FLORIDA STATE UNIVERSITY

COLLEGE OF COMMUNICATION & INFORMATION

EXPLORING THE DATA MANAGEMENT AND CURATION (DMC) PRACTICES OF SCIENTISTS IN RESEARCH LABS WITHIN A RESEARCH UNIVERSITY

By

PLATO L. SMITH II

A Dissertation submitted to the
School of Information
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Degree Awarded
Summer Semester, 2014

Plato L. Smith II defended this dissertation on June 23, 2014.

The members of the supervisory committee were:

Paul Marty
Professor Directing Dissertation

Helen Burke
University Representative

Stvilia Besiki
Committee Member

Lorri Mon
Committee Member

The Graduate School has verified and approved the above-named committee members, and certifies that the dissertation has been approved in accordance with university requirements.

This dissertation is dedicated to my God, grandfather (Plato Smith), grandmother (Lula Smith), mother (Joyce C. Smith), son (Daryl), and grandson (Lil Daryl).

ACKNOWLEDGEMENTS

I thank God for my Lord and Savior, Jesus Christ, through which His Holy Spirit equipped me with the strength, knowledge, skills, abilities, and fortitude to complete this dissertation.

The completion of this dissertation was made possible by the participation of scientists at several high-profile research laboratories at the Florida State University and scientists associated with the National Science Foundation (NSF) EarthCube project.

The following individuals have been significant in their encouragement of my efforts to complete this dissertation and for that I am forever humbled, grateful, and thankful. I especially thank Dr. Christie Koontz and Dr. Corrine Jörgensen who encouraged me to apply to the doctoral program in the School of Library & Information Studies at Florida State University. I also like to thank Dean Larry Dennis, Dr. Kathleen Burnett, Dr. Paul Marty, Dr. Lorri Mon, Dr. Besiki Stvilia, and Dr. Helen Burke for providing the necessary academic support, guidance, and technical acumen required for successful doctoral matriculation. The following is a fractional representation of the list of individuals over the course of my academic and professional career that have contributed to my success in various facets of my life whether directly or indirectly: Dr. Greg Newby, Dr. Reginald Stevens, Dr. Althea Jenkins, Dr. Hannelore Radar, Dr. A.K.S.K. Prasad, Dr. Cheryl Ward, Dr. Peter Lazarevich, Dr. Lawrence Abele, Dr. Dominic Farace, Dr. Ann Thistle, Dr. Sylvia Spengler, Irene Lombardo, Dr. Barbara Ransom, Dr. Bob Chadduck, Dr. Marco Tedesco, Dr. Joel Cutcher-Gershenfeld, Mimi McClure, Dr. Ruth Swan, Jeneice Smith, Brenda Wright, Lula Smith, Regina Harris, Teresa Sargent, Diana DeBoer, Ken Baldauf, Geoffery Miller, Dr. Judith Devine, Dr. Nancy Marcus, Dr. Melissa Johnston, Dr. Gary Burnett, Chuck Thomas, Giesele Towels, Willa Patterson, Dr. Alicia Viera, Vanessa Reyes, FSU Libraries DLC OPS students, Dr. Laurie Taylor, Mark Sullivan, Jeanette Dummer, Michael Kaplan, Wei Dei, Sharon Schwerzel, Tamara Weatherholt, Priscilla Caplan, Dr. Robin Sellars, Dr. Jenny Ma, Dr. Benjamin Speller, Susan Nutter, John Little, Dr. Courtney Ferguson, Carmen Dorsey, Prof. Shearer, Lionell Parker, Prof. Chapman, Dr. Gibbs, Mary Lou Norwood, Chaplain Barry, Mark Lazarus, Dr. Amoateng, Phyllis Carrington, Portia McQueen, Shanthi Natesan, Dwayne McCoy, Tabatha, Harry James, Tangy (Big Sis), Cherry Richardson, Delores Littlejohn, my dad, and Loretta Rhyne.

