Organizing scientific data sets: Studying similarities and differences in metadata and subject term creation

Hollie C. White

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the School of Information and Library Science (Information Science).

Chapel Hill
2012

Approved by:

Jane Greenberg

Brian Heidorn

Robert Losee

Richard Marciano

Javed Mostafa

ABSTRACT

HOLLIE C. WHITE: Organizing scientific data sets: Studying similarities and differences in metadata and subject term creation
(Under the direction of Jane Greenberg)

BACKGROUND: According to Salo (2010), the metadata entered into repositories are "disorganized" and metadata schemes underlying repositories are "arcane". This creates a challenging repository environment in regards to personal information management (PIM) and knowledge organization systems (KOSs).This dissertation research is a step towards addressing the need to study information organization of scientific data in more detail.

METHODS: A concurrent triangulation mixed methods approach was used to study the descriptive metadata and subject term application of information professionals and scientists when working with two datasets (the bird data set and the hunting data set). Quantitative and qualitative methods were used in combination during study design, data collection, and analysis.

RESULTS: A total of 27 participants, 11 information professionals and 16 scientists took part in this study. Descriptive metadata results indicate that information professionals were more likely to use standardized metadata schemes. Scientists did not use library-based standards to organize data in their own collections. Nearly all scientists mentioned how central software was to their overall data organization processes. Subject term

application results suggest that the Integrated Taxonomic Information System (ITIS) was the best vocabulary for describing scientific names, while Library of Congress Subject Headings (LCSH) was best for describing topical terms. The two groups applied 45 topical terms to the bird data set and 49 topical terms to the hunting data set. Term overlap, meaning the same terms were applied by both groups, was close to 25% for each data set (27% for the bird data set and 24% for the hunting data set). Unique terms, those terms applied by either group were more widely dispersed.

CONCLUSIONS: While there were similarities between the two groups, it is the differences that were the most apparent. Based on this research it is recommended that general repositories use metadata created by information professionals, while domain specific repositories use metadata created by scientists.

Salo, D. (2010) Retooling libraries for the data challenge. *Ariadne 64*. Available at: http://www.ariadne.ac.uk/issue64/salo/

In honor of Dr. Deborah Barreau, a dedicated teacher whose influence can be seen throughout my work.

ACKNOWLEDGEMENTS

A dissertation is never written in solitude. I would like to thank my husband, Curt Brinkmeyer, for patiently and continually encouraging me thorough the entire doctoral process.

I am grateful for the guidance provided to me by the faculty and staff at the School of Information and Library Science, especially my committee members: The Robert Losee, Javed Mostafa, and Richard Marciano. Plus, special thanks to Bryan Heidorn for taking the time to serve as my external member. I also received helpful advice and support from members of the Dryad Data Repository team, including Ryan Scherle, Craig Willis, Elena Feinstein, and Todd Vision.

In conclusion, I would like to acknowledge and sincerely thank Dr. Jane Greenberg for the constant counsel and advice.

TABLE OF CONTENTS

LIST OF TABLES

Table

LIST OF FIGURES

Figure

1. INTRODUCTION

Digital scientific data sets are information objects that are kept, shared, and reused by scientists and information professionals in repositories, digital collections, and libraries (Wallis, Mayernik, Borgman, & Pepe, 2010). Developing best practices for maintaining and organizing these information objects is a growing area of study and research (Big Data, 2008; Heidorn, 2008; Salo, 2010).

Salo (2010) raises many issues about digital data repositories maintained by information professionals. One area addressed is information organization. Salo (2010) comments that the data being deposited into repositories "tend to be disorganized, poorly described if described at all, and in formats poorly suited to long term reuse". Salo's observation suggests that information professionals should organize data in more forward thinking and sustainable ways. Yet, later in the same piece, Salo (2010) also comments that, "libraries can no longer cling desperately to decrepit, arcane, inward-focused standards such as MARC, not if the ultimate goal is to be part of a great global sea of data." This comment expresses Salo's belief that traditional knowledge organization systems used by libraries do not represent scientific data accurately or well. While influential and likely relevant, Salo's article presents the current reality of maintaining and curating scientific digital data sets, but does not include supporting data for many of her statements. Research examining the organizing issues related to research data is greatly needed. This research is a step towards addressing this need

2. PURPOSE

The purpose of this dissertation is to study how information professionals and biological researchers use information organization techniques, specifically metadata creation and subject term application, when working with scientific data sets. By examining and comparing the organizing behavior and output of these two groups, recommendations are made, within the context of this study, to improve repositories designed to accommodate the special needs of scientific data sets.

This dissertation begins by defining five terms that are central to the research. Next, a literature review will address research about personal information management and knowledge organization that supports this type of research. Following the literature review, an overview of research related to the Dryad Repository is presented. Preliminary pilot and exploratory studies that were conducted as a foundation for this research are discussed next. After the discussion of foundational research, the methodology is presented and followed by results. The discussion section addresses findings from the results within the context of this study, which is followed by the conclusion.

3. CLARIFYING TERMS

Five terms integral to understanding this research are defined in this section. These five terms are “organizing”, “scientist”, “information professional”, “metadata”, and “subject terms”. These five terms will be used throughout the paper to describe specific concepts central to the research discussed in this study. Use of these terms without clarification may be confusing. For that reason, ample clarification is being provided for the benefit of the reader.

For the purpose of this dissertation, the term “organizing” is used a signifier for all activities related to the areas of information organization and knowledge organization. Organizing in this sense includes cataloging, classification, metadata creation, subject term application, and arrangement. Information organization and knowledge organization are discussed in more detail in Section 4.

For the purpose of this research, the term ‘scientist’ is used as a signifier for those participants that conduct research in bioscience areas. Scientists who participate in this study work in either academic or research positions where they interact with, create, and reuse digital data sets in the biosciences. These areas include, but are not limited to biology, botany, chemistry, marine sciences, genetics, and paleontology. The purpose of this definition is to give context to the use of the term “scientist” within this research. A more detailed discussion found in the Results section gives more insight into scientists as a participant group.

The term "information professional", for the purpose of this research, is used as a signifier for library and/or information scientists who work in library or repository positions and either work with or have the potential to work with scientific data. Information professionals will typically, but may not always, hold a post-baccalaureate degree in library and information science. Information professionals have experience working with standards in an information science environment. The purpose of this definition is to give context to the use of the term "information professional" within this research. A more detailed discussion found in the Results section gives more insight into information professionals as a participant group.

For the purpose of this research, the term "descriptive metadata" is used as a signifier for a type of organizing output that will be analyzed. Metadata, within this study, are basic observable elements, such as title, description, author, and date. These properties are not subject or aboutness related. The purpose of this definition is to give context to the use of the term "descriptive metadata" within this research. The types of metadata that will be collected and analyzed in this study are addressed more thoroughly in the Results section.

The term "subject terms", for the purpose of this research, is used as a signifier for another type of organizing output that will be analyzed. Subject terms are words that are selected to describe the aboutness of scientific data sets. Subject terms, within this study, will include topical, geographic, and taxonomic descriptors. The purpose of this definition is to give context to the use of the term "subject terms" within this research. The use and analysis of subject terms is discussed in more detail in the Results section.

4. LITERATURE REVIEWS: KO AND PIM

Two research areas influence this study: knowledge organization (KO) and personal information management (PIM). My perspective on this topic is that scientific data repositories present a unique environment where knowledge organization and personal information management converge. Information professionals create repositories that are based on traditional knowledge organization theories and schemes. Data repositories are different from personal information management systems. A scientific data set is created by a scientist or scientific team and is organized in a way that reflects the needs of those individuals. The data set reflects the personal organization scheme of the scientist creator. The section that follows includes literature reviews about knowledge organization and personal information management.

4a. Knowledge Organization (KO)

Researchers and theorists, such as Hjorland (2007, 2008), Dupre (1993, 2006), Miksa (1998), and Svenonious (2001), have dedicated careers to trying to understand and explain the importance and various nuances of knowledge organization and the more tangible functions of information organization. From early on, much of the research and theoretical advancements explored in this area has focused on trying to define what knowledge organization is, what the area entails, and how knowledge organization impacts the world of information science and beyond (Dupre, 1993, 2006; Hjorland, 2007, 2008; Huberthal,1998). Through this body of work it can be concluded that

knowledge organization is both a theory and a practice. The literature review that follows will discuss theory and practice separately.

4.a.1 Knowledge Organization in Theory

In this section the definitions and theory that explain the research area known as knowledge organization are examined. First, this chapter explores the philosophical and definitional underpinnings found in the two separate areas of knowledge and organization. Then, the literature review moves on to look at the historical and traditional perspectives and theories found in the information and library science community about knowledge organization. Lastly, this piece discusses emerging theories in knowledge organization that are changing how knowledge and information could be presented to users of libraries and other information environments.

4.a.1.a Foundations: Organization and Knowledge Defined

The literature examined in this portion of the review presents knowledge organization as having foundations in both the study of organization and knowledge. Thoughts and research surrounding the understanding of both organization and knowledge are divergent and intriguing.

4.a.1.a.1 Organization

Researchers in a variety of domains indicate that organization occurs internally as well as externally — expressed through physical actions as well as mental or psychological understanding (Zerubavel, 1991; Hunter, 2002). Physically, organization is being performed constantly through acts of spring cleaning or sorting (Jones, 2007a; Barreau, 2008). Mentally, ideas are divided into like and unlike (Hunter, 2002).

Organization occurs whenever there is a choice to be made. The literature highlights that a variety of words, such as, 'dividing', 'matching', 'sorting', and 'weeding', are used to describe activities related to the act of organizing. Researchers use a variety of terms to describe organization as well as to explain various manifestations and underlying meanings behind organization.

Zerubavel (1991), a sociologist, presents organization in understandable terms, explaining that organization, "the way we divide our surroundings, for example, determines what we notice, what we ignore, what we eat and what we avoid eating". While his research is mainly concerned with the social consequences of organization and its functions, Zerubavel (1991) discusses organization in terms of 'making distinctions', 'boundaries', and 'divisions'. He comments that, "separating entities from their surroundings is what allows us to perceive them in the first place. In order to discern any 'thing', we must distinguish that which we attend from that which we ignore" (Zerubavel, 1991). In this definition, organization allows us to create order out of chaos by creating divisions that create groups of like things that excludes unlike things. Organization becomes a choice about what people include in their lives. Zerubavel's (1991) examination of organization shows that the existence of organization embraces dichotomy because organization itself can be both trivial and important by stressing that each decision has an impact either big or small on the outcomes of everyday life. While Zerubavel's (1991) findings highlight the 'distinctions' or 'differences' perspective of organization, other researchers (Lakoff, 1987; Hunter, 2002) focus on the connection between like things.

Another perspective of organization comes from Lakoff (1987), a cognitive linguist, who describes organization in terms of categorization. His research articulates that the connections between like things, items that have things in common, are the beginnings of categorization. Lakoff (1987) believes that movements, speech, and understanding are "automatic and unconscious" reflections of categorization and explains that, "an understanding of how we categorize is central to an understanding of how we think and how we function and as such is therefore central to an understanding of what makes us human". This conclusion is worthy of note because organization becomes an essential characteristic of human existence. This definition of organization shows how integrated organization and organizational functions, such as thought, movement, and speech, are in our everyday world.

From more of an information and library science perspective, 'classification' is the term used by Hunter (2002) for organization. He believes that, "in essence the process of classification simply means the grouping together of like things according to some common quality or characteristic" (Hunter 2002). Classes are distinguished by including things that are the same and excluding things that are different. Hunter's definition shows an integration of both Lakoff's (1987) and Zerubavel's (1991) perspectives.

4.a.1.a.2. Knowledge

Knowledge organization not only has its origin in the understanding of organization, but has been developed through the study of knowledge. For this reason, the discussion of knowledge can be approached and researched from two different perspectives in relation to ILS: (1) philosophy and (2) information and library science.

These perspectives work together to form an intellectual grounding for knowledge organization research in library and information science. In this section, knowledge is discussed from a philosophical perspective and then the information science perspective is examined.

4.a.1.a.2.a. Philosophy's Perspective of Knowledge

Understanding knowledge has been the subject of research and theoretical discourse for many generations of philosophers. This is the domain of epistemologists, philosophers who study, theorize, and write about knowledge. Plato, Aristotle, Kant, and Locke, are some of the more well-known philosophers whose efforts went towards defining the nature of knowledge with consideration to its relationship to human thought and existence. The questions, 'what is knowledge?' and 'how do humans gain knowledge?', are central to these philosophical writings that have been reproduced both in print and electronically for many centuries.

The questioning of knowledge starts with philosophical thought and writing. In the *Theaetetus*, Plato (369BC/1999) introduces and then debunks three views of knowledge. He reasons that knowledge is neither perception, true judgment, or true judgment with an account. While Plato (369BC/1999) does not define what knowledge is, he does set the stage for future definitions of knowledge. Aristotle, a student to Plato, continues this discussion of knowledge. A philosopher of many things, in the *Metaphysics,* Aristotle (1st Century CE/199x) discusses knowledge as describing a thing—what is present when knowledge is present and what occurs when knowledge is absent. These questions about knowledge begin a much larger discourse on the nature of knowledge and human beings that continues even today. These early epistemologists

and the questions they ask have been highly influential in developing the foundation of how knowledge has be analyzed and studied by other philosophers and even information scientists. These concepts of knowledge have had a tremendous impact on subsequent epistemologists in the enlightenment period.

Later enlightenment philosophers delve deeper into trying to understand and define knowledge. Locke (1690/2004), in *An Essay of Humane* [sic] *Understanding*, posits that all humans are born a blank slate. Through this work, Locke (1690/2004) reasons that humans gain knowledge over time through experience and that complex ideas, or knowledge, can only come from the combination of simpler ideas. Knowledge in this sense is outside of the human mind—an external force that has to be digested (Locke, 1690/2004).

Kant discusses knowledge as well, but his vision of the human mind and knowledge is different from Locke's(1690/2004). In the *Metaphysics*, Kant (1781/2003) argues that the mind is not a blank slate. Instead, Kant (1781/2003) proposes that there are two types of knowledge: the *a posteriori*, synthetic knowledge, which a person gains from experience; and the *a priori*, analytic knowledge, which humans have independent of experience. The mind gives shape to information that is found in the outside world—employing both *a priori* and *a posteriori* knowledge. Kant's (1781/2003) definition explains that knowledge does not exist in the outside world, but is created by the mind. This approach goes against Locke's (1690/2004) blank slate idea.

By no means are these four philosophers alone in their discussion of knowledge. Other enlightenment philosophers, and then modern philosophers, continue this debate over the true nature of knowledge and the role of the human mind. Though they are not

discussed in this literature review in depth, these philosophers also made contributions to the understanding of knowledge and how it is now discussed in the information science and library community. Overall, the discussions of epistemologists are foundational to the information science view of knowledge organization because it is important to understand the first conceptions of knowledge before creating systems of knowledge.

4.a.1.a.2.b. *Library and Information Science's Perspectives of Knowledge*

As the field and study of knowledge organization has developed within information science, researchers in our field have also asked questions about knowledge. The literature shows that many of these questions revolve around the question of 'what is knowledge?' and 'what is information?' and 'how does knowledge and information interact?'. This definition of research in terms of knowledge and information occurs in a variety of sub-domains within information science.

Ackoff (1989), a researcher in the knowledge management community, identifies five levels of content in the human mind: data, information, knowledge, understanding, and wisdom. Data is functional and can be transformed into information, while information is moved by instruction to knowledge. Knowledge is the product of learning and increases with efficiency. In this way, knowledge is considered "know-how" and is obtained "either by transmission from another who has it, by instruction or by extracting it from experience" (Ackoff, 1989). Ackoff's (1989) account shows that understanding is developed through knowledge. While knowledge and information may expire, wisdom (and sometimes understanding) is permanent. This interpretation shows that knowledge is part of a complex chain where wisdom and understanding are impossible without knowledge

Similar to Ackoff's studies (1989), the research of Meadow, Boyce, Kraft, & Barry (2007), who study the concept of knowledge from an information retrieval perspective, look into defining terms like knowledge, information, data, wisdom, and intelligence in ways that can inform and enhance information retrieval. Meadow et al. (2007) define knowledge in terms that mimic Plato, explaining that knowledge is justified true belief that depends on community acceptance. Meadow et al.'s (2007) definition of knowledge is a collection of information from multiple sources. Knowledge, in Meadow et al.'s estimation, relates to databases and retrieval because of the underlying knowledge base.

In one of ILS's seminal works, Buckland (1991) outlines his study and definition of information in relation to knowledge---one of three ways in which he interprets information. The other two ways include: information as process and information as thing. Buckland (1991) explores the concept of "information-as-knowledge", by explaining the intangible nature of knowledge and how it represents itself through surrogacy to become information. Therefore, any information (coming from knowledge) in a tangible form becomes "information as thing". Buckland's perspective is elaborated on by Svenonius (2001), who conceives of knowledge in relation to information, data, library units, and documents. Through her research, Svenonius (2001) acknowledges that the concept of knowledge can be confined to "facts or true beliefs". Svenonius's definition of knowledge shows that knowledge is information dependent. The discourse within the information and library science research framework has the potential to impact ILS work. The definition of knowledge itself directs how knowledge and knowledge products can be organized and arranged.

The area of knowledge organization has a rich history coming from both the tradition of looking at knowledge and organization separately. Combining the two terms creates the discipline of knowledge organization that is studied by a variety of researchers in and outside of library and information science. Knowledge organization has a solid history in the library and information science community, but is a topic that also exists in the study of personal information management (PIM). While knowledge organization's relationship to PIM is discussed in more depth in a subsequent chapter, the next section of this literature review examines the varied historical and traditional perspectives and theories about knowledge organization found within the information science community.

4.a.1.b. The Development of Knowledge Organization Theory in Libraries

Knowledge organization is an area of theoretical and practical research in library and information science. Historically, is seems obvious, that as libraries began building collections, the need to structure that knowledge in logical ways became essential (Strout, 1969; Tait, 1970). Knowledge organization is called many things within the library and information science community. Some terms that are used include classification, documentation, and information organization (Olsen, 1998; Svenonius, 2001; Hunter, 2002; Briet,1951/2006). Information scientists debate about the purpose and scope of knowledge organization and look at the historical or social development of knowledge organization as ways to better understand the knowledge organization systems (KOS) that are used within the information and library science community. The efforts to define and theorize about knowledge organization have been numerous, so only a select few are examined in this literature review. The definitions and theories explored here are

included because they focus on the areas of classification and cataloging systems found in libraries.

Using the term "classification", Olsen (1998) discusses knowledge organization in relation to the Dewey Decimal System. Her approach to knowledge organization is more socially based and provides a different perspective than that presented by philosophers and historians. In her discussion, Olsen (1998) characterizes classification as systems of knowledge that have boundaries that often marginalize individuals because they are built upon certain socio-cultural foundations. In this sense, knowledge organization is a social construct that, "reflect the relationships perceived in the wider society". Olsen's (1998) discussion points out how theories and movements to classify the world's knowledge have tried to be holistic yet contain bias that limit the effectiveness of these systems. She states that, "no classification will ever be inclusive" and says that the classificationists who create these theories will always reflect the bias of the time period and place in which the information is taking place (Olsen, 1998).

Miksa (1998), another researcher interested in the Dewey Decimal System, also discusses the 'classificationists of knowledge', but takes this discussion in a different direction by taking a historical perspective instead of a social one. In his examination of the movement to classify knowledge, Miksa (1998) discusses a connection between the movement to classify the sciences and the movement to classify knowledge. He states that the drive to classify science was started by individuals who wanted to research and study phenomena from a more measured, logical approach – people like Tommaso Campanella, Carl Linnaeus, and Francis Bacon-- and soon became the domain of the encyclopedists to record (Miksa, 1998). Even with this strong emphasis on people who

looked at the world of knowledge organization and tried to create an system of knowledge for use by all, Miksa (1998) concludes that there is no literal connection between the library classificationists and the work done by knowledge scholars in the 16th, 17th, and 18th century. He states that the work being done by Dewey, the Library of Congress, and Cutter was not thoroughly documented nor are there any strong connections between theses knowledge organization systems and the classification that was going on by scholars outside of libraries during that time.

Babb's (2005) interpretation of this historical time period and the connections between the systems of knowledge created by classificationists in and outside of information and library science is slightly different from that explored by Miksa (1998). Babb (2005), a practicing cataloger, states that, " 18th century scholars sought to classify knowledge; 19th century library theorists integrated scientific concepts of classification […] into the bibliographic realm." This statement shows a belief that the research conducted in the 18th century by people classifying science directly lead to the knowledge organization systems that were created in the 19th century.

I believe both Miksa's (1998) and Babb's (2005) statements about the development of knowledge organization to be true. Miksa (1998) points out that there is no solid, concrete written proof that connects the thoughts and knowledge organization systems of the information classificationists to the work of the scientific classificationists, though he admits there is a connection. In contrast, Babb (2005) establishes a direct link between these two groups—stating it as almost a cause and effect equation. These two historical approaches when incorporated with the social perspective examined by Olsen (1998) show a more realistic pattern of how knowledge organization research formed

knowledge organization theory and schemas found in information environments. While Babb's (2005) and Miksa's (1998) interpretations of history may not agree about the links between ILS classificationists and the classificationists of science, both scholars point to the library science theorists of the very late 19th and 20th centuries as establishing the "knowledge foundations" on which most of information and library science current conceptions of knowledge is now based. Some of the underlying theory found during the late 19th into the 20th century can be found in the written works of Otlet (1903/1990), and Richardson (1930).

Otlet's (1903/1990) essay on bibliography, while focusing on practical subjects like how to catalog and classify books in a library setting, has its foundation in the idea that all human knowledge is unified into one large scheme where each piece of information has a distinct place in one large hierarchy of understanding. Rayward (1975) discusses the theoretical underpinnings of the International Institute for Bibliography (IIB), a society created by Otlet. In this discussion, Rayward (1975) characterizes Otlet's opinions on the need for formalized knowledge organization schemes as a,

> "incontestable necessity for a universal bibliographic repertory. Such a repertory, he [Otlet] believed, could properly be conceived of only as universal in scope. However many were the divisions and subdivisions of human knowledge, functionally, essentially it was a unity. No more than in Nature could there be found within its campus isolated, absolutely independent facts".

This explanation of Otlet's theory of knowledge emphasizes the point introduced earlier that all knowledge is one large unified structure. Miksa (1989) agrees with the points made by Rayward (1975) and even emphasizes that Otlet's view (1903/1990) is shared by formalized knowledge organization systems (KOS) like the Dewey Decimal System—

a system created in the late 19th century that was eventually adopted and endorse by Otlet. Documentalists of the 20th century, like Briet (1951/2006), maintained this idea of unity of knowledge as explained by Otlet (1903/1990) through the organization of archives and libraries.

Another researcher and theorist, Richardson (1930) explicitly discusses unity as the underlying framework for the purpose of bibliographical classification. He begins by connecting the order of the sciences with the order of things. In this sense things can be ideas, and therefore knowledge. Richardson (1930) explains, "the end is a whole, the process is a defining of classes and the binding of these classes together as a whole". Miksa (1998) explains Richardson's theory by stating, "while the classification of knowledge necessarily meant the identification and arrangement of ideas as categories, such ideas must in actuality correspond to "things"—that is to objects and, ideally, to all objects, real or imagined, material or immaterial, in existence". In both Richardson's (1930) lectures and Miksa's (1998) evaluation of Richardson, the underlying theory of unity is laid out in explicit terms showing that every idea, thing, or book has one specific place within a highly structured knowledge organization system—every piece together forms a piece of the bigger whole.

In the 21st century, different models and theories about knowledge organization are being presented. One more recent definition and theoretical explanation has been put forward by Hjorland (2007, 2008), who considers knowledge organization from two different perspectives. Hjorland (2007, 2008) discusses knowledge organization in both narrow and broad senses. Specifically, Hjorland (2007) states "KOS in a narrow, IS-oriented sense are those systems related specifically to organizing bibliographical records

(in databases), whereas KOS in a wide, general sense are related to the organization of literatures, disciplines, and people in different cultures". Hjorland's (2007,2008) definitions are analyzed in more depth in a later literature review chapter, but should be of note here because they show a solid foundation of how KOS is interpreted in theoretical, conceptual ways, as well as practical.

Knowledge organization continues to exist as an expanding and evolving area based on research and theory development done not only inside the information and library science community, but outside the field as well. With the development and promotion of international knowledge organization societies, emerging diverse definitions and theories about the way knowledge should be organized are being explored.

The following section of this literature review examines new theories that diverge from the traditional notion of the unity of all knowledge. The discussion found in the next section points to scholars outside of the information and library community, specifically in the sciences and interdisciplinary studies, who are creating emerging theoretical models that could potentially influence the information science perspective of knowledge organization and change the way knowledge organization systems are perceived and created.

4.a.1.c. New Theories about Knowledge Organization

With the emergence of Web 2.0 technologies and the proliferation of global connectivity, the traditional conception of what is knowledge organization is changing (Lopez-Huertas, 2008). A number of knowledge organization theorists and researchers in information science, the sciences, and interdisciplinary studies have been looking at

alternative knowledge organization schemas after finding that traditional knowledge organization methods, like the Linnaeus' *Systema Natura* in biology, or the traditional knowledge organization schemes, like Library of Congress Classification schemes, are not representing the realism of how the world's knowledge is organized. A number of theorist outside of the tradition library environment, like Dupre (1993, 2006), Huberthal (1998) and Weinberger (2007), are promoting a new theory of knowledge that is more social and interconnected than the unified approach discussed above.

In his theoretical scientific work, Dupre (1993) discusses his theory on the limitations of traditional knowledge organization schemes using the terms "plurality" and "unity" to describe the different approaches to organizing knowledge. "Unified" approaches are very similar to traditional, hierarchical systems created and used by libraries, while "plural" approaches are more social and flexible in function. Dupre (1993) explains that for centuries, the majority of modern science fully supported the theory of "a deterministic, fully law-governed, and potentially fully intelligible structure that pervades the material universe" (Dupre, 1993). This theory establishes that in a knowledge organization scheme each idea/piece of information would have a single place within one large structure: a hierarchy of kinds of objects, a hierarchy of theories, and what kinds of things the world contains.

Traditional approaches to organizing knowledge rely on unity and order, but there has been increase support in the idea that disorder, as opposed to order, is a more realistic representation of the way the real world is organized (Dupre, 1993, 2006; Huberthal, 1998; Weinberger, 2007). As discussed previously, traditional knowledge organization schemes are hierarchical and organize information linearly. Concepts are broken down in

parent to child type relationships as subjects move from broad to more specific. Topics can be related through broader or narrower terms and synonym—to just name a few of the relationships available. Everything in these systems is closely considered, well-thought out, and precise. In relation to unity, it " is often supposed that classification, and especially scientific classification, distinguishing between different kinds of things in the world, is an activity that has revealed, or can be expected to reveal, an orderly, unique, and perhaps hierarchical arrangement of things" (Dupre, 1993). Many interdisciplinary scholars and scientific philosophers reject the theory of unity. While unity may be easy to display in hierarchical KOS and knowledge representational diagrams, it does not reflect true life, nature, or reality. The theory of pluralism does (Dupre, 1993).

Pluralism is the ability of one concept to be located in many places within a knowledge organization scheme. Though this concept may be called different names in different subject domains, it has a place in each and is not claimed solely by one discipline. In pluralism, as opposed to unity, "there are many equally legitimate ways of dividing the world into kinds" and that there are many divisions that exist simultaneously (Dupre, 1993). Accepting the pluralism theory and breaking away from hierarchical, unified schemes, actually reflect the "true ontological complexities of the world" (Dupre, 1993). Dupre's theory (1993, 2006) has been highly influential in the sciences and for those people interested in new models for organizing knowledge. Librarians, like Jones (2009) equate these pluralities—as seen through the digital expansion of the internet—as being a symptom of the Web 2.0 and social networking phenomena.

