of Interview Data and Unstructured Questionnaire Responses.....26

3. List of Finalized Clusters and Codes for Analysis of Interview Data and Unstructured Questionnaire Responses.....29

4. Data Sharing Summary Table of Scientists at the Liquid Crystal Institute.....42

**Acknowledgments**

I want to thank my advisor, Dr. Karen F. Gracy for her support and positive attitude during the time I worked on my thesis. I want to thank my committee members, Dr. Marcia Lei Zeng and Dr. Catherine L. Smith, for their comments and help. I want to thank you my family and friends for being wonderful.

# Chapter I

## Background

The use of computers, digital data, and networks has revolutionized how research is accomplished in many scientific disciplines. As a result of these advancements, many of the small-sciences have become data-intensive, meaning that large quantities of data are being produced from the research conducted by the scientists .[1] On the other hand, there are other small-science disciplines that are not as data-intensive, producing much smaller quantities of data. Despite the differences in the scale of the research conducted, the research data produced in such small-science disciplines needs to be described, curated, and made available.

Learning about specific requirements of small-science disciplines such as data characteristics and the scale of data is important for data curation research. As the nature of the scientific discipline often determine the types and quantities of data produced, identification of particular data characteristics could help information professionals develop an understanding of the scientists' perspectives on their data management needs and data practices within the discipline.

Also, with increasing commercial exploitation and ambitious international experiments tackling grand research challenges, research data produced within academic institutes is becoming too expensive or impossible to replicate. Therefore, benefits of activities such as data sharing and reuse, and collaborative research are now more appreciated among research groups working on similar problems. When research groups working on similar problems share data with one another,

---

[1] Small sciences are typically described as hypothesis-driven research led by a single investigator or small research group that generates and analyzes their own data.

[2] Cyberinfrastructure refers to integrated systems that include hardware, software, and human

redundancies in data collection can be reduced; such collaborative activities also create larger study populations and datasets, thus increasing the statistical power of analyses and the reliability of research findings. To address the need to share and reuse data, often in collaboration with others, and to take advantage of the benefits provided by the new research environments, scientific communities must embrace advancements in computing technology through the development of centralized infrastructure.

Up until now, it has primarily been the larger stakeholders—government-funded agencies and research institutes such as Fermilab and Lawrence Livermore—that have contributed to the development of data curation infrastructure as well as best practices. Typically, these institutions have had sufficient funding to recruit staff and build infrastructure needed to manage their own data without assistance from outside.

While large research institutes can often rely on institutional resources to develop data curation programs, the researchers who work in smaller academic research institutes do not often have such infrastructure to meet their curation needs. In such smaller institutions, scientists generate or gather data into privately held sets or collections that they analyze locally. Also, typically their research funding can be limited, and the day-to-day conduct of research is often dependent on a few graduate students, who carry out much of the data collection, and manage and process these datasets during the course of a project (Cragin, Palmer, Carlson, & Witt, 2010, p. 4024). Researchers in academic research institutes often work on

several small and interrelated research projects simultaneously, many of which may be either not federally funded or not funded at all (Akers, 2013).

As a consequence of the resource disparity between the two types of institutions, it is mainly the larger research institutes that have had opportunities to study the issues surrounding building and extension of cyberinfrastructure (CI) and data management workflows.[2] As a result, most data curation best practices and recommendations generated by big-science research programs have not been applicable to university-based research institutes with fewer resources to develop such curation infrastructure. To address this gap, more studies and research are required to investigate disciplinary requirements for development of campus-wide data curation plans to deal with the limited resources, funding and expertise of such organizations. Additionally, identification of the practices, norms, and conditions of small-science at the disciplinary level is necessary for developing effective data services that support research processes within academic institutions.

**Problem Statement**

This study was designed to address the need for additional study of data curation requirements of scientists in small-science disciplines. The primary purpose of this study was to improve understanding of data curation requirements of one academic research institute (an academic research institute may be single or

[2] Cyberinfrastructure refers to integrated systems that include hardware, software, and human resources to support research requiring high-performance computing for modeling, simulation, prediction, and data mining; data management and visualization; virtual organizations (VOs); and educational enhancement (Conte et al., 2010).

multi-disciplinary), through investigation of data characteristics as well as data management perspectives and practices of the researchers. The investigator believed that the data curation model developed from the study's findings could determine the relevancy and applicability of those recommendations and best practices originally developed for the larger big science-oriented research institutes to a smaller academic research program.

Thus, the investigator undertook a case study of the data curation perspectives and current activities of the researchers at Kent State University's Liquid Crystal Institute (LCI). The Institute was selected as it was considered a suitable representative of academic research institutes, where scientists do cross-disciplinary research on liquid crystals. It typifies the small-science research unit, as defined above, where the scale of data produced is substantially smaller and where resources to develop codified data management policies and procedures are limited.

**Background of the Liquid Crystal Institute.** In 1965, the Kent State University Board of Trustees authorized the formation of the Liquid Crystal Institute under the direction of Dr. Glenn H. Brown, a faculty member in Kent's Chemistry Department and Regent University. Dr. Brown served as LCI director until his retirement in 1983. Other scientists at Kent joined the Institute, which sought funding for liquid crystal research. Major grants came from the National Institute of Health (NIH), the National Science Foundation (NSF), and the agencies of the United States Department of Defense. Research at the Institute, in collaboration with the Departments of Chemistry and Physics, helped establish the field of liquid crystals as an active area in both of these disciplines.

The LCI is a research-intensive institute where faculty members, doctoral students, and post-doctoral employees collaborate, with research interests varying from basic science to the development of practical application such as displays. The Institute includes a number of classrooms, 14 individual research laboratories, clean rooms, offices, a display manufacturing line, and associated service and support facilities.

# Chapter II

# Literature Review

## Research Objectives of the Study

This study sought to investigate various aspects of data curation in small-science disciplines, including disciplinary requirements for data management, data characteristics, researcher perspectives on their data management needs, as well as the researchers' actual data management practices. Thus the investigator reviewed the literature that provided her with insights into data curation activities and research, particularly for small-science research. Included in this review are disciplinary curation case studies, data curation plans, and pertinent NSF reports on development of CI.

## Data Curation and Understanding of Data Practices

Development of effective data curation plans requires that domain-specific researchers and librarians find a mature understanding of data curation requirements, practices and procedures. To stress this requirement, Haas and Murphy (2009) remind librarians that it is necessary for them to get to know the data options and obligations of the disciplines they serve and get ready to facilitate this communication before partnering with existing data sites, societies, research units, and publishers. Higgins, in her article "Digital Curation: the Emergence of a New Technology" (2011), reflects that technical development and a mature understanding of data practices and procedures is necessary for ensuring access, use, and reuse of digital data during its lifecycle. Yakel (2007) emphasizes the importance of data management in her definition of data curation: "digital curation

is the active involvement of information professionals in the management, including the preservation, of digital data for future use" (p. 335).

**NSF Data Management Plans (DMPs)**

Data management has been recognized as an important part of research activities. As a consequence, funding agencies such as the NSF are requiring investigators to present their DMPs along with their grant proposals. This raises the need to develop a good data curation plan within the scientific community, which wouldn't be possible unless collaboration between domain-specific scientists and librarians is established. In 2007, NSF mandated that all science and engineering data generated with NSF funding must be curated and preserved, and made broadly accessible and usable (NSF, 2012).

To help scientists at Purdue University meet the NSF requirements for development of DMP, the University's Libraries began an initiative that engages the faculty on the topics of data discovery, management, and organization. The librarians found that researchers uniformly expressed a need for organizing, describing, managing, archiving, and accessing data (Brandt, 2007). Starr, of the California Digital Library, mentioned that as more and more funding agencies require researchers to include DMPs with their grant proposals, "investigators are looking are looking to libraries to help with various aspects of research data management and curation, from creating DMPs to archiving and providing access to their research data" (Starr, Willet, Federer, Horning, & Bergstrom, 2012, p. 109).

**Data Management for Research Data Produced at University Settings**

DMPs (also known as data curation plans) have also been developed by universities to handle the research data generated by various academic units. Data curators often find that data management cannot be easily untangled from other research-related activities. In 2011, Fear conducted a data curation study entitled "You Made It, You Take Care of It: Data Management as Personal Information Management." In her research, she studied investigators' data management practices as well as the factors that motivate or inhibit changes to practices. Fear found that data management is "strongly connected with researchers' daily work" (p. 74), and it is confusing and counterproductive to separate data management activities from other research activities. She further argued that researchers are "involved in a range of data management practices that vary over the course of data's lifecycle" (p. 74) based on other information and documents investigators use to produce their work. She defined data management as a continuum of activities that starts with a grant proposal at one end and a publication or other final products at the other end.

Highlighting the importance of developing data curation plans in university settings, Walters (2009) described a university-based (at the Georgia Institute of Technology) digital curation program that offers a model to help librarians with the development of data curation programs. Walters explained that "the main characteristic of the program is that it is devoid of top-level mandates and incentives, but rich with independent, bottom-up actions" (p. 83). He clearly noted that a need for data curation plan has been put forth by researchers across various scientific disciplines. His study also suggested that collecting resources for

developing data curation programs at the institutional-level is challenging and incremental.

Moreover, Harvey (2010) has emphasized the necessity for development of digital curation plans within academic research units. He noted that it is very crucial that scientists provide access to their data over time by addressing issues such as intellectual property rights, and the lack of well-constructed metadata, as well as the threats such as media decay and hardware and software obsolescence that offer little hope for longevity of data.

As data management is dependent on creation of necessary infrastructure, the initial NSF CI report (which is also known as the Atkins' report) focused on developing CI for curation and management of scientific and engineering data, and emphasized five crucial aspects of curation. One of the aspects was that CI should provide "multidisciplinary, well-curated, federated collections of scientific data" (p. 7). The report acknowledged that, "absent systematic archiving and curation of intermediate research results (as well as the polished and reduced publications), data gathered at great expense will be lost" (p. 11). Additionally, it emphasized that "acquisition, curation, and ready access to vast and varied types of digital content provide the raw ingredients for discovery and dissemination of knowledge" (p. 44).

Effective data management supports data sharing and reuse, and is an integral part of data curation plans. Fry et al. (2008) stressed the importance of effective and efficient management and reuse of research data and considered curation practices as a key component in the knowledge economy in the years to come. Management of scientific data is essential for effective conduct of research as

well as its dissemination, use, and reuse. Fry et al. stated that the potential benefits of open sharing and reuse of research data include "maximized investment in data collection; broader access where costs would be prohibitive for individual investigators or institutions; potential for new discoveries from existing data, especially where data are aggregated and integrated; reduced duplication of data collection costs and increased transparency of the scientific record; increased research impact and reduced time-lag in realizing those impacts; new collaborations and new knowledge-based industries." (p. iv)

Creating data profiles/ data descriptions, which are basically detailed data about scientific data, is a very important part of the data management process. Witt, Carlson, and Brandt (2009) created a *data curation profile* "to provide detailed information on particular data forms that might be curated by an academic library." These "data forms are presented in the context of the related sub-disciplinary research area, and they provide the flow of the research process from which these data are generated" (p. 93). These profiles can be created to evaluate data curation from the perspective of data producers using their own language.

Collaboration is an important component of data curation. Harvey (2010) stated that collaboration is firmly embedded in digital curation practice. Yet, full development and deployment of CI requires significant funding levels and expertise. To accomplish the goals of developing resources, expertise, and support for e-science infrastructure, government agencies and the scholarly community have encouraged active and ongoing collaboration among many large organizations such as NSF, the National Science and Technology Council, the National Academy of

Sciences, and the Data Curation Centre in the United Kingdom (Atkins et al., 2003). Michael Day (2007) studied the reasons why collaboration is becoming increasingly important in support of scientific data curation practice and development of Institutional Repositories (IR). He recognized scientific research collaboration as a key part of the scientific research process, including the curation of the data produced by observation and experiment.

**Data Curation Needs of Small-Science Disciplines**

Data curation research is very crucial to supporting data practices of the researchers working in academic research institutes. Because of the unique qualities of researchers' data, their data management requirements must be designed according to types and uses of the data. The variation in the nature of data from discipline to discipline will necessitate tailoring of data curation services, while keeping development in alignment with the growing global e-research and curation infrastructure (Cragin et al., 2010). In their data curation work at Mount Holyoke College, Goldstein & Oelker found that effective and efficient curation of scientific data in smaller academic research units may require institutions to adopt a policy of cooperation and collaboration (Goldstein & Oelker, 2011). Building trust and establishing new relationships may aid these units in moving forward with an institutional approach to digital curation planning (Schmidt et al., 2010). Establishment of software infrastructure (Institutional Repositories, or, IR) in collaboration with other larger institutions is the most important task to be addressed in an effort to reduce the huge costs involved in such projects.

Although a number of research studies have been focused on research data produced within universities, data management (curation) remains ad hoc in practice. Most datasets produced in scientific laboratories across universities lack descriptive metadata; thus, they remain hidden within the institutions (Cragin et al., 2010). These research institutions tend to produce detached, diverse datasets that are required to be synchronized into something that is greater than the sum of its parts (data federation). Also, in contrast to large datasets, which are more likely to be highly standardized and automatically accompanied by metadata, smaller datasets lack adequate documentation (Akers, 2013). Thus, efficient curation planning for campus-wide institutes should mainly focus on developing approaches to documentation, organization, preservation, and dissemination of datasets that have no permanent home outside of the laboratories and offices in which they were created (Akers, 2013).

**Data Sharing Standards and Reuse**

To emphasize the importance of data sharing as one of the functions of data management, Harvey (2010) mentions that "active management of data for current and future use relies on effective sharing of data, which in turn, relies on agreement on and adoption of standards" (p. 96).

In the *Revolutionizing Science and Engineering through Cyberinfrastructure* report, Atkins states that in the near future data curation experts predict that mature CI will enable scientific communities to share raw and processed data easily, not only within a research group or institution but also among scholars in related scientific disciplines and locations around the world. In the report it is also

mentioned that establishment of a sustainable CI "is an exciting opportunity to share insights, software, and knowledge, to reduce wasteful re-creation and repetition" (Atkins et al., 2003, p. 12). Cragin et al. (2010) emphasizes the development of a wider range of curation services to support deposit of data into shared repositories. According to the authors, data sharing will require research communities to adopt uniform or widely applied data standards as well as disciplinary repository services.