The following organizations have been instrumental in my professional development: The National Science Foundation (NSF), Florida State University School of Information, Florida A&M University (FAMU) Libraries, Digital Curation Centre (DCC), International Conference on Grey Literature (GreyNet), UNC DigCCur, JCDL Doctoral Consortium, Cornell University Library/ICPSR Digital Preservation Management Workshop, UIUC Summer Institute for Data Curation in the Humanities, Association of Research Libraries (ARL), North Carolina State University Libraries, Duke University Libraries, Emerald Publishing, FSU Libraries, the Council on Library and Information Resources (CLIR)/DLF Postdoctoral Fellowship Program and The University of New Mexico College of University Libraries & Learning Sciences (UL & LS).

"In their hearts humans plan their course, but the Lord establishes their steps." – Proverbs 16:9

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Beginning January 18, 2011, proposals submitted to The National Science Foundation (NSF) must include a supplementary Data Management Plan (DMP) of no more than two pages. The NSF DMP requirement has significantly redefined the role of scientists, researchers, and practitioners in the United States of America (USA) by presenting the opportunity to engage in effective data management planning and practices for current and future use. In order for data to be useful to research, science, scholarship, and education, data must be identified, described, shared, discovered, extended, stored, managed, and consulted over its lifecycle (Bush, 1945; Lord & Macdonald, 2003; Hunter, 2005; JISC, 2006; UIUC GLIS, 2006/2010; NSF, 2011).

Within the scope of this research study data management planning is defined as the planning of policies for the management of data types, formats, metadata, standards, integrity, privacy, protection, confidentiality, security, intellectual property rights, dissemination, reuse/re-distribution, derivatives, archives, preservation, and access (NSF, 2011). The management of data includes analog [physical], digitized [made electronic] & born digital [no physical surrogate] data. NSF's data management plan requirements have incentivized the development of a multitude of programs, projects, and initiatives aimed at promoting and providing data management planning knowledge, skills, and abilities for NSF data management plan requirements compliancy. Without the specification, clarification, & definition of key concepts; assessment of current data management practices, experiences, & methods; interrelationships of key concepts; and utilization of multiple methodological approaches, data management will be problematic, fragmented, and ineffective. The accomplishment of effective data management is contingent on funders, stakeholders, and users' investment and support in *Infrastructure, Cultural Change, Economic Sustainability, Data Management Guidelines, and Ethics and Internet Protocol* (Blatecky, 2012, p. 5) across organizations, institutions, & domains.

One of the goals of the researcher "is to select a theory or combine [multiple theoretical perspectives] so they resonate with the guiding research questions, data-collection methods, analysis procedures, and presentation of findings" (Bodner & Orgill, 2007, p. 115) within a conceptual framework that "places its assumptions in view for practitioners" (Crotty, 1998). The introduction of the Conceptual Framework for Analyzing Methodological Suppositions (Burrell

& Morgan, 1979: Morgan & Smircich, 1980: Morgan, 1983, Solem, 1993) to gather competing approaches and paradigmatic assumptions for multiple paradigm *integration* and *crossing* via interplay (Schultz & Hatch, 1996) is an attempt by the researcher to build theory from multiple paradigms through Metatriangulation (Lewis & Grimes, 1999), a theory-building approach. Within this study, the Data Asset Framework (DAF) is framed as a sequential mixed methods explanatory research design (Creswell and Plano Clark, 2011) and applies social science research to facilitate scientific inquiry.

The purpose of this study is to investigate the data management and curation practices of scientists at several research laboratories at the Florida State University and select scientists associated with the National Science Foundation (NSF) EarthCube project. The goal of this research is not to provide extensive literature review to prove the need for effective data management practices but to provide empirical evidence to support current data management and curation practices. Within the scope of this dissertation, data management and curation practices will be generally defined as the effective aggregation, organization, representation, dissemination, and preservation of data. Data refers to analog and digital objects, databases, data sets, and research data. For purposes of discussions in this study, data is both singular and plural.

Data management and curation practices include four key concepts: (1) data management planning, (2) data curation, (3) digital curation, and (4) digital preservation. Literature review suggests that these key concepts when applied with relevant standards, best practices, and guidelines can assist scientists in ensuring the integrity, accessibility, and stewardship of research data throughout its lifecycle.