Expanding upon this theory and similarly putting it back into the information science domain, Weinberger (2007) believes that the physical limitations of materials in

the past has restricted how knowledge could be organized and, now the new digital environment has allowed organization to go beyond the physical. He elaborates that "as we invent new principles of organization that make more sense in a world of knowledge freed from physical constraints, information doesn't just want to be free. It wants to be miscellaneous" (Weinberger, 2007). The concept of miscellaneous order emphasizes the ideas of disorder or multiple, simultaneous ways of organizing—this is very similar to the discussion of plurality introduced by Dupre (1993, 2006). Weinberger (2007) opposes the idea of organization as "standardization in order to drive out efficiency"— a concept that opposes the chain of the human mind introduced by Ackoff (1989) where knowledge embodies efficiency. The idea of things being classified in many different ways is supported by core knowledge organization scholars, like Langridge (1992).

Interdisciplinary studies is another area that is often frustrated with the limitations found in traditional knowledge organization schemes. Huberthal (1998), a German professor in interdisciplinary studies, explores the unity and plurality theories on knowledge organization, commenting that,

> "hopes should not be pinned to an approach which attempts with a single blow to establish unity among sciences on a meta-theoretical level, for example, the thesis that there is a method monism among all sciences or an ultimate unity of all scientific topics. Apart from the fact that meanwhile the sciences are too diverse and complex, such a theory would inevitably be much too general to provide concrete directives for the procedure of subject-overlapping research on a specific topic".

Huberthal (1998) expresses the view that pluralism involves *overlapping* between things and not hierarchy. She emphasizes how a "one plan fits all" model does not necessarily work for domain level knowledge organization. According to another researchers interdisciplinary studies, Klein (1999),

> "the metaphor of knowledge as a foundation or a linear structure has been replaced by images of a network, a web, and a dynamic system. Comparably, the metaphor of unity, with its accompanying values of universality and certainty, has been replaced by metaphors for plurality and relationality in a complex world".

If plurality reflects worldly complexity, then it is necessary for information science to be aware of this reality.

Opinions from outside of the knowledge organization community are now becoming more central to the International Society for Knowledge Organization Systems (ISKO) cause. One researcher gaining the attention of the ISKO community is Szostak (2004), a Canadian economist, who argues that the way knowledge is currently organized by the information sciences is not useful to the expert researcher and that new approaches should be taken. Szostak and other researchers in the knowledge organization community, such as Gnoli and ISKO Spanish chapter, are now working on new systems that include focusing on organizational schemes that are more interdisciplinary. Specifically, Szostak (2007) proposes organizing knowledge based on phenomena, method, and theory—where "a new KOS should allow users to shift from one perspective or viewpoint to another, thus reflecting the multidimensional nature of complex thought". These ideas were embraced by the Italian chapter of the ISKO in a document entitled "The Leon Manifesto"[1], and are considered a relevant proposal for the future of knowledge organization. The knowledge organization community, specifically the Networked Knowledge Organization Systems/Services (NKOS) group, has discussed the relationship of social tagging (a pluralistic approach) to knowledge organization and is still considering the consequences of combining these approaches (Nielsen, 2008).

[1] The Leon Manifesto: http://www.iskoi.org/ilc/leon.php

4.a.1.d Summary

Examining the theory that makes up the field of knowledge organization shows that this area of study is in a unique situation—a place of transition moving from tradition notions of structure and relationship to new visions of multiple connectivity. Rich with new ways for interpreting concepts and for influencing the design and implementation of information systems, knowledge organization and the theories that underlie set a foundation for research in the field of information and library science. Many of the concepts discussed in knowledge organization theory revolve around bibliographic objects. These theories have never been applied to data objects. The research takes these theories into consideration when researching organization and data.

4.a.2. Knowledge Organization in Practice

In the previous literature review on knowledge organization theory, the concept of plurality is discussed as a new approach for organizing knowledge in more practical, non-linear ways reflected in the real world. This approach is contrasted with traditional, unified knowledge organization schemes found in libraries and other information centers. Libraries and other information science based institutions engage in knowledge organization through organizing information. In many cases these information organization practices and products are traditional knowledge organization approaches that have developed, maintained, and evolved over hundreds of years (Strout, 1969; Tait, 1970; Babb, 2005).

Professional librarians work to bring "like" things-- documents, information objects, or materials-- together and to differentiate them from the "unlike" (Hunter, 2002). Information and library science has a rich history full of influential knowledge

organization system creators, like Bliss, Cutter, Dewey, Panizzi, and Ranganathan, who created classification and/or cataloging schemes that were meant to be used in libraries and other information environments to assist in the description and access of materials. As discussed in the previous chapter on theory, knowledge organization is a means for representing the knowledge of mankind. According to Babb (2005), beginning in the 19th century, "as natural scientists explored and expanded upon their knowledge of the world, cataloging and classifying their exploration, so libraries made place for these findings, both physical and theoretical". This statement points out that knowledge organization is not just a theoretical concept, but also the physical and tangible activity of information organization. The techniques typically used by libraries to organize information are referred to as cataloging, classification and metadata.

The three areas of cataloging, classification, and metadata are information organization practices that have developed in libraries and other information environments to provide access to information needed by information patrons. As physical manifestations of knowledge organization, these three information organization practices provide a strong foundation for research about knowledge organization theoretical approaches that can better assist information system users.

This literature review examines knowledge organization research found in information and library science professional practice. The first portion of this literature review gives an overview of Hjorland's (2008) analysis of the knowledge organization sub-areas that form information organization. Based on his discussion, a framework mapped to the concepts of unity and plurality, is introduced that can be applied to categorizing recent knowledge organization research. Next, this literature review

discusses recent knowledge organization based research performed in cataloging, metadata, and classification communities that can be categorized as either a unified or plural approaches. This is followed by a discussion on the future of hybrid approaches—those systems that use both unified and plural methods. In conclusion, the last section elucidates the links between knowledge organization research and theory.

4.a.2.a. Hjorland's analysis of research in knowledge organization

One of the most prolific theorists and researchers in the area of knowledge organization is Hjorland and Nissen Pederson (2005) and Hjorland (2007, 2008) whose body of research has been influential in shaping the current state of knowledge organization research, as well as offering a reflective account of the development and continuation of knowledge organization as a field of study. While Hjorland's work is discussed briefly in the previous literature review on knowledge organization theory along with a survey of other theorists, one of Hjorland's (2008) most recent publications characterizes the way that knowledge organization is perceived, studied, and defined within the knowledge organization research sub-community in ILS.

In this research article, Hjorland (2008) presents both a narrow and a broad definition of knowledge organization. Narrowly defined he describes knowledge organization as " activities such as bibliographical databases, archives and other kinds of 'memory institutions' by librarians, archivists, information specialists, subject specialists, as well as computer algorithms and laymen" (Hjorland, 2008). This narrow definition soundly places knowledge organization in the realms of library science and only barely hints at the ability of laymen to be able to perform or engage in knowledge organization tasks. Hjorland's (2008) broader sense has a more global view, expressing that "KO [is]

about the social division of mental labor […] about how knowledge is both socially organized and how reality is organized". This broader sense introduces an interesting dichotomy of knowledge organization being either intellectual or social in essence. Hjorland (2008) elaborates that this broader definition of knowledge organization can only be successful "by the single sciences", meaning by studying one specific domain. Some may perceive this definition as limited because of its obvious bias towards domain analysis, an area of knowledge organization research, which focuses on the intense study of only one subject area instead of taking a more general view. This broader definition of knowledge organization eliminates the idea of interdisciplinary knowledge organization and limits the impact that can be made by knowledge organization products that aim to organize multiple domains.

Hjorland's (2008) discussion changes when he states that "there exists no closed 'universe of knowledge' that can be studied by KO in isolation from all the other sciences' study of reality". Thus, Hjorland (2008) places the study of knowledge organization in an interdisciplinary perspective, while limiting the ability of people to study knowledge organization by encouraging a focus on one specific domain. These statements seem somewhat contradictory if not convoluted. Yet, the definitions still give a nice context to the challenges and approaches to researching and studying knowledge organization. Hjorland's (2008) broader definition shows both a practical and a theoretical perspective of what knowledge organization research can entail. It also elaborates the usefulness of knowledge organization is not only intellectual sense but also in a social sense. He also strengthens his definitions by an accompanying discussion of approaches to knowledge organization.

Very few knowledge organization researchers have tried to classify knowledge organization to the same depth that Hjorland (2008) has attempted. Even with the domain analysis bias, Hjorland's article (2008) is an important contribution to the research and development of field of knowledge organization. Hjorland (2008), not only defines knowledge organization, but attempts to identify the ways in which knowledge organization has been researched and analyzed. Hjorland (2008) outlines seven different ways in which knowledge organization has been studied. These seven approaches include, traditional; facet-analytical; information retrieval; user-oriented/cognitive; bibliometrics; domain analysis; and other approaches. These approaches are either the social organization of knowledge or intellectual/cognitive organization of knowledge. The dichotomy introduced by Hjorland (i.e. social and intellectual) appears orthogonal to the concepts of plurality and unity discussed in some depth in the previous chapter. In this sense, plural approaches are more social in nature, while unified approaches rely more heavily on intellect. By being both unified and plural in nature, knowledge organization is both a mental activity as well as a community activity—a multi-faceted physical practice as well as a mental exercise.

4.a.2.b. Approaches to cataloging, classification, and metadata

Knowledge organization is a central concept to the field of information and library science. In practice, libraries use knowledge organization systems "to describe schemes that arrange information in such a way to facilitate search and retrieval" (Hodge, 2000). In the real world of library practice, knowledge organization manifests in a variety of ways, include cataloging, classification, categorization, ontology, taxonomy, indexing, controlled vocabularies, and subject headings (Hodge, 2000; Shiri & Chase-

Kruszewski, 2009). These terms are often used interchangeably to name different techniques and approaches for studying knowledge organization in the information and library science community. The research base on these areas is dense and numerous. In the previous literature review chapter on theory, a dichotomy is introduced between the theories of unification and plurality. Using these theories in relation to Hjorland's analysis (2007, 2008) as discussed above, I have created a framework for analyzing knowledge organization work in evolving library and information science environments that implement different underlying knowledge organization approaches. The purpose of this section of the literature review is to analyze and discuss recent research in the areas of cataloging, metadata, and classification by examining this research foundation as either a unified or plural. After this comparison in approaches, a brief discussion introduces the concept of hybrid approaches as a future research area worth investigating.

4.a.2.b.1 Unified Approaches

Unified knowledge organization approaches in practice, for the purpose of this literature review, are based on schedules, resources, or schema developed by information professionals for the purpose of categorizing and linking the universe of knowledge known by man. These approaches manifest themselves in library and information science as traditional classification schemes, controlled vocabularies, and other information organization products, like the Library of Congress Subject Headings or the Dewey Decimal Classification, created and maintained only by information professionals. The historical development and definition of unified approaches is discussed in more depth in the previous section on knowledge organization theory.

One study that looked at knowledge organization systems in digital collections is conducted by Shiri and Chase-Kruszewski (2009). According to their study of 269 online digital libraries in North America, the two most common underlying knowledge organization schemes are locally developed taxonomies, used by 113 libraries, and LCSH, used by 78 libraries, in some variation or modification (Shiri & Chase-Kruszewski, 2009). This research shows that traditional, unified knowledge organization approaches—specifically taxonomies and controlled vocabularies-- are frequently implemented in information systems meant for virtual and remote access. This statistic on the prolific use of unified knowledge organization techniques in digital libraries shows a continuing commitment to unified knowledge organization approaches in evolving library environments.

A large portion of knowledge organization research done in library and information science revolves around the library catalog. The subject access based catalog has become a popular topic for many researchers in both digital and traditional libraries. Historically, knowledge organization structures like classifications would go ignored by the underlying information retrieval systems because these systems relied heavily on automated pattern-matching and traditional keyword-based indexing (Papadakis, Kyprianos, Mavropdi, & Stefanidakis, 2009). The idea of a subject-based catalog or a subject-based information retrieval system that works in digital libraries or library catalogs seems to have had a recent renaissance for researchers. This subject-based approach is referred to as the classified catalog or subject search (El-Sherbini 2008; Bland & Stoffan 2008; Papadakis et al., 2009). These knowledge organization based approaches focus mainly on how to use cataloging and classification knowledge

organization products like subject headings or call numbers to describe and access materials.

The research done by Papadakis et al., (2009) examines the digital library environment to try to extend basic syndetic structure found in LCSH to an Web Ontology Language (OWL)-based ontology in order to create better retrieval for end users. Syndetic structures are found in thesauri and relate to the concepts of broader term, narrower term, and related term. The research done by Papadakis's group looked at the 'semantic extend' to which these thesaural relationships matched the LCSH subdivision-based relations of topical, form, geographical, and chronological subdivisions, then mapped those relationships to OWL. After linking these relationships to OWL, creating a user interface, and testing this new catalog system, the researchers concluded that using the intellectual knowledge organization approach enhances the user's "congnitive learning, since they were able to discover which subject headings corresponded to their information needs" (Papadakis et al., 2009). Overall, this research concludes that unified approaches, in this instance subject headings and thesauri, can be an essential part of an information system.

Researching pursued in a traditional library environment, Bland and Stoffan (2008) focus on similar issues to that of Papadakis et al., (2009), but take a different approach by looking at the library catalog and how to use classification or call numbers to enhance user experience. Bland and Stoffan (2008) claim that the most common use of cataloging systems take little account of the classified approach to cataloging by only giving users minimal access to call numbers. These classified catalogs use interfaces that show Library of Congress classes to users in order to allow them to easily navigate

and find information that is needed. These findings indicate that unified approaches to knowledge organization can be key educational tools for guiding and immersing the information system user into the chosen subject area. Again, this research asserts that unified approaches are relevant and vital to the sustainability of access to system content.

Schwartz (2008) goes on to comment that the future of cataloging and classification systems is in "guided navigation". Guided navigation systems would help users interact with already established thesaural relationships found in controlled vocabularies, like LCSH. The interaction with these vocabularies from an information architecture and human computer interaction design perspective would give new life to library cataloging. The OPAC would become a teaching tool for users who are searching for specific terms (Schwartz, 2008). Schwartz's conclusions and her evaluation of this approach are similar to those findings and developments being made by the Helping Interdisciplinary Vocabulary Engineering (HIVE) project.

The HIVE project being created by the University of North Carolina at Chapel Hill takes a guided navigation approach by dynamically integrating multiple controlled vocabularies to be used for resource description and access. Using the HIVE system for the description of resources allows both professional and non-professional metadata creators to see how choice terms relate to other terms when comparing multiple vocabularies. Though research on the effectiveness of this system is limited at the moment, the potential for greater unified knowledge organization approaches could be valuable. Projects, like HIVE and classified catalogs, are taking innovative approaches to applying unified knowledge organization techniques to the library cataloging and classification environment.

Unified approaches have a rich tradition in the library and information knowledge organization areas of cataloging, metadata, and classification, but new plural approaches are now drawing increasing amounts of attention in the ILS community.

4.a.2.b.2 Plural Approaches

Plural approaches in practice, for the purpose of this literature review, are based on the vocabulary choices, semantic relationships, and electronic links that society or groups of individuals make either consciously or unconsciously between terms or concepts to create an information space that has its own knowledge. The definition of plurality is discussed in more depth in the previous literature review on theory, but in relation to cataloging, metadata, and classification, pluralistic approaches manifest themselves through more social means—specifically in the area of democratic or social tagging and folksonomy.

Everyday users of the internet seem to prefer plural approaches like folksonomy over the unified approaches used by libraries when organizing and sharing content online. Schwartz (2008) argues that,

> "study after study tells us that users turn first to Google as a source of information, and that even when they can be persuaded to use library-supplied indexing services, their search behaviors do not directly make the most of the controlled structures we [information scientists] labor to provide. And now users are doing their own indexing, primarily for rediscovery of known objects in personal information spaces, but with the side effect of finding how others have used the same terms, or how others have indexed the same items."

Because of user enthusiasm, the social networking phenomena of tagging has recently been embraced by the library environment in ways that reflect the plurality theory in the organization of information and knowledge.

Folksonomy, is a more recent way of organizing information that is increasingly popular in the online environment. In folksonomy (a combination of the terms "folk" and "taxonomy"), "there is no hierarchy, and no direct specified parent-child or sibling relationships between […] terms. […] These folksonomies are simply the set of terms that group of users tagged content with, they are not a predetermined set of classification terms or labels" (Mathes, 2004). Folksonomies are a compilation of tags that are created by individuals to create an uncontrolled vocabulary (Bland & Stoffan, 2008). These collections of tags are basically user created flat taxonomies that are flexible and often seen as contradicting the formalized hierarchical and enumerative structures found in controlled vocabularies and thesauri.

The research area of PIM teaches us that everyone creates metadata in their daily lives (Jones, 2007b) and now with the emergence of folksonomy the idea of anyone being able and allowed to create metadata has become a normal assumption. Before the phenomena of tagging occurred, there were two assumptions about why untrained professionals wanted or would create metadata. The first was that content creators would create metadata to provide better access to their material of the web. This instance is referred to as author created metadata. The second assumption was that metadata creation by non-professionals and non-authors was done by 'community or subject enthusiasts' who wanted to provide metadata because of certain amount of expert subject knowledge (Greenberg, 2010).

Metadata creation in the form of tagging, social bookmarking, or folksonomy has shown that neither of these assumptions are necessarily as true as they were in the past. At this point, in the use of social networking sites such as Furl, Del.icio.us, and Flicker,

to name only a few, include subject metadata created by individuals who may have little to no expertise in the subject area they are creating metadata about and may often rely on community concensus—not formalized training—for help describing material on the web.

The rise of social tagging in the online community should indicate to the information community that the average public and even scholars are no longer satisfied with traditional, unified knowledge organization structures. As mentioned previously, Weinberger (2007) believes that tagging has resulted in people "rapidly miscellanizing our world, breaking it out of their old organizational structures, and enabling individuals to sort and order them on the fly. This goes far beyond simply organizing your information so you can find it again". Pluralistic approaches to organizing information provide users of information a way to interact and organize information in ways that physical limitations never allowed before.

Library and information science researchers have started experimenting with plural approaches to organizing information. Rafferty and Hidderly (2007) discuss a "new way of indexing" in relation to the traditional way of subject indexing. In traditional knowledge organization environments indexing is typically controlled by the professional indexer who uses one traditionally unified system for organizing materials. Rafferty and Hidderly (2007) comment "that the meaning of documents derives from the interaction of the document and the reader [...] there are interpretations of the document rather than a single authoritative interpretation". Their study of social tagging in the Flickr environment shows that the individual views of people outside of the information

profession is an important idea to think about when considering knowledge organization systems.

According to Shirky (2005), many problems that occur in unified knowledge organization approaches are not present when using plural approaches. After analyzing both ontological structures (here cited as a unified approach) and tagging (the example of a plural approach), Shirky (2005) concludes that folksonomy or tagging allows individuals to create value through organization for themselves and for others. In his opinion, this break from unified approaches is preferred because it is not a system that is "forced onto" its users.

While information professionals who are inclined to use more unified approaches to knowledge organization may argue that folksonomies could not have relevance in the library community of cataloging and classification, a study conducted by Spiteri (2007) found that this is not the case. Spiteri (2007) analyzes tags used in Del.icio.us, Technorati, and Furl and found that many of the tags "correspond closely to a number of the NISO guidelines pertaining to the structure of terms, namely in the types of concepts expressed by the tags, the predominance of single tags, the predominance of nouns, the use of recognized spelling, and the use of primarily alphabetic characters". This finding, as well as those by Shirky (2005) and Rafferty and Hidderly (2007), show that social approaches to cataloging and classifying materials could have a beneficial impact in traditional library settings. Further studies, performed by Kipp and Joo (2010) and Kipp (2011a, 2011b), have supported these earlier conclusions. This being considered, the idea of integrating or replacing traditional, unified approaches with social, pluralistic

approaches could potentially be daunting to information professionals who are strongly embedded in library tradition.

4.a.2.b.3. Future solution: Hybrid Approaches

The use of unified and plural approaches need not be an 'either/or' scenario (Gordon-Murnane, 2006). Some researchers have been examining unified approaches, like controlled vocabularies, in comparison to plural systems, like tagging, and have found a compromise or middle ground of sorts. The creation of hybrid approaches--those systems that use both unified and plural methods for creating knowledge organization systems—have, in some cases, been discussed as being the best approach for meeting user needs in terms of access and discovery (Gordon-Murnane, 2006; Noruzi, 2007; Schwartz, 2008; Bruce, 2008).

Hybrid approaches to knowledge organization are research projects or other knowledge organization-based implementations that combine aspects of both unified and plural approaches. Hybrid approaches can have strong bibliographical control foundations while combining in elements of the more social aspects of democratic tagging. Hybrid approaches can use both traditional library means for approaching knowledge organization while implementing elements of folksonomy.

An example of a hybrid approach to knowledge organization can be found in the study conducted by Bruce (2008). In this study, journal articles that were indexed by both ERIC thesaurus terms and CiteULike tags were retrieved and analyzed. The study looked for matches between both the thesaurus terms found in ERIC and the uncontrolled folksonomy tags used in CiteULike. The results of the study concluded that tagging could be a useful supplement to traditional controlled vocabularies because it "provide[s]

a means for personal organization outside the framework" of traditional, unified systems. This conclusion is intriguing because it shows the potential for a mixed methods approach to organizing information and knowledge. The study also concluded that "users do not use the same terminology as subject specialists", which points to the need for future study and research into how to integrate traditional knowledge organization approaches with more personal, social approaches.

Research and development of hybrid approaches for creating knowledge organization products imply that the way the library and information community is thinking about cataloging, metadata, and classification is evolving, so the approaches and implementation of these techniques should be evolving as well. Combining techniques, using both plural and unified knowledge organization approaches, allows for richer resource descriptions and greater user access to information systems. Yet, the area of pluralistic research is admitted recent and the implementation of these more socially-based research approaches is burgeoning as well. This being stated, it should come to no surprise that hybrid approaches while being of increasing interest to many researchers, does not have such an expansive research base at this time.

One of the key tenants of knowledge organization in practice has been to establish how to create knowledge organization systems that help the general user and allow the public to have access to information that is most important to his/her life (Cutter,1904/1985; Strout, 1969; Babb, 2005). Based on the discussion earlier in this paper, both unified and plural approaches have been used to research cataloging, metadata, and classification in ways that successfully create innovative systems that promote use and access to information. In addition, the hybrid approaches that mix both

plural and unified knowledge organization approaches have great appeal because they use traditional knowledge organization foundations while integrating more social and unstructured user generated content. Much of this research shows promise for the future development of knowledge organization as an essential part of any underlying information system.

4.a.2.c. Summary

In this section, Hjorland's analysis of research in knowledge organization is used as a basis for creating a framework that examines unified, social and hybrid approaches in the context of cataloging, metadata, and classification research. This analysis and examination of research has reasserted my own conceptualization about the bonds between research and theory in the field of knowledge organization. Based on the research done in this chapter and the examination of theory in the chapter before, I believe that theory and research in knowledge organization are a cyclical process. Theory needs research in order to grow, expand, and to be applicable in the world at large. Research needs theory to help direct and refine future discovery efforts and growth. The research that has been conducted in knowledge organization has not been applied to scientific data. Current projects that study scientific data, addressed in Section 5: Exploring Scientific Data Projects, are only beginning to incorporate knowledge organization perspectives into their research design.

4.b. Personal Information Management (PIM)

The field of personal information management is,

> "both the practice and the study of the activities people perform to acquire, organize, maintain, retrieve, use, and control the distribution of information items such as documents (paper-based and digital), Web pages, and email messages for everyday use to complete tasks (work-related or not) and to fulfill a person's various roles[...] (Jones and Teevan, 2007)."

This definition points out the mediums of PIM research, paper and digital; as well as, the activities that are performed during PIM: acquiring, organizing, maintaining, retrieving, use, and control; and how all inclusive PIM environments are, including work-related and personal situations. Researchers studying PIM include information scientists, psychologists, computer scientists, and lately, even domain knowledge experts, such as engineers (Lansdale, 1988; Jones, 2007b; Hicks, Dong, Palmer, & Mcalpine, 2008). Considering the definition introduced above, scientific data sets can be considered a type of personal information. They are products of people's research and are organized using metadata (White, 2010a). Before discussing scientific data sets in more detail, the relationship between PIM and organization will be explored.

PIM examines the personal while also looking at the work related, but it should be noted that while PIM is personal, it is not private. Lansdale (1988) expresses this difference between personal and private by explaining, "this is personal information not necessarily in the sense that it is private, but that we have it for our own use". This concept of "our own use" is central. By adding in the idea of 'personal work' items, a new range of information can be included in what is studied in PIM.

A sub-area of personal information management is group information management (GIM)--the PIM of groups. The research in this area focuses mainly on how

collaborative groups manage their personal information together (Erickson, 2006). In the sciences, GIM can be seen through collaborations that are "shaped by social norms of practice, the structure of knowledge, and the technological infrastructure of the scientific discipline" (Hara, Solomon, Kim, & Sonnenwald, 2003).

While scientific data sets are not a main focus of PIM related research, the area of organizing personal information is extensively researched in PIM. Organizing research in PIM can be divided into five categories of study conclusions:

- Organizing provides context
- Organizing reminds
- Organizing is visual
- Organizing assists keeping and finding
- Organizing is unnecessary

The discussion that follows elaborates on these five conclusions.

4.b.1. Organizing reflects context

An essential area of research on organizing has been in the theory of context. Research surrounding the relationship between context and organization is an essential part of understanding people's personal information spaces. Researchers, such as Barreau (1995), Kwasnik (1989), and Malone (1983), found that organizing in PIM depended on context. Their studies have shown that context "is the situation in which an event occurs" including "all aspects of a person's experience" as well as being a "factor in human behavior" (Barreau, 1995).

In earlier paper-based studies, Malone's (1983) findings show how well people organize based on the context of their offices—in this situation context determines

organization. One of his conclusions is that the location of physical files often indicates the importance of those files in a person's personal space of information. Further studies confirmed that this type of organization is classification beyond document attributes (Kwasnik, 1991). Similarly in a physical office environment, Kwasnik (1991) investigates how individuals organize and classify in their own work spaces. Her research findings suggest that context is continually at play when organizing within a personal space of information.

Additional research found that context has an affect on organization in both the physical and digital environment—revealing that context occurs in almost any personal information collection and is vital within a personal space of information. Barreau (1995) emphasizes that organizing can provide context to how the document was either created or acquired in the digital environment as well. Her study looks at what factors influence classification and has results similar to those found in Kwasnik's study despite the difference in PIM medium. Barreau (1995) concludes from her research that, in information storage and retrieval systems' "classification of work products and processes rarely fit neatly into document-specific categories such as subject and form", the act of organizing allows a classification that reflects more than a typical knowledge organization system would typically contain (Barreau, 1995).

Context is a notable finding for PIM because it shows a distinction between what has been assumed in developing traditional knowledge organization systems (i.e. that subject and form are most important when making classifications) and the real organization processes that individuals undertake in day-to-day life situations. This

illustrates how quantity and complexity have more influence in PIM organization--a point that traditional knowledge organization systems often fail to acknowledge.

In many ways, context is like Dervin's (1992) sensemaking. Both Spurgin (2006) and Jones (2007b) have made this link between sensemaking and context. In the discussion of his research, Jones (2007b) elaborates that "people often structure and organize information as part of a process to make sense of the information and to make sense of the situations where it will be used". This perception of organization creates a form of classification that goes beyond traditional library approaches like "aboutness" and subject analysis. With context, research shows that organization serves a more practical function than what is recognized by knowledge organization specialists and adds a new layer of understanding about the importance of organization within in personal work environments.

4.b.2. Organizing Helps Remind

The theory of organization as a reminding function of PIM is another notable contribution. In certain studies related to context, it has been found that the way documents are organized can remind users about tasks that need to be performed, as well as indicate the personal importance of documents (Cole, 1982; Malone, 1983; Barreau, 1995, 2008).