**Cyberinfrastructure**

Data management (curation) plans cannot be implemented without establishment of required infrastructure. As noted above, Atkins et al. (2003) define CI as consisting of "software programs, services, instruments, data, information, knowledge, and social practices applicable to specific projects, disciplines, and communities of practice" (p. 5). The *NSF Advisory Committee for Cyberinfrastructure on High Performance Computing (HPC)* report suggests that CI is not all about advancement of high-performance computing (NSF, 2011b). Not only should it focus on creating capabilities that support sharing and efficient reuse of data across science research communities, but it should emphasize other important goals such as acquiring new applications and standards that promote interoperability and that can be incorporated across institutions and disciplines. Thus, a well-planned CI will ensure accessibility and availability of data acquired at great expense for future generations, and enhance research collaborations over distance and across disciplines. NSF has played a major role in the development of best practices for establishment of CI and development of data curation for scientific communities through its Advanced Cyberinfrastructure Program (ACP). "The ACP offers a

significant opportunity for research into the more effective applications of information technology and opportunities for identifying and refining its supporting CI" (Atkins et al., 2003, p. 14).

Software and data are the most important components of CI. Data and software are interdependent. To access and understand the data, software is needed. Therefore, when there is talk of data, it must be accompanied by a discussion of software and codes required to access and understand the data. The NSF, in its solicitation and funding guidelines for Sustainable Digital Preservation and Access Network Partners (also known as DataNet) envisions the creation of new types of organizations that integrate library and archival sciences, CI, computer and information sciences, and domain science expertise (NSF, 2012). The *NSF Task Force Report on Software for Science and Engineering* states that "software is a critical and pervasive component of the cyberinfrastructure for science and engineering" (2011a, p. 4). It is also mentioned in the report that development of software is of paramount importance since "it binds together the hardware, networks, data, and users such that new knowledge and discovery result from CI" (p. 4). According to the report, development of software infrastructure is critical for supporting knowledge generation in a shared and collaborative system. Good software infrastructure should be capable of meeting present and future or unexpected needs of its community.

**Data Curation Case Studies**

The Digital Curation Centre (DCC) has emerged as a critical source of information about curation practices in research environments. Through their

Disciplinary Approaches to Sharing, Curation, Reuse and Preservation (SCARP) Project, the DCC conducted immersive case studies that focus on studying disciplinary differences in attitudes and approaches to data deposit, data sharing and reuse, and curation and preservation. The main goal of the project's case studies is to study investigators' perspectives and practices, and the tools and methods they use in curating of research data in order to identify and promote good data curation practice. The disciplines that were studied for the purposes of the project include art and humanities, social sciences, life sciences, and physical sciences. Within the physical sciences, astronomy, climate science, and crystallography were selected for further investigation (Key Perspectives, 2010). Below, two of the SCARP case studies, as well as another case study conducted at Johns Hopkins University, were examined for their relevance to this study on curation practices in smaller institutions.

The case study of Edinburgh Division of Psychiatry's Neuroimaging Group reflected on how the nature and size of the datasets and their uses by investigators influenced curation practices and policies. The study demonstrated that "how investigators and investigators heedful attention to each other's data underpins curation" (p. 5). Because of the nature of this field of research, neuroimaging demands continuous care of increasingly large and dynamic datasets. Some of the data is being shared through e-science projects "aiming to provide federated data storage and improve data integration" (p. 5). Data reuse is very important in the areas where novel analysis techniques are used to identify patterns in images or in the associated clinically-related and demographic data on subjects. To facilitate

effective reuse of the data, “documentation and metadata on research subjects and on analytic protocols is encouraged through curation practices" (Whyte, 2008, p. 8).

The SCARP case study in engineering research field provides some additional insights about data deposit, sharing, and reuse. Studies on communication patterns in this area revealed that “the care and use of knowledge, information, and data are embedded in the social processes that engineers use to do their work” (Ball & Neilson, 2010, p. 11). Data sharing in the engineering field is conducted through personal contacts and informal channels. Due to the nature of engineering data, rationale, methods, and analysis used to produce data are as important as the data itself and must be documented and considered for sharing along with the data. Intellectual property rights may be an issue in this field, as the methods used in engineering research and design, in commercial context may be considered intellectual property and as a result impose some limits on data sharing. Moreover, in the engineering field, asset management is considered a very important practice and needs to be addressed through data curation practices (Ball and Neilson, 2010).

The SCARP case study findings suggested that curation practices vary in different disciplines and that curation tools such as the Digital Curation Lifecycle Model provides useful models for good data curation practices. The most critical finding of these studies was that a generic approach to data curation is not applicable to all disciplines; therefore, it is crucial that each research community develop its own domain-specific strategies based on the local investigation of the investigators’ needs, expectations and data practices (Key Perspectives, 2010).

A third relevant digital curation study, conducted at Johns Hopkins University, investigated the Institutional Repository (IR) developed as a component of overall digital library architecture to support preservation of data (Choudhury, 2008). As Harvey notes, "repository software is aimed at allowing more efficient management of institutional assets and providing fast, easy access to its contents from remote locations" (2010, p. 192). The IR represents services that support data curation as part of evolving CI featuring open modular components. The Johns Hopkins study showed that establishment of IRs is essential for data management efforts to be successful and they must be integrated into a larger landscape of repositories that serve a distributed and diverse academic community.

This review of literature pertinent to data curation in small science settings provided the investigator with a broad understanding of data curation activities, as well as the significance of practices such as data sharing and reuse, and collaboration for scientists across various disciplines. The NSF reports gave her insights into the infrastructure required to support such activities. In addition, the case studies provided an overview on the importance of disciplinary data curation research within various disciplines. The studies demonstrated that each discipline or a group of disciplines working on similar problems have their specific requirements that need to be investigated closely. The next section discusses how this context relates to the objectives of the study.

# Chapter III

# Methodology

The main goal of this study was to study the data requirements of researchers at the LCI in order to develop a data model that illustrates how disciplinary requirements for data management, shaped by characteristics of the data, affect the researchers' perspectives on their data management needs as well as their data management practices. Kent State University's LCI was chosen as an appropriate site for the purposes of conducting this case study because it meets the definition of a university-based institution, in terms of its goals and objectives, infrastructure, and data practices. This study enabled the investigator to:

- Articulate the perspectives of LCI researchers on data management (curation) concepts;
- Identify and understand discipline-specific data curation practices;
- Develop a clearer understanding of the data in this institutional setting; and,
- Develop a model of the data world of liquid crystal scientists.

A combination of data collected via a questionnaire instrument (six researchers participated in the questionnaire) and interview instrument (five researchers were interviewed) yielded relevant data bearing on the central research questions of the study:

- What are the disciplinary requirements for data production, management, and preservation that influence data curation activities within the LCI research laboratories?

- What are the LCI researchers' perspectives toward data curation practices such as data deposit, sharing and reuse, curation, and collaboration?
- How do LCI researchers manage their data within their laboratories? What are the established practices for data management, if any?

Answers to the above questions were sought by collecting and analyzing empirical data on the researchers' data management requirements, perspectives, and practices via interviewing and online questionnaire techniques, as detailed below.

**Overview of the Study's Methods**

This research employed a case study approach to study and learn about discipline-specific data curation requirements, and perspectives and practices of the researchers in one small science branch. Therefore, the LCI was selected as an exemplar of academic research institutes where scientists from multiple small-science disciplines conduct research.

Interviewing and online questionnaire techniques were utilized to collect data for the purposes of this study. The investigator designed an online questionnaire for the purposes of collecting preliminary data that could provide background information and context for follow-up interviews. The interviews were carried out to further explore issues and concerns about data curation that had not been addressed fully in the questionnaire responses.

The participants in the study were recruited from a population of 23 faculty members who were primarily principal investigators of research at the LCI, and 57

graduate students and post-doctoral employees. To recruit participants for this study, the investigator released the questionnaire to the twenty-three faculty members and three graduate students on August 16th, 2013. The investigator also sent a follow-up email to generate additional responses a week later. After the initial data collection via the questionnaire had concluded, the investigator contacted LCI faculty members to schedule follow-up interviews. At the end of this phase, the investigator had gathered questionnaire data from six researchers and had conducted five interviews. Following the conclusion of data collection, the interviews were transcribed, and all questionnaire and interview data were coded, categorized, and analyzed using qualitative data analysis software. Further details about each aspect of the methods employed by the investigator are provided below.

**Data Collection Methods**

The investigator employed a two-pronged strategy involving both Web surveying and interviewing techniques to collect empirical data on the curation practices and perspectives of researchers at the LCI. After identifying willing participants for the study, she used an online questionnaire conducted via the online survey service Qualtrics to gather preliminary information from employees at each research unit and used the results of the questionnaire to inform follow-up interviews with those researchers. These interviews further explored issues and concerns about data curation that had not provided in questionnaire responses. By employing these data collection techniques, the investigator had two sources of information to help her understand the differences between recommended best practices and actual data management/curation practices in researchers'

workplace. The investigator used the questionnaire instrument developed for this purpose (see Appendix A).

The interviews were conducted based on a set of pre-determined open-ended questions to gain a sense of typical data curation practices and needs within the institute. All interviews were recorded and took about 30 to 45 minutes. The investigator used the interview guide developed for this purpose (see Appendix B). The interviews were informal and open-ended, and carried out in a conversational style. The investigator also took notes in conjunction with the recordings, to have a backup in case of equipment malfunction.

Both the interview and questionnaire questions were designed based on the investigator's understanding of digital curation processes, as gained from a review of the literature on data management and digital curation. To ensure that the instruments included appropriate and effective questions, all questionnaire and interview questions were pretested on several graduate students within the Institute prior to the formal launch of the questionnaire and conduct of interviews with senior investigators.

**Recruitment of Participants**

As noted earlier, this study was designed to investigate data curation practices and perspectives of the researchers at the LCI. Liquid crystal science and technology has always been driven by interdisciplinary research efforts involving expertise across the boundaries of biology, chemistry & materials science, physics, mathematics, and device engineering. "The activities at the Institute are focused on the forefront of this cross-disciplinary area, transcending research from individual

scientific fields and streamlining inquiry, from basic science to the development of practical applications" (Liquid Crystal Institute, 2013). Therefore, the results obtained from this research were valuable to data curation research, as previously published case studies were not focused on data curation requirements for a multi-disciplinary domain.

The participants in this study were recruited from the roster of researchers employed at the LCI. The populations from which individual participants were recruited included 23 faculty members (who were mainly the Principal Investigators of research at the LCI), as well as 57 graduate students and doctoral employees. The initial objective for recruiting participants was to survey all the faculty members and interview up to three researchers from at least half of the fourteen laboratories at the LCI. Most of faculty members have their own laboratory, in which 5 to 10 graduate students, doctoral employees, and post-doctoral employees work, on average. Each of the participants belonged to a laboratory or research project within the institute.

The faculty members and doctoral students who participated in the study represented six out of fourteen laboratories within the LCI. The investigator recruited no participants from the same laboratory so that she could have samples from a wide spectrum of research areas, from theoretical to applied physics and chemistry. Table 1 summarizes the participation of researchers in the interview and questionnaire phases of this study.

**Table 1. Summary of Researcher Participation in the Interview and Questionnaire Phases**

| Participant | Title | Completed Questionnaire | Interviewed | Represents a Laboratory |
|---|---|---|---|---|
| 1 | Faculty member | No | Yes | No |
| 2 | Faculty member | No | Yes | No |
| 3 | Faculty member | Yes | Yes | Yes |
| 4 | Faculty member | No | Yes | Yes |
| 5 | Faculty member | Yes | Yes | Yes |
| 6 | Faculty member | Yes | No | Yes |
| 7 | Doctoral student | Yes | No | Yes |
| 8 | Doctoral student | Yes | No | Yes |

## Data Collection Activities

The investigator started data collection by releasing the questionnaire to faculty members of the LCI on August 16, 2013. This initial invitation resulted in the completion of the questionnaire by one faculty member. To generate additional participation, a follow-up email was sent on August 24, 2013 to the rest of the faculty members and several graduate students who the investigator perceived to

be more knowledgeable about the data practices of the laboratories where they conduct research. After the second invitation, three principal investigators (PIs) and two doctoral students participated in the questionnaire, bringing the total number of participants to six. The doctoral students who participated in the research were independent researchers working on their own research projects. When the questionnaire participants were contacted to schedule a follow-up interview, only two of them agreed to participate. The rest of the researchers expressed concerns about the length of the questionnaire, so they declined to take part in the study.

To recruit more researchers, the investigator decided to meet those individuals who had not agreed to participate at their offices, in order to solicit recommendations for other faculty members who might be amenable to being interviewed for the study. This technique was successful, resulting in three additional researchers agreeing to participate in a full interview. Moreover, she had two short conversations with two more researchers that resulted in usable data. The data gathered through these conversations corroborated many of the findings about data management practices and perspectives that the investigator documented in those full interviews with participants. At the end, the investigator was successful in collecting data from a total of five interviews with faculty members and six completed questionnaires from faculty members and graduate students (as noted in Table 1).

**Maintaining Confidentiality of Participants**

Questionnaire data was collected and maintained on the server of the Qualtrics surveying service. For the interview data collection phase, the investigator

stored the data on her personal computer, and all folders that contained the data were password-protected to ensure that the data was not accessible by unauthorized users. In the transcription phase, the investigator excluded any identifying information to avoid breaching the confidentiality of the participants. This report on results of the study uses pseudonyms for the names of all researchers in both paraphrasing and direct quotations from transcripts.

**Data Analysis**

For analysis of the questionnaire data generated by the Qualtrics survey, the investigator exported the data from Qualtrics server to Excel and created summary tables for each question. For any open-ended questions, NVivo was used to categorize and analyze the responses along with the interview data. The survey data was primarily used to make profiles for each participant (see appendix C for complete list of participant profiles).

For analysis of the qualitative data collected in the interview phase of the study, the investigator transcribed the interviews so that the typed text could be easily imported into NVivo. The software helped the investigator organize (categorize using codes) and analyze (find patterns and relationships among concepts) non-numerical or unstructured data to generate the study results.

**Development of a data model for data curation research.** After the transcribed data was imported into the NVivo software, the investigator started to code data, using the 28 predetermined headings and themes (see Table 2). The themes and headings for initial data analysis drew on concepts from two sources:

(a) the literature review; and, (b) the research questions discussed in the Research Objectives section above.