The combination of the conceptual framework for analyzing methodological suppositions (Burrell & Morgan, 1979; Morgan & Smircich, 1980; Morgan, 1983; Solem, 1993), Metatriangulation (Lewis & Grimes, 1999), and the Data Asset Framework (DAF) (JISC, 2009) contributes to the development of an interdisciplinary conceptual framework model concept capable of addressing the data management and curation issues common across disciplines. For the purpose of this dissertation "research data are being understood as both primary input into research and first order results of that research[1]" (ESRC, 2010, p. 2).

[1] Sustainable Economics inspired the definition for a Digital Planet: Ensuring Long-Term Access to Digital Information. Final Report of the Blue Ribbon Task Force on Sustainable Digital Preservation and Access, February 2010.

CHAPTER ONE
INTRODUCTION

> *"If we are to do an adequate job of conceptualizing, we must do more than just think up some definitions, any definitions, for our concepts. We have to turn to social theory and prior research to review appropriate definitions. We may need to distinguish sub-concepts, or dimensions, of the concept. We should understand how the definitions we choose fits within the theoretical framework guiding the research and what assumptions underlie this framework" (Schutt, 2006, p. 93.).*

The challenge of how to effectively manage research data affects all scientists and researchers within and across multiple domains. Managing research data through data management and curation (DMC) services (including data management planning, data curation, digital curation, and digital preservation) is a common topic in the Academic Research Libraries (ARL) and Library and Information Science (LIS) literature. Multiple perspectives on how researchers from different domains store, manage, and use research data are necessary to understand the challenge of managing research data effectively—a problem common to multiple disciplines.

> *Researchers report that they struggle unsuccessfully with storage and management of their burgeoning volume of documents and data sets that they need and that result from their work. While some universities have devised new services to better manage data and other information derived from research, many researchers flounder in a disorganized way and rising accumulation of useful findings may be lost or unavailable when conducting future research.* (Kroll & Forsman, 2010, p. 5)

1.1 Statement of the Problem

The challenge of effectively managing data within and across disciplines, is compounded by the additional research problems of: (1) definitional confusion and lack of clarification of key concepts and (2) lack of a systematic approach for the development of theory of data management and curation services. Literature review suggests (Merton, 1968) that the identification, clarification, and linking of the underlying suppositions of key concepts are necessary to develop theory. Sometimes key concepts are not identified, clarified, or linked. Data

management key concepts are not consistently identified, clarified, and linked. Data management and curation key concepts are frequently used interchangeably, which creates confusion, inconsistent application, and disjointed learning outcomes.

Digital preservation is one of the four key concepts of data management and curation introduced in this dissertation. The other three key concepts include (1) data management planning, (2) data curation, and (3) digital curation. These key concepts represent various stages and processes in the storage, management, and preservation of data over its lifecycle. A theory of digital preservation or digital curation should begin with the identification, clarification, and linking of these key concepts within an interdisciplinary conceptual framework model. Currently there is a need for interdisciplinary conceptual framework models (Whyte, 2012; Parsons et al., 2011) that clarify, identify, and link the underlying suppositions of these key concepts to build data management and curation theory. The identification, clarification, and linking of these key concepts were addressed by the data management and curation framework (see Fig. 1) developed from the preliminary study. Metatriangulation, a multiple paradigm theory building approach, was introduced in this dissertation to contribute to data management and curation theory.

1.2 Research Purpose

The purpose of this research study is to (1) identify and clarify the key concepts in data curation, (2) link those key concepts of data curation into a framework, (3) integrate the underlying suppositions of those key concepts within a paradigm conceptual model, and (4) use the data asset framework (DAF) methodology to explore the data management and curation practices of scientists at research labs. The implications from this study can lead to recommendations to improve the current data management and curation practices of participants.

1.3 Research Questions

The following research questions were developed to gather information on how data is currently stored, managed, and preserved by scientists at select research labs at FSU. In order to effectively manage research data, scientists need to identify, classify, assess, and organize their data. The Data Asset Framework (DAF) methodology is one approach for auditing data assets to improve current data management and curation practices. These research questions were answered using the DAF methodology as an approach to perform an assessment of data assets. It

is assumed that the adoption of relevant data management standards, best practices, and guidelines, where appropriate will lead to improving the data management and curation practices of scientists at research labs at FSU.