As with many areas of PIM, organization as reminding has been successfully studied and observed in both physical and digital information spaces. Research by Malone (1983) and later Barreau and Nardi (1995) determine that file placement serves an important reminding function, specifically when it is intentionally used as a way to

remind people of things that need to be done. Malone (1983) concludes that "reminding is a subtle but very important aspect of desk organization".

In physical environments, and then more prominently in digital environments, reminding has been studied as an organizational method used to assist in finding (Cole, 1982; Barreau, 1995, 2008). Though later sections of this piece will discuss the link between organization and finding in more depth, it is imperative to also acknowledge it in this section as well. Barreau's (2006) research elaborates that in the digital environment, reminding is not only about finding or search, but about "triggering memory, managing tasks, and learning from experiences" because people forget where things are located. The concept of forgetting is key because forgetting is a human flaw that information systems cannot rely on without help from organizational tools. Research has found that *organization as a reminding function,* "helps us to make connections between things that we have forgotten are there, synthesize information from diverse sources, identify undesirable clutter, or remember why we have the files in the first place" (Barreau, 2006). Barreau's point is telling because it shows how essential the action of organization is to the PIM environment as a whole, as well as, emphasizing the essence of organization beyond traditional keyword searching mechanisms. The reminding function of organization points out the connections that can be made between ideas—a very knowledge organization based concept that is found even in the area of PIM research.

Research has shown that organization is seen through spatial location on people's desks and can be represented by size, location or color. While visual concepts and organization are presented later in this literature review, these visual characteristics help remind a person about important tasks. Reminding goes beyond physical space. Taking

on a new importance in the digital environment, research that looks at tool design have concluded that organization is an important reminding tool in electronic interfaces as well. (Robertson, Czerwinski, Larson, Robbins, Thiel, & van Dantzich, 1998).

4.b.3 Organizing is a Visual Concept

Relating to the idea of reminding, another popular research area for organization is the idea that organization is a visual concept. As with research in reminding, the visual importance of organization has been successfully studied in both physical and digital environments. In the physical environment, Miller's (1968) research shows that "people like to locate information spatially, and that fact tells us something important about the way man and information interact". Miller's comment points to the essential link between organization and information—again establishing organization as an essential and singular part of the PIM process. Because of this, Miller (1968) argues that "the priority of space as an organizing principle is so compelling that we frequently take information that is really not spatial in character and give it a spatial representation just so we can think about it more clearly and remember it more accurately". Miller's argument takes organization beyond reminding in the realm of the visual. Miller's (1968) argument asserts that organization as a spatial/visual concept represents clarity of subject as well as form, and calls to mind the studies of Jahoda, Hutchens, and Galford (1966) that also indicate the importance of subject and form within PIM organization situations.

Research about the visual importance of organization goes beyond the physical desk and has positioned itself onto the internet and personal desktop—items that are used everyday to perform PIM functions. In the digital environment, organization as a more visual concept is considered a tool for desktop design, tool development, and even web

interfaces (Robertson et al., 1998). The way information is organized in a user interface can impact the effectiveness of web and electronic desktop tools. Research into organization as a visual concept goes beyond tool development and web interfaces. Visual organization is also employed in personal digital desktops that are used everyday. For personal use, organization can be seen through folder creation and hierarchy (Jones, 2007b). How these folders are named, moved, renamed, and deleted are part of understanding organization. These folders are arranged and maintained in certain ways so that they can be effectively used by the creator. Organization has an impact beyond the pure act of organizing and has become an essential part of everyday tasks in order to successfully use information and materials.

4.b.4. Organizing Assists Keeping and Finding

As previously mentioned Jones (2007b) study of personal information management broke down PIM activities into three areas: finding, keeping, and meta-level. Much of the discussion of organization in this literature review has focused on the meta-level activities concept of organizing. Yet, organization is not only studied as its own step in the PIM process, but as an assisting function that interacts with the keeping and finding activities as well.

In the area of keeping, organizing is discussed in the context of file folders. Many researchers lump organizing with the act of maintaining--- both keeping and organizing as maintenance activities. Research on organization as a function that assists in keeping activities began by looking at the workplace environment. Barreau's (2008) research elaborates on three types of information found in the workplace that were originally introduced by Barreau and Nardi (1995). In this research, the three types of information

are archived information, working information, and ephemeral information. Archived information is typically organized information. It is often a "completed work—a finished paper or project or report, for example that may be carefully labeled and placed in a folder or subdirectory" (Barreau 2008). This information is supposed to be kept for an extended period of time because it has long-term value and placing it in some type of organizational structure allows it to be maintained for a longer period of time. Archived information involves organizing for the purposes of historical record. It is organization for the sake of potential long-term use.

Finding, refinding, and reminding are often linked concepts found in PIM studies (Jones, 2007b). Finding, also known as information retrieval, and its related functions are actions that are seen to benefit from organizing activities. The way in which organization is analyzed and studied in PIM revolves around research in finding, refinding, and reuse of information and, ultimately, data. There are two ways in which studies look at organization in relation to retrieval functions: finding and refinding.

Early research by Lansdale (1988) pointed out the recall and recognition efforts of organizing. Through the use of keywords and knowledge organization structures, organization assists with search. The other way in which organization is studied is mentioned earlier in the reminding portion of this piece. Reminding or remembering to look or find serves the function of organization as assisting re-finding. According to Barreau (1995, 2006), "people organize information so that they can find it later".

Research on organization in the finding sense again characterizes the action as an afterthought, like spring-cleaning. As an afterthought helping with finding and reuse, organization is still an essential part of any information system. Yet, even with these

conclusions on organization's value as both a singular meta-level activity and as an assisting function, some researchers in PIM have been known to downplay the significance of organizing actions (Elsweiler, Ruthven, & Jones, 2005;Whittaker, Bellotti, & Gwidka, 2006).

4.b.5. Organizing is Unnecessary

For as many of the PIM articles that try to point out how unique and important organization is, those same articles cite organization as being unused and unnecessary. Some theories put forward organization as a way of hindering PIM, specifically in the area of PIM tool development. Elsweiler et al.'s (2005) research on PIM tools has negative findings in regards to organization. The study concludes that the organizational methods used in tools and user interfaces of PIM force users to conform to hierarchical organizational schemes and therefore placing a burden on the user's memory. Instead of hierarchical information structures, he demonstrates a use of tags as assisting PIM. Yet, Elsweiler et al. ignore the fact that tagging and folksonomy are organizing activities as well. This perception ignores the full scope of what can encompass an organizing action.

Other studies have indicated that in certain environments, like e-mail, the act of classifying or organizing documents is totally eliminated from the PIM process and not needed (Whittaker, Bellotti, & Gwidka, 2006). In an earlier article, Barreau and Nardi (1995) determine that schemes to organize and keep archival information are not used or relevant to many people. In these articles, search is cited as a process that excludes the need for organization of any kind.

4.b.6 Summary

The personal information management (PIM) area of study gives a different perspective on the idea of organization. Unlike knowledge organization where organization is theoretical, physical and, at times, spiritual, organization in PIM is solidly grounded in the realm of actions and activity.

PIM research on organization is the way real people, as opposed to trained information professionals, organize within their own chosen environments. This perspective on how real people organize their own things is an important consideration for the future of library and information science systems that try to appeal to a large community of diverse users. Studying personal organization has shown that people organize for a variety of reasons and are influenced by their environment and chosen medium, as well as subject and format. Most of the research looking at PIM has shown that organizing, either by itself or by helping other functions, is an essential part of information.

In the knowledge organization sections of this dissertation a discussion of new conceptualizations of knowledge organization looks in depth at the definitions and theories of organization that are currently at use within the sciences, interdisciplinary studies, and information science. While the idea of organization in PIM does not really challenge the notion of organization set forth by knowledge organization research, my assertion is that organizing activities deserve research beyond information retrieval. The research areas of context, reminding, and visual presentation within the PIM research literature create an argument for looking at everyday organizing. This concept needs to

be explored outside of the PIM environment and be considered in relation to knowledge organization systems that are used for scientific data sets.

The previous literatures examine the concept of organization in the areas of knowledge organization in theory, knowledge organization in practice, and personal information management. From these three reviews, a few points can be concluded. First, it can be concluded that organization is both a conceptual issue as well as a practice one. Second, research on and questions about organization can occur in a variety of subject domains. And lastly, that organization plays a vital role in information science and libraries. In this growing body of research about knowledge organization and personal information management, the research being presented addresses resources, but not data sets. This research suggests that the need to do research in an area where knowledge organization meets personal information management will be helpful in figuring out ways to improve information systems designed for accommodating electronic scientific data sets.

4.c. Methodologies Used in KO and PIM

The purpose of this section is to examine the multitude of research methods that have been used in knowledge organization theory, knowledge organization practice, and personal information management. The following literature review on methodology begins with an examination of research methods used to study knowledge organization theory. Next, the focus of the piece changes to discuss the methods used in knowledge organization research. After analyzing knowledge organization, research in personal information management will be examined. Discussion then turns to the benefits and limitations of the research methods being conducted in these three areas. Lastly, this

literature will evaluate the effectiveness of all these methods and suggest the most successful method for evaluating scientific personal organization practices in comparison to traditional knowledge organization schemes found in libraries. The terms used to describe methods in this literature review come from the works of Pickard (2007), Powell (1991), and Busha and Harter (1980).

4.c.1 Methods used in Knowledge Organization Theory

Researchers examining the field of knowledge organization theory use the historical research and critical analysis methods in order to create definitions, descriptions, and characterization of what encompasses the area of knowledge organization theory. For the purpose the literature reviews presented here, theory is presented as a type of research—an intellectual output meant to enrich the world of library and information science. Evaluating the research methods in articles and books about theory can be a challenging task due to the stylistic differences employed when writing about theory. These articles are structured differently from traditional scientifically inspired research articles that report methods used and findings explicitly. In theory, the conclusions or ideas themselves are the focus of the article and the research or systematize approaches for reaching those conclusions are barely mentioned.

The research presented in the previous literature review on theory can be divided into categories based on the types of writings being used to analyze the topic. The first type of writing discussed is based on already established knowledge organization systems. These researchers include, Olsen, Miksa, and Babb. The second set of researchers is more focused on creating their own knowledge organization systems and these researchers include, Otlet, Richardson, and Dupre. A third set of researchers, Hjorland,

Weinberger, Klein, and Huberthal span both of these groups because they are both observing knowledge organization schemes from a historical or content analysis perspective plus creating their own interpretation of knowledge organization schemes by adding to the various perspectives and structures. The most commonly used methods included historical research and content analysis. These methods were used either individually or in a mixed methods approach.

Historical research is an application of the scientific method with a historical focus. The steps involved in conducting historical research involve: the identification of a historical problem; a gathering of information; the formation of a hypothesis linking certain variables; organization and collection of evidence; drawing conclusions; and record these conclusions (Busha & Harter, 1980).

Miksa (1998), Olsen (1998), and Babb (2005), three researchers discussed in more detail in the earlier knowledge organization theory literature review, use historical research in creating their conclusions about the influence of the classificationists of science in the formation of traditional knowledge organization schemes. Each of these three researchers is looking at the same historical problem, yet come up with different conclusions in regards to scientific classificationists and knowledge organization. This dissimilarity in results implies that historical research may not be the best approach for examining influences for knowledge organization system development.

Yet, historical research is also used by interdisciplinary scholars, like Huberthal (1998) and Klein (1999), in their characterizations of images of plurality and interdisciplinarity in real life situations. Klein (1999) researched university departments—specifically the scope of departments and how those departments would merge or diverge over time. She

was able to come to a successful conclusion using historical research methods about the image of unity as presented by traditional knowledge organization schemes when comparing this image to pluralistic models of knowledge organization. Historical research approaches seemed to be more successfully applied by interdisciplinary scholars than when they are applied by researchers in library science.

Content analysis is another method employed in theoretical research and development of knowledge organization by information scientists and others. Content analysis studies, sometimes referred to as critical analysis (Pickard, 2007), use methods that analyzes the contents of different media (i.e. information objects, articles, etc) in a systematic and quantitative way in order to form meaningful conclusions (Busha & Harter, 1980). An example of a less rigid, informal application of the content analysis method can be seen in Hjorland's (2008) classification of different approaches to knowledge organization. In this application of content analysis, Hjorland (2008) looked at the variety of information objects being created by researchers and practitioners in the field of knowledge organization. After analyzing the contents of these knowledge products, Hjorland (2008) classified them into seven different information organization categories.

Weinberger (2008) also takes a content analysis approach when he discusses the three orders of order when classifying different types of organization. Dupre (1993, 2006) is another research who applies a content analysis approach when developed his theories on plurality and unity.

Content analysis seems to be employed successfully in the area of knowledge organization theory, yet there are some limitations to the method. Categorization and

classification is a key component of the content analysis process. Categorization, even when employed in quantitative studies, often involves subjective, human driven process and is less controlled than more experimental methods. This means that results for content analysis studies can be somewhat biased. Bias can been seen in Hjorland's (2008) analysis of the different types of knowledge organization When he places domain analysis as a main approach to knowledge organization theory. Yet, content analysis does offer a less qualitative means for examining and characterizing research content. Historical research and content analysis are two research methods used frequently in the development, creation, and evaluation of knowledge organization theory. Due to knowledge organization theory's roots in epistemology, it is not surprising that qualitative approaches are used more prevalently in the conceptualization and research of this topic.

4.c.2. Methods used in Knowledge Organization Research

As discussed in previous literature reviews, knowledge organization is both a conceptual process as well as a physical act. While research in knowledge organization theory often revolves around the way ideas or concepts linking together and the best way to represent those connections, research in the physical act of knowledge organization (through information organization) revolves around the way knowledge organization systems (KOSs) interact with and assist people and information systems in libraries or other information institutions. Research methods used to study the physical manifestations of knowledge organization in information systems found in libraries and the internet seem to employ variations of experimental research and content analysis methods.

Powell (1991) and Pickard (2007) observe that the experimental research approach is rarely used successfully in library research involving human subjects. Experimental research is best performed when researching information systems where human are not involved because the cause and effect relationship between predefined variables can be controlled. Humans can be involved in experimental approaches, but it is often difficult to maintain control when humans are involved. Yet, information retrieval studies can effectively use the experimental method because of the amount of control of independent variables possessed by the researcher (Pickard, 2007). One example of an experimental study that uses information retrieval techniques to examine knowledge organization is conducted by Papadakis et al. (2009). In this study, the links between a controlled vocabulary, specifically the Library of Congress Subject Headings (LCSH), and ontologies were explored in a system that housed thesis and dissertations. While the researchers categorized this approach as a "case study"—a term used very liberally and discussed in more depth later in this literature review-- the inherently experimental design to the study is undeniable as the system was testing the effectiveness of a particular procedure and predicting the applicability of such procedures for library system users.

Another method that is of particular note in this section is the quasi-experimental approach. The use of experimental research design is often limited to systems research and it is recommended to avoid using human subjects when using a experimental design. In libraries and information systems, where humans are often an essential consideration in design and implementation, the quasi-experimental research method is employed when examining the relationship between systems and humans in field-focused experiments

(Powell, 1991; Pickard, 2007). A variation of the experimental research method, the quasi-experimental method allows for some experimental methods to be employed for more organizational, behavioral, or socially focused outcomes (Pickard, 2007). The potential for quasi-experimental methods in information systems, especially as a foundation of a mixed method approach, is great. Often during examination of the literature discussed here, quasi-experimental methods laid the foundation for a more in-depth application of content analysis.

The content analysis method, discussed in more detail in the section above on research methods employed in studies of knowledge organization theory, is also used when evaluating the effectiveness of certain elements of KOSs used in the creation, development, enhancement, or evaluation of information systems. One example of the content analysis approach being used in the research of knowledge organization can be found in the recent work of Shiri and Chase-Krusezewski (2009). Shiri and Chase-Krusezewski (2009) evaluate what type of underlying knowledge organization system is employed in digital libraries by reviewing a 200+ sample of electronically available digital collections produced by North American institutions. The result, discussed in more detail in the previous literature review on knowledge organization research, of this content analysis was a categorization of the most popular knowledge organization methods used by digital libraries. Other researchers who use content analysis inspired approaches are Spiteri (2007) and Bruce (2008) who examining various aspects of either unified or plural knowledge organization systems in the context of real life implementations.

In his application of content analysis methods, Spiteri (2007) compares similar characteristics between NISO standards and tags found in three popular tagging sites. This comparison, discussed in more detail in the previous chapter on knowledge organization research shows how the content analysis approach can be powerful tool for comparing and categorizing two very similar occurrences in the digital world that may have at other times have been thought of as not related at all. The content analysis methods used by Bruce (2008) is very similar to the way it is employed by Spiteri. Bruce's study (2008), discussed in more detail in the previous literature review on knowledge organization research, used content analysis to conclude that hybrid approaches – those approaches that integrate unified and plural systems—could be beneficial in future system development. Examining these two research studies together shows how similar methods can be used in almost the same situation to come up with differing, yet complimentary results. Unlike the historical research methods performed by Miksa, Babb, and Olsen, that looked at the same time period, with almost the same research question, and used similar methods, but ended up with different results, this instance of research is different. In the studies examined in this section, Spiteri (2008) and Bruce (2008) employ the same method and look at similar topics within knowledge organization but come up with two different, yet complimentary results. This implies that content analysis is potentially successful method for studying knowledge organization research in real life situations.

With its reliance on real world system evaluation, the research of knowledge organization uses methods similar to those found in knowledge organization theory, yet have slightly more emphasis on experimental based research methods. The reason for

this difference is because of the need for real world knowledge organization systems to be grounded in practical outputs.

4.c.3 Methods used in Organization Research of Personal Information Management (PIM)

Studies in personal information management (PIM) are concerned with how individuals conduct everyday organizing activities in their home or work environments. The previous literature review on personal information management briefly discusses some of the dialog that has occurred between PIM researchers about the best way to conduct a study on PIM related topics. The consensus on studying organization in particular in relation to PIM is that this topic is more personal, highly individualistic, and difficult to study (Kelly, 2006; Jones, 2007b). For these reasons, many more naturalistic methods are used to PIM. Specifically, the application of ethnography and case study are most popular when studying organization in PIM situations, yet identifying the application of these two methods can be challenging when evaluating PIM studies.

In the world of research, identifying the differences between case studies and ethnography is difficult for a few reasons. The term 'case study' is often applied erroneously for situations that are ethnographies or in some cases the term 'case study' is treated more as a catchall phrase for research that does not quite fit nicely in any other category (Pickard, 2007). Because of this reason and the confusion that occurs with these two terms, I have chosen to use Pickard's (2007) application of the terms 'case study' and 'ethnography'. According the Pickard (2007), 'ethnography' is a research method that describes and interprets a cultural and social group by observing participants to collect information in an exploratory way. There is a certain amount of integration and

assimilation into that culture that may take place. A 'case study', according to Pickard (2007), is used to develop an in-depth analysis of a single case by visiting a case site multiple times at regular intervals. These studies are often observational.

Many PIM studies are done in field-type research approaches, where researchers go to office or other work environments and observe a select number of pre-arranged participants (Cole, 1982; Malone, 1983; Kwasnik,1989; Barreau, 1995, 2008; Barreau & Nardi, 1995) . During these observations, the researcher typically interacts with the participant by introducing task scenarios, like sorting physical documents, navigating through email, or describing electronic desktop arrangement, or by asking interview-type questions (Kwasnik,1989; Barreau, 1995, 2008; Barreau & Nardi, 1995). This type of research design works well in both case study and ethnography based studies.

Early PIM research on organization done by Cole (1982) and Malone (1983), have more of an ethnographic approach to their design due to the pioneering and exploratory style of the information being gathered. In these studies, the researchers examined general organizing behaviors and conclusions were drawn based on the observations. The studies performed by Kwasnik (1989), Barreau (2008), and Barreau and Nardi (1995) use more of a case study type approach with multiple visits to the same site over a period of time. Kwasnik's research (1989) involved visiting office workers and observing mail sorting behavior. In subsequent visits to the case study site, Kwasnik was able to mimic the sorting behavior in ways there were satisfactory to the participants of the study. In the case of Barreau (1995) and Barreau and Nardi (1995), these studies originally seem to mimic an exploratory ethnography examining office workers' use of the new electronic desktop environment. Yet, in later research publications, Barreau

(2008) follows up and conducts further research with the same participants involved in the 1995 studies, thus transforming the method used into a more case study-like approach.

As mentioned previously, the methods of ethnography and case study are frequently employed when researching of organizing behaviors in PIM research. While these two methods allow for more in-depth research of a small sample of participants, there are some limitations. These methods are often misapplied and it can be difficult to locate a pure application of either study in library and information science (Pickard, 2007).

4.c.4. Summary

This literature review focuses on the research methods used in knowledge organization theory, knowledge organization research, and personal information management. According to the literature examined here, qualitative analysis methods were used regularly in both theoretical and research oriented knowledge organization studies. In studies that looked at human interaction with knowledge organization systems, the method of quasi-experimentation is considered more effective than pure experimental designs because of the inclusion of human elements. For PIM studies, ethnography and case studies are preferred methods. Looking at the methods evaluated in knowledge organization theory, research, and personal information management it can be determined that a diverse range of methods has been used when studying aspects of organizing. This review provides a useful means for considering research approaches, and prompts the question: what is the best method to use while conducting research

involving scientists and information professionals when comparing knowledge organization schemes and personal organization?

This literature review section points to the benefits and limitations of methods used in knowledge organization theory, knowledge organization research, and personal information management. Given what is known about the strengths of current methods used in examining these three areas, it can be concluded that a mixed methods approaches are the best for examining topics in these areas. For that reason, a concurrent triangulation mixed methods approach was chosen for this dissertation study. This method allows for both a quasi-experimental study design while also incorporating more qualitative questionnaires. Data analysis involves examining quantitative findings reported by participants and analyzing narrative text from responses.

5. Dryad Research

The goals of this dissertation have been shaped by working with Dryad and researching topics associated with managing scientific data sets. The Dryad Repository Project is an internationally funded repository project that houses data underlying publications in the biosciences. One component of this project, led by the Metadata Research Center[2], involves investigating metadata, vocabulary control, and data life-cycle issue. Research approaches for the Dryad repository have focused on investigating the practicality of using a metadata application profile in a practical DSpace environment (Greenberg, White, Carrier, & Scherle, 2009), investigating metadata practices of scientists (White, 2008), and looking at metadata life cycle issues (Greenberg, 2009). Researchers in this group have used standard methods including surveys, experiments, and intensive interviewing techniques. Metadata has been a primary area of focus for this group since the project began.

Dryad is not the only project researching scientific data sets. There are a variety of approaches currently being used to study how to organize data with the intent for reuse and long term maintenance. Other projects, including the Center for Embedded

[2] Metadata Research Center: http://ils.unc.edu/mrc/

Networked Sensing (CENS)[3] and the Science Data Literacy Project[4] use a range of approaches for examining issues surrounding digital scientific data sets. The issues examined by these groups include data management, data storage, information organization, and education about data. Over the last few years, much of the publications about these projects and project updates have concluded that more research still needs to be done (White, Carrier, Thompson, Greenberg, & Scherle, 2008; Mayernik, 2010a, 2010b; Wallis et al., 2010).

What follows is a discussion of Dryad's research projects that have influenced the work of this dissertation. Contents of this section include a study of demographic information on scientists, a controlled vocabulary study, an exploratory ethnographic study, my pilot research, and concludes with a summary.

5.a. Demographic Information on Scientists

A survey was conducted by the Dryad repository development team to understand more about the importance of data and data sharing in the evolutionary biology community. This survey, reported on the Dryad wiki, confirms that demographically, evolutionary biology is made up of a variety of professionals who study topics such as neurobiology, genetics, morphology, ecology, paleontology, systematics, and physiology. The largest percentage of people who responded about data sharing (N=453) studied genetics (specifically, phylogenetics/systematics or genomics) making up 43.5% of the total respondents. The most popular type of data the respondents created and shared were

[3] Center for Embedded Networked Sensing (CENS): http://research.cens.ucla.edu/

[4] Science Data Literacy Project: sdl.syr.edu

experimental data in tabular form and DNA, RNA, or protein sequences. Greenberg, reporting on this data as well, notes that,

> "evolutionary biologists enthusiastically engage in data sharing with over half the participants indicating having been asked to share data, and accommodating such requests 75% of the time. Additionally, 69% of participants noted that they have requested data from other scientists/researchers. These initial survey results, and the fact that 70% of the participants indicated their data was in digital format bring metadata management needs to the forefront" (Greenberg, 2009). "

These survey findings, emphasized by Greenberg (2009), show the importance of organization issues surrounding data-sharing and reuse by others plus the information organization issues for those information institutions that will eventually be responsible for maintaining them. Understanding these practices is essential for information communities that assist in the creation of systems that facilitate access of data.

5.b. Dryad Vocabulary Study

In 2007, the Dryad Repository team conducted an informal vocabulary analysis study to see which vocabularies were most appropriate for describing publications in evolutionary biology journals. The team sampled 600 keywords from Dryad partner journals and divided those keywords into facets, such as, taxon, geographic name, and time period. These terms were then mapped to 10 different controlled vocabularies. Some of the vocabularies used include the National *Biological Information Infrastructure Thesaurus*, the *Library of Congress Subject Headings,* and the *Getty's Thesaurus of Geographic Names*. Terms were mapped to vocabularies according to categories: exact match, partial match, no match. Preliminary findings from this study are also reported by Bedford, Greenberg, Hodge, White, and Hlava (2010).

Though further research is needed, findings from this study indicate that there is currently no single vocabulary that is adequate for describing an interdisciplinary field like evolutionary biology. Specific findings show that the Library of Congress Subject Headings, a large, whole world approach vocabulary, had a 22% exact match and only a 25% partial match. The findings also introduce other questions about vocabularies and how representative they are for subject terms used by scientists and information professionals.

5.c. Controlled Vocabulary with Scientists and Librarians

Huang's research examines the usability issues surrounding the Helping Interdisciplinary Vocabulary Engineering (HIVE) project. The populations examined for this usability study are information professionals and research scientists. Huang's findings (2010) regarding these two user groups involve aspects of curation, including time it takes to conduct a search and knowledge about controlled vocabularies for science. Findings from this study are also reported by Bedford et al. (2010).

Huang's (2010) research shows that information professionals typically take longer when choosing subject terms to describe scientific material. Also, this research found that both scientists and information professionals seem unaware of the type of controlled vocabularies that are available for describing scientific material. Examining these findings and comparing them to the findings of the vocabulary studies performed by Dryad show that there is much room for research..

5.d. Ethnographic Research

White (2010a, 2010b) conducted an intensive ethnographic study that examined the way that researchers in evolutionary biology organized data that supported publications. The methodology of this study relied on traditional PIM best practices and involved intensive, focused interviews where both the scientist and information science researcher reviewed and analyzed the research practices. Findings from this research study show that for their own data, scientists do create their own personal organization schemes to organized and eventually share data (White, 2010b). Even so, it is likely that the ethnographic approach used in this work does not give a complete picture of organization issues associated with scientific data sets.

The findings from White's study are supported in part by some of the research conducted by the CENS research group. While a portion of this research has found similar findings as White (2010a), the methodology used by CENS researchers is different (Mayernik 2010a, 2010b). The research methodology used by this group relies less on naturalistic means and has created a lab based study that uses both videotaping and think aloud protocol more reminiscent of the work done by Suchman when studying air traffic controllers. Yet, lab-based studies are often criticized by PIM researchers (Kelly, 2006; Jones, 2007b). Much of the PIM literature cites the need for research in PIM related areas to be done in more naturalistic settings. Arguments by Jones and Kelly have emphasized the need to do ethnographic or case study based research in narrowly defined fields.

At the moment, there is no established 'best practice' for researching scientific data sets. Many of the approaches listed above have looked to lexicography,

anthropology, and psychology in order to find and experiment with appropriate methods for studying scientific data communities.