**Table 2. Proposed Initial Code List for Analysis of Interview Data and Unstructured Questionnaire Responses**

| Headings and Themes |
| --- |
| Data Curation |
| Organizing |
| Archiving |
| Data sharing |
| Collaboration |
| Preservation of reliable and authentic electronic records and digital objects |
| Collecting resources for developing data curation programs |
| Application of technological standards |
| Intellectual property rights |
| Social processes |
| Data description/metadata |
| Data storage |
| Data |
| Large-scale data federation across multiple disciplines |
| Data verification, validation, sustainability and reproducibility |
| Data discovery |
| Data modeling, representation, and exchange |
| Accessibility and availability of data |
| Knowledge discovery |
| Inherent fragility and evanescence of media and files |
| Effective reuse of the data |
| Cyberinfrastructure |
| Computational simulation |
| Shared vocabularies |
| Application of specific data formats |
| Consistent open policies |
| Open software |
| Rapid obsolescence of hardware and software |
| Inherent fragility and evanescence of media and files |
| Institutional repositories |
| Software infrastructure |

Relying on these initial codes helped with retrieval and analysis at the beginning of the analysis phase. Since the investigator had planned to alter those categories during focused analysis, as the coded data accumulated the investigator added 38 more codes as needed, and renamed some codes to more accurately reflect language used by researchers to refer to certain data management concepts. The more the investigator proceeded with the data analysis, the more she adapted her coding strategy. She found numerous instances where codes with similar meanings needed to be merged and where related codes needed to be grouped and organized hierarchically.

After assigning all the codes to the data (including the initial and new ones) and reviewing each one closely for relationships and potential overlap among them, some of the codes were merged together. As a result, the number of codes was reduced to nineteen. Later, the coded data was grouped into three categories: (a) Researchers' perspectives on data characteristics, (b) Researchers' perspectives on data management, and (c) Researchers' data management practices. After classifying the data into the three categories, the investigator tried to link classes with simple statements that expressed the connections. By defining the data categories and establishing the relationships among them, the investigator thus developed an explanatory model for the data world of liquid crystal scientists.

Through close investigation of the categories and their relationships, the investigator realized that these three categories could influence each other in certain ways. She found that the data characteristics are dependent on the discipline where the data is generated. She noticed that because of the nature of liquid crystal

discipline, the data produced by the scientists have specific qualities such as quantifiability and replicability. She also found that the liquid crystal research is not very data-intensive. Working with the data gathered under the cluster of Researchers' Perspectives on Data Management helped the investigator realize that the data characteristics could have a mutual influence on the researchers' perceptions on their data management needs. In other words, the data management requirements of the researchers are determined by the quality of the data produced within their discipline. Finally, she realized that data management practices of the researchers are informed and influenced by both the data characteristics and the researchers' perceptions on their data management requirements. All these connections led to the development of the model for the data world of liquid crystal scientists.

While Table 2 included the initial headings and themes that were employed by the investigator at the beginning of data analysis, Table 3 contains the finalized codes that were derived from both the initial codes (developed at the proposal phase) and the main themes in the transcribed interviews. Some of the codes in this table reflect the major concepts in data management (curation) that were explored by the investigator.

**Table 3. List of Finalized Clusters and Codes for Analysis of Interview Data and Unstructured Questionnaire Responses**

| Clusters and Codes |
|---|
| Perspectives on Data Characteristics |
| Data quantifiability |
| Data Replicability |
| Data Reusability |
| Perspectives on Data Management |
| Controlled Vocabularies |
| Data Deposit |
| Data Sharing |
| Data Licensing |
| Data Preservation |
| Intellectual Property Rights |
| NSF Data Management Plans |
| Data Management Practices |
| Controlled Vocabularies |
| Data Deposit |
| Data Documentation and Metadata |
| Data Licensing |
| Data Preservation |
| Data Retrieval |
| Data Selection and Appraisal |
| Data Sharing |
| Intellectual Property Rights |
| Security Measures |

**Definitions of the codes in Table 3.** As mentioned earlier, some of the codes in Table 3 were created based on the main data curation concepts that appeared in the transcribed interviews. As developing an understanding of these concepts or codes are essential to understanding of the significance of this study, the investigator has provided the readers with some explanations of why these concepts

were studied and definitions for these codes. Codes may appear in more than one cluster, as they may be used in reference to either researchers' perspectives on a particular concept or to the actual practices of the researchers as they were reported to the investigator.

**Controlled vocabularies** are used to increase consistency in data so that there is a shared and common language among the researchers working on similar problems. Using controlled vocabularies also enhance quality of metadata. Therefore, in a standardized data sharing, controlled vocabularies ought to be developed and used among researchers, and collaborators within and across disciplines. Achieving agreement over standards, especially metadata vocabularies such as controlled vocabularies and ontologies, is arguably the greatest challenge of all disciplines.

**Data deposit practices** encompass the LCI researchers' actions to preserve their data by uploading to an IR. These practices are very important as they support accessibility and availability of data over time. In standardized DMPs, research data are deposited into an IR. The NSF (2011a) reports that the development of digital repositories is integral to data deposit practices as "it binds together the hardware, networks, data, and users such that new knowledge and discovery result from CI."

**Data documentation and metadata** "is an essential ingredient of the management (curation) process. Descriptive information and classification labels that group related items provide context and facilitate the reuse of specified research output"

(NISO, 2013). Data documentation will enhance all the data management stages. Without describing and contextualizing research data, it becomes challenging to retrieve, access, share, or make sense of research data. Data should be accompanied with metadata when it is first created and stored. The researchers, who have responsibility for storing the data, should take the significance of this stage into account.

**Data Management Plans.** A primary responsibility for data librarians in the digital era is to provide a range of services associated with management of data within and outside their institutions. Recently, there has been an increasing demand for librarians to advise researchers and to provide practical help with DMPs. This evidence confirms the NSF statement (2007) that the development of DMPs will not be possible unless collaboration between domain-specific scientists and librarians can be established.

**Data quantifiability** refers to the quality when data can be reduced into functions, equations and summaries. This definition emerged from the data collected for the purpose of this study rather than being predefined in the literature (see next chapter for detailed explanation).

**Data preservation** is mainly dependent on whether data is stored and managed according to the existing best practices and standardized procedures. Therefore, data deposit and data preservation concepts are very dependent on each other in

data curation activities. Preservation planning has become more prominent and essential as part of data management processes, as digital assets are exposed to several threats such as technology and format obsolescence, and media decay that may endanger the existence of data.

**Data replicability** refers to the quality when data can be regenerated by the computer codes built by scientists.[3] This definition emerged from the data collected for the purpose of this study rather than being predefined in the literature (see next chapter for detailed explanation).

**Data retrieval:** One of the most critical functions of Institutional Repositories is to organize and facilitate timely access to data. Most of these systems are meant to provide users with enhanced searching and data retrieval capabilities.

**Data selection and appraisal** involves the development of criteria for determining which data should be kept for long-term and which data should be discarded.

**Data sharing** is important as it is a prerequisite to research data use and reuse, and as Harvey (2010) mentions "active management of data for current and future use

[3] A code was defined as a "program" written to simulate a system, solve some equation numerically, or to analyze the data they have taken via an experiment.

relies on effective sharing of data." In data sharing practices, some tools (such as emails, digital repositories, etc.) are used to share data within and across disciplines.

**Data storage** is one of the most critical stages of data management in a sense that other data management practices (such as documentation, preservation and dissemination) are highly dependent on how data are stored. Akers (2013) explains that it is critical to develop approaches to documentation, organization, preservation and dissemination of datasets that have been generated as a result of research in small-science disciplines. Quality, sustainability, accessibility and availability of data in the future are directly associated with how research data are stored and managed during the course of research. As Atkins (2007) in his report emphasizes that management is dependent on creation of necessary infrastructure, developing and storing data in an IR assists scientists to better organize, describe, preserve and provide access to their data.

**Data reusability**. The point of curating data is that they remain available for use and reuse by legitimate users. In this study reusability is defined in relation to characteristics of the data such as quantifiability and replicability. The participants explained that the as the data can be quantified and replicated, it would not be reusable to other researchers.

**Intellectual property rights and data licensing** are two of the main issues that must be addressed in data sharing. As data cannot be copyrighted, scientists often

find it difficult to share the data that they perceive of as their intellectual property. Data licensing refers to the process through which a license, that is a legal instrument for a rights holder to permit a second party to do things that would otherwise infringe on the rights held, is obtained for data sharing and reuse. Data licensing is a need that arises directly from this trend towards the planned release of research data. Licensing data can be complicated by the fact that different aspects of data, such as field names, the structure and data model for the database, visualization and reports derived from the data, may be treated quite differently.

# Chapter IV

# Research Findings

In this chapter, the participants' opinions on characteristics of their data and their data management needs as well as their actual data practices are discussed. Early on in this section, the issues surrounding the data sharing concept are reviewed, followed by data deposit and data preservation issues. The investigator has created pseudonyms (R1, R2, …) to refer to the participants instead of using their actual names (please consult Appendix C for more detail on the researchers' backgrounds).

## Data Management Perspectives and Practices

As characteristics of data vary significantly among disciplines, there is considerable variance in data management requirements. In other words, the unique qualities of data within a specific discipline could result in different data management perspectives and practices. The findings of this study indicated that unique characteristics of the data have an impact on perspectives of the researchers on data management concepts, and both the data features and the perspectives of the researchers have affected their data management practices.

## LCI Researchers' Perspectives on Data Characteristics

The analysis of the data collected for this study determined that the non-observational data produced as output of the research within the LCI has specific characteristics. These characteristics are different than those of observational data produced within disciplines such as astronomy, astrophysics, and oceanography. Many of the researchers explained how their data practices have been affected by

the type of data they produce. Quantifiability was named as one of the data characteristics, that is, that the data can be reduced to equations and summaries. As a result, the researchers do not find their research to be data-intensive. In addition, replicability was mentioned as a unique quality of the data, meaning that the data can be recreated by using the computer codes the researchers build. The participants believed that these two characteristics of the data have reduced the reusability of their raw data. As a result, researchers do not value their raw data, because it is simply output from their codes. They also explained that they have no need to save it, as the data can be recreated. The researchers value the computer codes and articles they write, but they do not see these as data. Also, they do see the value of organizing and storing their research notes, even though they do not see them as data. The perspectives of LCI researchers on the data itself color their beliefs and practices in all aspects of data management, as will be seen in the discussion below.

**Data Management Perspectives and Practices of LCI Researchers**

In the following section, the researcher's data sharing perspectives, practices, and the issues surrounding data sharing are discussed in detail.

**Data sharing.** Sharing data is an essential prerequisite to its use and reuse. Therefore, it was very important for the investigator to study the LCI researchers' data sharing perspectives and practices. The interviews revealed that researcher data sharing practices are affected by the features of non-observational data and the researchers' perspectives on data sharing. As the researchers believe that the data they produce can be quantified and used in their articles they write, they see no

value in sharing it. The next two excerpts show how researchers' perceptions of the nature of the data have affected their perceptions on data sharing practices. R4 felt that there is a difference between the disciplines that produce data through observations and the ones such as liquid crystal science that generate data though models. He believes that the simulation data LCI researchers generate through their computer codes has no value in itself. Rather, these simulations are the means of generating results, as they are reducible to mathematical functions or algorithms, and the summaries are then used in publications. He emphasized that these data characteristics could influence scientists' data sharing practices within each discipline:

> In sciences such as oceanography, it would be difficult to quantify the results so succinctly, so, you cannot toss all the data. You have to keep the data. You wouldn't toss the data [that] you spend [a] million dollars of the NSF money, and you want to make them available to people. We are not in that kind of deal.

Replicability was also mentioned as another feature of the data produced within the LCI. In liquid crystal research, data can be regenerated as long as the computer codes and software that produced the data are available. This means that the exact same data can be replicated whenever a program is run by a researcher. This quality of the data has had a major impact on the researchers' views on data sharing,

> Q: Are you willing to share your data with other researchers working on similar projects?

> R4: A lot of data [generated in my field] is simulations, based on the software. We can share the software that generates the data. And so, somebody else can regenerate data by changing the algorithms. So it's the simulation software that is of value to share. We can always regenerate the data.

R2 noted that she writes her own codes to generate data. She explained, "My research is very data-intensive in a sense that when we write our code and run it, it is just numerical data. Then the numerical data has to be visualized." She views the data as something that is not worth sharing. This perspective on data sharing is directly dictated by the data characteristics.

R4 believes that if data are to be shared, its usefulness should be taken into account; otherwise, there is no point in sharing it. He explained that because in his field data can be regenerated by the computer codes and because the most significant data is used in his articles, even if the data is made accessible, it would not be of any use to other researchers. He noted:

> Sharing my data depends on their usefulness. If somebody tells me that I should share the first two drafts of my book, sharing my data would be that kind of thing. All the data we generate is not something we would like to track.

The researchers interviewed for this study expressed that data sharing is not very common within their discipline, because they do not see the simulations generated from the codes as data. Therefore, asking him about his data sharing

practices seems like an error. In the following excerpt R1 explains how data sharing is viewed in his field:

> It is such a weird hypothetical question. It is like asking me, "would I share my tooth brush with somebody else?" In theoretical physics it is weird to share data. In other branches of science it may make more sense to share their data. I would share my data with the public only if I was required by funding agencies. However, there is no such requirement yet …

Many of the researchers felt that they had proprietary rights to their data. R3 explained that "I do not want to share it [my data] with other researchers, because I want to use them myself in other projects."

Since the researchers do not perceive their data to be valuable, or do not perceive value in sharing it, they believe that only the results drawn from the data should be made available to other researchers. When the investigator asked the researchers about the most desirable time at which they would prefer to share their data, all of them responded that they do not like to share their data at any point before the publication of results.

R5 explained that, if they were required to share their data, "we will be willing to share our data only after the paper is published. After the data is published I have no problem to share the data with the public." He believes if data are to be shared, it would be better to be made available by the publisher as supplementary material. He gave an example of one of his articles published by Elsevier. He noted that, "the Article of Future format of Elsevier allows users to

access data points directly from the published graph," which facilitates data sharing at the publication level. R5 has developed a DMP. In his plan, he mentions that:

> Additional data [the data that is not used in the articles] generated during the course of this project will be made available to the general public within six months after publication by archiving it to storage in the cloud (Google Drive or Dropbox) in folders with free public access.