1. How do researchers create, manage, store, and preserve research data?
2. How can the identification and clarification of key DMC concepts be resolved within and across disciplines?
3. What are some of the theories, practices, and methods disciplines use to address research data management in your discipline?
4. How can multiple paradigms' perspectives on data management and curation practices within and across disciplinary domains contribute to building DMC research & theory?

The research questions were addressed through the use of Metatriangulation for data management and curation theory development and the DAF methodology operationalized within a conceptual framework for integrating multiple paradigm perspectives (see Fig. 7). The results from this dissertation study has met the goal of developing learning outcomes with implications that articulate the significance in improving the data management and curation practices where feasible, permissible, and beneficial.

1.4 Significance of Research

The significance of this research includes: (1) identification and classification of research data assets, (2) articulation for the need to address data management challenges, (3) development of data management and curation services policies and procedures, (4) improvement of current data management practices, and (5) contribution to the future development of data management plans that meet and/or exceed funding agencies' data management plan requirements. Also, this research introduces the key concepts of data management and curation for concept analysis and data management and curation theory development considerations.

1.5 Assumptions

The research project assumed: (1) there is a need for scientists at research labs at FSU to identify and clarify their research data, (2) there is a need to improve current data management and curation practices of scientists at research labs at FSU, (3) there is a need to develop a theory

of data management and curation based on theory and practice, (4) there is a need for the clear articulation and introduction of the key concepts of data management and curation to disciplinary domains beyond the dominant disciplinary domains in which most of the current data management and curation research takes place (i.e. ARL, LIS), and (5) scientists who participated in this research study will reevaluate their data management and curation practices.

1.6 Definitions

The identification, clarification, definition, and concatenation of underlying suppositions of the four key concepts of data management and curation are fundamental to this research study. As Denzin noted: "Herbert Blumer traced sociologists' inability to develop sound theory to a misunderstanding of concepts" (Denzin, 1970/2009, p. 33). Theory is more than simply identifying and defining key concepts but also requires the linking of underlying suppositions (Merton, 1968). Within the scope of this study, the definitions of the key concepts are:

1. ***Data management planning*** is a data lifecycle management process comprised of departmental, institutional, or organization policies and procedures governing the creation, organization, dissemination, preservation, and comprehensive lifecycle management of research data and information in accordance with relevant standards, best practices, and guidelines. Data management planning includes data curation, digital curation, and digital preservation. Data management and curation services include data management planning.
2. ***Data curation*** is a data lifecycle management process of providing descriptive, annotative, and representative information for research data through metadata.
3. ***Digital curation*** is a data lifecycle management process of storing, managing, and storing curated research data within a repository or digital content management system.
4. ***Digital preservation*** is a data lifecycle management process of maintaining the authenticity, integrity, and security of curated research data within a standards-based repository or digital content management system for long-term archival preservation.

The key concepts are defined in literature as follows:

1. **Data Management Planning [DMP]** is the planning of policies for the management of data types, formats, metadata, standards, integrity, privacy, protection, confidentiality,

security, intellectual property rights, dissemination, reuse/re-distribution, derivatives, archive, preservation, and access (NSF, 2011);

2. **Data Curation [DaC]** is the "active and ongoing management of data through its lifecycle of interest and usefulness to [research], science, scholarship, and education" (UIUC GSLIS, circa 2006) (includes analog, digitized, & born digital research data);
3. **Digital Curation [DiC]** is the "maintaining and adding value to a trusted body of digital information for future and current use; specifically, the active management and appraisal of data over the entire life cycle" (JISC, 2006) (includes digitized & born digital research data);
4. **Digital Preservation [DP]** is "the series of technical, strategic, and organizational actions and interventions required to ensure continued and reliable access to authentic digital objects for as long as they are deemed to be of value" (JISC, 2006) (includes digitized & born digital research data).