5.e. Pilot Research

Understanding the need to look beyond the traditional ethnographic approaches for researching issues related to scientific data, a study was conducted using a mixed methods approach. The purpose of the study was to examine similarities and differences in the organizing output of both scientists and information professionals. The study also helped determine the effectiveness of the mixed methods approach and served as a pilot study for the dissertation research. The pilot methodology used two questionnaires (a preliminary questionnaire and a follow up questionnaire). The study also included a basic comparison for similarities and differences in terms of metadata, data arrangement, and subject term applied by each group.

In this study three scientists and three information professionals were given the same data set and asked to simulate integrating that data set into their own collections. Participants were sent, via email, the study materials and had a time limit of two weeks to complete the simulation. The simulation and answering of questionnaires were performed at the participants' own leisure at their work or office environments. The principal investigator was available via email to answer any questions during the process. Over the course of two weeks, all participants completed the simulated data integration and answered the two questionnaires before sending all of their responses to the principal investigator.

The pilot was conducted to determine the viability of using a mixed methods approach to collect data and to get a sense of what type of answers participants would

give in a simulation. To accomplish these two goals, the data was examined for similarities and differences within and among the two groups. The researcher also assessed the timing, completion rate, and followed-up with participants via informal interviews to consider the effectiveness of the methods used.

Results from the pilot data suggested that the methodology was successful. All six participants were able to complete the quasi-experimental simulation of data integration and organization. All participants completed the simulation and questionnaires within the two-week period. Minimal questions were asked during that time. While questions were asked of the PI during the study, these pertained to the clarification of terms used in the study. Specific clarification related to the terms "metadata" and "surrogate".

Informal post-interviews with select participants also indicated that the methodology was successful for gathering data about organizing practices. Participants found the simulation enjoyable and not overly taxing. The main suggestion from participants was to change terms like "metadata" and "surrogate" to terms that were more understandable.

Findings from this pilot study can be summarized in a few points:

- Information professionals create more surrogates and forms of metadata than scientists
- Scientists could use metadata, but did not necessarily create it for this scenario. (see White, 2010a for more information on scientists and metadata creation)
- Scientists change the data set arrangement more than information professionals.
- Scientists used arrangement of data as a way to formulate research questions.

The study suggested that these groups are similar, but there were some observed differences.

The findings from this pilot suggested the need for further study in the area of organizing output. As reviewed earlier, the organizational underpinnings of repositories are based on assumptions held by information scientists. The true need is for scientific data repositories built on demonstrated research as opposed to rationalistic hearsay. This dissertation study is a contribution towards demonstrated research.

6. Methods

6.a. Research Purpose

The purpose of this research was to study how scientists and information professionals use information organization techniques, specifically descriptive metadata creation and subject term application, when working with scientific data sets. By examining and comparing the organizing behavior and output of these two groups, recommendations can be made to improve repository systems designed to accommodate the special needs of scientific data sets. A concurrent triangulation mixed methods approach was used to fulfill this study's research goals.

6.b. Research Goals

The research goals of this study were,

- To characterize similarities and differences between how information professionals and how scientists organize scientific data sets.
- To interpret how descriptive metadata and subject terms were used in both communities.
- To make recommendations for repository development based on research findings from comparative study.

As shown in the Literature Review section, organizing has a substantial body of research. Only a portion of this topical area can be examined by this dissertation. The

research question presented later in this dissertation posits a overarching question and a series of sub-questions that examine the organizing output of scientists and information professionals when working with scientific data sets. Areas not addressed by this research include social tagging, automatic metadata generation, and the cognitive science rationale for organizing. These topics are of value and could provide excellent foci for future studies.

6.c. Research Question

The research question below is one broad question with five sub-questions. Six questions, presented as follows, were developed to guide this study.

Research Question: In the context of scientific data sets, what types of distinguishable similarities and differences exist between the ways researchers in the biosciences who use research data and information professionals who curate research data create metadata and apply subject terms?

Descriptive Metadata (about a resource, with exclusion of subject metadata which is covered in questions 3-5).

1. What types of formal/standard metadata are currently being applied by both groups?
2. What types of personal metadata are currently be applied by both groups?

Subject Terms

3. Which controlled vocabularies map best to subject terms created by both groups?
4. What is the extent of overlap in subject term application between the two groups?
5. What is the extent of divergence in subject term application between the two groups?

These research questions are answered using the research methods described in the section that follows.

6.d. Introduction to Methods

The methodology section of this dissertation begins by giving a high level overview of the methods. Then, the rationale for using this methodology is discussed. Following the rationale is a more detailed account of the implementation of the methodology. The procedures discussion is divided into sub-sections about data collection; population and subject recruitment; and data analysis.

6.e. Methods Overview

A concurrent triangulation mixed methods design was used in this study. The methods that was employed used both quantitative and qualitative approaches for data collection and analysis. Standardized methods that influence the study's methods were a quasi-experiment design; questionnaires; and qualitative and quantitative data analysis techniques. The qualitative data analysis techniques were influenced by grounded theory approaches (Glaser & Strauss, 1967; Strauss & Corbin, 1990; Chamez, 2006), while the quantitative data analysis techniques were counting and mapping (Xu & Lancaster 1998; Greenberg, Pattuelli, Parsia, & Robertson, 2001; Greenberg, Spurgin, & Crystal, 2006). Below are introductory summaries of the methodologies and techniques used during this dissertation.

Concurrent triangulation mixed methods: The study was designed to concurrently collect and analyze qualitative and quantitative data. Qualitative and quantitative results were merged together for interpretation and to develop recommendations for repository systems. All recommendations are given within the context of this study.

Quasi-experimental influenced set up: Research subjects were divided into two groups. The two groups consisted of scientists who use data and information professionals who curate scientific data. These two groups were given the same data sets and task scenarios in the form of data integration simulations. After completing this task, participants were asked to deposit both data sets into a controlled system to compare elements of metadata creation in a more controlled environment. Recruitment and participant demographics are discussed in more detail in the Results section.

Questionnaires: Participants were given two questionnaires. The first questionnaire was completed before the experiment began and collected basic demographic information about each participant's experience when working with research data. After the data integration simulation task, the second questionnaire was filled out. This second questionnaire asked questions about the actions that were performed during the simulation and the type of organizing output that was created from it.

Qualitative data analysis: Grounded theory influenced data analysis was used to analyze data produced from the questionnaires. An inductive analysis coding process, consistent with principles discussed by grounded theory researchers (Glaser & Strauss, 1967; Strauss & Corbin, 1990; Chamez, 2006), was used for this study.

Quantitative data analysis: The quantitative processes of counting and mapping were also used to analyze the metadata and subject terms deposited into Dryad. Mappings

were created between subject terms and established knowledge organization-based controlled vocabularies.

6.f. Methods Rationale

Due to the emerging nature of research on the convergence of scientific data sets, NKOS, and organizing behaviors, a concurrent triangulation mixed methods approach was used for this study. This application of the approach used a quasi-experimental study design and grounded theory influenced analysis techniques. It blended the need for control while still maintaining an element of the naturalistic.

A concurrent triangulation mixed methods design combines both quantitative and qualitative techniques. Supporters of this method claim it leads to balanced research studies that yield richer results (Creswell & Clark, 2006). Another proponent of triangulation claims that combining quantitative and qualitative methods minimizes the bias associated with using a single method (Kennedy, 2009).

A quasi-experimental influenced approach was part of the study's design structure. The portions of the study that reflect this quasi-experimental influence were the structure of the participant organization and the data collection. The quasi-experimental influenced approach is "used in natural settings, when some control over the experimental conditions can be exerted, yet full control is either not possible or not desirable" (Hank & Wildemuth, 2009). Experimental methods have historically been used to study areas of indexing in information systems (Cleverdon, 1970; Svenonius, 1986; Fidel, 1992; Rowley, 1994). While it is acknowledged that the quasi-experimental method is not *perfectly* naturalistic, this method is considered more naturalistic in approach than other experimental methods (Hank & Wildemuth, 2009). The quasi-

experimental method was used for this research to maintain a certain amount of control while still trying to include elements of the naturalistic. This emphasis on the balance between control and naturalistic is part of the reason a mixed methods approach was chosen for this study.

PIM research has often relied on more naturalistic approaches—specifically case study or ethnography—for gathering research data (Cole, 1982; Kwasnik, 1989; Barreau, 1995, 2008; Barreau & Nardi, 1995; Whittaker, Bellotti, & Gwizda, 2006; Hicks et al., 2008). The use of questionnaires and surveys for gathering research data is consistent with naturalistic approaches. The questionnaires used in this study prompted the participants to respond with short answers; narrative; and list-based answers. Grounded theory analysis captures narrative data from these questionnaires and gives the responses context and meaning (Chamez, 2006).

Part of this study used a grounded theory approach to analyze responses to questionnaire prompts. These answers were written in response to questions about descriptive metadata creation and subject term application. The grounded theory influenced data analysis approach provides "systematic, yet flexible guidelines for collecting and analyzing qualitative data to construct theories grounded in data themselves" (Chamez, 2006). Being influenced by grounded theory, this research does not deduce a hypothesis from existing theory. A grounded theory influenced design allows for rich qualitative research by "examining processes, making the study of actions central, and creating abstract interpretative understanding of the data" (Chamez, 2006). The processes focused on for this study were actions related to descriptive metadata creation and subject term application.

Researchers use a variety of data analysis approaches to study metadata, NKOS quality, and information organization. These approaches include surveys, counting metrics, and vocabulary mapping techniques (Xu & Lancaster 1998; Greenberg et al., 2001; Greenberg, Spurgin, & Crystal, 2006). Both qualitative and quantitative approaches have a history of use in researching this area. Researchers recognize that more studies on metadata and its applications are needed in order to determine the most appropriate research methodologies (Greenberg, Spurgin, & Crystal, 2006). There is no single "best practice" for researching metadata, indexing, or information organization.

For this reason and other rationale included in this section, this study used a concurrent triangulation mixed methods approach incorporating influences of quasi-experimental, surveys, qualitative and quantitative data analysis methods.

6.g. Methods

The Methods and the Procedures sections describe the methods and steps used in this research study. The concurrent triangulation mixed methods approach used included: a quasi-experimental research design, two questionnaires, and two types of data analysis techniques (qualitative and quantitative). This process is outlined in the seven steps that follow.

1. Scientists and information professionals were divided into two groups.
2. Participants were sent study instructions, two data sets, and two questionnaires.
3. Each participant answered the basic demographic questionnaire.
4. Each participant simulated integrating the data set in to his/her own collection.
5. Each participant answered a questionnaire describing the organizing output process performed in step 4.
6. Each participant sent questionnaires, data sets, and any other output created in Step 4 and 5 to the researcher.
7. Each participant submitted data sets into Dryad and created descriptive metadata and subject terms.

The Procedures section that follows discusses these steps in more detail.

6.h. Procedures

To fully implement this study there were seven steps that need to be completed. The following is a descriptive break down of each of these steps.

6.h.1 Step 1: Information professionals and scientists were placed into groups.

Once recruited and consent forms were completed, each participant was assigned to either Group L or Group S. This grouping follows a quasi-experimental methodology where the two populations are divided into naturally occurring groups. Group L members were information professionals and Group S members were scientists. Group names did not affect the study implementation, but were used to generalize research results during analysis. Group assignment was reflected in the unique identifier assigned to participants in order to maintain anonymity throughout the study. Unique identifiers for information professionals had the base name of "mrcinfoprof" and were followed by a three digit number. An example is "mrcinfoprof789". Scientist participants received a similar unique identifier with the base name of "mrcscientist" followed by a three digit number. An example is "mrcscientist789". Participant recruitment is described in more detail in the Results section.

6.h.2. Step 2: Participants were sent study materials.

Participants were emailed electronic copies of the study instructions, two data sets, and two questionnaires. A discussion of these study components follows.

6.h.2.a. Study Instructions

Each participant was given study instructions to guide his/her progress for completing the study. The study instructions included an order of operations guide and

the unique study identifier that needed to be used in the Dryad system. Due to its length, the study instructions are included in the Appendix.

6.h.2.b Data Sets

Each participant was given two data sets to work with during the experiment. The purpose of giving every participant the same data sets was to maintain the type of control that is necessary in quasi-experimental designs. The data sets selected came from the original Dryad Digital Repository[5]. This repository uses a creative commons license that allows for the sharing of data without direct contact or personal permission for the creators. The data used in this study are included below using the Dryad citation practice:

> Symonds MRE and Tattersall GJ (2010) Data from: Geographical variation in bill size across bird species provides evidence for Allen's rule. Dryad Digital Repository. doi:10.5061/dryad.1421
>
> Price SA and Gittleman JL (2007) Data from: Hunting to extinction: biology and regional economy influence extinction risk and the impact of hunting in artiodactyls. Dryad Digital Repository. doi:10.5061/dryad.82

The first data set, Symonds and Tattersall (2010), is an Excel file of bird beak measurements that includes Excel fields for bird beak length, geographic location, and bird body mass. This data set was chosen because it is easily understandable to the new users and has a certain amount of universal appeal to a variety of scientists working in the biosciences.

The second data set, Price and Gittleman (2007), is an Excel file of species names that includes fields for hunted status, group size, day length for developmental stages, and

[5] Dryad Data Repisotory: http://datadryad.org/

mating season. A second data set was not used in the pilot test. The addition of the second data set was based on feedback on the pilot study design. Feedback from the study indicated that one data set limited the ability of results to be generalizable and for comparison between results.

Every participant used the same two data sets in order to maintain control, yet accommodate for subject area expertise that could create bias. The rationale was that having two data sets as a control would increase the ability to generalize about the two populations during data analysis.

6.h.2.c. Questionnaires

Each participant was sent two questionnaires. These questionnaires are discussed in more depth "step" descriptions that follow.

6.h. 3 Step 3: Each participant answered demographic questionnaire.

Before beginning the simulation each participant answered a basic demographic questionnaire. This questionnaire asked seven questions about professional title; years of experience; frequency of data use; experience with data reuse; educational background; data management training; and data deposition experience. The purpose of these questions was to provide context for the data and metadata that was collected later. A more detailed discussion of the data analysis that this questionnaire underwent is found in Section 7: Results. The questionnaire used for this study is included in the Appendix. Both groups were given the same questionnaire to answer as a means of control for comparing results. A similar questionnaire was used during pilot testing. In terms of this questionnaire, changes that were made to the pilot and the actual study are the addition of the fourth, sixth, and seventh questions.

6.h.4. Step 4. Each participant simulated integrating the data set.

As per the directions of a quasi-experimental methodology, each group was given a different treatment. In this case, the treatment takes the form of different task scenarios simulating the way the data sets can be integrated into both communities' collections. The description of the task scenario for Group L (information professionals) is included in the Appendix.

The Group L Task Scenario focuses on the information professional as someone who curates data sets. The purpose of this scenario is to have the information professional simulate the steps he/she performs when curating digital data sets.

Following a quasi-experimental influenced design, the Group S Task Scenario was expectedly different from the task scenario for Group L. The description of the task scenario for Group S (scientists) is included in the Appendix.

The Group S Task Scenario focuses on organizing issues in relation to a data-reuse scenario. This simulation takes this into account and looks to create a semi-structured exercise.

Following the instructions given in the task scenarios shared above, each participant simulated integrating the data sets into her/his own collection. This process takes place in the individual's own office / work environment.

6.h.5 Step 5. Each participant answered follow up questionnaire.

After each participant completed the data set integration simulations, she/he completed the Follow-Up Questionnaire. This questionnaire consists of four main questions with sub-questions that prompt the participant to reflect on the data set integration process. The Follow-Up Questionnaire is included in the Appendix.

The purpose of the Follow-Up Questionnaire was for participants to reflect on and describe their organizing process and output. Results from these questionnaires are discussed in the Results section.

The Follow-Up Questionnaire used in this dissertation study is similar to the one used in the pilot study. There are a few differences between the two questionnaires, specifically:

- A different question order to improve flow.
- Confusing terminology from the pilot questionnaire was changed and revised for clarity in the version. Specifically, the term "metadata" was replaced by "keywords".
- More emphasis was placed on questions asking about formalized guidelines and standards.

6.h.6 Step 6. Study materials sent back to researcher.

Each participant was asked to send the researcher the Demographic Questionnaire, the Follow-up Questionnaire, and any organizing output created during the Data Set integration simulation. The type of organizing output sent back to the researcher included metadata records, lists of keywords, additional articles, and graphs. The Results section discusses these items in more detail.

6.h.7 Step 7: Data sets submitted into Dryad Instance.

After sending both questionnaires back to the researcher via email, each participant was asked to deposit the two, original data sets into the Metadata Research Center's instance of the Dryad Repository[6]. The Dryad Digital Repository is a

[6] Dryad MRC instance: http://mrc.dryad.org

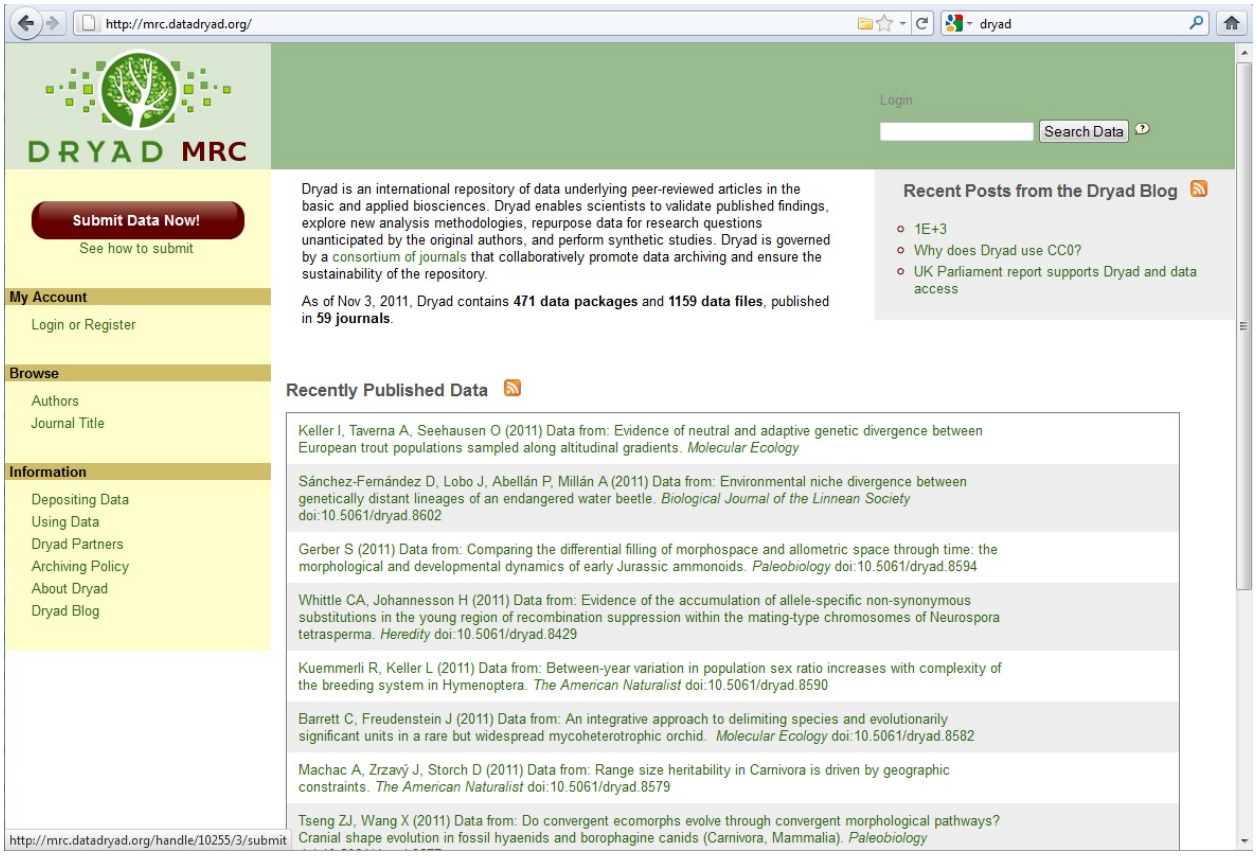
repository for data supporting publications in the biosciences using DSpace architecture and an application profile knowledge base (White et al., 2008). The Dryad Digital Repository Development team created an instance of their system to be used by Metadata Research Center affiliates. The purpose of this instance was to allow Dryad-based metadata research while still preserving the data integrity of the current production system.

The MRC instance of the Dryad Digital Repository was used in this study is to create a controlled system for analyzing organizing output. Dryad already had a mechanism for collecting data sets, specific types of descriptive metadata, and subject terms. Information professionals often use DSpace repositories, like Dryad, to maintain information objects. Scientists in the biosciences are often encouraged to deposit data into repositories systems. As an established and currently-used repository system in the sciences, an instance of Dryad fit the needs of this research study.

The following text is a description of how the MRC instance of Dryad was used within the context of this study. This description includes screenshots to convey the process each participant went through when depositing data sets and creating metadata. For the duration of this study, access to this instance and all the data submitted into was limited to the doctoral candidate, the doctoral candidate's advisor, and Dryad Repository architects.

Below is a screenshot of the MRC instance of the Dryad Repository Home page.

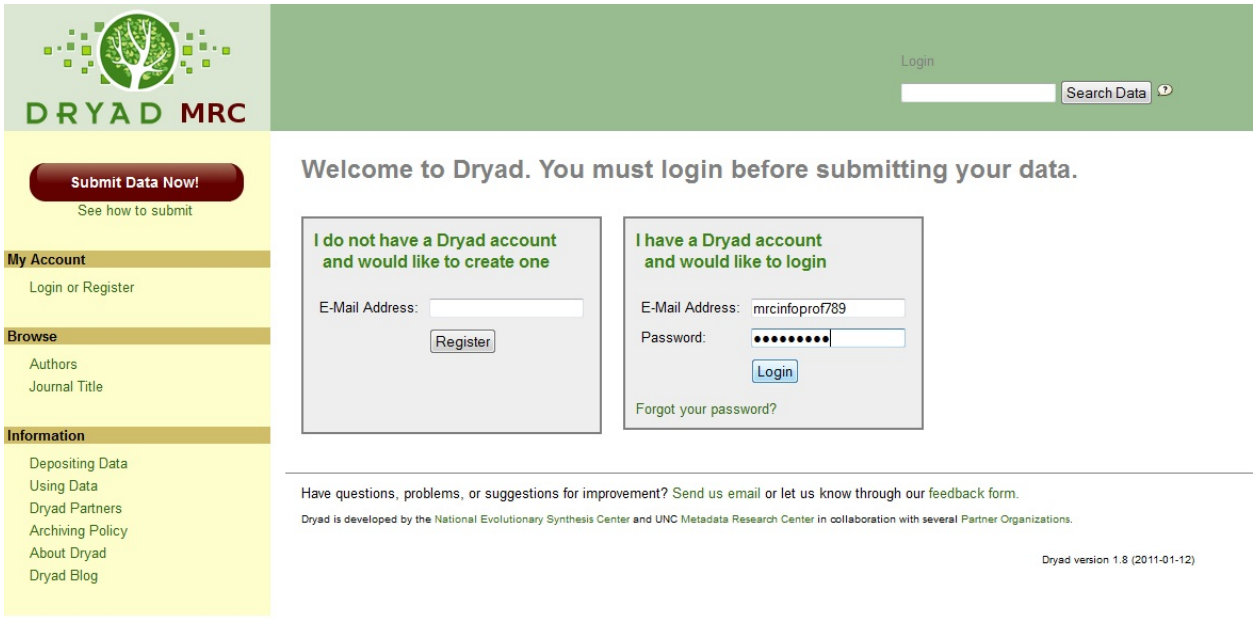
Figure 1: Dryad homepage



This instance is an exact copy of the real Dryad repository. The MRC instance uses all the same interfaces and underlying metadata schemes that are in the production version of Dryad.

In order to complete Step 7, participants were each given a personalized email address and password to use for logging into the MRC instance of Dryad. Each email address was created by the researcher and included the participant's unique identifier before the @ sign. Email addresses were registered with either Gmail or Yahoo. Having an email address is a requirement for logging into all Dryad systems. Email addresses were assigned to each participant in order to track results in an anonymous way.

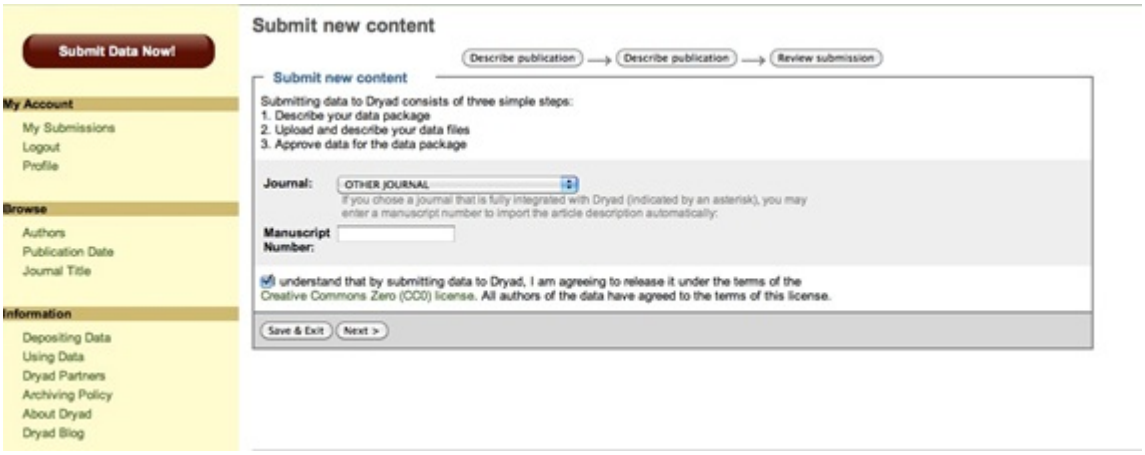
Figure 2: Dryad log-in screen



After logging into the system, participants entered the data set and metadata into the Dryad instance. The data entry process consists of three screens: a general entry screen, a publication metadata submission screen, and a data file metadata submission screen.

The Dryad "Submit New Content" screen is captured below.

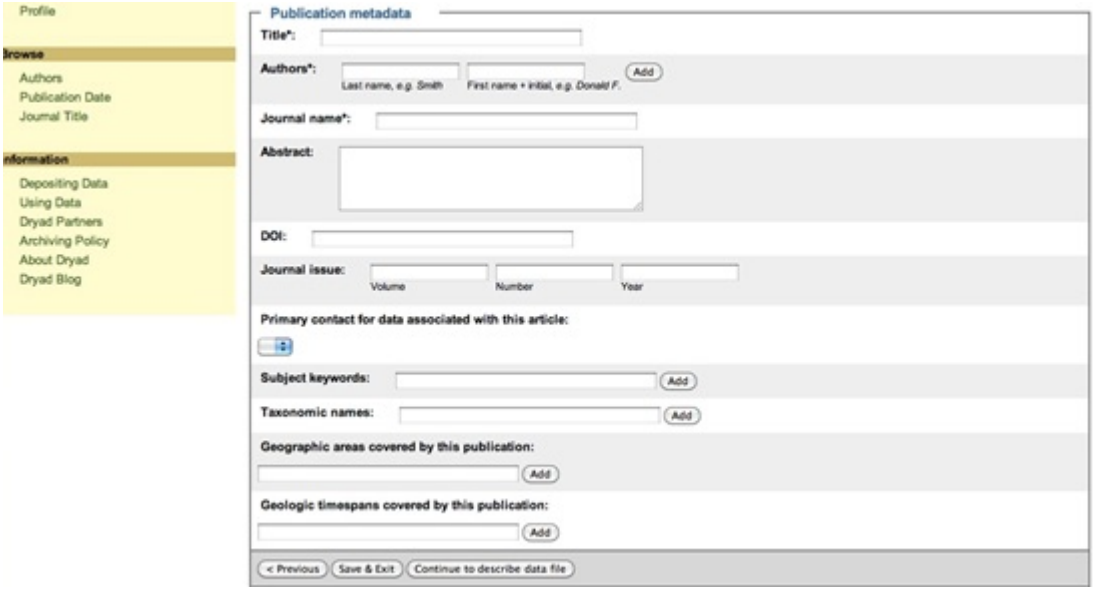
Figure 3: Dryad submit content screen



This screen explains the process for depositing data sets.