R1 stated that the most valuable gain of their research is their articles, not the data. He reported that,

> In my field, the great significant science is getting into articles. From my point of view, articles stand as last thing recorded. I recognize that there are other scientific fields where that's not the case, [meaning] there are some raw data produced that does not get into articles and needs to be preserved.

He believes data sharing is more common within the disciplines where all the data produced during the course of research is not used in the publications. He explained that as all the data produced in his research is used in articles, sharing data at the publication level is what he would prefer to do:

> [In my research,] data is usually shared at the publication level … I can hardly remember a situation where I was asked to share my data. We put lots of effort to write a paper. So, our results are the most valuable thing that is produced. If I want to go back to find something, I would go back to look at my articles rather than the notes.

Given the nature of the data, several researchers view their data sharing activities exclusively as sharing the articles with the public through publications.

***Data sharing in practice*.** The data collected for the purpose of this study suggests that the LCI researchers primarily use cloud-based, password-protected systems such as Dropbox, Evernote, and Google Drive to share their codes and data with their collaborators and students. R2 explains that, "I have shared even my codes, but usually with the people with whom I collaborated." Although most of the interviewees agreed verbally that data should be made available to the public, in reality, they believe that the situation is more complicated. R1 produces some of his data by hand, using pen and paper. To share it with his students, he scans all the papers and puts them on Evernote. He has also used email and Dropbox as ways of data sharing. R2 uses a combination of Dropbox and Google Drive to share her data. She explained that "we put our data on KSU Dropbox, especially if we want to share them with a colleague. We sometimes use Google Drive to share our data."

Studying the data sharing practices of the researchers revealed that the researchers are hesitant to share their data and codes with the public, except for the results derived from the data. The researchers are usually in the habit of sharing data with their students and collaborators using cloud-based storage and email.

The following table summarizes how different research outputs could influence the researchers' perceptions of data sharing and as a result, their data practices.

**Table 4. Data Sharing Summary Table of Scientists at LCI**

| | Research Outputs | | | |
|---|---|---|---|---|
| DM Perspectives & Practices | Codes[4] | Output data from the codes | Notes | Results (Publications) |
| Intellectual Property Status | Belongs to the researcher/can be patented and sold | Belongs to the researchers/can be licensed if required to be shared | Belongs to their students/No comments on their IP status | Belongs to the publisher |
| Value to the Researchers | Valuable | Not as valuable, as can be replicated by the code/reduced into summaries used in articles | Valuable | Valuable |
| Data Sharing Practices | Not shared with the public/only shared with students and collaborators | Not shared with the public/can be shared if required by funding agency | Usually kept as research notebooks/ if scanned can be made available via cloud storage | Should be shared through the publisher |

***Intellectual property rights.*** Within the LCI, the researchers use the codes that they have written themselves to generate data (simulations). As they put a lot of energy and intellectual efforts in writing these computer codes, they believe that they are the ones who must be the true owners of the codes and data. Many of the researchers revealed their concerns about the legal issues that come with sharing

[4] A code was defined as a "program"written to simulate a system, solve some equation numerically, or to analyze the data they have taken via an experiment.

the raw data. Although all of the researchers expressed their agreement that the results of publicly funded research should become public property and be shared, they were still hesitant about sharing the data.

The interviewees addressed the issue at the data level as well as the publication level. R2 explained that:

> Yes, I agree that the work that is generated with public dollars should belong to the government and the people who are members of that government. However, there are many commercial products and intellectual properties produced at the university; spin off commercial products for sell[ing] and business.

R6 stated that, if data is to be shared with the public, then those who take advantage of other researchers' data should be required to give credit to those who actually own the data. To address the legal issue of data sharing, he suggested that he would request co-authorship for those responsible for the generation of the data.

The interviews revealed the researchers' concerns over the legal and commercial issues of sharing their data. R2 differentiates between the computer codes and the data generated by the codes. She believes because she develops her codes, they are considered her intellectual property; therefore, she only shares them with her collaborators:

> Let me distinguish between these two things. I have the code that is developed by me, and the data that is generated by the code. The code itself potentially has commercial value. We describe the way the code works (the algorithm) in the paper that we publish, but the code itself,

> we could patent and sell it later. I usually share my codes only with my collaborators. For instance, I had a colleague at the University of Massachusetts at Amherst; she wanted to use my code to do some calculations. So, I sent one of the students to visit, so that they work on it together. So, sometimes the human element accompanies the data and says, "let's set this up for you."

R4 believes that all the good data produced by his computer codes is used in his articles, therefore, he does not usually share his data. He explained that the only way of giving access to his data is through his published articles. That is why he only worries about the issue of intellectual property rights at the publication level:

> If I want to share the results, it is the journal's decision [whether or not they want to share it]. That's a copyright issue. If the paper is published by the journal, they own the copyright, so it would be up to them to decide to make it public. If the journal reviews the paper and makes it better, that process belongs to the journal. If I want to share the paper I can only share the version before submission to the journal.

***Data licensing.*** As the researchers expressed their concerns over the issues of copyright and intellectual property rights, the investigator was interested in learning about the researchers' data licensing perceptions and practices. Many of the researchers explained that they are not concerned about development of specific procedures for licensing their data, as they do not share any data with the public. Although the researchers do not have data licensing procedures in place,

most of them felt that they may be required to develop such policies and procedures as part of their data management practices in the near future. R1 explained his view on data licensing as described in the following excerpt:

> I had never thought about data licensing. I have thought about it the other way around: me making use of other people's data. I am very careful to always give credit to people in my work even if I am only making a Powerpoint talk and I am using a picture from Wikipedia. I use the URL to that page [to give credit to the author of the article].

He believes that in the disciplines where there are lots of patent royalties at stake, people have to be careful with legal requirements. He noted that since there is not a requirement in his field, data licensing has not been considered such a critical issue. In contrast to what R1 believed, R2 noted that because of the NSF requirement for data sharing, "we may have to license our codes and data in the future." She explained that if such a requirement is mandated by funding agencies in the future, she will certainly develop procedures to license her data. R6 stated that if data is to be redistributed, it should be made available with a proper license.

**Use of controlled vocabularies.** As a tool that improves the quality of data sharing, the investigator tried to learn if controlled vocabularies have been developed and used by the researchers. The responses suggest that none of the researchers have ever felt a need to develop a controlled vocabulary, because the benefits that they can receive through application of common vocabularies are still unknown to them and have not been recognized by the researchers.

> Q: Have you ever felt the need to develop a shared vocabulary for increasing consistency for better data sharing?
>
> R5: We have not developed any controlled vocabulary yet, [and] I do not think if a controlled vocabulary was developed, it would help with data sharing, because my data are composed of basically codes written by myself.

Given researchers' data sharing perspectives, it is unsurprising that they do not see value in developing controlled vocabularies for their data management purpose.

**Data documentation and metadata.** To learn about the researchers' data deposit practices more in detail, the investigator inquired about how they describe their research data. R5 explained that all the data stored in folders contains a machine-generated metadata file detailing the specifics of the digital data stored in the file (metadata is generated automatically at the time the data is first saved in the folders). As he uses cloud-based storage spaces for his data deposit, he uses the tagging system provided by the service to label his data. R1 explained that he uses different ways to describe his data. He believes that in disciplines where various kinds of data are produced, sticking to a standardized method of describing data is difficult.

> There are branches of science where the kind of data that you are accumulating is always the same, the way that I perceive it…. But, here we keep doing different things, and so I don't have a standardized, computer[ized] way to describe my data. The manual note-taking,

> making up descriptive file names, and having headers in files are the way we describe our data.

R2 explained that she does not have any standardized procedures for describing her data either, and that her data is only accompanied by machine-generated metadata:

> We do not describe our data. When data are generated in XYZ format for visualization, since it's a very specific format it tells you detailed information about the composition of data, but it doesn't tell you anything about the content of the composition. There are multiple output files that come with the code, the XYZ files, which are exclusively for visualization; there are other data that are generated that might be ways of analyzing and characterizing the data.

As above discussion reveals, computer-generated metadata usually contains technical information about data. If the information about the content of data is to be provided, it must be supplied by the scientists themselves.

R5 stated that the metadata produced by the computer cannot be used for retrieving the data, as it is generated and stored in a separate folder. He defined metadata as information about the conditions of the experiments, rather than information about the data files. He described his views on the application of metadata: "Metadata is important for the data itself, but it is not important after we have analyzed the data. After we do base-line corrections, the metadata becomes important. It is very important when you want to do something with data."

R1 stated that "we do not describe our data. We store our data in the basic formats of the software in our computers." R6 does not believe in the provision of

metadata for his data, he explained, “because the meaning of my data sets is clear in my research; they do not need to be described.” The most important function of metadata is that it contextualizes the research data. Problems may arise later, if the researchers do not capture the correct context as they record their experiments and acquire the data.

***Note-taking perspective and practices.*** In most of the scientific laboratories, note-taking is considered to be an important research activity. Scientists are advised to keep constant track of their data. The notes taken by researchers contain a lot of information about the data and are used in the data quality control process. R1 explains why note-taking is considered an important activity:

> New students may not take good notes of their data, so as a result the data may get lost. For instance, in my research I would write down which calculations are associated with which output data. The beginning students think just they can remember it, that's the situation where it is sometimes a problem.

R2 explained that recording the metadata about the codes is an essential part of scientific research. She has a notebook to record all of the information about the data she produces:

> We are generating the data that can be used in publication, then we need to document them. We need to write down the version of the code. I used to keep all these information in a notebook on paper. Now I need the electronic version of the notebook. Because papers can be lost, burnt, stolen, or left behind.

R5 noted, “If we don’t write down the metadata, then it would be hard to trace the data back.” He explained that in his laboratory metadata is generated both by researchers via the note-taking process and by the machine (the computers). He actually showed his laboratory notebook to the investigator and mentioned that it is really important to keep it in a safe place.

R1 explained that one of the problems with novice researchers is that they do not know how to take good notes. New researchers are constantly encouraged to develop good note-taking habits and document sufficient information about their data and research processes. One of the graduate students who participated in the questionnaire explained that researchers should document information such as where the data is coming from, in which experiment and the setting where the data is produced, what material or tools are used to produced data, for what purpose data is produced, and at which stage of research the data is generated.

**Data storage and preservation.** The investigator already knew that the Institute had not developed a digital repository prior to beginning the study. Based on that, she was specifically interested in discovering the researchers’ perspectives on developing a data repository within their institute. Most of the researchers who participated in the survey expressed the opinion that if their institute had had a central repository system, they would contribute data to it. They named some of the advantages of depositing data in a repository, including increased security of data, increased ease in finding and understanding data, enhanced compatibility with federal agencies’ requirements, and facilitation of data maintenance. R1 had a different view than rest of the researchers about the development of a repository

within the Institute. He noted that because his field is not data-intensive, there is no need to develop an IR for data storage.

> I don't think that's necessary [to develop an IR]. I think that it's something that varies a lot from one scientific field to another and in our field, I think we can get by with sort of standard software packages such as Evernote and Dropbox to maintain records, but I could see how there could be other scientific fields where massive collections of data are produced that required special services.

He believes, however, that "storing data in a repository more complies with the NSF rules for data management and it would be a reasonable way to organize data." He noted that he would deposit data in a repository only if he was required to do so. However, he doubted that in his field a data repository is necessarily required. Instead, he suggested that development of a repository that holds articles makes more sense in his field. This view draws from his belief that the data he produces is not as important as the results.

***Data deposit practices.*** The study shows that researchers' data deposit practices are not very complicated and are limited to storing data on cloud-based storage, personal computers, hard drives, USB memories, and/or hard discs. R1 explained that, based on the content of his data, he chooses different storage spaces. He explained,

> Evernote and Dropbox, those are the two that I mainly use as my storage spaces. It depends on the type of data. If it is graphical files or PDF, Evernote

is a better option, but if the data is numerical, it does not work very well on Evernote, but it does on Dropbox.

R2 expressed her concerns over hardware failures. She described that she used to store her data on a computer until she encountered a hardware failure. Since then, she has been using Dropbox to keep a copy of her data other than the copy stored on her personal computer. She perceives that the cloud-based storage spaces are reliable. When the investigator asked her about storing data in an IR, she responded that she would use a repository if the Institute had already established one.

R2: I am very much concerned with the risks that threaten the existence of my data such as device failure. That's why I have Dropbox, I trust to this kind of business. Even if they were to go out of business tomorrow, I would at least have a few copies on my laptop.

[…]

Q: Would you tell me why you have trust in the cloud-based storage such as Dropbox?

R2: I do trust Dropbox, because it has not messed me up yet.

Similarly, R3 and R4 revealed that they do not have any standardized procedures for storing their data. R4 explained that in his research, only software and the algorithm used to produce data are of value to be stored. He believes that data can be produced at any time if the software gets archived.

***Data retrieval practices.*** R1 utilized Evernote to store and provide access to his data. He explained Evernote and Dropbox searching systems as follows:

> With Evernote, there are different kinds of searching. I put descriptive tags based on what the project was involved, which student was involved, and in what year some work was done. With Dropbox it's more like exploring directory structure. There are directories and subdirectories. Work can be easily organized using the directories.

The researchers are primarily taking advantage of the searching capabilities offered by the cloud storage systems to access their data. However, unlike most of the researchers who use cloud storage searching systems to retrieve their data, R2 noted that she uses the Secure Shell (SSH) protocol[5] to move files from one system to another, and to search for and retrieve her data.

***Access management practices (security).*** The researchers who participated in this study had different opinions on how to manage data accessibility, and have employed different security measures to control and limit access to their data. Most of them either do not have any security policies or use the normal username and password system needed to log in to their computers or storage systems. R1 explains how he limits access to his data: "I have whatever password system that is built in those systems. I guess I would say I would trust the security measures that Evernote and Dropbox provide." R5 who uses cloud-based storage such as Evernote and Dropbox to store, and manage access to his data, said that, "these cloud-based systems allow easy upload and download of all project data by participating researchers with password protection." On his laboratory network, he does not have any security control to limit his students' access to the data. He noted that "we do

[5] Secure Shell (SSH) is a cryptographic network protocol for secure data communication.

not manage access to data. Everybody can access- whoever who has access to the lab drive."