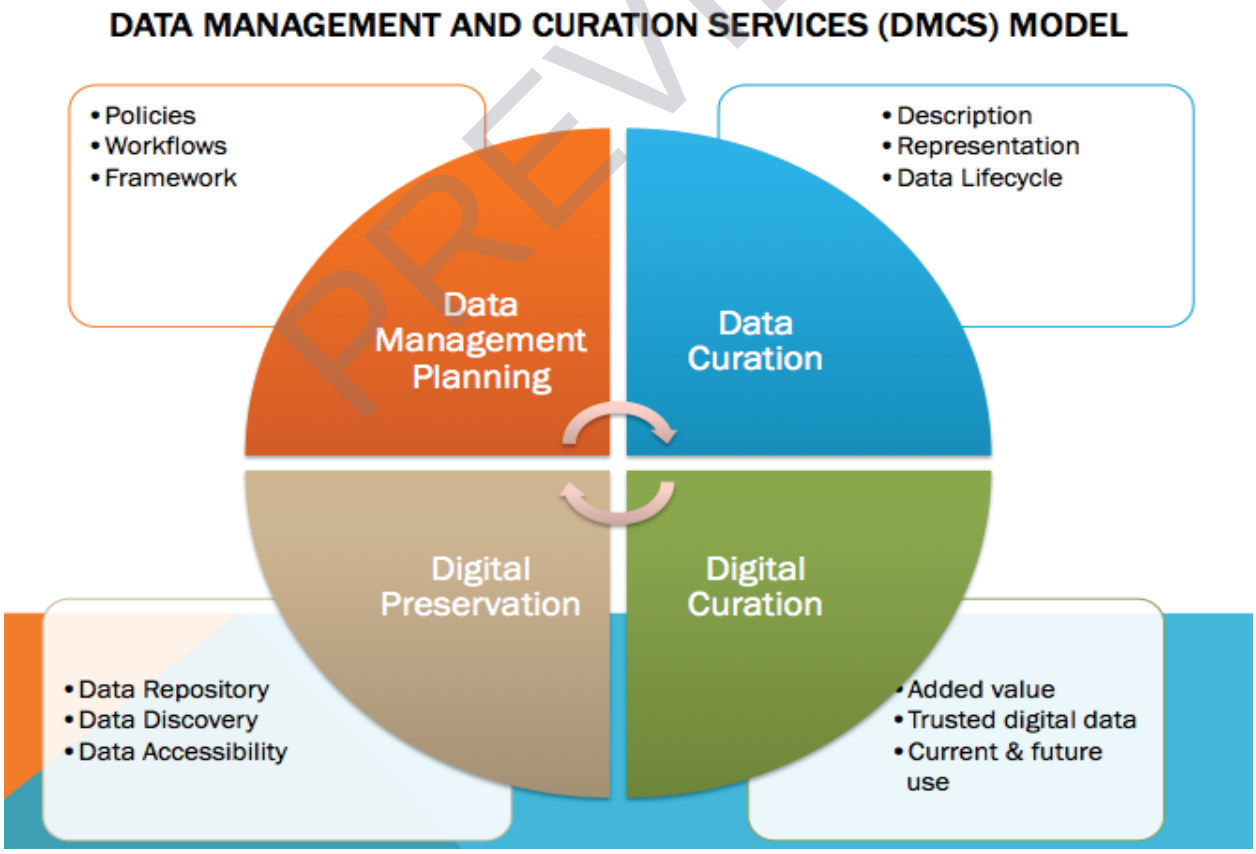

Fig. 1 – Data Management and Curation (DMC) Key Concepts

Additional concepts and terms relevant to this research are defined in literature as follows:

1. **Concept** – "A mental image that summarizes a set of similar observations, feelings, or ideas" (Schutt, 2006, p. 92);
2. **Cyberinfrastructure** – The integration of personnel, services, organizations, computing hardware, data and networks, digitally enabled sensors, observatories, and experimental facilities with base technology (computer and information science and engineering-CISE) and discipline-specific science (NSF CI Council, 2006, p. 6; Atkins et al., 2003) for "supporting and enabling large increases in multi-disciplinary science while reducing duplication of effort and resources across scientific domains" (Bowker et al., 2010, p. 100);
3. **Data** – "A reinterpretable representation of information in a formalized manner suitable for communication, interpretation, or processing" (CCSDS, 2002/2012, p. 20);
4. **Methodological** – "This assumptions holds that a qualitative researcher conceptualizes the research process in a certain way. For example, a qualitative inquirer relies on views of participants, and discusses their own views within the context in which they occur, to inductively develop ideas from particulars to abstractions" (Creswell, 1994; Creswell, 2007, p. 248);
5. **Paradigm** – The philosophical stance of a researcher whereby basic set of beliefs guide action (Denzin & Lincoln, 1994; Creswell, 2007, p. 248). "A paradigm is a set of generalizations, beliefs, and values of a community of specialists" (Creswell & Clark Plano, 2011, p. 39);
6. **Model** – "A model is a representation of an idea, object, event, process, or system, which concentrates attention on certain aspects of the system – thus facilitating scientific inquiry" (Briggs, 2007, p. 73);
7. **Reference Model** – "A reference model is an abstract framework for understanding significant relationships among the entities of some environment, and for the development of consistent standards or specifications supporting that environment. A reference model is based on a small number of unifying concepts and may be used as a basis for education and explaining standards to a non-specialist" (CCSDS, 2002/2012);
8. **Theoretical perspective** – The philosophical stance informing the methodology and providing a context for the process and grounding its logic and criteria (Crotty, 1998);

9. **Theory** – "A theory must contain a set of propositions [stated relationship] or hypothesis that combine descriptive and relational concepts" (Denzin, 1970/2009, p. 34).

1.7 Overview of Theory

Metatriangulation (Lewis & Grimes, 1999) is a theory-building approach that builds theory from the integration of multiple paradigms perspectives (Gioia & Pitre, 1990) through conceptualized triangulation (Denzin, 1970) that includes (1) multiparadigm reviews, (2) multiparadigm research, and (3) metaparadigm theory building based on multiparadigm exemplars. Building on previous research in organizational theory, Lewis & Grimes developed a meta-theory building approach that is most suited for disciplinary domains with inconsistent theories such as the field of data management and data curation within the ARL and LIS disciplines. DMC research in the LIS field is primarily dominated by ethnographic and pragmatic studies as evidenced by the preliminary study. A multiple paradigm perspective theory building approach such as Metatriangulation is enabled to facilitate an improved understanding of the significance of data management and curation common to multiple disciplinary domains.

1.8 Overview of Conceptual Framework

The conceptual framework for analyzing methodological suppositions (Burrell & Morgan, 1979: Morgan & Smircich, 1980: Morgan, 1983; Solem, 1993, p. 595) (see Fig. 5) is a general metatheoretical framework (Solem, 1993, p. 594) that provides a two-dimensional perspective for the methodological analysis of social theory (Solem, 1993) within a domain to facilitate scientific inquiry. This framework contains major suppositions such as the ontology, epistemology, frame of reference, concepts, and methods processes involved in the analysis of a problem. This conceptual framework accommodates the analysis, comparison, and integration of multiple perspectives in the investigation of a problem common across multiple domains.

1.9 Overview of Method

The DAF is a methodology composed of survey and interview research methods. The purpose of the DAF is to identify, assess, classify, and organize data assets in order to develop recommendations that will improve research data management. An adapted DAF methodology was operationalized within this study as a two-phase sequential mixed-methods explanatory

research design (Creswell and Plano Clark, 2011). The mixed-methods consisted of survey questionnaires administered in Phase 1 and semi-structured interviews conducted in Phase 2 in this study. It was assumed that the application of social science research methods to the DAF methodology would produce empirical data with implications to improve the data management practices of scientists of research labs at FSU and the wider DMC, ARL, and LIS communities.