The next page participants encountered was the publication metadata screen.

Figure 4: Dryad publication metadata fields screen

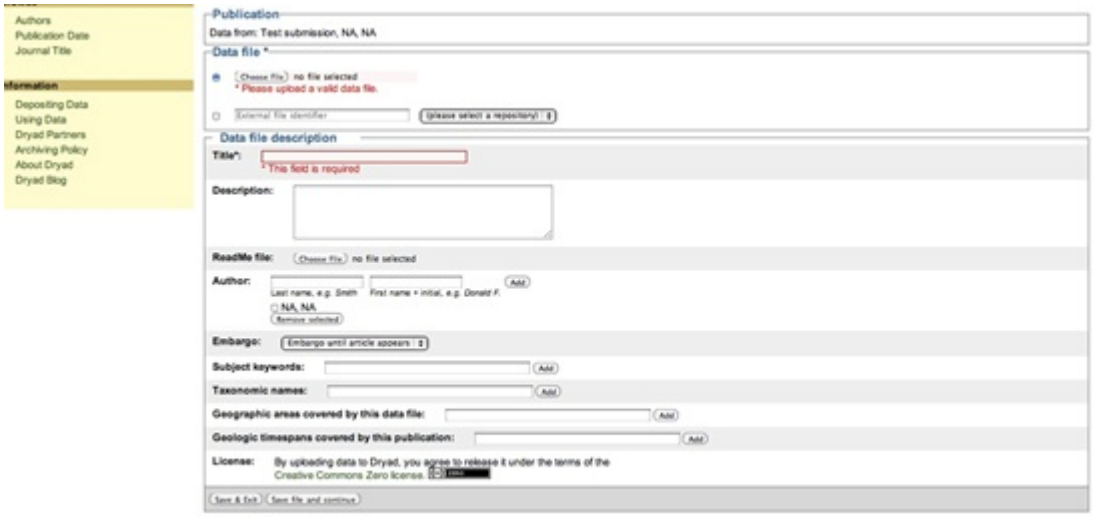


On this screen, participants entered in data about the original publications associated with the data sets. Participants have data about the original publication from the Instruction Sheet they were sent after agreeing to participate.

The type of metadata collected on the entry screen includes information about title, author, journal name, an abstract (written in free text), doi, journal issue, and a variety of subject terms (specifically, topical, taxonomic, spatial, and temporal terms). After filling out these metadata collection fields, the participants will click the "Continue to describe data file" to proceed to the next screen.

The data file description page is the next data entry form with which participants interacted.

Figure 5: Dryad data description screen



In this data entry screen participants uploaded data sets in the "Data File" section. Next, the participants created descriptive metadata for elements such as, title, description (using free text), author, and a variety of subject terms (specifically topical, taxonomic, spatial, and temporal terms). Before ending this portion of the study, participants saved the data file metadata. They were then prompted to review the created metadata in a separate screen in order to make any changes before exiting the system. Data collected from the questionnaires, the descriptive metadata, and subject terms input into the Dryad instance

were analyzed using qualitative and quantitative data analysis approaches. Data Collection and Analysis are discussed further in Sections 6.i and 6.j.

6.i. Data Collection and Participants

Data was collected through questionnaires, data set deposition, and metadata field completion. Questionnaires were used to gather information about participant demographics and data set organizing processes. Data set deposition and metadata field completion occurred in the Dryad instance.

6.i.1 Recruitment rationale

A total of 27 participants completed this study. Section 7a. Participants discusses participant recruitment in more depth. While recruiting for a previous study, White (2010b) found that scientists preferred to participate in studies where they could accomplish tasks and answer questionnaires without having to schedule face-to-face meetings with researchers. For this reason, the study reported here relied on remote completion. Questionnaires, task scenarios, and instruction sheets were emailed to the participants. They were asked to complete all activities within two weeks of agreeing to participate. Completing the task scenario, answering the questionnaires, and depositing the data sets into the Dryad instance was estimated to take no more than an hour and a half total. At any time, participants were welcome to email or call the researcher if they have questions. The Appendix contains an example of how this information was conveyed to the participants.

6.i.2 Participant Incentives

Participants were a $20 Amazon.com e-gift card for participating in the study as an incentive. Participant incentives were given out based on the level of task completion. The tasks are considered "fully completed" when both questionnaires were returned to the researcher and both data sets were deposited into Dryad with accompanying descriptive metadata and subject terms.

6j. Data Analysis

Data analysis of this research study focused on two components:

1. Qualitative data analysis: Using a grounded theory influenced analysis for both demographic and process oriented questionnaires.
2. Quantitative rubric: Analyzing the metadata submitted into Dryad.

Specific areas of each component were examined in detail. A description of the data analysis techniques used to examine each component is described as follows.

1. Qualitative data analysis

   The questionnaire answers were analyzed using grounded theory data analysis techniques. The data collected from the questionnaires took the form of short answers; narrative, paragraph-like descriptions; and procedural lists. An example of these responses is seen in Section 7 Results. To begin the inductive analysis process, the answers from questionnaires were reviewed and saved as RTF files. These files were opened in the qualitative data analysis software, Atlas.ti[7]. After an initial review, descriptive codes were created to highlight salient information

[7] Atlas.ti: http://www.atlasti.com/

organization processes and decisions. Codes were created to highlight similarities and differences related to metadata creation and subject term application. Twenty-two codes were created during this process. These codes are shared in the Results section.

Under the consultation of P. Mihas (personal communication, December 17, 2010; personal communication, October 17, 2011), Odum Institute's Coordinator of Education and Qualitative Research Consultant, a reviewer was selected to independently verify the codes within Atlas.ti. The codes were verified and returned. These codes were compared with each other using the Coding Analysis Toolkit (CAT)[8] All coding was also reviewed by the G. Liu (personal communication, December 15, 2011) Empirical Research Associate at Duke Law School. Multiple reviewers of the coding were used in an attempt to eliminate bias. Coding results are discussed in more detail in Section 8. Results.

2. Quantitative data analysis

Quantitative analysis was performed using traditional indexing methods of counting and mapping. The purpose of using these basic methods was two-fold:

- To characterize the information organization behavior of both scientists and information professionals in a controlled system.
- To learn more about the metadata fields that are frequently used by each group as well as a comparison of the level of subject analysis ability associated with each group.

[8] Coding Analysis Toolkit (CAT): http://cat.ucsur.pitt.edu/

To perform this data analysis, counting and mapping methods were used to show specific connections between descriptive metadata elements used, preferred subject terms, and established controlled vocabulary. The Quantitative Analysis Rubric, presented in the Figure that follows, shows the specific data that was analyzed.

Figure 6: Quantitative analysis rubric

--Counting—

Metadata

-- What Dryad metadata fields were used?
-- What fields were not used?
-- Which fields were most frequently used?

Subject terms

-- Number of subject terms total
-- Average # of terms chosen by each individual,
-- Average # of terms chosen by each group.
-- Create a tag cloud to visualize the subject terms used and their frequency
-- Out of the 4 subject descriptors available: general subject, scientific names, temporal, and spatial -- which area is used most frequently?

--Mapping—
Subject terms

--Map each subject term used to 4 of the vocabularies, some of which are cited in the Dryad vocabulary study from 2007. Vocabularies included are: Library of Congress Subject Headings(LCSH); Medical Subject Headings (MeSH); National Biological Information Infrastructure Thesaurus (NBII); and Integrated Taxonomic Information System (ITIS). All of these vocabularies were available using the Helping Interdisciplinary Vocabulary Engineering (HIVE)[1] demonstration project.

--Frequency of specific vocabulary terms mapped to which population

-- Compare terms used to original terms created by the data set author; which group has more overlap of terms.

This rubric highlights questions and tasks that related to the research questions presented in Section 6c. Based on the analysis performed in this rubric the following outcomes will be presented in the Discussion section.

For Question 1: Propose metadata for data repository use

For Question 2: Report on current metadata used by information professionals

For Question 3: Recommend controlled vocabularies for data repository use.

For Question 4: Report similarities in subject term use

For Question 5: Report differences in subject term use.

As shown by the outcomes above, each question relates to a specific outcome presented in Section 8. Discussion. These outcomes are based on results presented in Section 8. Results.

7. Results

Using the methodology discussed in the previous section, data about the information organization approaches of information professionals and scientists was collected. These results relating to information organization and scientific data sets are reported in detail. The results section begins by reporting on participant breakdowns. Findings from both the PIM task scenario as well as the Dryad task scenario portions of this study are reported next. The PIM task scenario portion of this study simulates what individuals would do when integrating datasets into their own collections. The Dryad task scenario portion of this study simulates what individuals would do when integrating datasets into a shared public system. For each portion results describing the information professional population are presented first, while results describing the scientist population are presented second.

The results reporting of the PIM task scenario of this study presents descriptive statistics from two questionnaires. The first questionnaire collected demographic data. Results from the first questionnaire reports type of position, the years each participant was in his/her current position, areas of expertise, education, and experience using data. The second questionnaire collected each participant's reaction after performing the PIM-influenced task scenario. Results from the second questionnaire include the type of changes each participant made to the data set, the type of metadata used to create read me files or records, the type of keywords created to describe the data sets, and the type of guidelines used for creating metadata. Participants also had the opportunity to return

both the modified data sets and all additional documentation created during the task scenario. Results from these materials are reported as well.

The results portion of the Dryad task scenario presents descriptive statistics from metadata and subject terms deposited in the Dryad system. Participants were asked to deposit two data sets in the MRC Test instance of the Dryad repository. During deposition, participants had to apply both descriptive metadata and subject terms to describe each data set. The descriptive metadata applied included information including title, author, and description. The subject terms applied include spatial, temporal, topical, and scientific terms. A discussion comparing these two participant groups is included in Section 8: Discussion.

7a. Participants

Participants were recruited from April to August 2011. The proposed plan was to include 30 participants: 15 information professionals and 15 scientists. A total of 207 people were contacted via personal email as potential participants in this study. Of the 207 people, 164 potential scientist participants were contacted from a list of scientific researchers interested in scientific data management issues. The remaining 43 potential information professional participants were recruited from a list of information professionals interested in metadata and scientific data.

A total of 11 participants withdrew early from the study. The participants who withdrew included seven information professionals and three scientists. When a reason was cited for withdrawal, the most frequently used reason for withdraw was a lack of time or being too busy than originally thought. Time related issues cited included a finals/final grades being due, busy conference schedule, doing field research in an area

with little to no internet service, and not enough time between field research and other travel. Upon analysis, due to the time period (late spring through summer), it is not surprising that more academically included participants withdrew.

By study completion, 27 participants were successfully recruited for this study. Out of those 27, 11 were information professionals and 16 were scientists. All 27 participants completed the personal organization task component. The majority, 25 participants, completed the Dryad task component. One participant out of each group (one information professional and one scientist) did not complete the Dryad portion. The scientist participant cited confusion on how to export data from a certain software program into Dryad. Even though asked, the information professional participant did not respond to questions about why the Dryad component was not completed.

7b. PIM-influenced portion

The results reporting of the PIM task scenario of this study presents descriptive statistics from two questionnaires. The first questionnaire, included in the Appendix, collected demographic data. Results from the first questionnaire reports type of position, the years each participant was in his/her current position, areas of expertise, education, and experience using data. The second questionnaire, included in Appendix, collected each participant's reaction after performing the PIM-influenced task scenario. Results from the second questionnaire include the type of changes each participant made to the data set, the type of metadata used to create read me files or records, the type of keywords created to describe the data sets, and the type of guidelines used for creating metadata. Participants also had the opportunity to return both the modified data sets and

all organizing output created during the task scenario. Results from these materials are reported as well.

7b1. Demographic Questionnaire

Group characteristics reported in Demographic Questionnaire are reported in following order: type of position, the years each participant was in his/her current position, areas of expertise, education, and experience using data.

7b1a: Position/Years in position

Each participant was asked to list his/her current job title. Results from this question are reported for both information professional and scientist participants.

Information Professionals

Information professionals participating in this study came from all levels of professional library and information management structure. Position titles included a Postdoctoral Researcher, an Associate Dean, and various levels of librarian status. Only one of the participants had a title that specifically included the phrase "Data Services". Information professional participants had 3 different types of position: library related, research-related, and technology-related. Table 1 groups the position titles.

Table 1: Information professional positions and titles

Library-related positions	Research-related positions	Technology-related positions
Associate Librarian Associate Dean of Libraries Head, Metadata & Cataloging Science Data Services Librarian Science Librarian	Research Assistant Post-Doctoral Researcher	Data Program Manager Principal Computer Scientists Technical Information Specialist

Each participant was asked to choose one of four categories to describe the number of years she had been in her position. Information professionals chose from four categories based on the year ranges of: 0-3, 4-7, 8-11, and 11 or more. Eight participants indicated that they had been in their current position for 3 or less years. The other three participants each chose one of the other categories.

Scientist

Scientist participants represented a variety of positions in the academic teaching and research community. Participants ranged from professors to post-doctoral fellows. Positions held by scientists could be categorized as either professor or research positions.

Table 2: Scientist positions and titles

Professor positions	Research positions
Assistant Professor of Biology	Natural Resources Specialist
Assistant Professor of Entomology	Plant Systematist and Herbarium Curator
Assistant Research Professor	Postdoctoral Fellow
Associate Professor	Postdoctoral Research Fellow
Associate Professor	Research Geologist
Professor	Research Scientist & Curator
Professor and Chair	Research Zoologist/Curator
Visiting Assistant Professor of Biology	Senior Research Entomologist

Looking at this group a little more closely, three biological specimen curators participated in this study as scientist participants. Two postdoctoral fellows participated as well.

Each participant was asked to choose one of four categories to describe the number of years she had been in her position. Scientist chose from four categories based on the year ranges of: 0-3, 4-7, 8-11, and 11 or more. Two participants (both

postdoctoral fellows) had 3 or less years in their positions. Seven participants had 4 to 7 years in their positions. Two participants had 8 to 11 years in their positions. Five participants had served more than 11 years in their positions.

7b1b: Area of expertise

Both groups of participants were asked to list areas of expertise or specialization. Participants were given an unlimited number of expertise types to list in free text. The number of specializations listed ranged from one to four.

Information Professional

Nine information professionals listed library science, information science, or informatics as an area of expertise. The other two participants had subject expertise in computer science and linguistics respectively. Five information professionals listed areas of expertise in the sciences. Scientific areas are listed in bullets as follows:

- Astronomy
- Biogeochemistry
- Biology
- Botany
- Environmental science
- Physics
- Plant pathology
- Toxicology

Scientist

Participants were asked to list their areas of subject expertise and all 16 participants listed at least one. A total of 15 scientist participants listed a type of biology as an area of expertise. Types of biology listed included aquatic, evolutionary, population, and pollination. Other biology-related areas of expertise listed were

biological oceanography, biodiversity, and biochemistry. Areas of expertise that were not biology are listed in bullet points as follows:

- Chemistry
- Ecology
- Entomology
- Genetics
- Molecular ecology
- Paleontology
- Palynology
- Parasitology
- Zoology

7b1c: Education

Participants were asked to list the type of degrees obtained. Respondents listed the type of degree and in which subject that degree was obtained.

Information Professional

Participants were given the opportunity to list the educational degrees each had obtained. The number of degrees listed by information professional participants ranged from one to four. The majority of participants indicated they obtained two degrees. One participant indicated having received a PhD. All 11 participants listed either completing or being the process of completing at least one master's degree. Eight participants indicated having received a master's degree in library and/or information science. Six participants included another master's degree in a topic other than library or information science. Nine participants noted having a bachelor's degree.

Scientist

Participants were given the opportunity to list the educational degrees obtained. A total of 14 scientist participants had obtained a PhD in a scientific sub-domain. The highest degrees obtained by the other two scientist participants were master's degrees.

7b1d: Experience using data

Participants were asked a series of questions related to data. The three areas examined include experience using data created by others, data management, and data deposition into repositories.

Information Professionals

Using data created by others

Experience with using data created by others varied among information professionals. Seven information professionals indicated that they had experience using data created by another person or organization. Two participants indicted that they had no experience using data created by another person or organization. One participant indicated both with a caveat that it depended on the definition of data being used.

Data management

The type of data management training also varied among information professionals. Three participants indicated having participated in formal data management training. Five participants mentioned that they had informal data management training. Two participants indicated having no data management training at all. One participant did not answer that question.

Data deposition into repositories

Eight participants had experience depositing data in a data repository. Six participants include specific subject repositories where they had deposited data. Table 3 lists these five repositories and the web address at which each can be located.

Table 3: Repositories used by information professionals

Repository Name	**Web Address**
Dryad	http://datadryad.org/
Knowledge Network for Biocomplexity (KNB)	http://knb.ecoinformatics.org/index.jsp
National Library of Medicine (NLM) Repository	http://collections.nlm.nih.gov/muradora/
National Snow and Ice Data Center	http://nsidc.org/
Planetary Data System	http://pds.nasa.gov/

Two participants listed institutional repositories as sites of deposit. A personal website was also listed as a place of deposition by one of the participants.

Scientists

Using data created by others

A total of 15 scientist participants had at one time used data created by another person. Only one scientist participant had not used data created by another person.

Data management

The types of data management training varied among scientists. Two participants have had formal data management training. Five participants indicated having informal data management training. Nine participants had no formal data management training.

Data deposition into repositories

A total of 15 scientist participants noted having deposited data in some type of repository. Table 4 lists these five repositories and the web address at which each are located.

Table 4: Repositories used by scientists

Repository Name	Web Address
Dryad	http://datadryad.org/
Genbank	http://www.ncbi.nlm.nih.gov/genbank/
Global Biodiversity Information Facility (GBIF)	http://www.gbif.org/
Treebase	http://www.treebase.org

Nine of participants indicated depositing data in Genbank. Four scientists had deposited data into Treebase. Two participants each have used Dryad and GBIF repositories. Other places used by scientists for depositing data include society websites and institutional repositories.

7b2. Follow-Up Questionnaire

The second questionnaire prompted participants to reflect on the PIM-influenced data set simulation task scenario. This section of the results will report findings from both the yes/no responses and the narrative coding related to specific areas of descriptive metadata creation and subject term application. Results reporting begins by presenting changes participants made to the dataset. This is followed by reporting on descriptive metadata choices including metadata surrogates, metadata schemes, and metadata creation guidelines. Next, controlled vocabulary use is presented. The section ends by reporting on data set organization procedures that were performed by each participant.

7b2a: Data Set Changes

Participants were asked about the changes they had made to the accompanying data sets. Types of changes made included saving data in a different format and the rearrangement or deletion of data within the spreadsheets. Some participants changed the data sets in multiple ways. These changes are discussed in more detail in the section below.

Information Professionals

Seven of the eleven information professional participants reported making changes to the data sets. The most frequently made changes were to data arrangement; file format changes; and deletion of columns/rows. Changes to data arrangement were made by three different participants, while file formats changes and deletion of columns/rows were made by four participants

Scientists

Thirteen out of 16 scientists reported making changes to the data sets. The most frequent changes made were to file format, rearrangement of rows/columns, and actual changes to data set coding. File format changes and data rearrangement were reported by eight participants, while four participants discussed making changes to data set coding.

7b2b: Descriptive Metadata Use

Participants were also asked to report descriptive metadata use. Based on this reported data, specific codes were used to analyze a) the creation of metadata surrogates, b) the use of formal and informal metadata schemes, and c) the use of standardized guidelines as aids for describing information objects. Results from these three areas are discussed in the following section.

7b2b1: Metadata Surrogates

Metadata surrogate creation refers to when an accompanying document is created that describes the content of the data set. Examples of metadata surrogates are read me files in text format and standardized metadata records in Dublin Core.

Information Professionals

Eight of the eleven participants reported creating some type of surrogate to describe the data sets. For the purpose of this study, a surrogate was considered to be any type of external descriptive file or notation that was used to identify and/or describe the data sets. Types of surrogates included formal metadata records, read-me files, and citations. One participant indicated that a read me file or records was not created, but this participant indicated that a citation within a personal repository was used. Another participant indicated that because the data sets were only part of a much larger collection of materials, they did not receive records or other metadata, but indicated that the collection the data set belonged to would receive a metadata record. These two instances are examples of participants who self-identified as "not creating metadata", but their narratives/short answers indicated otherwise. In total, ten out of the eleven information professional participants used some type of descriptive metadata to identify data sets in their own workflow. Only one participant who indicated not making changes to the dataset had a narrative that supported that claim.

Scientists

Five out of the sixteen scientists created some kind of descriptive metadata about the data sets. These metadata records were read-me files only. One scientist indicated creating a type

of surrogate, but it turned out to be a reference list. This piece of additional documentation was excluded in the five listed earlier.

7b2b2: Metadata Schemes

The use of standardized, local, and personal metadata schemes was also examined for both groups. Formal standardized metadata schemes are national or international metadata or cataloging standard. Examples include Dublin Core[9] or the Ecological Metadata Language[10]. Local metadata schemes are created by the individual's institution, lab, or library. Personal metadata schemes are based off of personal preference or long term habit.

Information Professional

Eight participants indicated they used metadata. Five of these eight participants listed the metadata schemes that were actually used. Three participants indicated using a locally developed, yet standardized scheme for scientific data. Two participants mentioned using Dublin Core. One of the participants who used Dublin Core also mentioned using a scheme related to the software he/she was using. One other participant mentioned using the Ecological Metadata Language.

Scientists

None of the scientist participants listed any formal metadata scheme being used. Only one scientist indicated using a type of personal scheme.

7b2b3: Metadata creation guidelines

[9] Dublin Core: http://dublincore.org/documents/dces/

[10] Ecological Metadata Language: http://knb.ecoinformatics.org/software/eml/eml-2.1.0/index.html

The use of metadata creation guidelines by both information professionals and scientists was also reported and examined. The purpose of looking into this area of metadata practice is to understand the overall principles that guide metadata creators and depositors in personal and more formalized systems.

Information Professional

Participants were asked if any guidelines or rules were used to help in creating descriptive metadata. Four participants cited specific standardized rule sets. These rules included the Anglo-American Cataloging Rules, Second Edition (AACR2)[11]; the United Kingdom Data Archive File Format Guidelines[12]; and Microsoft Excel documentation guidelines[13]. One participant cited using local guidelines for preparing the data.

Scientist

As with metadata schemes, none of the scientist participants listed use of any type of metadata creation guidelines. The results did indicate that scientists were creating metadata and altering data sets in order to conform to certain software packages. A discussion of this occurrence among scientist participants is included in the Discussion section.

7b2c: Subject term application

Subject term application, (keyword creation) was also examined. Subject terms are words that are selected to describe the aboutness of scientific data sets. Participants were asked

[11] AACR2: http://www.aacr2.org/

[12] United Kingdom Data Archie File Format Guidelines: http://www.data-archive.ac.uk/create-manage/format/formats

[13]Microsoft Excel information and help: http://office.microsoft.com/en-us/excel-help/

if they created keywords or subject terms to describe the data set. Subject term, if created, were characterized as either standardized controlled vocabularies or personal/folksonomy terms. Standardized subject terms come from a national or international controlled vocabulary like the Library of Congress Subject Headings (LCSH)[14] or Medical Subject Headings (MeSH)[15]. Personal tags or folksonomies are subject terms (or tags) created from personal preference or local vernacular.

Information Professionals

Eight of the eleven information professional participants indicated that they created keywords to describe the data sets when simulating integration of those data sets into their own collections. Four participants noted that they drew from standardized controlled vocabularies for their subject terms. The controlled vocabularies reported as used included the Medical Subject Headings (MeSH), Library of Congress Subject Headings (LCSH), National Biological Information Infrastructure Biocomplexity Thesaurus (NBII)[16], and Global Change Master Directory Keyword list (GCMD)[17]. One participant indicated using both MeSH and NBII. Two participants noted using MeSH subject headings. One participant each indicated using LCSH, NBII, and GCMD.

[14] LCSH: http://www.loc.gov/aba/cataloging/subject/

[15] MeSH: http://www.nlm.nih.gov/mesh/meshhome.html

[16] NBII: http://www.nbii.gov/portal/server.pt/community/biocomplexity_thesaurus/578

[17] GCMD: http://gcmd.nasa.gov/

Local controlled vocabularies or keyword lists were also reported by participants. Two participants indicated using locally developed and controlled terms lists. One participant reported a combination of a local list and MeSH.

Two participants reported that the subject terms came "from within the file" itself. One participant did not specify from where the subject terms they used came. Three participants indicated that they did not create subject terms to describe the files.

Scientists

Two of the sixteen scientist participants reported creating keywords for the data set. Only one participant mentioned using a form of vocabulary control even though this person did not create keywords during the task scenario. The sources for vocabulary control were Phylogeny[18] and Wikipedia[19]. While not formal controlled vocabularies or thesauri, these two resources still served as a source of vocabulary control.

7b2d: Procedures for organizing and describing datasets

Participants were asked to write a small narrative describing the procedures that were used to simulate integration into their own collections. Narrative answers were given in the form of longer narratives, short answers, and lists. Examples of these three types of answers are included below.

[18] Phylogeny: http://www.ucmp.berkeley.edu/exhibit/introphylo.html

[19] Wikipedia: http://www.wikipedia.org/

Narrative

> Qualification - We do not have readme files for this level of granularity. What you call a data set, we call a product. We collect products into a "collection", for example the 19K images taken by the Viking Spacecraft of Mars. Collections are collected into Archive Bundles which is the archival package. I expect that we would create one collection for your two data sets and one archive bundle. The collection and bundle could have an optional aareadme [sic] file. (infoprof014)

Short answers

> "No guidelines or rules followed, just personal practice" (mrcinfoprof001)

Lists

> "1. I copied the species names from Data set 2 to another column
> 2. I deleted the specific epithets for the first column and the genera from the second column of mammal data just described so that the genera and specific epithets would be in their own columns in line with the bird data.
> 3. I moved the body masses from the mammal data to the same column (D) as in the bird data after seeing that the units were the same.
> 4. I moved the remainder of the mammal data set to the right of where all the bird data would be (column M).
> 5. I moved the column headings from the mammal data to the appropriate columns, made them fit into one row and word-wrapped them.
> 6. I added all the remaining column headings from the bird data to the mammal data.
> 7. I changed all the font and cell styles to be the same." (scientist001)

Every participant returned some type of narrative, short answer, or list response. It was determined that including every response in this results section would be cumbersome and tiring for the readers. For this reason, segments of the narrative responses are included in the Section 8 in order to enhance and inform the Discussion.

For data analysis, narratives were entered into the Atlas.ti software, version 6.1. A total of 22 codes were created during an inductive coding process to describe results from the Follow-Up Questionnaire. Codes were used to highlight salient information organization issues

mentioned in narrative and short answer responses from the Follow-Up Questionnaire. Narrative coding data was not included in quantitative analysis. Two coders applied the codes using the code explanation table included in the Appendix.

Using the Coding Analysis Toolkit[20](CAT), coding from both coders was compared for basic frequency, exact match, overlap, and Kappa Scores. Table 5 highlights the results of the CAT output. The first column lists the codes used to mark up the narratives. The second column lists the number of time the corresponding code was applied by the first coder. The third column lists the number of times the corresponding code was applied by the second coder. The fourth column lists the number of exact matches for the corresponding code between Coder1 and Coder2. Exact match includes not only the use of the code, but the exact highlight length used on a sentence or word that was coded. The fifth column lists the number of times coding overlapped between the two coders for the corresponding code. Overlap is the not an exact match, but part of the two coders highlights overlapped when assign that specific code. The sixth column lists the Kappa score for each code. Kappa is measured in percentages. The last column lists the Kappa score with overlap as a consideration.