R2 stated, "I do not think anyone wants to steal my data," to emphasize that limiting access to the data is not her main concern. R4 stores his data on his laboratory and office computers. He noted that he uses his computers' password systems to protect his data:

> We do not have any security measures in place to protect our data except for normal passwords. But for example, if I am sharing data with students, I don't know if they have [a] password on their computers or not. I don't care!

R6 believes that ideally there should be a policy that manages access to data; he feels, however, that imposing such policies will obstruct free flow of data. He expressed his opinion that "because data in research activities vary so widely […] it seems rather unrealistic to formulate a detailed policy."

***Data preservation perspectives and practices.*** This study showed that the researchers do not have specific procedures for preserving their research data. When the investigator asked the researchers about their data preservation practices (data preservation refers to the processes of maintaining accessibility to digital objects over time), all the researchers believed that it is the publishers' role to take care of their articles and their associated data. R1 explained his view on data preservation in the following way:

> In terms of long-term preservation, the scientific articles or what goes into scientific articles should be preserved permanently by the publisher. Articles are the main products of the research. Most

> publishers provide ways to support supplementary information in different file formats, and I would believe that that should be permanently preserved. Both of these jobs should be done by the publisher.

Thus, the researchers believe that the publisher should preserve raw data associated with their articles as supplementary material. Similarly, R5 recognizes publishers as responsible bodies for preserving research results: "publishers should be responsible for preservation and sharing our results." R1 noted that only their results should be preserved permanently, and other data produced in the course of research should be kept for a short period of time. He said, "the data that are [part of] larger simulations, those things should be preserved for a few years, like for a maximum of five years." Similarly, R5 stated that raw data should not be preserved permanently: "we want to save the data for five years after we published the results. We don't need the data after the results are published."

The data collected for this study informed the investigator that the researchers have not employed any data preservation strategies other than making back-up copies of the data. Also, the researchers are not aware of the potential physical and technological threats that may shorten the lifecycle of the data. When the investigator asked R1 about the way he handles the issue of format obsolescence, he explained,

> I understand the problem, but I don't really have a good answer for that. I am not going to worry about this problem. I am going to trust these publishers

> that continue to have systems of transferring things into other formats as [the] Internet evolves.

The responses suggest that since the researchers do not see a need to preserve their data for a long time, they do not really bother with the impacts of technological changes on their data. However, this may not be the case in the disciplines where, due to the nature of data sharing, use and reuse of data is very critical to the scientists who do not actually own the data.

The interviews with the researchers revealed that the primary preservation strategy employed by the researchers to provide short-term access to the data (the codes and simulations and other types of data) is to make back-up copies of their data on multiple storage spaces. The researchers do not have any plan for long-term preservation of their data, as they believe that only their publications should be preserved permanently. They do not perceive their data as permanent assets. R3 reported that he only make back-up copies of his data to secure them against any possible threats. Similarly, R1 requires his students to back up their data on a regular basis,

> Once a year, I give a little speech to my students and emphasize to those who work for me that they are responsible for having back-up copies of their data they get paid to generate. If they have a hard drive crash, I will not be sympathetic to them for having lost something, I will be angry at them for having failed to have two back-up copies. So, I give this speech to our student group meeting once a year.

R2 explained that to secure her data against potential damage such as hardware failures, she makes redundant copies of her data on cloud-based storage spaces. She stated, "I make back-up copies of my data on a cloud-based storage. If something happened to my data on the computer, my data still will be sitting in the cloud." R4 has developed his own unconventional way of preserving his publications. He believes that research results should be preserved by the publishers permanently. However, as paying publishers for preserving journal articles requires spending huge amount of money, he has been using his Website for storing and preserving his articles. He addressed the issue of technology obsolescence in his references to data preservation:

> I use MatLab [to produce my data]. So our data format doesn't get obsolete, because it's such a huge industry. I keep the old version of the software as well. Sometimes I upgrade the older version of the file format, but I usually archive the old version of the computer program.

R6 noted that other than publishing all pertinent data in articles and storing the data on the department network, he has not developed any procedures for long-term storage of data. The researchers' responses to the questions about data preservation imply a majority of them do not have a clear understanding of data preservation.

**Data appraisal and selection.** Although the researchers do believe that the research data should not be preserved for the long run, most of them noted that they have procedures for data selection and appraisal. That is because of this belief that appraisal and selection is a process that is embedded in research, meaning that

during the course of research, some data are automatically qualified and selected, and some are disqualified and omitted from the process based on the criteria determined during the research. Some of the researchers expressed this opinion that good data would assist scientists to proceed toward research objectives. R1 explained data selection process in his field thus:

> Q: Do you have any procedures for section and appraisal of your data?
>
> R1: I generally try to understand the same thing through theoretical approaches and compare them to see similar kinds of results from different approaches to understand why. It is a kind of quality assurance not really at data level, but at the scientific conclusion level.

R2 described that during the process in which a code is built, not all of the data that is generated is good, thus not all of the data will be published. Therefore, the flawed data is usually excluded from the rest of the data. She further explained that there are a lot of broken versions before there is a fixed version. When the fixed version is built, the data used in the paper (the "good" data) can be generated. She said, "we threw these [data produced by the broken versions] away once the code is running and we are generating the data that can be used in publication. But sometimes we hold on to it, because it was part of the building process."

As indicated above, research data is normally checked for its quality and usefulness during the research procedure. One of the graduate students who participated in the survey describes how data is checked in his laboratory:

> This [data selection and appraisal process] depends on the kind of data we collect. After certain experiments, we have to analyze and evaluate data to

see if they are of any use to our goals. If yes, we categorize and mark them. R5 sees the data selection and appraisal process as a teamwork activity, and believes that the decisions about which data is qualified and which is not are made during the research process: "I have discussions with my students to see if certain data is relevant. Bad data is omitted from the data file. Clearly only bad data (not every experiment works) is discarded."

R6 believes that selection and appraisal of research data could be an unmanageable task. He explained that the data produced in his field is uncontrolled with unpredictable content and representations, and therefore selection and appraisal of the data could be a cumbersome manual step that takes a long time. He suggested that an automatic way to check the relevancy and quality of the data should be developed.

**Data Management Plans (DMPs).** The interviewees included some comments on the quality and usefulness of the NSF's data management plans during their interview. As most of the researchers are not knowledgeable about developing an effective DMP, they agreed that collaboration between scientists and librarians is necessary, and could assist them to develop better and more comprehensive plans. R5 had developed his first DMP in collaboration with one of the librarians at the KSU Libraries. He noted that because most scientists do not know what should be included in a plan, collaboration between the scientists and the librarians would definitely be beneficial: He explained,

> It would be good if there is collaboration between the LCI researchers and the university library to help with developments of DMPs … It took way more

time that it should have. I asked somebody from the library to develop one. He helped me to come up with good ideas.

Most of the researchers, such as R4, believed that what has to be included in a plan really depends on the unique features of the data produced within the discipline. He explained that in sciences where it is difficult to quantify data, it is necessary to precisely store, preserve, share, and reuse the data. He concluded that DMPs suitable for the type of data produced at the LCI requires procedures different than the ones used in the data-intensive disciplines.

The researchers who had developed DMPs expressed their uncertainty about the quality of their plans. R2 noted that she is not very advanced in developing a framework for an appropriate DMP for the LCI. She had developed a DMP in collaboration with other LCI researchers. She expressed her feelings as, "I'm not certain that the NSF's DMP is good enough." She suggested that the NSF could assist with the development of discipline-specific DMPs by designing DMP frameworks in collaboration with researchers from each area.

**Data Analysis Findings**

In this section, the investigator provides more detail on the data analysis phase of her research. The research questions were designed so that the investigator could gather information on the following themes: the researchers' data management practices, the researchers' perspectives on data management, and disciplinary requirements for data management. After reviewing the survey responses, transcribing the recorded interviews, and analyzing the textual data, the investigator found that these larger themes that were identified as potential

categories prior to data collection were in fact reflected in the data gathered in the course of this study. Thus, the investigator felt confident in using these themes as a starting point for organizing and analyzing the data; ultimately, clusters of coded data were developed based on these identified themes (see Table 3 in chapter 3).

After conducting the initial coding and analysis, and then organizing codes into clusters of related codes, the investigator examined relationships among the codes more closely and began to develop a model for describing the data world of scientists at the LCI that incorporated their perspectives and practices relating to data curation. As an aid to understanding this model, she created a graphical representation to illustrate the relationships among these perspectives and practices. This model reflects the worldview of the scientists who participated in the study—the ways in which they thought about data management. The metacategory of "Data World of Liquid Crystal Scientists" is organized into three smaller clusters of related concepts: "Data Management Practices," "Researchers' Perspectives on Data Characteristics," and "Researchers' Perspectives on Data Management." Each cluster is further divided down into associated codes. Each code within the clusters represents specific phenomena and concepts within the data gathered for this study.

In the models developed to represent data world of the liquid crystal scientists, different types of arrows are used to show the relationships between the codes and clusters and to emphasize the differences in the nature of the interactions that occur among the codes and clusters. While the solid arrows show the influence of one entity on another entity, the dotted arrows show the mutual effect of entities on each other. Also, since the codes in different clusters have different natures, and

some of the codes appear in more than one cluster, the investigator has employed different types of shape outlines for each type to emphasize the differences among them.

**Relationships and interactions among phenomena and concepts represented by codes.** The concepts and phenomena represented by the codes within a specific cluster often influence one another. The "Researchers' Perspectives on Data Characteristics" cluster incorporates the codes that represent specific qualities of the data (i.e., the nature of the data being collected in the conduct of liquid crystal-related research activities). The investigator identified quantifiability, replicability, and reusability as prominent data characteristics of the data addressed by the interviewees. She determined that there are certain relationships among the codes within the cluster. She realized that features such as quantifiability and replicability reduce reusability of the data. For instance, in liquid crystal research, the data can be quantified into summaries and equations, thus, the scientists believe that it would be of no use to any researchers to keep, share, and reuse the raw data resulting from the computer simulations.

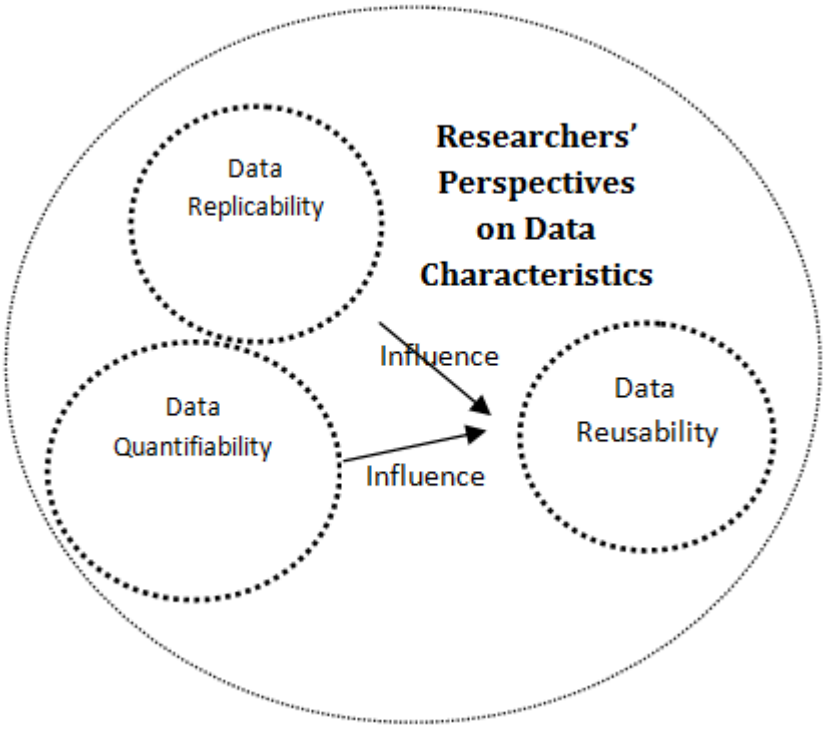


**Figure 1. The interaction of the codes within the cluster of Researchers' Perspectives on Data Characteristics.**

The cluster of Researchers' Perspectives on Data Management contains the codes associated with the data gathered on researchers' perceptions of different concepts of data management. After working with this data, the investigator found that there are mutual effects among certain codes within the cluster, meaning that the researchers' views on one concept could affect their perceptions of the other ones. For instance, the researchers' views on data sharing could influence their views on data licensing practices. Although some of the participants believed that they do not require data licensing procedures because data sharing is not a very common data activity within the discipline, they felt that in the future they might be required to share data and develop the procedures that support such activities. Similarly, the investigator realized that there is a relationship between the

researchers' perspectives on the NSF' Data Management Plan and their data deposit activities. Most of the researchers noted that it is not really necessary to develop a repository for data storage purposes, however, they believed that developing and storing data in a repository would comply more closely with the NSF rules for data management.