1.10 Summary

This dissertation proposed the use of Metatriangulation of data management and curation as a theory-building approach, operationalization of the Data Asset Framework (DAF) methodology as an approach to investigate data management practices, and integration of a conceptual framework that explored the methods, concepts, frame of references, and how multiple disciplinary domains manage, store, and preserve their research data. Scientists from several research labs at Florida State University and select scientists associated with the National Science Foundation (NSF) EarthCube project were purposively selected to participate in this dissertation research. The data generated from this research study will serve as references for exploring current and future data management and curation practices. The learning outcomes from this study included research, practical, and social implications resulting in recommendations that contribute to improving the ways scientists manage, store, and preserve research data. The preliminary study[2] (see 3.11.1 Preliminary Pilot Study Results) on the data management and curation opinions of multiple stakeholders across multiple disciplinary domains conducted in December 2012 preceded and contributed to the development of this dissertation.

[2] GL15 Conference Proceedings. (2014). https://easy.dans.knaw.nl/ui/datasets/id/easy-dataset:57106

CHAPTER TWO
LITERATURE REVIEW

> *"Information science would do well to develop more and better theories, for without adequate theoretical support, we may do a technically brilliant job of solving the wrong problems.*" (Dow, 1977, p. 323)

2.1 Data Management and Curation (DMC)

2.1.1 Data Management Planning, Data Curation, Digital Curation, and Digital Preservation

A major aim of the data management and curation services community is to improve the management of research data for current and future use. Within the scope of this study, "research data" is data that you currently hold, that has been collected and/or used in the course of your research. Research data can be primary data collected by you or your research group or secondary data provided by a third party. It may be quantitative or qualitative (e.g. survey results, interview transcripts, databases compiled from documentary sources, images or audiovisual files) (McGowan & Gibbs, 2009). Currently, literature review of research data management and digital preservation/curation discourses promote the need for theory development in the area of data management and curation practices within the disciplinary domain of library and information science (LIS). Recently, there have been national and international calls for proposals for systematic and interdisciplinary approaches to improve data management, data curation, research on data, and digital preservation within and across disciplinary domains. These recent calls for proposals for theory of digital preservation and theory of digital curation research represent a growing trend. Digital preservation, theory of digital preservation, data curation, theory of data curation, data management and curation developments, and conceptual framework models/tools for integrating all of these concepts within and across multiple disciplinary domains continue to be phenomena of interest in the profession, academia, and data management research. Multiple disciplinary domains, particularly data intensive scientific disciplinary domains, will benefit from the development of a theory of digital preservation and theory of digital curation that embraces multiple paradigmatic perspectives such as the development of data management and curation (DMC) theory proposed in this dissertation. Even though the origins of data curation and digital curation started in the

academic research libraries, data management and curation, e-science, and LIS research communities, the scope and reach of data management and curation now includes all disciplinary domains. To better address the data management and curation research issues common across multiple disciplinary domains, the main key concepts of data management and curation must be identified, clarified, defined, articulated, linked, and associated with an established theory (i.e., social theory, organizational theory, systems theory, etc.) for improved understanding, comprehension, and knowledge transfer within and across multiple disciplines

The four key concepts of data management and curation (DMC) are:

1. Data Management Planning (Entire data lifecycle – DCC Curation Lifecycle Model)
2. Data Curation (Level 1 Curation - Traditional academic information flow)
3. Digital Curation (Level 2 Curation - Information flow with data archiving)
4. Digital Preservation (Level 3 Curation - Information flow with data curation) (Lord and Macdonald, 2003)

DMC practices include four major data lifecycle management processes that:

1. Fulfill departmental, institutional, organizational policies & data management requirements;
2. Provide data creation (primary, secondary, tertiary data), data publication, minimal data description;
3. Facilitate added value (metadata), management & storage of archived data over data lifecycle;
4. Integrate a series of technical & strategic actions and consultations to ensure continual data authenticity.

2.1.2 Data Management Planning

Progressive research and development of data management plans began to permeate throughout the wider research and learning communities with the announcement of the National Science Foundation (NSF) data management plan (DMP) requirement that became effective January 18, 2011. In brief, the NSF DMP requirement calls for proposals seeking NSF funding to provide a supplemental document, not to exceed two pages, detailing the dissemination and sharing of research results congruent with the NSF policy on managing research data as outlined