[20] Coding Analysis Toolkit: http://cat.ucsur.pitt.edu/

Table 5: CAT analysis of Atlas.ti codes applied by two coders

Code	Coder1	Coder 2	Exact Match	Overlap	Kappa	Kappa (inc. Overlap)
Best practice	7	1	0	1	0.00	0.13
Choice not taken	2	14	2	1	0.14	0.21
Choice taken	2	8	0	1	0.00	0.10
Choices	7	5	1	0	0.09	0.09
Data set changes	19	35	9	3	0.20	0.27
Data set process	26	22	3	7	0.07	0.24
File format	16	36	2	7	0.04	0.18
Local guidelines	5	8	3	2	0.30	0.50
Naming	5	13	0	2	0.00	0.11
No data set changes	6	8	0	6	0.00	0.43
No guidelines	8	9	3	4	0.21	0.50
No metadata	1	0	0	0	0.00	0.00
Personal guidelines	5	12	2	2	0.13	0.27
Personal metadata	15	5	0	1	0.00	0.05
Personal subject terms/folksonomy	8	4	1	2	0.09	0.27
Repository	7	18	2	3	0.09	0.22
Sense making	5	23	0	2	0.00	0.07
Software	23	29	2	17	0.04	0.64
Standard guidelines	3	3	0	1	0.00	0.17
Standardized metadata	10	14	0	9	0.00	0.38
Standardized subject terms	9	8	5	3	0.42	0.92
Time	2	2	0	0	0.00	0.00
Totals	191	277	35	74	0.08	0.26

The codes highlighted in Table 5 were used to gain more insight into participant descriptive metadata creation and subject term application. Coding in relationship to narrative responses are discussed in more depth in the Discussion section.

7b2e: Additional Documentation and Data Set file formats.

Participants were asked to return any additional documentation that was created during the PIM-influenced task scenario. Additional documentation was collected in order to

characterize the types of records, metadata, and supplemental material that were created by each user group when working with scientific data sets.

Information Professionals

Nine of the eleven information professional participants returned some type of additional documentation. The most additional documentation that participants sent was three separate pieces of material. The types of materials sent include notes; explanatory emails; processing steps; metadata records; keywords; and references.

Six out of eleven information professional participants returned a version of the data sets to the principal investigator. Three participants returned data sets saved in multiple formats. Another three participants returned data sets saved in only one format. Data sets altered by information professionals were saved in Excel (xls), comma separated value (csv), portable document format (pdf), and Text (txt) formats. Four of the six participants who returned data sets saved the files in comma separated value (.csv) file.

Scientists

Five out of sixteen scientist participants returned some type of additional documentation. The most additional documentation that was sent by a scientist participant was six pieces of material. The types of materials sent include the instruction sheets with edits; procedures; file maker pro files, and different excel file outputs plus graphs.

Four out of the sixteen scientists returned a version of the data sets to the principal investigator. One participant returned data sets saved in multiple formats. Data sets returned by scientists included Excel (.xls), Nexis (.nex), Text (.txt), and File Maker Pro (fp7).

As well as file format, scientists also reported the type of software used to create these files. The type of software used was varied. Software programs listed as being central to the data set organization process and the frequency of use are included in Table 6.

Table 6: Software and use by the scientist participants

Software	Frequency of use
Filemaker	3
R	3
JMP (SAS product)	1
Microsoft Excel/dBasel	1
Microsoft Access	1
SPSS	1

Two other scientists, not include in the chart above, mentioned database software, but did not list specific names.

7c. Dryad portion

Results from the Dryad task scenario include descriptive statistics from descriptive metadata and subject terms deposited in the MRC-Dryad instance. Participants were asked to deposit two data sets in the MRC Test instance applying both descriptive metadata and subject terms describing to each data set. Dryad's descriptive metadata and subject terms is represented by metadata element properties. These properties (like title or author) are presented in fill-in fields. Figure 7 is a screenshot of the fill in fields used to collect descriptive metadata and subject terms in Dryad.

Figure 7: Repeat of Publication Metadata Screen

The descriptive metadata applied included information including author, corresponding author, description, publication name, and title. The subject terms applied include spatial, temporal, topical, and scientific name fields.

This portion of the results section begins reporting participant information and then elaborating on data set deposition for this task scenario. Dryad descriptive metadata field usage will then be reported for both groups. Reporting on the Dryad task scenario concludes with a description of the subject term application that each group of depositors used within the MRC-Dryad Instance

7c1.Participants and Data Set Deposition

A total of 25 participants completed the Dryad task component. Ten information professionals and 15 scientists completed the Dryad portion of this study. Even though asked, the information professional participant did not respond to questions about why the Dryad

component was not completed. The scientist participant cited confusion on how to export data from a certain software program into Dryad.

As mentioned previously, each participant was asked to deposit two data sets into the MRC test instance of the Dryad Repository, yet some participants only deposited one data set. Nine of the ten information professional participants who completed the Dryad portion deposited two data sets. One information professional deposited a single data set. Eleven out of 15 scientist participants deposited two data sets into Dryad. Four scientist participants deposited a single data set into Dryad.

7c2. Descriptive Metadata

The Dryad repository uses a Dublin Core application profile approach for describing metadata about data sets. The types of descriptive metadata collected by Dryad include author, corresponding author, description, publication name, and title. Figures 8 and 9 presents the metadata field usage by participant type for Data Set 1 and 2.

Figure 8: Dryad Metadata Field Usage for Data Set 1

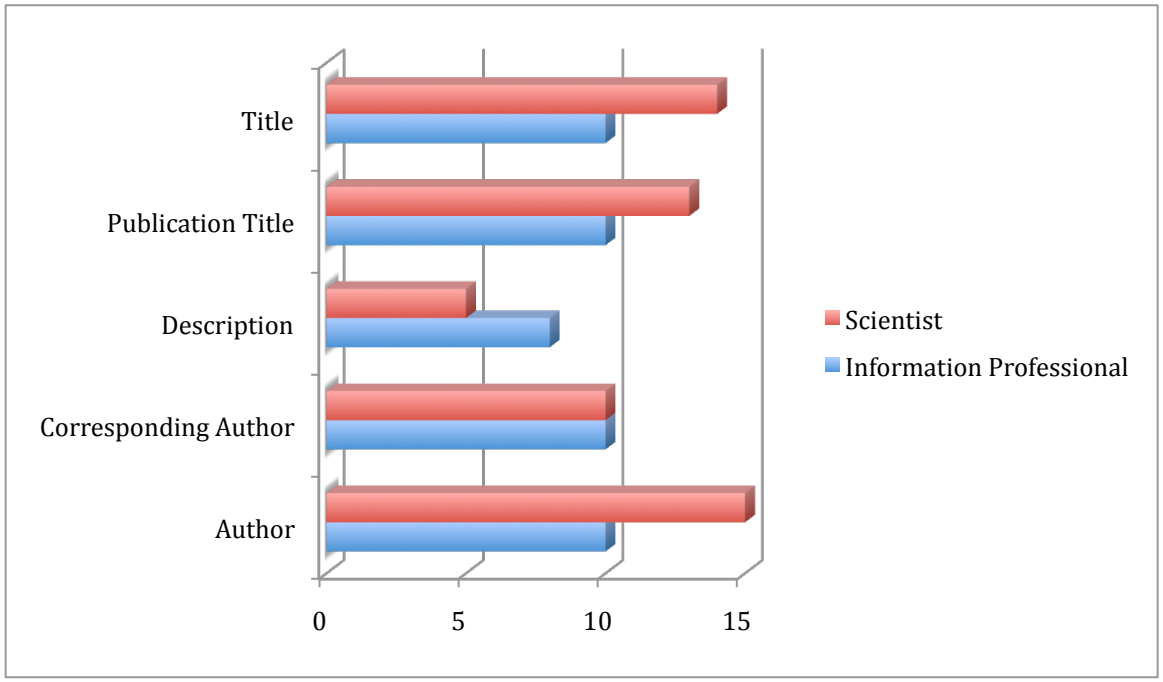
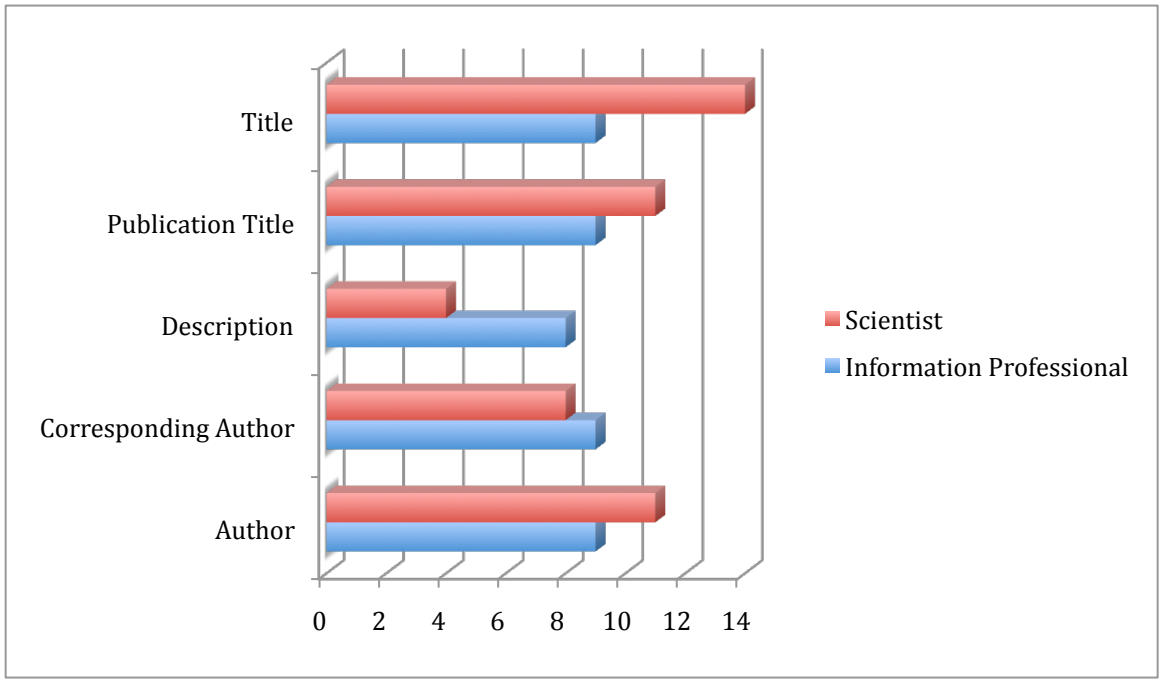


Figure 9: Dryad Metadata Field Usage for Data Set 2



Information Professionals

All information professional participants who completed the Dryad portion of this study entered in metadata for the author, corresponding author, publication name, and title fields. Only eight participants entered in descriptions for data set 1.

All information professional participants who deposited data set 2 entered metadata into the author, corresponding author, publication name, and title fields. Only six information professional participants used the description field.

Scientists

All scientist participants entered in metadata for the author, publication name, and title fields. Only ten scientist participants entered in corresponding author information for data set 1. Five scientist participants entered information into the description field.

All scientist participants who deposited data set 2 entered metadata into the author, publication name, and title fields. Only eight scientist participants used the corresponding author field. Four scientist participants used the description field for data set 2.

7c3. Subject terms

Dryad collects four types of subject terms during data deposition: spatial, temporal, topical, and scientific. Participants had a choice as to whether or not they used certain types of subject term fields. The most frequently used subject term field was the topical field with ten information professionals and ten scientists filling out this field for data set 1. Nine information professionals and eight scientists filled out this field for data set 2. The temporal field was the least used field out of all subject term types. Only one information professional and four scientists using the field for data set 1. One information professional and two scientists used the

field for data set 2. The section that follows presents findings from the four subject term areas collected during the Dryad task scenario.

7c3a.Spatial terms

Spatial terms describe the geographic location associated with a data set or article. Content that is appropriate in this field include geographic coordinates; city, town, country, and continent names; and broad labels such as, "worldwide" or "global". None of the participants in this study used geographic coordinates.

Table 7. Spatial terms divided by group type and data set

	Information Professional		Scientist	
Spatial	Data Set 1	Data Set 2	Data Set 1	Data Set 2
# of participants who used	**5**	**3**	**7**	**4**
% of participants by group	**50%**	**30%**	**47%**	**36%**
Total # of terms applied by group	**10**	**3**	**8**	**3**
Avg # applied	**4**	**1**	**2**	**1**

Information Professionals

For data set 1, five out of ten information professional participants (50%) used the spatial metadata field. One of these five participants used the field improperly and included a scientific name. Information professional participants applied ten unique terms. The five most frequently used terms were "Australia", "Africa", "Antarctica", "South America", and "New Zealand".

For data set 2, three out of ten of information professional participants (30%) used the spatial metadata field. Each participant entered a single term and there was no overlap between terms. The three terms applied were "global", "unknown", and "halocene". The term "halocene" was applied inappropriately since it is a temporal term.

Scientists

For data set 1, seven out of 15 scientist participants (47%) used the spatial metadata field. Eight unique terms were applied. The term "global" was used three times. All other terms were used once. One participant used the field incorrectly and added annotations

For data set 2, four out of 11 scientist participants (36%) used the spatial metadata field. Three participants used the term "global". The other participant used the field incorrectly and added annotations. One scientist participant filled in the spatial metadata field, but used as an annotation field as opposed to actual subject terms. For data set 1 this particular participant wrote "a wider range than I have time to enter". For data set 2, "data not in spreadsheet" and "may be in original paper(s)".

7c3b. Temporal terms

Temporal terms describe the time period in which the data sets were collected. Both participant groups used temporal metadata fields less than any of the other three subject metadata fields.

Table 8. Temporal terms divided by group type and data set

	Information Professional		Scientist	
Temporal	Data Set 1	Data Set 2	Data Set 1	Data Set 2
# of participants who used	**1**	**1**	**4**	**2**
% of participants by group	**10%**	**<10%**	**27%**	**18%**
Total # of terms applied by group	**1**	**1**	**1**	**1**
Avg # applied	**1**	**1**	**1**	**1**

Information Professionals

Only one information professional participant applied temporal metadata to data set 1. That one participant used the term "holocene" to describe data set 1. Only one information

professional participant applied temporal metadata to data set 2. This participant used the term "anthropocene" and it was the only term applied for the temporal metadata field.

Scientists

Scientist participants used temporal metadata more frequently than the information professional participants. For data set 1, four out of the 15 scientist participants (27%) applied temporal metadata. All four participants used the term "recent". For data set 2, two out of eleven participants (18%) used the same temporal term "recent".

7c3c.Topical terms

Topical terms describe the topic, main theme, or "aboutness" of the data set. The topical term field was the most frequently used of all the subject terms for both information professional and scientist participants.

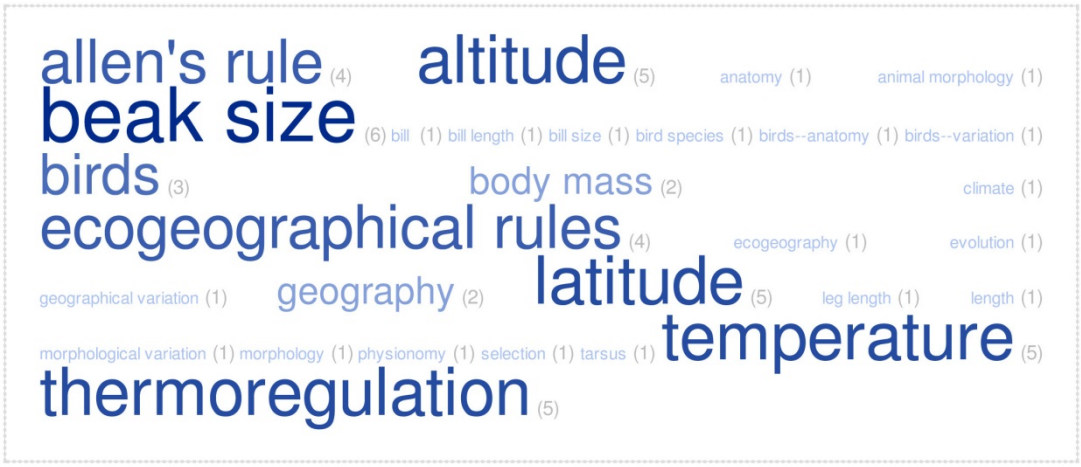
Table 9. Topical Terms Divided by Group Type and Data Set

	Information Professional		Scientist	
Topical	Data Set 1	Data Set 2	Data Set 1	Data Set 2
# of participants who used	**10**	**9**	**10**	**8**
% of participants by group	**100%**	**90%**	**66%**	**72%**
Total # of terms applied by group	**27**	**23**	**30**	**38**
Avg # applied	**6**	**5**	**5**	**6**

Information Professionals

Information professional participants applied topical terms more frequently than the other three types of subject terms. All (100%) of the information professionals who deposited data sets used topical terms. Figure 10 shows a tag cloud of all the topical terms information professional participants applied for data set 1.

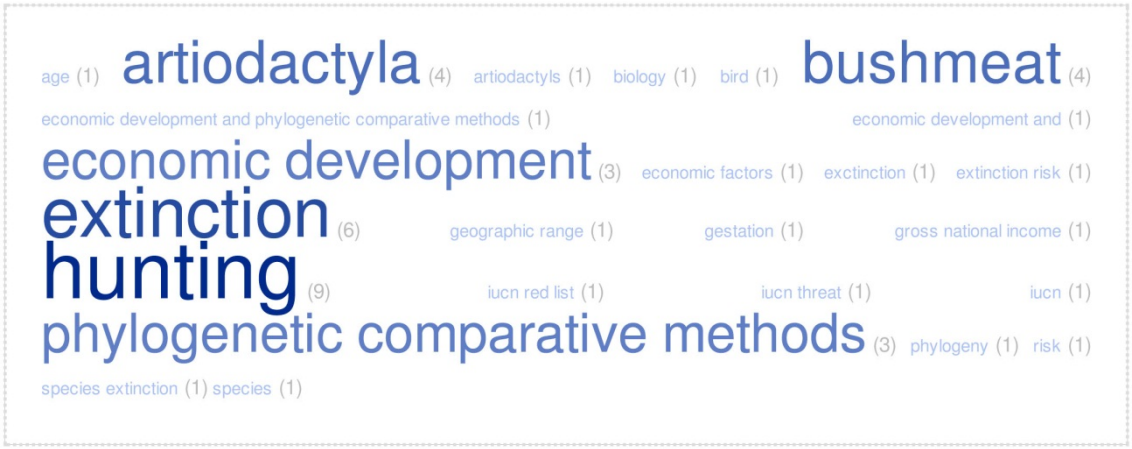
Figure 10: Information Professional Data Set 1 Tag Cloud



For data set 1, the five most frequently used terms were “beak size”, “thermoregulation”, “temperature”, “latitude”, and “altitude”.

Figure 11 shows a tag cloud of all the topical terms information professional participants applied for data set 2.

Figure 11: Information Professional Data Set 1 Tag Cloud

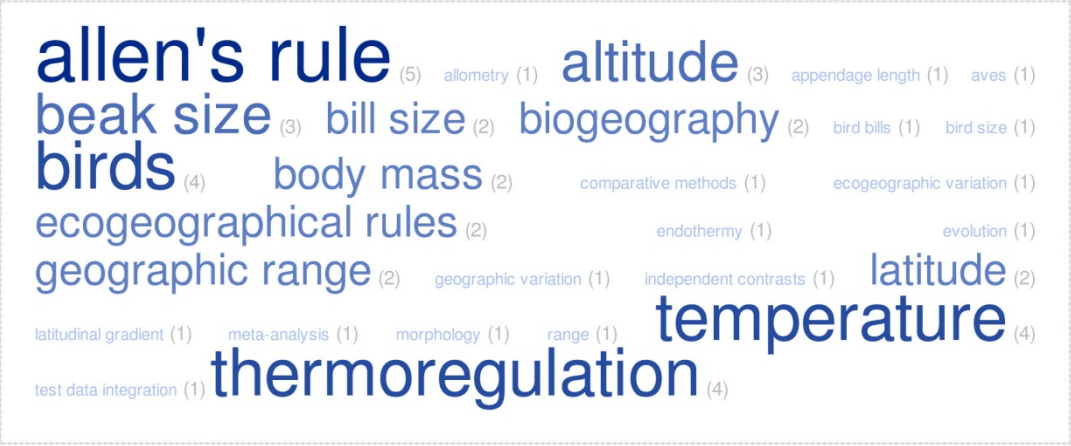


The five most frequently used terms were “hunting”, “extinction”, “artiodactyla”, “bushmeat”, and “economic development”.

Scientists

Scientist participants also applied topical terms more frequency than the other three types of subject terms. For data set 1, ten out of 15 scientist participants (66%) applied topical terms. Figure 12 shows a tag cloud of all the topical terms scientists participants applied for data set 1.

Figure 12: Scientist Data Set 1 Tag Cloud



The three most frequently used terms for data set 1 were temperature, thermoregulation, and birds.

For data set 2, eight out of 11 scientist participants (72%) applied topical terms. Figure 13 shows a tag cloud of all the topical terms scientists participants applied for data set 2.

Figure 13: Scientist Data Set 2 Tag Cloud

The three most frequently used terms were “hunting”, “extinction”, and “economic development”.

7c3d. Scientific

The scientific term field is used to describe standardized scientific names of biological species mentioned in the data set. Participants applied more scientific terms than they did any other terms. The range for scientific term application was very large because only a few participants applied a large number of terms.

Table 10. Scientific terms divided by group type and data set

	Information Professional		Scientist	
Scientific	Data Set 1	Data Set 2	Data Set 1	Data Set 2
# of participants who used	**6**	**7**	**9**	**6**
% of participants by group	**60%**	**78%**	**60%**	**55%**
Total # of terms applied by group	**236**	**210**	**224**	**209**
Avg # applied	*	*	*	*

*Average for Scientific names was skewed because one to two participants applied over 200 terms, while others only applied one

Information Professionals

For data set 1, six out of ten information professional participants (60%) used the scientific metadata field. The six most frequently applied terms were "Galliformes", "Psittaciformes", "Estrildidae", "Laridae", "Lybiidae", and "Sternidae".

One participant used the scientific terms field incorrectly and applied common names instead of scientific names. Examples of common term used are "toucan" and "penguins". One participant applied 219 scientific terms to data set 1.

For data set 2, seven out of nine information participants (78%) used the scientific metadata field. Participants either applied one or 209 terms. Five participants applied a single scientific term. Four participants applied the term "Artiodactyla". One participant applied the term "Addax nasomaculatus". Two participants applied the same 209 scientific terms. These 209 terms were listed as headings in data set 2.

Scientists

For data set 1, nine out of fifteen scientist participants (60%) used the scientific term field. 224 unique scientific terms were applied by scientist participants. The terms "Galliformes", "Laridae", "Lybbidae", "Psittaciformes", "Raphastidae", "Spheniscidae", and

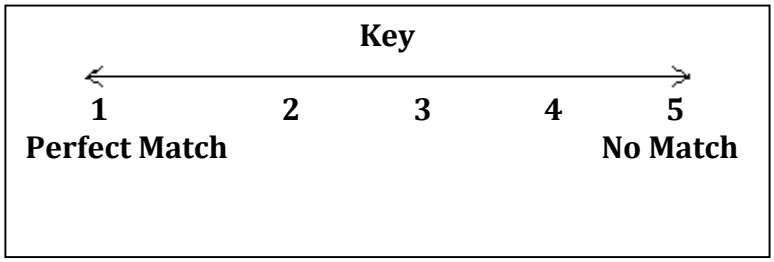
"Sterniscidae" were each applied four times. "Aves" was applied twice. All other terms were applied once.

For data set 2, six out of eleven scientist participants (55%) used the scientific term field. The most frequently applied term was "Artiodactyla" with five uses. All other terms were applied once. One participant applied 219 scientific terms. Another participant applied two terms. All other participants applied only one term.

7c3e. Vocabulary Mappings

All the terms created by the information professionals and scientists were mapped to already established controlled vocabularies. The vocabularies used for this study were: Library of Congress Subject Headings (LCSH); Medical Subject Headings (MeSH); National Biological Information Infrastructure Thesaurus (NBII); and Integrated Taxonomic Information System (ITIS). Certain types of subject terms were mapped to facet appropriate vocabularies. Due to its length, coding instructions and mapping criteria are included in the Appendix. Table 11 shows which type of subject term was mapped to which vocabulary and gives the average code applied for that vocabulary based on Coder.

Table 11: Vocabulary mapping averages by coder



	Coder 1		Coder 2	
	Info Prof	Scientist	Info Prof	Scientist
Spatial terms				
LCSH	3.6	3.5	2.7	2.5
NBII	4.7	5.0	4.3	4.5
TGN	2.7	2.0	2.3	1.0
Temporal terms				
LCSH	4.3	5.0	4.0	5.0
MeSH	5.0	5.0	5.0	5.0
TGN	5.0	5.0	1.0	5.0
Topical terms				
LCSH	3.2	3.4	3.6	2.9
MeSH	3.9	3.9	4.3	3.4
NBII	3.3	3.5	4.3	3.4
Scientific terms				
LCSH	3.8	3.9	3.7	3.8
MeSH	4.8	4.9	4.9	5.0

NBII	5.0	5.0	5.0	5.0
ITIS	1.2	1.1	1.3	1.3

Coder 1 mapped every code to the vocabularies specified above. Coder 1 mapped 1,024 terms. Coder 2 was a volunteer verification coder and mapped a subset of terms chosen from every fifth term of the larger 1,204 group. Coder 2 mapped 209 terms. Based on these codes, averages were taken for each vocabulary.

The goal was to have a score closest to "1", which is a perfect match according to the scale presented in the Appendix. Due to the small number of terms used for spatial and temporal terms, the research was unable to make any general statement about the averages. A more in-depth analysis of the topical and scientific terms is presented in the Discussion section.

7c3f. Comparing Subject Term Application

Out of the four sub-areas of subject terms used in this study (spatial, temporal, topical, and scientific), the topical terms were used most frequently by both groups. There were 12 terms in each data set that were used by both group. In relation to similarity in term application between the two groups, 27% of the terms applied by both groups overlapped for data set 1 and 24% of the terms applied by both groups overlapped for data set 2. The overlapping topical terms used by both information professionals and scientists were:

Terms from Data Set 1

Allen's Rule
Altitude
Bill size
Bird
Birds
Body mass
Ecogeographical rules
Geographic variation
Latitude
Morphology
Temperature
Thermoregulation

Term	Information Professional	Scientist	Term	Information Professional	Scientist
Allen's rule	4	5	Ecogeographical rules	4	2
Altitude	5	3	Geographic variation	1	1
Bill size	1	2	Latitude	5	2
Bird	1	2	Morphology	1	1
Birds	2	3	Temperature	5	4
Body mass	2	2	Thermoregulation	5	4

From Data Set 2

Age
Artiodacytls
Artiodacytla
Bushmeat
Economic development
Extinction
Extinction risk
Geographic range
Gestation
Hunting
Phylogeny
Phylogenetic comparative methods

Term	Information Professional	Scientist	Term	Information Professional	Scientist
Age	1	1	Extinction risk	1	2
Artiodacytls	1	2	Geographic range	1	1
Artiodactyla	4	2	Gestation	1	2
Bushmeat	4	1	Hunting	8	6
Economic development	3	4	Phylogeny	1	1
Extinction	1	4	Phylogenetic comparative methods	2	2

Looking at the differences, this means that information professionals applied 15 (33%) terms that were different from what scientist applied (i.e. unique to their group) for data set 1 and 11 (22%) terms for data set 2. Scientists applied 18 (40%) terms that were different from what information professionals applied (i.e. unique to their group) for data set 1 and 26 (53%) terms unique to their group for data set 2. A more detailed discussion of these results can be found in the Discussion section.