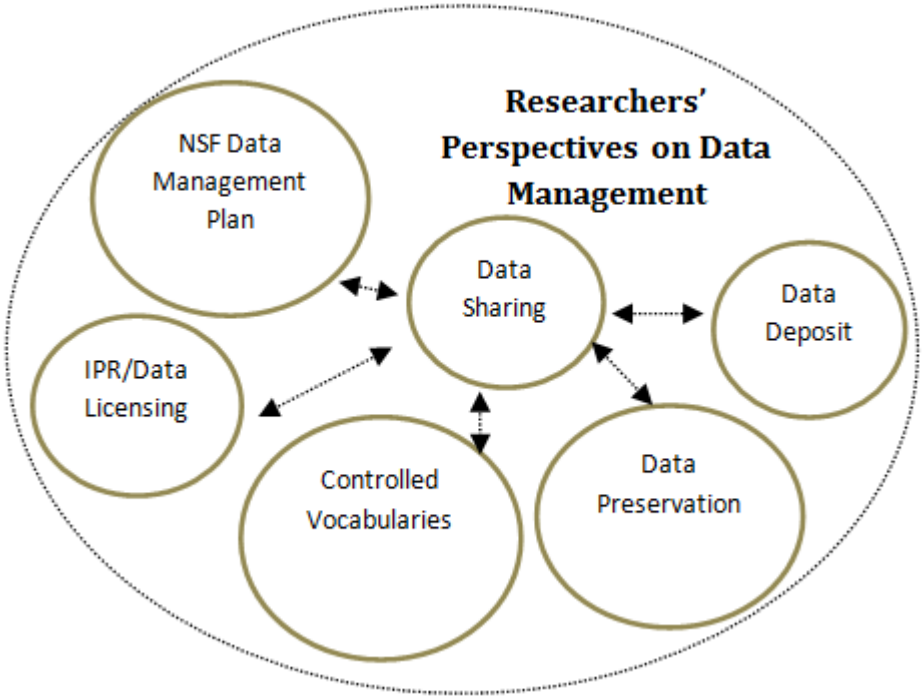


**Figure 2. Interactions between the codes within the cluster of Researchers' Perspectives in Data Management.**

Within the cluster of Data Management Practices, the investigator identified two major code groupings, gathering together those codes associated with data sharing and those associated with data deposit. The investigator decided to investigate the codes such as data deposit, data preservation, data retrieval, security measures, and selection and appraisal as one category. Similarly, she grouped together codes such as data sharing, data licensing, intellectual property rights, and

metadata and controlled vocabularies. She realized that codes within each subcategory have reciprocal impact on each other and on the codes within the other category. For instance, the researchers' data deposit practices influence all the practices associated with data deposit (including data documentation, data preservation, etc.) as well as the practices in the data sharing category. Repositories provide a platform for data sharing practices of scientists. Many of the researchers do not see a need to develop a repository where data can be shared and preserved, because they think after they published their results, they do not want to keep the data for a long time.

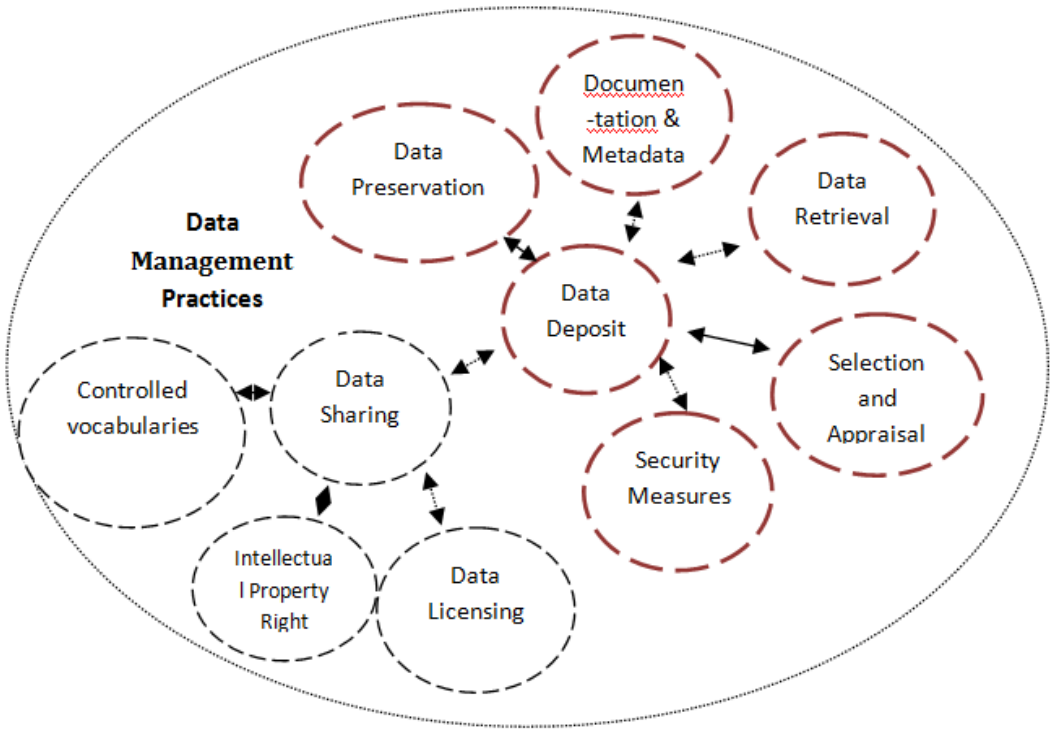


**Figure 3. The interaction between the codes within the cluster of Data Management Practices**

## The Data World of Liquid Crystal Scientists

As it is evident in Chapter II (Literature Review), data curation research is very crucial to supporting data practices of the researchers conducting studies within academic research institutes (small-science disciplines). To develop effective data management (curation) procedures for each discipline, characteristics of the data produced within the discipline should be investigated, in addition to current practices and perspectives of researchers. The investigator suggests that the data model developed from this study's findings could be used to study disciplinary requirements of researchers across other scientific disciplines and develop or enhance their data procedures. This model would inform librarians and scientists about how they should approach disciplinary data curation research.

**Development of the data model for liquid crystal scientists.** When the investigator tried to identify the connections between these clusters, she realized that there is a reciprocal relationship between researchers' perspectives on data characteristics and their views on their data management needs. Also, she noticed that the researchers' perceptions of both data characteristics and data management have affected their data management practices within the discipline.

After considering the various concepts and phenomena represented in the code clusters, and the interactions and influences among all of these entities, the investigator generated a graphical representation of the data practices and perspectives of the LCI group that captured the various relationships noted above (see Figure 4).

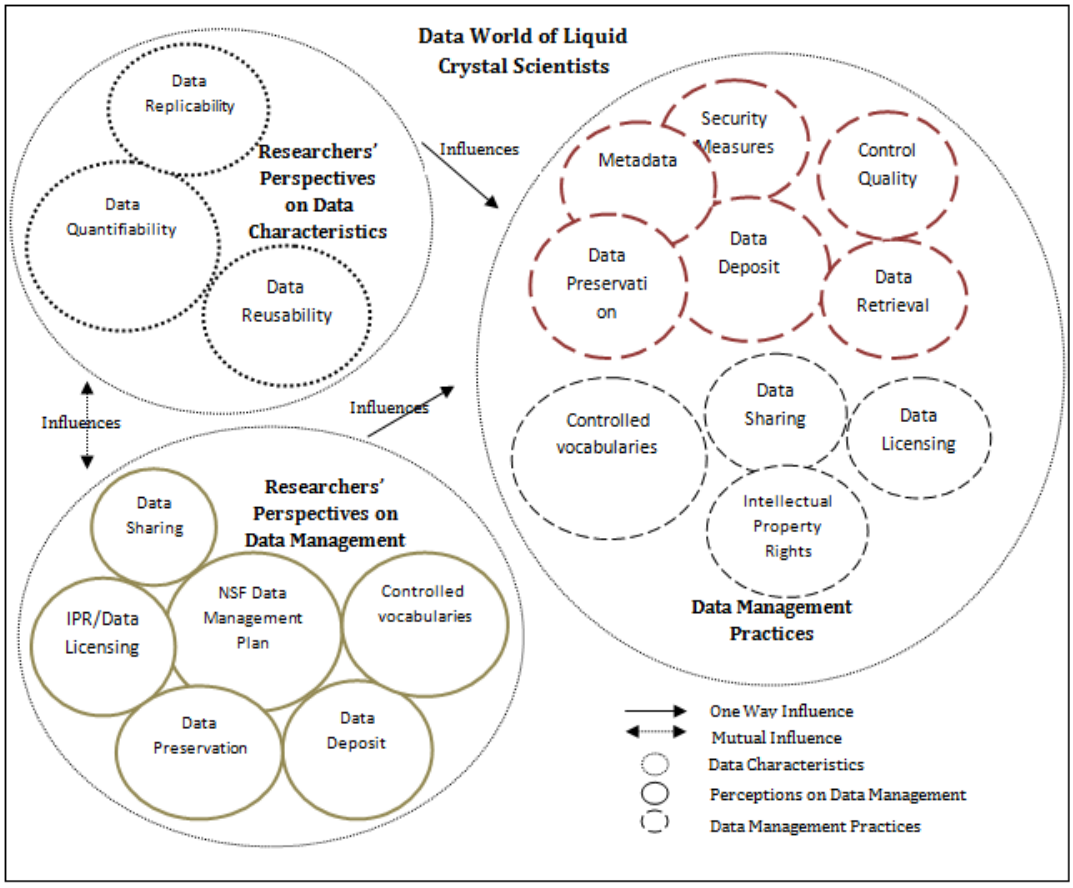


**Figure 4. Graphical representation of the clusters and the interactions among them within the data world of liquid crystal scientists.**

**Examples of interactions in the world of liquid crystal scientists.** The following example shows how the Researchers' Perspectives on Data Sharing entity interacts with the Data Quantifiability and Data Sharing Practices entities.

The researchers believe that the most important science generated within their field is summarized in their publications; therefore, they consider their articles as the most important output of their research and do not see any value in data

sharing. As a result, their data sharing practices are limited to sharing their publications, rather than sharing raw data directly with other researchers.

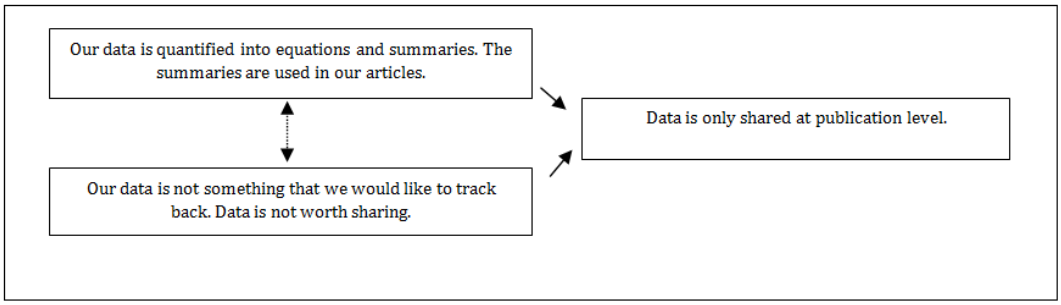


**Figure 5. Interaction between perspectives on data quantifiability and data sharing, resulting in data sharing practice.**

The following example illustrates the interaction of the concept of replicability (defined as one of the qualities of the data) with other entities located within the other clusters. The figure demonstrates that both the replicability of data and the researchers' perceptions of data sharing inform and affect their actual data sharing practices, as the two entities have a mutual effect on each other. Also, it depicts that the researchers' perspectives on data sharing and on intellectual property rights mutually influence each other (these two entities are encircled in the same clusters) and finally inform and affect the researchers' data sharing practices (resulted in the computer codes generated through LCI research projects only being shared with the researchers' collaborators).

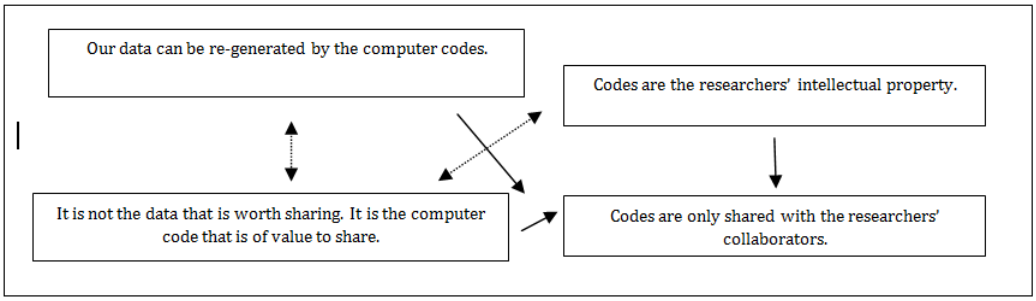


**Figure 6. Interaction between perspectives on data replicability and reusability, affecting data sharing practices.**

The above examples illustrate that the interaction between the researchers' perspectives on the data characteristics and their perceptions of their data management needs determine their data management practices. Most importantly, the data model of the liquid crystal scientists confirms that studying of disciplinary requirements prior to any data curation planning is very essential. In other words, the findings provide sufficient evidence that the data management (curation) recommendations developed by large national research institutes may not be suitable for small-science disciplines such as liquid crystal science, as researchers' disciplinary needs may be distinctly different in terms of how data is created, used, shared, and retained.

**Research Questions and Findings**

To summarize the findings of this research the investigator will review the results as they relate to each one of the research questions. As it was discussed earlier in the Discussion section, categorizing the data into clusters of codes and drawing the relationships among them helped the investigator develop an

explanatory model to answer the central research questions. This study was designed to answer three central questions:

- *What are the disciplinary requirements for data production, management, and preservation that influence data curation activities within the LCI research laboratories?*

The investigator found that the qualities and uniqueness of the data produced within the LCI have influence on the researchers' perceptions of different aspects of data management and consequently, their actual data management practices. For instance, because the type of data can be replicated by the computer codes originally developed by the researchers, they would not see any value in sharing, and reusing of the data. They simply share the codes with their collaborators if the same data needs to be generated.

Similarly, quantifiability was considered to be one of the specific features of the data. The interviewees emphasized that their research is not very data-intensive, meaning that they produce small-scale data. They further explained that as the data can be quantified into summaries and equations, and all the summarized and condensed data will be shared through their research publications, it would of no use to any of the researchers to store and preserve the data for long-term accessibility. They believe if any researchers would like to access the essence of their data, they can refer to the publication instead of the raw data.

- *What are the LCI researchers' perspectives on data curation practices such as data deposit, sharing and reuse, curation and collaboration?*

The interviews show that the researchers' views on different aspects of data management have been affected by their perceptions of the important qualities of the data. Most of the researchers believed that as their data are quantifiable. That is, the most significant summaries of the data are used in the research publications. As a consequence, from their point of view only the results are worth sharing with the public. Many of the participants emphasized that it is very crucial that the computer codes and software originally developed by them are only shared with their collaborators and students in order to protect intellectual property rights of the researchers and the institute. They stated that if data is to be shared with the public, it should be made available as supplementary material through the publishers.

A majority of the researchers agreed that if the Institute had already established a repository, they would deposit the output of their research in it. They believed that contributing data to an IR would make them more compliant with NSF rules for data management. The researchers do not see a need to preserve the data for a long period of time. Instead, they noted that a period of five years would be enough for preserving the kind of data that they generate in their research. Furthermore, as results are considered the most prominent gains of the research, they noted that those results should be stored and preserved permanently. They recognized the publishers as the entities responsible for long-term preservation of their publications and associated data, rather than the Institute, or University Libraries.