8. Discussion

Understanding more about metadata creation and subject term application performed by information professionals and scientists could have an impact on repository systems. In this section, findings reported in the results section are discussed in more detail. Specific quotations from the qualitative narratives are integrated into this section in order to give more context to the results being discussed.

The Discussion section begins by examining the descriptive metadata. PIM-influenced and Dryad task scenarios results are discussed and compared. This is followed by a discussion of guidelines and software. Next, subject term application is examined. This section also discusses and compares results from the PIM-influenced and Dryad task scenarios, plus an analysis of controlled vocabularies for describing scientific data sets. The next topic covered in this section is the question of "who should create metadata?" and the role of the information professional. The Discussion section concludes by presenting some of the limitations of this dissertation study.

8a. Descriptive Metadata

Descriptive metadata was collected using two task scenarios: PIM-influenced and Dryad. The sections that follow analyze and discuss findings from these task scenarios in more depth.

8a1. PIM-Influenced Task Scenario

The PIM-influenced task scenario prompted participants to reflect on their own processes and to try to describe them. As mentioned earlier, quantitative statistics were collected from reported participant behaviors, while qualitative descriptions and analysis are based off of participant narratives.

For the PIM-influenced scenario, participants reported organizing behaviors and wrote narratives to describe them. Based on these narratives, three codes were used by the researcher to highlight mentions of metadata in the narrative/short answers. The three metadata-related codes were: standard metadata, personal metadata, and no metadata. The codes were used in this study to highlight relevant explanations that support the descriptive statistics reported in Section 7. Results.

Almost all information professionals (10 out of 11) reported creating a type of metadata surrogate to describe scientific data sets.

Standard metadata is a national or international metadata or cataloging scheme, like Dublin Core[21] or Ecological Metadata Language[22]. All standard metadata usage was conducted by information professionals, as reported in the Metadata Schemes section.

In contrast, scientists and information professionals reported using personal metadata schemes. A personal metadata scheme is metadata application behavior based off of personal preference or long term habit. Some examples of comments and descriptions of personal metadata used are included below:

[21] Dublin Core: http://dublincore.org/documents/dces/

[22] Ecological Metadata Language: http://knb.ecoinformatics.org/software/eml/

- There are 5 elements that are necessary: uniqueID, title, creator, date, contact information, abstract (optional), keywords (optional) (infoprof007)

- Title: Morphological and geographical data for bird species (data set 1).
  These data support the results and conclusions of Symonds, M.R.E. and Tattersall, G.J. (2010) Geographical variation in bill size across bird species provides evidence for Allen's rule. American Naturalist 176: 188-197. doi:10.1086/65366
  Methodological details available in the publication cited above.

  References used to compile the data set are listed below. Numbers in parentheses match numbers found in the "References" column of the data set:

  (1) Agreda, A. and D. J. Anderson. 2003. Evolution of single-chick broods in the swallow-tailed gull Creagrus furcatus. Ibis 145:E53-E58.

  (2) Barrett, G., A. Silcocks, S. Barry, R. Cunningham and R. Poulter. 2003. The New Atlas of Australian Birds. Royal Australasian Ornithologists Union, Hawthorn East.

  (3) Dunning J. B., Jr. 2007. CRC Handbook of Avian Body Masses. 2nd edition. CRC Press, Boca Raton. (infoprof008)
- Created separate records for each data set and included the title, author, publication information, used publication type of "Other", and selected the institutional affiliation of the author(s). (infoprof011)

Each bullet point represents a different response to this topic. These variations, all provided by information professionals, have many similarities to the Dublin Core Metadata Initiatives Core Elements. These responses may show that information

professionals' descriptive metadata practices rely on creating simple personal schemes that pull out title, author/creator/ and other publication information. While information professionals typically used more simple schemes, scientist participants (five out of 15) created some type of "read-me" file. One example of a more complex personal metadata schemes was used by a scientist and is included below as an example:

Figure 14: Data Set Metadata Example

Data Set 1:

Family
Genus
Species
Body mass (g)
Bill length (mm)
Tarsus length (mm)
Middle Toe length (mm)
Tarsometatarsus+middle toe length (mm)
Midpoint latitude (degrees from equator)
Midpoint altitude (m)
Minimum temperature at midpoint of range (degrees Celsius)
References

Data Set 1 from:

Symonds MRE, Tattersall GJ (2010) Geographical variation in bill size across bird species provides evidence for Allen‚ÄôÄôs rule. American Naturalist 176: 188-197. doi:10.1086/653666

Data Set 2:

Species
Hunted?
2006 IUCN threat (criterion A)
Adult Body Mass (g)
Age at First Birth (days)
Age at Sexual Maturity (days)
Gestation Length (days)
Weaning Age (days)
Inter-Birth interval (days)
Group Size Area (km2)
Geographic Range
Home Range (km2)

Maximum Longevity (months)
Population Density (km2)
Mean Human Population Density (km2)
2003 GNI (US $ millions)
Mating System
Diet
Habitat Breadth

Data Set 2 from:

Price SA, Gittleman JL (2007) Hunting to extinction: biology and regional economy influence extinction risk and the impact of hunting in artiodactyls. Proceedings of the Royal Society B 274: 1845-1851. doi:10.1098/rspb.2007.0505

This example highlights that the metadata created by scientists seems to have a different focus from that created by the information professional. The system is more complex and granular that focuses specifically at data set content. As opposed to looking at "aboutness", like the information professionals' schemes, this scientist's scheme is focused on the details.

Another scientist participant was particularly interested in recording and preserving the provenance information of the data set. The participant's exact phrasing was "I created a field in my database to note where each record came from [because…] I assume that these will not be the only two datasets I'll need to incorporate for an overall analysis of this type of data. So I need to keep track of where the different data sets come from." (mrcscientist013).

Only one participant, a scientist, specifically stated that there was no need to create metadata. That scientist pointed out that the description could easily be obtained from a Methods section of a manuscript (mrcscientist015). This participant's response is very similar to the one scientist from White's earlier ethnographic study (2010). In that study, a single scientist participant did not create metadata for scientific data sets or

collections. All descriptive information was kept in previous publications. Participants for these two studies did not overlap.

8a2. Dryad Task Scenario

As reported earlier in Section 7c2: Descriptive Metadata, both groups of participants were more likely to create metadata for author and title. In contrast, other metadata elements like "description" were used more by information professionals than scientists. Out of all the metadata elements used, description was the one least applied by both groups. Author and title seem to be important to both user communities, while description (what a data set is about) may not be. These findings indicate that, it may be important for those who design metadata schemes to consider using descriptive metadata that is easily applied by all communities. Considering the PIM-based information organization behaviors of each group and mimicking their PIM behaviors in the metadata schemes that are being used could improve the metadata being created.

In terms of descriptive metadata and information organization, one point that was made clear in the narrative responses was that there is a difference between what would be done in personal collections and what is done in formalized collections like Dryad. For example, one participant stated," for me personally, there is a difference between what I put in my 'personal collection' and what I'd put into Dryad for permanent archiving" (infoprof005). The same participant emphasized,

> "Note, I did this task from the perspective of me integrating the datasets into my personal workflow for preserving them for MYSELF rather than for anyone else. If I were making them openly available and/or getting them ready to submit to Dryad I would do different things (save them as csv rather than Excel, put the README in a separate file, etc.)"

This sentiment was seen in the actions of scientists as well. For the scientist researcher, the division of what was created for his/her own system and what was created in a more formalize setting like Dryad. This was seen in the increase of use of descriptive metadata when depositing data sets into Dryad as presented earlier in Section 7 Results. It can also be seen in the discussion of when metadata should be used as included previously with the scientist who did not create metadata.

The PIM literature emphasizes this division between what is meant to be shared and what is not (Lansdale, 1988; Jones and Teevan, 2007). Data sets fall in an odd place in this division. While originally meant for personal or group work, these data sets are eventually shared in the large community through repositories. The PIM vs. Sharing tension was present for both information professionals and scientist participants.

Another reason for this difference in metadata creation between the PIM-influenced task scenario in comparison to the Dryad-based simulation could stem from different domain emphasis on information organization and data management training. Information professionals have opportunities to participate in organization courses offered or required in library and information science programs or on the job. This is supported by eight out of the eleven (73%) information professional participants having some type of data management training. In contrast, seven out of fifteen (46%) scientists indicated having some type of data management training. Five of those scientist participants indicated the training they had received was informal. A study by Qin and D'Ignazio (2010a, 2010b) reports that it was difficult to get science majors to enroll in courses about managing scientific data. This shows that perhaps the sciences are not emphasizing data management training early enough in the educational process. While

more research is needed in this area, repositories should take this into consideration when designing metadata schemes to assist in data management. Data management and metadata creation are often outlined in guidelines. The use of guidelines was also examined by this study to give context to the descriptive metadata that was collected.

8b3. Guidelines, rules, and standards

Reported results indicate that both information professionals and scientists are influenced and guided by certain universally standardized policies when performing information organization techniques. The narrative responses support this data and assist in categorizing the guidelines being applied. Guidelines can be formal standards created by large institutional or academic associations; a lab policy or procedure manual; or even personal rules created for individual use. Each type of guideline helps form information organization consistency in practice, but has a different intended audience.

Codes Related to Guidelines

As reported earlier, four information professionals and one scientist reported specific guidelines that influenced the metadata creation process. The use of guidelines was examined in more detail by inputting the narrative/short answer responses into Atlas.ti version 6.1. Four codes (standard guidelines, local guidelines, personal guidelines, and no guidelines) were created to indicate the types of guidelines that were most central to this study. Definitions for these codes can be found in the Appendix. Standard guidelines are a software based, national, or international guideline as a guiding principle for the information organization process. Examples include Excel

documentation, Taxonomic Rules, or Anglo American Cataloging Rules (AACR2). Below are a few comments relating to guideline use.

- with the content rules being AACR2ish if created by a cataloger (i.e. name forms, etc., would follow AACR2, although certainly not all AACR2 data would exist in simple DC). (An author/user might have done an initial deposit into the repository with a DC-based webform and following no content standard, which would then be upgraded by a cataloger to AACR2ish by staff.) (mrcinfoprof009)

- "Guide for file formats is based on UK Data Archive" (mrcinfoprof010)

These two examples show two different approaches to guidelines for metadata creation about data sets. The first example shows how traditional standards, like those criticized by Salo (2010) are being adapted by information professionals for work in the digital environment. The second example shows how national groups are creating standards that go on to inform practice.

Local guidelines are institution, lab, or library-based rules that are established to support information organization tasks performed for a select group of people. In order to protect the anonymity of the participants, three examples of local guidelines could not be included because they linked to institution-based urls. Below is an example of a participant's institutional based guideline that does not link to a specific url:

> "Purpose: provide user with enough information to determine the usefulness of the data set.
>
> Should start with a topic sentence, describing what information is in the data set. Good to include parameters, location, temporal coverage info in first few sentences so users can get at-a-glance idea of what this is.
>
> If possible, should contain a maximum of approximately 15 lines of information that may or may not be found in other fields.
>
> Acronyms should be expanded to provide understanding, but keep length in mind." (infoprof006)

Local standards are only temporary measures or made to fill in the gaps where larger standards leave off. When local standards are not available, personal preference or long term information organization habits (based on PIM practices) are used.

Personal guidelines were also mentioned as being used. Personal guidelines are based on personal preference or long term habit. An example of a personal guideline is included below.

> "Knowledge that each record should have a unique identifier (it was not initially assumed that each taxon was unique)
> Changed some of the field names to not include punctuation or characters that might impede calculations on those fields.
> Was not as strict in this pass in field name designation (would normally use camel case and/or underscores between words)" (mrcscientist010)

Personal guidelines become an important part of the metadata process. As mentioned previously, there is a distinction that is made between the "personal" and those meant for a larger audience.

The use of no guidelines was also mentioned in the narratives. Out of the eight instances where participants commented on not using guidelines, only one gave a reason. The reason stated was "Used a narrative format and did not follow any rules (did not know of any in particular)" (mrcscientist010). It is likely that participants did not use standards because there are currently only a few available that specifically address the needs of scientific data sets (Salo, 2010). Communities and organizations, like the Data Observation Network for Earth (DataONE)[23] or the Association of Research Libraries (ARL)[24] have only just begun drafting or revising such standards. In the meantime, individual information professionals and scientists involved in the long term organization

[23] DataONE: https://www.dataone.org/

[24] ARL escience: http://www.arl.org/rtl/eresearch/escien/nsf/

and maintenance of data sets must patch together a group of standards based on previous cataloging and archival standards, personal preference, and community established "best practice" to try to fill the gap until new data set-centric guidelines can be put into place.

8b4. Software influencing information organization

As stated by mrcscientist009, "[m]y guidelines are 'provided' by limitations to MS Access Databases". In essence, the software acts as a "guideline" or "rule" in determine the type of information organization behaviors that occur. A total of 14 scientist participants reported using a type of software program to organize the data sets. Scientist participants appear to be highly influenced in their information organization habits by the chosen software they use for performing scientific work.

As mentioned previously in the Methods Section, the data sets were given to the participant in the form of Excel files. Excel had been the chosen format of the original authors of the data sets and the format that had been deposited into the Dryad system. When working with the data set, scientists' comments about software, file format, and organization. Below are three examples presented in bullet list form:

- "I imported them into R and rearranged the data […] I brought them into a readable format and rescalled the data for analysis.(mrcscientist003)
- "imported them into JMP (a SAS product) to facilitate data examination without this sort of rearrangement" (mrcscientist005)
- "I exported a tab delimited text file from Excel and tried to import it into the Nexus format of Mesquite" (mrcscientist007)

Software formatting and the guidelines underlying them are central to the scientists' information organization process. Below is an example of a narrative that shows how important software can be the scientific process:

> "My first inclination with this data set is to put it into a phylogenetic framework. Of course, I don't know if a phylogeny has been hypothesized for these taxa, but frankly I would not find it appropriate to examine these data outside of a phylogenetic context, as there are certain to be evolutionary/historic influences on these traits. I generally use the phylogenetic software Mesquite (www.mesquiteproject.org) to store and explore characters on phylogenetic trees, so I wanted to get the taxonomic data and associated character states into Mesquite (which uses a Nexus type of data matrix format, .nex, file included). I started by trying to get Excel to export the data in a format that Mesquite would import, but I'd never done this before. I thought perhaps a tab or comma separated list would work, and it did import the data into Mesquite, but the genera and species were not linked. As a result of this, Mesquite returned an error due to duplicate taxon names. After about a half an hour of trying to export different formats and/or repair things on the Mesquite end I was still unsuccessful. Were this a real data set that I needed to explore, I would likely write a short script in Python to concatenate the genus and species names in the tab delimited text file that I exported from Excel, and hopefully that would allow each taxon to have a unique name in Mesquite. I have included the files I exported from Excel (data set 1.csv, .txt). I would also need to find/import phylogenetic hypotheses for these taxa into Mesquite." (scientist007)

Software guidelines and considerations, while prevalent in the scientist group, where not confined to scientists. Software considerations were also mentioned by information professionals. Below are three examples listed in bullet list form:

- "This was based on the requirements of the repository software environment" (mrcinfoprof002)
- "though the Morpho wizard prompts for useful information and is nicely integrated with the NBII Biocomplexity Thesaurus, so there is some guidance within the tool" (mrcinfoprof003)
- "followed procedures […] according to the Microsoft Excel documentation" (mrcinfoprof011)

These results indicate the software employed during the organization process actually influences the resulting organizing processes and formats. Software seems to have an unexpected influence on everyday practice and ultimately impact the way scientific data is collected, organized, used, and later reused by others. Unfortunately, there is no single software cited as central to all scientific work.

In both communities, there was a persistent and almost amusing dichotomy between an absolute insistence that that the data set is saved in a neutral file format, but a heavy reliance on specific proprietary software for processing those data sets. The impact of how software affects scientific work is not addressed fully in this dissertation, but is an area that could benefit from further study.

8b. Subject Term Application

Subject term application was observed in two situations: a PIM-influenced task scenario and the Dryad-based task scenario. The sections that follow analyze and discuss findings from these task scenarios in more depth.

8b1.PIM Influenced Task Scenario

To analyze the PIM-influenced task scenario, two codes were used to highlight subject term application in the narrative/short answers. The two codes used were: Standardized subject terms and personal subject terms/folksonomy. The code "no subject terms" was not used because none of the narrative/short answers mentioned not using subject terms. The codes are used in this study to pull out relevant explanations that support the descriptive statistics reported in the Results section.

Standardized subject terms are terms or tags that originated in a national or international controlled vocabulary that was used during the information organization process. Examples are Library of Congress Subject Headings (LCSH)[25] or Medical

[25] LCSH: http://www.loc.gov/aba/cataloging/subject/

Subject Headings (MeSH)[26]. Participants listed the various standardized controlled vocabularies that were used to create subject terms. Examples include comments such as "[n]ormally we use MeSH headings for cataloging the resources" (infoprof002) or "[a]dd keywords using NBII Biocomplexity Thesaurus (could've borrowed some MeSH terms from PubMed record for the paper as well)"(infoprof003). The use of standardized subject terms was only mentioned by information professionals.

Both information professionals and scientist participants mentioned using personal subject terms. Personal subject terms are those subject terms or tags that were created based on a personal preference or long term habit. Examples of narrative responses that discussed personal subject terms are included below.

- maybe a term or two not taken from a controlled vocabulary, if they seem like obvious choices. (mrcinfoprof003)
- own system (mrcscientist002)
- authors' abstract. In actual practice, would review publication and, in consultation with author(s), identify additional keywords, other descriptors (taxonomic, geographic, etc.).

Personal subject terms were something that participants from both communities saw the value of using. When not even promoted by the task scenario, the use of terms that were generated from free text was an added component to every day processes.

[26] MeSH: http://www.nlm.nih.gov/mesh/meshhome.html

8b2. Dryad Task Scenario

To review previously reported results, in the Dryad-based task scenario, subject terms were applied by six information professionals and ten scientists for data set 1 plus nine information professionals and eight scientists for data set 2. These results show that information professionals are more likely to create subject terms in their everyday work with data sets than scientists. It also shows that when prompted, scientists will create keywords.

8b3. Similarities and Differences

Based on the results presented by the subject terms in Dryad, some conclusions can be made about the similarities and differences that occur when information professionals and scientists are prompted to apply subject terms to scientific data sets. As a group, information professionals applied fewer terms (27 compared to 30) and had a greater agreement in term application than the scientist group. Information professionals were also more likely to use topical terms when depositing data into Dryad. These consistencies in this population could have resulted from a smaller group of participants or a reliance on standardized information organization training is frequently required to receive a Library and Information Science degree. Information professionals also used more controlled vocabulary systems in their everyday work. According to inter-indexing consistency studies that have been performed since the 1960s, using an indexing aid like a controlled vocabulary increases the consistency among indexers applying subject terms (Hooper, 1965; Reich & Biever, 1991; Sievert & Andrews, 1991; Medelyan & Witten 2006, 2008; Hughes & Rafferty, 2011).

This dissertation did not apply any of the standards of inter-indexing consistency measures to these results, such as those created by Hooper (1965), Rolling (1981), or Medelyan (2009), because of criticisms that have been brought up against the measures themselves. Researchers, such as Rolling (1981), point out that consistency should not be conflated with quality. The argument remains though that some type of measure should be created for future work in repository metadata creation and subject term application. These measures may contribute to a better sense of who should be performing curation duties in the repository. Comparing consistency with some of the techniques mentioned in the "metadata creation" literature reviewed in Section 4, could elucidate domain and repository specific results with practical (if not statistical) significance.

8b4.Controlled Vocabularies for Scientific Data Sets

As reported in Section 7c3e: Vocabulary Mappings, taking the spatial, temporal, topical, and scientific terms that were applied by each participant, mappings were created between those participant-created terms and already established controlled vocabularies. While overall results of the mappings for all four subject term areas are presented in Section 8: Results, the following section will put forward suggestions of what type of controlled vocabularies may best represent scientific data sets.

Before this discussion progresses, it should be acknowledged that the use of a controlled vocabulary in an information system is a topic of debate within the library and information science community. Information retrieval research has shown that use of controlled vocabularies does not improve precision and recall results (Cleverdon, 1970, 1984; Fidel, 1992; Rowley, 1994). In contrast, indexing researchers have criticized lab-based controlled vocabulary research (Svenonius, 1986) and shown that controlled

vocabulary usage can improve indexing consistency (Hooper, 1965; Leonard, 1977; Reich & Biever, 1991; Sievert & Andrews, 1991). While both sides of this debate bring up valid and research-based points, the use of controlled vocabularies to underlie information systems is still a persistent practice at institutions like the Library of Congress and other academically-focused libraries. For that reason the following discussion presents reasons why certain vocabularies were found to represent participant-created terms better than others. Discussion will be limited to the areas of topical and scientific terms. All discussion should be taken within the context of this study.

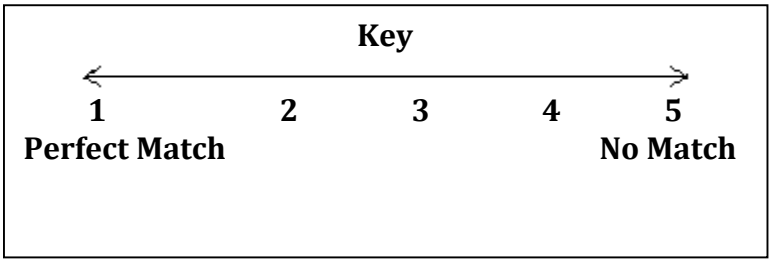
As presented in detail in the Appendix, topical terms were mapped to three controlled vocabularies: Library of Congress Subject Headings (LCSH), Medical Subject Headings (MeSH), and the National Biological Information Infrastructure's Biocomplexity Thesaurus (NBII)[27]. Scientific terms were mapped to four vocabularies: Library of Congress Subject Headings (LCSH); Medical Subject Headings (MeSH), the National Biological Information Infrastructure's Biocomplexity Thesaurus (NBII), and the Integrated Taxonomic Information System (ITIS)[28]. Both mappings were created using a 5 point scale: 1 being a perfect match and 5 being no match. An overview of each vocabulary and the coding scale used during the analysis are included in the Appendix. Details about how the mappings were conducted are included in the Methods section.

[27] NBII: http://www.nbii.gov/portal/server.pt/community/biocomplexity_thesaurus/578

[28] ITIS: http://www.itis.gov/

The codes for each term mapping were averaged within a given controlled vocabulary. This resulted in an average score being applied for each vocabulary. These scores ranged from 1 to 5. In this context scores of 1 were considered better because they indicate a larger portion of perfect matches. Table 12 shows a sub-set of the averages presented earlier in the Results section.

Table 12. Topical and scientific subset from Table 11



	Coder 1		Coder 2	
	Info Prof	Scientist	Info Prof	Scientist
Topical terms				
LCSH	3.2	3.4	3.6	2.9
MeSH	3.9	3.9	4.3	3.4
NBII	3.3	3.5	4.3	3.4
Scientific terms				
LCSH	3.8	3.9	3.7	3.8
MeSH	4.8	4.9	4.9	5.0

NBII	5.0	5.0	5.0	5.0
ITIS	1.2	1.1	1.3	1.3

Examining both coders' data, the average number calculated for each vocabulary was similar between the two participant groups. For topical terms, Library of Congress had the strongest mapping with scores of 3.2, 3.4, 3.6 and 2.9. Yet, a score of 3, according the scale found in the Appendix, only indicates a partial match. Library of Congress performed the best of the three vocabularies searched but by no means provided a perfect match for many of the terms applied by participants.

Looking at scientific terms as displayed in Table 12 above, ITIS had the strongest mapping with scores of 1.2, 1.1, 1.3, and 1.3. ITIS scores were all between 1 and 2, thus indicating that this is a strong choice for describing taxonomic names in a repository system. In contrast, the popularly used LCSH scored between 3 and 4 with scores of 3.8, 3.9, 3.7, and 3.8.

Library of Congress Subject Headings (LCSH) is a well-documented controlled vocabulary that has multiple guidelines created just to support the application of its terms. Mapping results from this study included some perfect matches and no matches, but consisted mainly of a variation of partial matches. The reason for this lies in the guidelines that support it. The first example comes from the Library of Congress Rules Interpretation, Appendix A: Section A.25:CSB49 Scientific names of plants and animals,

> "When two names are given enclosed within parentheses and separated by a colon, generally, capitalize both names since this method of presentation usually indicates names of divisions that are capitalized. (In transcription, retain the colon but without a space on either side.)"

The second example comes from the Subject Cataloging Manual, Section H 1332:

> "a. Latin or common name. Prefer the common name if it is in popular use and unambiguous, using as reference sources Web. 3, other general dictionaries or encyclopedias, recent textbooks, popular field guides, and lists of official common names issued by societies or government agencies. Prefer the common name for animals and plants of economic importance, such as pests or cultivated plants. Prefer Latin when the common name represents several levels (species, genus, family) or the term is not in general lay usage. In general, for organisms occurring only in foreign countries, prefer the Latin name unless an English common name is found in standard reference sources. However, a local common name may be used if it does not conflict with a common name from the United States. Do not begin a heading for the name of a plant or animal with the word common, unless the name appears in that form in Web. 3 or some other authoritative source"

LCSH rules are not created to represent scientific (binomial nomenclature) names. The creation and application of headings using LCSH emphasize the common names. One information professional participant (following the guidelines) did use the common name in the scientific name metadata field of Dryad.

According to PIM-influenced results, four information professional participants used controlled vocabularies, such as the Medical Subject Headings (MeSH), Library of Congress Subject Headings (LCSH), National Biological Information Infrastructure Biocomplexity Thesaurus (NBII), and Global Change Master Directory Keyword list (GCMD). Scientists did not use controlled vocabularies, but taxonomic lists. None of the participants noted using ITIS, a controlled vocabulary of taxonomic names. These results combined with the averages related to the coding results indicate that information professionals are not necessarily using the best vocabularies possible for things like taxonomic names.

The differences between how LCSH and ITIS mappings may relate to the difference in vocabulary type. The Library of Congress vocabulary is a general vocabulary, while ITIS is a specific vocabulary. Lancaster (1986) discusses the differences between these two types in relation to the information retrieval concepts of precision and recall. General vocabularies help with general search and minimize indexing inaccuracies, therefore improving recall (Lancaster, 1986). Specific vocabularies allow for an information object to be put into many small classes, therefore resulting in higher precision. It is recommended that repository developers should look more closely at the terms applied by their users and choose vocabularies that appropriately match user needs. They may also want to consider who is creating the metadata in order to ensure the appropriate granularity that would be required during metadata creation.

8c. Who Should Create Metadata

Reviewing both the descriptive metadata and subject term application results, the question of "who should create metadata?" was brought up at many points. This question seems to an important one and should be addressed during the repository design phase. Metadata schemes, controlled vocabularies, and deposition interfaces are all affected during this process. There are many approaches that have been taken to answer the question of who should create metadata.

8c1. The Larger Debate in Metadata Creation

Many suggestions have been put forward as to "who" is the best at creating metadata. These suggestions include professional created, author created and automatically generated options.

Researchers that promote the professionally created metadata cite the need for quality and consistency, but acknowledge that it is costly in both time and money (Currier & Barton, 2003; Bruce & Hillman, 2004). Researchers investigating the area of author-created metadata study examine whether authors, who know the material better than others, create better descriptions of their own pieces (Greenberg et al., 2006). Automatic indexing research has suggested that various statistical or rule-based methods could be used to assist or replace the manual indexing process (Sparck Jones, 1972; Salton,1975; Vleduts-Stokolov, 1982; Anderson & Perez-Carballo, 2001; Hlava, 2005; Coyle, 2008; Medelyan & Witten, 2008; Medelyan, 2009). The research area of clustering, in particular, could elucidate some helpful results in terms of automatic indexing options (Mostafa, Quiroga, & Palakal, 1998).