- *How do LCI researchers manage their data within their laboratories? What are the established practices for data management, if any?*

This study illustrates the researchers' specific data management practices. These practices are influenced by their perceptions of the characteristics of the data. Since the data produced within the LCI has its own specific qualities, the researchers have developed their own perspectives on data management. Most of participants explained that data sharing is not very common within the discipline. As a result, they only share their codes (with their collaborators) and publications. One of the researchers noted that he provides access to his articles by putting them on his personal website. Many of the researchers are taking advantage of tools such as email, Dropbox, and Google Drive as means of sharing data with students and employees within the Institute.

The interviewees stated that they would not need controlled vocabularies as tools for enhancing data sharing activities within their discipline, as their data is typically numerical. Similarly, as data sharing is not a common data practice within the LCI, the researchers have not seen a need for developing data licensing policies and procedures.

The researchers believe that the selection and appraisal of the data is a process that occurs as part of the research, meaning that the faulty data will be deselected as scientists proceed with the research to achieve the research objectives.

To store and secure the data (codes and simulations and other types of data produced during the course of research) the participants primarily use password-

protected, cloud-based storage systems such as Evernote, Dropbox, Google Drive, etc. They have not developed standardized ways of describing their research data, except for the notes taken during the course of research. In most cases, the researchers use their traditional paper notebooks to document information about their data and experiments.

In some laboratories, some of the data is produced by proprietary hardware and software. In such situations, the metadata is automatically generated along with the data as separate files and in separate folders. This kind of metadata cannot be easily used for data retrieval as the data is not encapsulated by the metadata. Additionally, some of the researchers reported that they also use the tagging systems provided by the cloud-based storage for labeling and access to the data. Some of the researchers store a copy of their data on their personal computers to have a back-up copy of their data.

Data is stored and preserved for a period of five years within the LCI. The primary preservation strategy used by the researchers is bit-stream copying (meaning that they make an exact duplicate of the digital objects on a regular basis). Multiple copies of the data are restored frequently on multiple storage devices such as cloud-based storage, personal computers, or hard drives.

# Chapter V

# Discussion

## Research Findings and Recommendations

In a recent issue of *Information Standards Quarterly* devoted to digital curation issues, guest editor Sarah Callaghan of the British Atmospheric Data Centre stated, "digital curation is an oft-neglected part of the research process, as most researchers have neither the time and funding, nor possibly even inclination to deal with it" (p. 2). That statement corroborates what this study found—that researchers at the LCI do not currently have incentives to develop collaboration and connections with data curation experts who could assist them in improving the quality of their data management practices. This lack of inclination could be due to the fact that they are more vigorously involved in their day-to-day scientific research and are less concerned about the advantages they may receive through sharing, reuse of, and long-term accessibility to their data. They do not currently see any tangible benefits in following best practices in data curation, as their focus is on managing data for current research projects.

The investigator found that although it is true that many small-science disciplines have become data-intensive, such as oceanography, there are also other disciplines, such as liquid crystal research, that are not as data-intensive. She also found that the study of data characteristics for particular disciplines is an essential first step in conducting data curation research within each discipline. Because the nature of data can be unique to a discipline, data curation requirements can be very specific to a discipline. The uniqueness of data characteristics for a particular field

will necessitate the tailoring of data management practices and plans to fit scientists' needs. These findings confirm that recommendations and best practices developed from perspectives of the larger big-science research institutes cannot be indiscriminately applied to small-science disciplines, due to the variation in data characteristics among the sciences.

The investigator also realized that there is a reciprocal relationship between the perspectives of the researchers regarding characteristics of the data and the researchers' perceptions of their data management practices needs. Replicability and quantifiability were mentioned as unique features of the data produced within the LCI. From the researchers' perspectives, these data qualities reduce the reusability of the data outside their research projects. As a result, the researchers would see no value in sharing and reusing of the data beyond their students and collaborators. Given their perceptions, it is not surprising that they have not developed data licensing policies as a way of addressing the issue of intellectual property rights. Similarly, based on their views on the nature and usefulness of the data, they do not see any need to develop controlled vocabularies for improving the quality of the data management practices.

The investigator found that the researchers' perspectives on different aspects of data management also mutually affect each other. For instance, the way the researchers view data sharing practices within their discipline has influenced their perspectives on data deposit and preservation. Many of the researchers believe that only the research results (in the form of published articles) should be stored and preserved permanently. This belief has resulted from their perception

that there is no value in sharing their raw data. Other types of data (such as the computer codes and simulations) produced during the course of research are only stored for a limited period of time (e.g., 5 years). Standard data management practices, such as data documentation (the provision of metadata), access management and retrieval, and preservation have been also undervalued by the researchers.

The investigator found that the researchers' data management practices have been influenced by their perceptions of the data characteristics as well as their perspectives on their data management needs. The results confirm that, based on the uniqueness of the data generated within the disciplines, data practices should be tailored to meet disciplinary requirements. For example, the computer codes that generate the data are developed by the researchers, and thus are considered intellectual property of the researchers. As a result, the codes are only shared with their students and collaborators. LCI researchers believe that only final results appearing in publications should be shared with the public, as they are the most valuable benefits of their research. They recognized the publishers as responsible bodies for sharing both their research results and the associated results.

To store, secure, and retrieve research outputs (including codes, software, research notes and simulations), the researchers commonly use their personal computers and cloud-based storage spaces such as Evernote, Dropbox, and Google Drive. Data preservation practices of the researchers are restricted to producing back-up copies of their data on their office and laboratory personal computers as well as on cloud-based storage locations. Storing a copy of the data on cloud-based

storage is regarded as a way to safeguard it against any possible hardware and media failures. The interviewees did not feel that the risks of data loss were serious enough for them to seek a more secure and permanent long-term solution for data management. Because they are more likely to view data management from the perspective of current research projects, they do not have any plan for long-term preservation of their research outputs into the future.

When LCI scientists were asked about development of an IR within their institute, however, they agreed that storing data in a repository would more likely comply with the NSF rules for data management. Based on what the participants mentioned about the data storage needs and practices during the interviews, it seems that establishing a repository system privately maintained by the Institute, but with similar functionality to the solutions such as Dropbox that they currently employ, might be an appropriate solution for the LCI.

**Merits of the Study**

This study provides current and future LCI researchers and curators with essential information on what should be included in data curation plans. Beyond the value for those participants at the research site, the study results will inform small data producers (mainly researchers at academic research institutes) of the importance of good data curation policies, and provide them with a model to inform their development of effective data curation practices to describe, preserve, share, and reuse their data. Although each discipline may have its own specific disciplinary requirements that need to be investigated closely, the model developed from this study may be useful for data curation studies in other domains.

Another particular merit of this study was the important finding that the the concept of data sharing was defined differently by participants in this study than the original preconception of the investigator. While the investigator viewed data sharing as a data management activity through which data is made available to the public (i.e., researchers outside the LCI), the interviewees could only perceive data sharing to be the practice of sharing their data within their own research group at the Institute or, occasionally sharing with research collaborators in other institutions. The investigator believes that the participants' views on data sharing are important for data curation research, as they are the ones who have a better understanding of the nature of the data they produce as well as their disciplinary needs.

**Limitations of the Study**

**Number of participants and degree of participation in the study.** One of the major limitations of this study was that the investigator could not interview all the participants who had participated in the questionnaire. Also, not all the interviewees participated in the questionnaire. In such situations, the investigator tried to combine some of the questions (those that were not repetitive) from the questionnaire with those of the interview instrument in an effort to collect as much data that she could during the interviews. Although in the proposal she had planned to include only questionnaire data collected from the faculty members, the investigator also incorporated data collected from two doctoral students as their data practices and perspectives contributed to a more complete explanation of the data world of LCI researchers.

**Scope of data collected.** Another limitation of the study is that, the investigator did not collect and review the policies, data curation plans, or other written documentation employed within the LCI. In the future studies, any institutional-wide policies or plans developed for regulating data procedures or management should be collected and analyzed for evidence of data curation policies and procedures.

**Techniques used to reduce bias.** Bias of investigator interests and objectives often can subtly influence data collection and analysis techniques used in qualitative research. To avoid collecting responses that reflected what the investigator hopes to find, she omitted leading questions that might have influenced how participants responded during her interviews. Results may have been affected negatively by untrue or partially true answers. To avoid such situations, the investigator asked for clarification where she suspected that answers may have not been clearly stated. In data analysis phase, the investigator tried to avoid situations where she may have interpreted participants' responses in a way that reflected her own ideas and knowledge of the field instead of the scientists themselves. Thus, she tried not to change the participants' words when transcribing the recording or interview notes. Also, the investigator tried to remain as objective as possible by using direct quotations from the transcribed interview text in this report whenever possible in order to remain faithful to the exact words, concepts, and sentences that the participants used to respond to the questions.

**Data validation.** As the scope and time of this study were limited, the investigator did not have a chance to validate the model that she had developed out

of data curation perspectives and practices of participants. In order for the model to be adopted for data curation research, the investigator would need to extend the current study to include participant validation of the model.

**Future Studies**

The results of this study have established the groundwork for future work to examine different aspects of data management (curation) processes within small-science disciplines. As each discipline could have its unique data management requirements, more data research would provide opportunities for data librarians or researchers to investigate various aspects of data management across various disciplines.

**Validation of the data world of liquid crystal scientists.** As a result of this study, the data model of the liquid crystal of scientists was developed based on the data characteristics, and data curation perspectives and practices of LCI researchers. However, because of limitations in time and scope of the study, the model has not been validated by the participants in this research. Therefore, validation of the model developed to explain the data world of small-science practitioners such as liquid crystal scientists is a major area of future research that could be pursed either by the investigator or another interested researcher.

**Data curation issues and limitations.** A major area of future research lies in investigating the issues, concerns and barriers that the scientists at small or medium-sized institutes are commonly facing and that prevents them from developing standardized data management programs. Also, future studies could

identify appropriate solutions to address the issues, concerns, and barriers of research institutions facing limited infrastructure and resources.

**Campus-wide research data services.** Another potential area that may need further investigation is the possibility of the establishment of campus-wide research data services. It is very important that such services are provided to academic researchers through some sort of information organizations such as academic libraries. Different collaborative models for providing the most effective data services to the academic community must be further investigated. The costs involved in such collaborative activities should also be also considered. Last, the incentives and motivations that could encourage scientists to be actively involved in data management activities should be studied.

**Data sharing and reuse.** As data sharing and reuse of research data are sometimes the least understood aspect of data management, it is potentially an area worthy of further study. It would be beneficial to compare data sharing and reuse activities across scientific disciplines to identify critical differences and areas of shared concerns. Such studies could later be used to inform data curation research aimed at providing higher quality on-campus research data services.

## Conclusion

The investigator focused her study on improving the understanding of data curation requirements of one academic research institute that does cross-disciplinary research on liquid crystals, through investigation of disciplinary data management requirements, the researchers' perspectives on their data management needs, as well as their data management practices. Over the past few

years, it has been primarily the larger big-science research institutes that have had enough funding to develop best practices and to build necessary infrastructure for data management (curation) practices. Thus, this study was designed to address the gap in the research required to investigate disciplinary requirements for development of campus-wide DMPs within a university setting.

The investigator employed a case study approach to collect empirical data on the data curation perspectives and practices of researchers at the Kent State University's LCI through application of both Web surveying and interviewing techniques. She used an online questionnaire to gather preliminary information from employees at each research laboratory and used the results of the questionnaire to inform follow-up interviews with those researchers. The interviews were carried out to further explore issues and concerns surrounding data curation that had not been provided in questionnaire responses.

Collecting and analyzing the data from the participants in this study assisted the investigator in the following: 1) developing a clear understanding of the nature of the research data generated within the LCI; 2) articulating the researchers' perspectives on data management concepts; 3) identifying and understanding established data management practices; and, 4) developing a data model that clearly illustrates the interactions among the data, the researchers' perspectives on data management needs and the actual data management practices.

Furthermore, the data model developed from this study confirmed that as data characteristics can be very specific to a discipline, data curation requirements may also vary across small-science disciplines. Therefore, the recommendations and

best practices developed from the perspectives of large big-science research institutes may not be indiscriminately applicable to all disciplines, and should be tailored to the particular data curation needs of each discipline.

# APPENDICES

# Appendix A: Questionnaire Instrument

- Please identify your role within the LCI Environment (select one):
    - Faculty member
    - Principal investigator
    - Post-doctoral employees
    - Graduate students
    - Other: _____
- What is the name of your laboratory? _____
- In what format (s) are the data stored? Are you using any open format? (Open formats provide access to the source code for file formats, and to the documentation about them. This allows the future possibility of developing tools to migrate to open formats (Harvey, 2010)). If so, please name them. (In the context of this research, data refers to any digital output of research. It could be generated in textual, graphical, numeric, tabular, database, 3-D, etc. forms).
- What tools, software, or hardware are used in generating the data?
- What tools, software, or hardware are required to utilize the data?
- What tools, software, or hardware are required to store the data?
- Do you have any procedures for data validation and verification?
- Are you using any open source software or open standards in your data curation process?
- Do you currently make back-up copies of your data?
- If you answered “yes” to this question, approximately how often do you make back-up copies?

- If your institution had a centralized institutional repository for scientific data, would you store your data in it?
- Have you developed best practices or procedures for data selection?
- Have you developed best practices to describe your data?
- Have you developed best practices for creation of datasets?
- Have you developed best practices for data quality control?
- Have you developed best practices for long-term storage of data?
- Have you developed best practices for licensing your data?
- Have you developed policies for access management to your data?
- Have you developed policies that support data sharing and collaboration among scientists?
- Have you undertaken a risk assessment of your digital content?

# Appendix B: Interview Instrument

**General Descriptive Characteristics of Data**

- Tell me a little bit about your research.
- How do you produce and record your data?

**Descriptive Standards (Metadata)**

- Do you make use of any descriptive standards to label or provide access to your data? If you have used any standardized forms of description, please identify them.
- If you don't use any descriptive standard to describe your data, how do you search or retrieve your data within the storage system?
- As a scientist have you experienced a situation in which you could not access specific data in your daily activities? If so, what was the reason (s) for that?
- Do you see a need to describe your data using a descriptive standard at all?
- If you could record information about the data, what information do you think should be recorded to make the data accessible and understandable?

**Data Curation Lifecycle**

- How are your datasets organized?
- Do you use any media or system for your physical data storage?
- Do you have security measures in place to control access to the data?
- Do you see a need for implementing a system to manage access to your data collection?

- Have you established procedures for preparing your data for storage? If so, what are they, and why were they established?
- Have you developed data quality assurance/control procedures for your data? (a set of processes to evaluate the quality of data before and after they are collected)
- I would describe a data management plan as a framework by which data are acquired / produced, maintained, and made available. Based on this description, would you believe that you have data management procedures in your laboratory?
- If so, please describe your data management plan briefly.
- Do you see a need for developing tools that support large-scale data federation within or across disciplines?
- As you may know, beginning January 18, 2011, proposals submitted to NSF must include a supplementary document of no more than two pages labeled "Data Management Plan" (DMP). This supplementary document should describe how the proposal will conform to NSF policy on the dissemination and sharing of research results. Have you ever been required to provide a DMP along with your grant proposal?
- Do you see a need for any future collaboration between yourself and the University Libraries for help with various aspects of research data management / curation?

- Do you agree with this statement: “if research is publicly funded, the results should become public property and therefore properly managed and preserved”?

**Data Sharing**

- Will you be able to share your data with other researchers working on similar problems?
- If you do not like to share your data, what are your reasons for not sharing your data?
- If yes, at which point of time in your research process would you prefer to share it?
- If you would be willing to share your data, with whom you would share it?
- Would you place any conditions on sharing your data with the groups or people you have identified (such as requiring some form of acknowledgement, etc.)? If yes, please explain those conditions.
- Have you developed any shared vocabularies or common keywords as part of your data sharing policies? If yes, do you use these vocabularies with your current data?
- If not, have you ever felt the need to develop a shared vocabulary aimed at increasing consistency for better data sharing?
- What value would the data have for these groups or individuals if you share your data with them?

# Appendix C: Participant Profiles

In this appendix, the profiles of the researchers who participated in this study are provided. Each researcher who is profiled below is associated with either a laboratory or project at the LCI. All of the participants already had procedures in place for data creation prior to the start of the study, as evidenced by the documentation they kept of their research activities.

**Researcher 1 (R1)** primarily uses preparatory software such as Mathematica to produce data. His research output is mainly numerical data and computer simulations. He does not have any established procedures to support description, storage and preservation, licensing and sharing as part of his data management practices.

**Researcher 2 (R2)**'s research output is numerical and graphical data as well as computer simulations. She and her research group use visualization software such as SimReplay and Visualization ToolKit (VTK) to generate visualization data. She has not established any procedures for description, quality assurance and control, preservation, licensing, and sharing.

**Researcher 3 (R3)**'s data is produced by means of proprietary software such as Microsoft Word, Microsoft Excel, Igor Pro, and Adobe Illustrator. His data is primarily numerical and computer simulations. He has not developed any data

policies on data creation, description, storage and preservation, licensing, and sharing to support his data management practices.

**Researcher 4 (R4)**'s research output is numerical data as well as computer simulations produced by originally developed software. He has not established any procedures for description, storage, preservation, licensing, and sharing for managing his data.

**Researcher 5 (R5)**: The data produced in his laboratory is primarily produced by proprietary software. His data output is mainly in numerical and graphical formats. In order to manipulate data in other software, he commonly converts data into other formats such as ASCII files. He has protocols in place for data creation and control quality, however, he has developed no policies on description, storage, and preservation, licensing, and sharing.

The following researcher declined to be interviewed, but provided responses via the questionnaire.

**Researcher 6 (R6)**'s research output is in numerical, tabular, graphical, and database formats. Data is produced in his laboratory using proprietary software such as Microsoft Word, Microsoft Excel, Igor Pro, and Adobe Illustrator. He has developed policies on data creation, selection, and sharing. He does see a need for developing procedures for data validation and verification, preservation, and licensing in the near future. He, however, does not believe data description, access

management, and risk assessment procedures are necessary for management his data.

**Doctoral student 1**'s research output is numerical, and, graphical. Data is produced in his laboratory using software such as LabView, Matlab, Fortran, Mathematica. He does see a need for developing procedures for data validation and verification, preservation, and data description. He, however, does not believe data licensing, selection and appraisal, access management, and risk assessment procedures are necessary for management his data.

**Doctoral student 2**'s research output is primarily in numerical, and, graphical formats. He produces data in his laboratory using software such as Labview, Matlab, C++ in Visual studio or specific software coming with a particular hardware device. He does see a need for developing procedures only for data selection and appraisal. He, however, does not believe in development of procedures for data description, licensing, sharing, long-term preservation, access management, and risk assessment procedures.

# Appendix D:

## Glossary of Cyberinfrastructure and Digital Curation Terms

**Academic research institutes** are always affiliated with universities and refer to the units where scientists generate or gather data into privately held sets or collections that they analyze locally. In such organizations, research funding can be limited, and the day-to-day conduct of research is often dependent on a few graduate students, who carry out much of the data collection, and manage and process those data during the course of a project (Cragin, Palmer, Carlson & Witt, 2010).

**Bit-stream copying** refers to the process of making an exact duplicate of digital objects. It is commonly practiced as backup and restore, that is, backing up (copying) computer files on a regular basis and restoring them if the data in the primary source is corrupted (Harvey, 2005).

**Case study** is an in-depth investigation of a discrete entity which may be a single setting, subject, collection, or event on the assumption that it is possible to derive knowledge of the wider phenomenon from intensive investigation of a specific instance or case (Gorman & Clayton, 2005).

**Computer code** is defined as a "program"written to simulate a system, solve some equation numerically, or to analyze the data they have taken via an experiment.

**Cyberinfrastructure** refers to integrated systems that include hardware, software, and human resources to support research requiring high-performance computing for modeling, simulation, prediction, and data mining; data management and visualization; virtual organizations; and educational enhancement (Conte, Glenn, Green, Lalwani, Martin, Ottaviani, & Song, 2010).

**Data** in this study refers to non-observational data produced as the outputs of computer codes originally developed by LCI researchers. The data primarily consists of numerical data generated from simulations.

**Data curation** refers to the activity of managing and adding value to data from its point of creation, to ensure it is available for discovery and reuse (Digital Curation Centre, 2013).

**Data-intensive** disciplines typically produce that "large-scale data," meaning that a scientist can work as an individual or in small lab and still have "large-scale data" that needs curation.

**Data licensing** refers to the process through which a license—a legal instrument for a rights holder to permit a second party to do things that would otherwise infringe on the rights held—is obtained for data sharing and reuse purposes.

**Data reusability**. The point of curating datasets is that they remain available for use and reuse by legitimate users. Data sharing is a prerequisite to their usability and reusability.

**Digital preservation** refers to "the processes of maintaining accessibility to digital objects over time" (Harvey, 2005, p. 161).

**Institutional Repositories (IRs)** are repository software that are "aimed at allowing more efficient management of institutional assets and providing fast, easy access to its contents from remote locations" (Harvey, 2010, p. 192).

**Large big-science research institutes** are usually not affiliated with universities and are often government agencies. They are considered large because of the immense amounts of data they generate and analyze, and the computational methods used to analyze this data. These institutions support project coordination, resource sharing and increasingly standardized information flow through development of DMPs. Typically, they have sufficient funding to recruit staff and build infrastructure needed to manage their own data without assistance from outside (Cragin, Palmer, Carlson & Witt, 2010).

**Metadata** literally means "data about data." It articulates a context for objects of interest. (DCMI, 2013)

**Open formats** provide access to the source code for file formats, and to the documentation about them that allows for the possibility of developing tools to migrate files to open formats. Open file formats and open source software assist curation by allowing curators to have control over software source code and to see how file formats are structured (Harvey, 2010).

**Small-sciences** are typically described as hypothesis-driven research led by a single investigator or small research group that generates and analyzes its own data (Ciciora, 2010). In small-science disciplines, a scientist can generate large-scale data or small-scale data.

# Appendix E: List of Acronyms Used in This Research Report

| | |
|---|---|
| **ACP** | Advanced Cyberinfrastructure Program |
| **CI** | Cyberinfrastructure |
| **DCMI** | Dublin Core Metadata Initiative |
| **DCC** | Digital Curation Centre |
| **DMP** | Data Management Plan |
| **IR** | Institutional Repository |
| **LCI** | Liquid Crystal Institute |
| **NSF** | National Science Foundation |
| **SCARP** | Disciplinary Approaches to Sharing, Curation, Reuse and Preservation |

# References


Akers, K.G. (2013). Looking out for the little guy: Small data curation. *Bulletin of the American Society for Information Science & Technology, 39*(3), 58-59.

Atkins, D.E., et al. (2003). *Revolutionizing science and engineering through CI: Report of the National Science Foundation Blue-Ribbon Advisory Panel on CI.* Arlington.

Ball, A. & Neilson, C. (2010). *Curation of research data in the disciplines of engineering*. Edinburgh: Digital Curation Centre. Retrieved from http://www.dcc.ac.uk/scarp.

Brandt, D. S (2007). Librarians as partners in e-research: Purdue University Libraries promote collaboration. *C&RL News, 68*(6), 365-396.

Callaghan, S. (2013). Letter from the guest content editor. *Information Standards Quarterly, 25*(3), 2-3.

Choudhury, G.S. (2008). Case study in data curation at Johns Hopkins University. *Library Trends, 57*(2), 210-220.

Ciciora, P. (2010, Sep. 21). *Small sciences could benefit from better data-sharing practices*. Retrieved from http://www.lis.illinois.edu/articles/2010/09/small-sciences-could-benefit-better-data-sharing-practices.

Conte, M., Glenn, J., Green, J., Lalwani, L., Martin, S., Ottaviani, J., & Song, J. (2010). *E-science Task Force report.* Research Cyberinfrastructure website. Retrieved from https://docs.google.com/document/d/19nP-5Uza4-GHNzGAsmmMXVqCFbmRIOOVdpZwsh6LFMs/edit?pli=1.

Cragin, M. H., Palmer, C. L., Carlson, J. R., & Witt, M. (2010). Data sharing, small science, and institutional repositories. *Philosophical Transactions of the Royal Society, 368*(1926), 4023–4038. doi: 10.1098/rsta.2010.0165

Day, M. (2007). Toward distributed infrastructures for digital preservation: The roles of collaboration and trust. *The International Journal of Digital Curation, 3*(1), 15-28.

Digital Curation Centre. (2013). *What is data curation?* Retrieved from http://www.dcc.ac.uk/digital-curation/what-digital-curation.

Dublin Core Metadata Initiative. (2013). *Background*. Retrieved from http://dublincore.org/metadata-basics/.

Fear, K. (2011). You made it, you take care of it: Data management as personal information management. *The International Journal of Digital Curation*, *2*(6), 53-77.

Fry, J., Lockyer, S., Oppenheim, C., Houghton, J., & Rasmussen, B. (2008). *Identifying benefits arising from the curation and open sharing of research data produced by UK Higher Education and Research Institutes.* Loughborough: Loughborough University; Melbourne: Victoria University. Retrieved from http://ie-repository.jisc.ac.uk/279/2/JISC_data_sharing_finalreport.pdf.

Goldstein, S. & Oelker, S. K. (2011). Planning for data curation in the small liberal arts college environment. *Sci-Tech News, 65*(3), 5-11.

Gorman, G. E. & Clayton, P. (2005). *Qualitative research for the information professional: A practical handbook*. Facet Publishing: London.

Haas J. & Murphy S. (2009). E-Science and libraries: Finding the right path. *Issues in Science and Technology Librarianship, 57.* Retrieved from http://www.istl.org/09-spring/viewpoint1.html.

Harvey, R. (2005). *Preserving digital Materials*. Berlin: De Gruyter.

Harvey, R. (2010). *Digital curation: A how-to-do-it manual*. New York: Neal-Schuman.

Higgins, S. (2011). Digital curation: Emergence of a new discipline. *The International Journal of Digital Curation, 6* (2), 78-88.

Key Perspectives Ltd. (2010). *Data dimensions: Disciplinary differences in research data sharing, reuse, and long term viability.* Edinburgh: Key Perspectives Ltd. Retrieved from: http://www.dcc.ac.uk/sites/default/files/SCARP%20SYNTHESIS_FINAL.pdf.

Liquid Crystal Institute. (2013). Retrieved from http://www.lcinet.kent.edu/research/index.php.

Marshall, C. & Rossman G. B. (2011). *Designing qualitative research*. Los Angeles: Sage.

National Information Standard Organization. (2013). Data curation. *Information Standard Quarterly, 25* (3), 1-40.

National Science Foundation. (2001). *Ensuring the integrity, accessibility, and stewardship of research data in the digital age*. Washington D.C.: The National Academies Press.

National Science Foundation. (2011a). *Advisory committee for cyberinfrastructure task force on software for science and engineering: Final report*. Arlington, VA: NSF.

National Science Foundation. (2011b). *Advisory committee for cyberinfrastructure task force on high performance computing: Final report*. Arlington, VA: NSF.

National Science Foundation (2012). *Sustainable digital data preservation and access network partners (DataNet).* Arlington, VA: NSF. Retrieved from http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=503141.

Schmidt, L., Ghering, C., & Nicholson, S. (2011). Digital curation planning at Michigan State University. *Library Resources & Technical Services*, *55*(2), 104-118.

Starr, J., Willett, P., Federer, L., Horning, C., & Bergstrom, M. L. (2012). A collaborative framework for data management services: The experience of the University of California. *Journal of eScience Librarianship, 1*(2), 109-114. DOI: http://dx.doi.org/10.7191/jeslib.2012.1014.

Walters, T. O. (2009). Data curation program development in U.S. universities: The Georgia Institute of Technology example. *The International Journal of Digital Curation*, *4*(3), 83-92.

Witt, M., Carlson, J., & Brandt, S.D. (2009). Constructing data curation profiles. *The International Journal of Digital Curation, 4*(3), 93-96.

Whyte, A. (2008). *Curating brain images in a psychiatric research group: Infrastructure and preservation issues.* Edinburgh: Digital Curation Centre. Retrieved from http://www.dcc.ac.uk/scarp.

Yakel, E. (2007). Archives and manuscripts digital curation. *OCLC System and Services International Digital Library, 23* (4), 335-340. DOI: 10.1108/10650750710831466