The research reported in this dissertation did not set out to answer the question about who should create metadata, but results found here indicate that this question is essential when considering repository workflow. Approaches in the repository community have varied from author deposition with little curation to hybrid author/curator metadata creation to full curator metadata creation. Within the context of this study it was found that information professionals had more experience using metadata, while scientists were better at applying scientific terms. There is a distinction in the roles of "scientist as depositor" and "information professional as curator". It is

recommended that this distinction be taken into consideration during repository creation and design.

8c2. The Role of the Information Professional in Data Curation

Scientific data management is an area of growth for many information institutions. Earlier, the descriptive statistics reported that eight information professionals have been in their current positions for less than three years. This indicates the changing nature of library and information science work. Professionals are still learning how to work with this new type of information object. In some cases new positions are being created to handle this. In other situations, this responsibility is being given to newly hired information professionals and even researchers.

While participating in this study, respondents had to reflect on their role as current (or potential) data curators. Information professionals had to determine what needed to be done to the data set and how it would be described in order to be used at a later time. After completing these tasks, information professional participants mentioned having to choose between multiple roles before being able to complete the tasks at hand. Examples of this are:

> "For this simulation, I considered two different roles and approaches to working with the data. In my role as a [participant title], I would attempt to integrate the data into my local collection primarily for secondary analysis. This is the role adopted above. I would reformat the data for import into a statistical package for analysis.
>
> Another role not adopted here is that of programmer. In this case, I would not use the data set for secondary analysis, but might use it as a sample data set in a larger collection of test data. For example, I am currently testing approaches to automatic indexing. In this scenario, I would use the original data files with no modifications. Instead of a readme, I would likely maintain a separate spreadsheet/table describing each data set included in the collection and its origins." (infoprof001)

The term "cataloger's judgment" is frequently used in library literature and practice to describe the analytical process that information professionals go through in order to appropriately create descriptive metadata and apply subject terms. In some cases, this judgment means considering future use and almost "role-playing" the user. An example of this was found in infoprof008's narrative, where an important part of this was, "Placing myself in shoes of researcher trying to use these data" (infoprof008). The role of the cataloger and the needs of the users are primary concern during this analytical process (Lancaster, 1986).

The exercise allowed information professionals to reflect about their library's own practices and how it effects cataloging behavior and policy. Below is an example of this reflection by another information professional.

> "We have no good methodology for linking the dataset to the original article. If we hold the full text, they would exist as two separate citations in the repository with possibly a full text note referring to each other. I would grumble about this and then get over it. and then complain to our repository committee. And we would discuss and then re-discuss again the downsides of DSpace. And then we'd probably not move forward with any changes/local development unless/until we had so much need to make a cost/benefit/ROI use case. Ideally, we'd want the dataset to stand separately from the article, but related to the article in a structured way." (infoprof009)

This is a detailed example of the process information professionals undergo when depositing data into their already established collections. Different points of interest include examples of cataloger's judgment and the analysis process the effects descriptive metadata creation and subject term application.

One information professional discussed a process that involved using a hybrid approach, combining standardized metadata schemes, controlled vocabulary subject terms, and personal, uncontrolled vocabulary terms.

> "Make working copies of the data sets and move originals to 'safe' folder. Looks like EML would be appropriate (sic) for these data sets so I make EML records.Look up citing pubs to get more info as needed for metadata. In real life, depending on the data set, I'd probably add a lot more detail about the data set to the metadata record (taxonomic info, geographic coverage, etc.). Add keywords using NBII Biocomplexity Thesaurus (could've borrowed some MeSH terms from PubMed record for the paper as well), and maybe a term or two not taken from a controlled vocabulary, if they seem like obvious choices. […] Note that we consider these best practices, we could integrate data into our repository without making any of these changes. " (infoprof003)

Personal metadata and information organization of scientists is discussed in more detail by White (2010a), where it was found that scientists labeled certain type of information organization behavior as either "good science" or "bad science" based on personal preference. This same behavior of labeling information organization behaviors and applying a value judgment was found to be persistent in the information organization participants of this study. Instead of using the term "good science", information professionals used the term "best practice" to describe this desired behavior.

Findings from this study indicate that information professionals see their roles as changing. Data sets and other "born digital" materials present a new set of challenges. Participants reflected on these challenges and often wonder about their current roles, as well as what their roles will be in the future. Participant infoprof002 commented that,

> "My duties relate to managing digitized resources (books and films, to date) in a repository which provides a set of "digital library" access services. I have no experience with curation and ingestion of research data or other types of "born digital" resources at this time, though we do anticipate putting such content into our repository in the future."

The "future" role of the information professional as data and digital curator causes questions to be raised in the ILS community.

This study and the results discussed above help to further the distinction between "information professional as curator" and "scientist as depositor". It suggests that each group has a distinct approach to information organization based on domain background, training, and traditions.

8d. Limitations of this study

The discussion presented above is a first step to understanding more about metadata creation and subject term application performed by information professionals and scientists. The aim of this study was to combine qualitative and quantitative findings to elaborate on the current understanding of information organization behaviors regarding scientific data sets. The approaches used in this study succeed to a certain point, but there are also limitations that should be recognized. This section identifies these limitations by giving attention to the research methodology used, sample size, and the challenges presented with generalizing the results.

The concurrent triangulation mixed methods approach excels at simultaneously examining quantitative and qualitative data, yet is not perfectly suited to studying PIM activities. Because of the influence of quasi-experimental data collection and pre-selected data sets, the need for control added in an artificial component that eliminated the ability to report on natural behavior. This lack of naturalistic approaches may have some bearing on the PIM conclusions and recommendations that can be made about either population. Although the method itself has noted limitations, it should also be

realized that every effort was made in this study to encourage naturalistic behavior. This is especially true among scientist participants for the controlled tasks.

The sample size used for this study may also be considered a limitation. Small samples do not always lend themselves well to quantitative analysis. To review, 11 information professionals and 16 scientists were recruited to participate in this study. The aim of this study was to recruit 30 participants, with 15 participants being information professionals and the other 15 participants being scientists. Recruitment was a time consuming activity, and given the consistency among data gathered, the participant numbers were viewed as sufficient for this study. The sample size does present a limited amount of descriptive statistics about organizing behaviors in order to provide insight into the problems being studied. Clearly a larger sample would help confirm the findings and lend more support to the conclusions.

The third noted limitation considered the challenge in generalizing the results. The use of an artificial data collection scenario, the small sample size, and the use of the Dryad system as a control for collecting metadata impact the generalizability of the results. While the use of an artificial data collection scenario and the small sample size are addressed earlier in terms of other limitations, they also create a challenge for generalizing of results. The use of the artificial data collection scenario means that the results cannot be generalized as "real" behaviors. The small sample size means that the two participant groups may not be perfectly representative of their larger populations. More participants would be needed to make sure the conclusions could be generalized to the overall populations. The use of Dryad as a control mechanism for metadata can also be seen as a limitation. Dryad is an example of a repository already in use by the

scientific metadata community. By using only Dryad for the study, the conclusions about metadata and subject term were directly linked to the usage of only that one system. Being aware of this challenge, great attention was given to selecting two sample data sets that could be applicable to multiple disciplines in the biosciences. Additionally, Dryad coverage does include basic and applied biosciences and is multi-disciplinary in this sense. The conclusions and recommendations may have some application to systems beyond Dryad, a point to consider when comparing results of future studies in Dryad or other systems covering data in basic and applied biosciences.

While limitations are inevitable in any study design, it is hoped that they can be recognized and be used to promote better research in the future. By improving upon the limitations of this study, the results of future research may be able to contribute even more to the area of metadata creation and subject term application for scientific repositories. Section 9b., which is presented later, elaborates on some future research plans that could potentially build on the findings presented in this dissertation study.

9. Conclusion

This study took a concurrent triangulation mixed methods approach for studying metadata creation and subject term application describing scientific data sets. The user groups examined during this study were information professionals and scientists: two groups that are embedded in the current repository development community. This approach allowed for both quantitative and qualitative questions to be answered and discussed. The questions guiding this study were:

Research Question: In the context of scientific data sets, what types of distinguishable similarities and differences exist between the ways researchers in the biosciences who use research data and information professionals who curate research data create metadata and apply subject terms?

Descriptive Metadata (about a resource, with exclusion of subject metadata which is covered in questions 3-5).

1. What types of formal/standard metadata are currently being applied by both groups?
2. What types of personal metadata are currently be applied by both groups?

Subject Terms

3. Which controlled vocabularies map best to subject terms created by both groups?
4. What is the extent of overlap in subject term application between the two groups?
5. What is the extent of divergence in subject term application between the two groups?

The following section will draw from the results and discussion sections to review the answers to these questions.

9a. Review of the Findings

The data examined in this study found that similarities and distinguishable differences exist between the information organization habits and approaches of information professionals and scientists. These similarities and differences can be seen in the way descriptive metadata is created and subject terms are applied. Below is a review of the five focused questions which highlight findings presented earlier in this dissertation.

Question 1: What types of formal/standard metadata are currently being applied by both groups?

Findings from the PIM-influenced portion of this study show that information professionals are more likely to use formal/standard metadata in their everyday work than scientists. Information professionals use standards like Dublin Core and the Ecological Metadata Language.

Surrogate creation was not a typical part of the scientist workflow. When prompted during the Dryad scenario, scientist participants did create metadata, but only within the Dublin Core-based application profile metadata form they were given. Software was used a guideline or standard for scientist participants when creating metadata for both shared and personal use. For example, a total of 14 scientist participants reported using a type of software program to organize the data sets. Those 14 participants listed six specific types of software used and two un-named programs. Based on the data collected in this study, software use appeared to be a central part of the scientists' organizing processes.

Question 2: What types of personal metadata are currently be applied by both groups?

Personal metadata schemes were used by both groups. Personal metadata schemes are defined above in Section 8: Descriptive Metadata. Information professionals used personal schemes that reflected their training in formalized metadata standards used in libraries. When comparing scientists' systems with information professionals', the scientists' systems were more specific and focused less on "aboutness". The PIM sense of "for my use" in relation to "to share with others" was also present during this study.

Question 3: Which controlled vocabularies map best to subject terms created by both groups?

Based on the four sub-types of subject terms examined by this study (spatial, temporal, topical, and scientific), conclusions can be made that Library of Congress Subject Headings (LCSH) had the best subject term coverage for topical terms. For scientific terms, the Integrated Taxanomic Information System represented scientific names very well and had the highest average score of all the vocabularies. Determining a strong vocabulary choice for spatial and temporal terms was not possible during this study because those types of terms were used less frequently (if at all) by participants.

Question 4: What is the extent of overlap in subject term application between the two groups?

Information professionals and scientists did show an overlap in subject term application. There were 12 terms that both scientists and information professionals applied for data set 1. In terms of percentages that would be 27% of the terms used

for data set 1 were applied by both groups. Another 12 terms were applied by both scientists and information professionals. This means that 24% of the terms used for data set 2 were applied by both groups. Compared to foundational inter-indexer consistency studies (Hooper, 1965; Leonard, 1977), these numbers are within range for multiple indexers describing the same information objects.

Question 5: What is the extent of divergence in subject term application between the two groups?

Information professionals and scientists also had some differences in the subject terms they applied. As a group, information professionals applied fewer terms than scientists. Information professionals had a total of 15 terms that were different from what scientists applied for data set 1. For data set 2, information professionals applied 11 unique terms that were different from what scientists applied. Scientists applied 18 terms that were different from what information professionals applied for data set 1. For data set 2, scientists applied 26 terms that were different from what information professionals applied.

Summary

The results of this study provide insight into how two different communities create metadata and apply subject terms. Based on the findings from this study, it is recommended that repository designers examine information organization processes of their chosen user group before creating underlying information structures. These findings suggest that considering who will be creating metadata could have an impact on the type

of metadata field choices as well as controlled vocabularies used. Deciding which metadata approach to take during repository development could also impact the types of metadata guidelines that can be created.

This dissertation research was heavily influenced by the perspectives found in the Personal Information Management (PIM) community. Typically, PIM-related studies use naturalistic approaches to examine how people work within their own environments. This study relied more on control by introducing artificial elements into daily workflows. Comments from individual participants showed awareness of the artificial nature of the two preselected datasets. The same participants explained their rationales for how they successfully worked with the dataset for this simulation and how they would have worked in a more "naturalistic" situation. The unique PIM-influenced portion of the study lends support to recommend that personal information management (PIM) practices of scientists be considered during the repository planning phase.

Results from this study, outlined in more detail earlier, indicate that the software packages being used by scientists to create and analyze data sets have an impact on the process of "science" itself. Considering the diversity of software packages represented by even the small sample studied in this dissertation, it is recommended that a repository designed to represent an interdisciplinary domain should take this into account before making metadata decisions.

While researchers, such as Salo, have pointed to problems in information organization in managing scientific data sets very little research has been done on this topic specifically. This study was meant as a step the right direction. Metadata creation and subject term application is only one portion of the data life cycle. More studies will

need to be done in order to fully understand the impact that these two types of information organization have on the larger process.

9b. Future research

This dissertation study addressed a small portion of the issues related to the topic of organization in repositories with a focus on scientific research data and PIM. As stated earlier, the current research on metadata creation and subject term application in repository systems is limited. Information professional practitioners and researchers are working towards best practices, but these best practices are not necessarily based on research. Further study is needed to fully answer the problems found in this area.

This dissertation research has inspired me to think of future research endeavors that would elaborate on findings from this dissertation study. These future studies have the potential to address the limitations of this dissertation study. While still focusing on the area of scientific data and organization, these future studies will take different approaches to examining the chosen populations (information professionals and scientists). In the sections below, three future studies are introduced. These studies are referred to as the scientist-focused study, the one-year study, and the multiple repository study.

9b1. Scientist-Focused Study

The scientist-focused study will use the same study instruments found in this dissertation study, but focus on the scientist community and apply more naturalistic methods. The purpose of this study will be to elaborate on the type of differences that may arise when more naturalistic approaches are used and give a better PIM grounding

for conclusions and recommendations. To conduct this study, scientists will be given a choice of 20 datasets. Scientists will choose two data sets to work with and then undergo the same PIM-like and Dryad scenarios presented in this dissertation. Results from the scientist-focused study will be compared to this dissertation study's results. This comparison will result in stronger conclusions about the PIM of scientists. Recommendations will then be made to the Dryad group about metadata and subject term use.

9b2. One Year Study

The one year study will use the same methods that were used in the dissertation study, but the recruitment period will last for one entire year. The purpose of extending this study will be to recruit more information professional and scientist participants. With a larger participant pool for the one year study, the results will be easier to generalize in terms of population characteristics and behaviors. For the one year study the goal will be to recruit 50 scientists and 50 information professionals. Since the recruitment will last for one full year, factors such as timing of recruitment and annual conflicts will no longer be a restriction to data collection.

9b3. Multiple Repository Study

The multiple repository study will use different methods and approaches than those used in the dissertation study. The goal of the multiple repository study will be to eliminate the bias created when using a single repository system. The purpose of this study will be to understand more about how scientists perceive library repository records. Also, it will aim to get a sense of how scientists believe metadata and subject terms

should be applied to scientific datasets. The procedures that will be used include having 10 scientists donate data sets from their own collections. Then those data sets will be given to 10 different repository librarians. The librarians will be asked to put those data sets into their library repositories. The next step will be to have the scientists review the metadata records created by the librarians. The scientists will give feedback as to which single metadata record best represented the dataset. The scientists will then be asked which parts of that record made it the best representation. The researcher will then compare the top metadata record choices in terms of metadata element usage, number of terms applied, types of metadata schemes used, and types of controlled vocabularies used.

The future research studies presented here are meant to elaborate on the findings presented in this dissertation study. The area of digital data management is already a topic of concern for both information professionals and scientists. Studying metadata creation and subject term application behaviors allows for repository design decisions to be based on research and not just speculation. The goal of this dissertation research and the future studies presented here is to work towards better designed repositories for scientific data sets.

APPENDICES

Appendix A: Instruction Sheet

Thank you for agreeing to participate in this research study that examines organizing output in relation to scientific data. There are three components to this study that you will need to complete in order to receive your participation incentive. All study components should be finished by <date>.
Depending on your personal organization these tasks are estimated to take a few minutes to hour and a half to complete. As an incentive for participation each participant will receive an Amazon gift card once study the study is complete.
The components of this study are outlined as follows:

Step 1: Demographic Questionnaire
The Demographic Questionnaire inquires about basic information in order to obtain contextual information from all participants of the study. These questions are meant to be less personal, yet descriptive in nature. While some parts of this information may be used in future publications, only group characteristics and statistics will be revealed. Please complete this questionnaire first before moving onto Step 2.

Step 2: Simulate Integrating Data into Collections
Please treat this portion of the study as you would any other part of your normal work day.
You have been sent two data sets and a task scenario about what you are suppose to do with these data sets. Each data set should be integrated individually. The purpose of this portion of the study is to find out how different groups organize the same data set. Integration of this data set into your collection does not have to be real. It is a simulation. For the purpose of this study, the term 'organize' is used to indicate the changes, additions, or deletions that may occur when trying to integrate a dataset into your chosen collection. Examples of "organizing" can include: changes made to the arrangement of actual data in the file, changing the data set's file format, creating a read me file to explain how the use the data set, creating a read-me file or record to describe the data set, and adding any keywords about the data set.
Please save items that are created from this section of the study because you will need to use them later in Step 5. Things that you should save include:

--the data set after it has been integrated into your collection
--any accompanying records you may have created about the data set (including surrogates, readme files, etc.)
--any accompanying metadata you may have created about the data set (including keywords or tags)

Please complete this simulation after answering the Preliminary Questionnaire and before taking the Follow Up Questionnaire
Step 3: Follow Up Questionnaire
The Follow Up Questionnaire asks you to reflect on many of the actions that you performed in step 2. Please answer each question as specifically as possible.

Step 4: Email Questionnaires back to Hollie White
After completing Step 3, please email both completed questionnaires (the Demographic Questionnaire and the Follow Up Questionnaire) to Hollie White at hcwhite1@email.unc.edu.

Step 5: Depositing Data Sets into Dryad.
Below are directions for using this Dryad Digital Repository Instance[29].

1. Go to http://mrc.datadryad.org/
2. Press the Submit Data Now button
3. Use the following unique email address[30] and password to log in:
    - Email address: mrcinfoprof005@yahoo.com
    - Password: mrcstudy1
4. In the Journal box, select Other Journal and leave the Manuscript Number blank. Be sure to check the waiver box and then press the "Next" button.
5. Now that you are logged into the system, deposit both data sets into Dryad. Each data set should be deposited individually.
6. As prompted by the Dryad system, create metadata and apply subject terms describing the data sets. Use the information provided in the task scenario to fill in information about the journal article. For Step 5 you are describing the original datasets (individually) and the article citation information included in the task scenario should help you finish adding details.

Please contact Hollie at hcwhite1@email.unc.edu if you have any questions about using the Dryad instance.

Completing the study
Please finish all 5 steps of this study by <date>. Only once all items have been verified as complete will each participant be sent his/her incentive for participation.
The following items are considered completed study components:
--Preliminary Questionnaire (completed and emailed to Hollie White at hcwhite1@email.unc.edu)
--Follow Up Questionnaire (completed and emailed to Hollie White at hcwhite1@email.unc.edu)
--Data Sets entered into Dryad at http://mrc.datadryad.org/
--Any accompanying records/surrogates/readme files (if created, emailed to Hollie White at hcwhite1@email.unc.edu)
--Any accompanying keywords/tags (if created, emailed to Hollie White at hcwhite1@email.unc.edu)

---

[29] Please note that this study uses a special Metadata Research Center instance of Dryad.

[30] Please use this email address and password in order to protect your anonymity during the study.

Appendix B: Demographic Questionnaire (for by both group L and S)

1. What is your professional title?

2. For how long have you held this position?

   ☐ 0-3 years ☐ 4-7 years ☐ 8-11 years ☐ 11 or more years

3. How frequently do you work with research data?

   ☐ daily ☐ a few times a week ☐ every few weeks ☐ every month

4. Have you used data created by another person or organization in your research?

   ☐ Yes ☐ No

5. Describe your educational background:

   - Area of Study:

   - Degrees obtained:

6. Have you had any data management training? If so, please describe this training.

7. Have you ever deposited data anywhere? If so, where?

Appendix C: Task Scenario: Information professionals

Group L Task Scenario:
In your professional position, you have just received the following data sets as new additions to your repository/library/digital collection. Please follow your normal processing procedures in order to simulate integration of these items into your collection. Information about the data set:

The data sets you will be using for this portion of the simulation are from the following publications:

> Symonds MRE, Tattersall GJ (2010) Geographical variation in bill size across bird species provides evidence for Allen's rule. American Naturalist 176: 188-197. doi:10.1086/653666
>
> Price SA, Gittleman JL (2007) Hunting to extinction: biology and regional economy influence extinction risk and the impact of hunting in artiodactyls. Proceedings of the Royal Society B 274: 1845-1851. doi:10.1098/rspb.2007.0505

The citations for the data, formatted in the preferred Dryad citation format, are as follows:

> Symonds MRE, and Tattersall GJ (2010) Data from: Geographical variation in bill size across bird species provides evidence for Allen's rule. Dryad Digital Repository. doi:10.5061/dryad.1421
>
> Price SA, and Gittleman JL (2007) Data from: Hunting to extinction: biology and regional economy influence extinction risk and the impact of hunting in artiodactyls. Dryad Digital Repository. doi:10.5061/dryad.82

Please remember:
For this simulation, please do not change your normal workflow or organizing processes in order to create more deliverables. I am more interested in seeing what you would normally create. While completing this task, please save changes including additions, and deletions made to the data set, as well as, any accompanying descriptive information or records that you may create in support of the data set.

Appendix D: Task Scenario for Scientists

Group S Task Scenario:
As part of a new grant project, you are doing some preliminary work in preparation for a publication. In order to complete this publication, you must integrate the attached data sets into your own data collection (as well as possible) and come up with appropriate research questions.

Information about the data sets:
The data sets you will be using for this portion of the simulation are from the following publications:

> Symonds MRE, Tattersall GJ (2010) Geographical variation in bill size across bird species provides evidence for Allen's rule. American Naturalist 176: 188-197. doi:10.1086/653666
>
> Price SA, Gittleman JL (2007) Hunting to extinction: biology and regional economy influence extinction risk and the impact of hunting in artiodactyls. Proceedings of the Royal Society B 274: 1845-1851. doi:10.1098/rspb.2007.0505

The citations for the data, formatted in the preferred Dryad citation format, are as follows:

> Symonds MRE and Tattersall GJ (2010) Data from: Geographical variation in bill size across bird species provides evidence for Allen's rule. Dryad Digital Repository. doi:10.5061/dryad.1421
>
> Price SA and Gittleman JL (2007) Data from: Hunting to extinction: biology and regional economy influence extinction risk and the impact of hunting in artiodactyls. Dryad Digital Repository. doi:10.5061/dryad.82

Please remember:
For this simulation, please do not change your normal workflow or organizing processes in order to create more deliverables. I am more interested in seeing what you would normally create. While completing this task, please save changes including additions, and deletions made to the data sets, as well as, any accompanying keywords or read-me files that you may create in support of the data set.

Appendix E: Follow Up Questionnaire

Follow Up Questionnaire (used by both group L and S)
When answering the questions below, please reflect on the process you just performed simulating the integration of the data sets into your collection.

1. Did you make any changes to the attached data sets?
   - ☐ Yes (please answer a & b only)
   - ☐ No (please answer c only)

   a. If yes, Describe the changes you made to the data sets?
   b. If yes, Please list any guidelines or rules that instructed you in how to change the data sets?
   c. If no, Please list any guidelines or rules that instructed you not to change the data sets?

2. Did you create a read-me file or record to describe the data sets?
   - ☐ Yes (please answer a & b)
   - ☐ No

   a. If yes, How many read-me files or records did you create?

   b. If yes, Please list any guidelines or rules that instructed you in creating this read-me file or record?

3. Did you create any keywords to describe the data sets?
   - ☐ Yes (please answer the sub-question)
   - ☐ No

If yes, Please list the source(s) these keywords came from?

4. Describe the process the data sets underwent (i.e. the actions you performed) in order for them to be fully integrated and accessible in your collection

Appendix F: Narrative Coding Instructions

There are 27 documents that need to be coded. Documents are a combination of narrative, short answers, and lists. The question prompt is included for each response. Questions are italicized and should not be coded. Only the plan font (no italics) should have coding applied to them.

How to apply coding

Each Word document represents one participant's response. Each response needs to be coded using the codebook provided.

Please use the Word Commenting feature to add a code. Highlight the section of text that the code will apply to, then under the Review tab, select New Comment. Write the appropriate code in the comment box. When finished coding a document resave it.

Please send all newly coded documents back to Hollie once you are done.

Codebook

A total of 22 codes have been used for this study. Codes can be used multiple times in one narrative. Codes applied should come from the codebook provided below.

Code/Tag	**When it should be applied**
Best practice	Participant mentions that a certain approach or behavior is a "best practice", standard, or preferred choice by the field or organization.
Choices	Participant discussed that there are multiple choices that can be made when dealing with data sets. The person will then elaborate on the choice that is or is not chosen. Parent tag for "choice taken" and "choice not taken".
..Choice taken	Participant mentions that a choice was made. This tag applies to the chosen option. Child tag of Choices.
..Choice not taken	Participant mentions that a choice was made. This tag applied to the option that was not chosen. Child tag of Choices.
Data Set Changes	Use to highlight all areas that describe the changes that a data set underwent. Changes include: file format revisions; column revisions/additions/deletions, etc.
Data Set Process	Use to highlight all areas that describe the organizing process that a data set underwent. Not necessarily changes to the data set, but the steps the participant performed to use the data set.
File Format	Participant mentions anything dealing with the file format.
Local Guidelines	Participant mentions using local (institution, lab, or library)- based guidelines to help guide the information organization tasks performed.
Naming	Participant mentions that the name of the file was changed to confirm with certain standards or personal preferred practice.
No data set changes	Participant mentions that no changes were made to the data set.
No guidelines	Participant mentions that guidelines were not used during the information organization process.
No metadata	Participant mentions that no extra metadata was created during the information organization process. This means no read me files, no cataloging record, and no citation.

Personal guidelines	Participant mentions that the guidelines that helped create the information organization process came from personal preference or long term habit.
Personal metadata	Participant mentions that the metadata created was based off of personal preference or long term habit.
Personal subject terms	Participant mentions that the subject terms (or tags) used were based off of a personal preference or long term habit.
Repository	Participant mentions a repository. This can be a specific repository or a vague reference to an institutional repository.
Sense making	Participant discusses how certain tasks help him/her "understand" the data better.
Software	Participant names specific software that was used. Do not include repositories.
Standard guidelines	Participant mentions using a software based, national, or international guideline as a guiding principle for the information organization process. Examples (but not limited to) are Excel documentation, Taxonomic Rules, or Anglo American Cataloging Rules (AACR2).
Standardized metadata	Participant mentions using a national or international metadata or cataloging scheme. Examples (but not limited to) are Dublin Core or Ecological Metadata Language.
Standardized subject terms	Participant mentions a national or international controlled vocabulary that was used during the information organization process. Examples (but not limited to) are Library of Congress Subject Headings (LCSH) or Medical Subject Headings (MeSH).
Time	Participant mentions how long the information organization process takes or that there is not enough time to do what would be a "normal" work process.

Appendix G: Mappings Instructions and Coding Key for Vocabularies

Created by Jane Greenberg
Modified by Hollie White
Using both the HIVE system and verifying in the original online vocabulary home, search each term in the specified vocabularies.

1.Topical
Vocabularies to search LCSH, MeSH, and NBII

2. Spatial
Vocabularies to search LCSH, NBII, and TGN

3. Temporal
Vocabularies to search LCSH, MeSH, NBII, and TGN

4. Scientific
Vocabularies to search LCSH, MeSH, NBII, and ITIS.

Vocabulary and Vocabulary Server addresses:
HIVE: http://hive.nescent.org:9090/home.html
LCSH: http://id.loc.gov/search/
MeSH: http://www.nlm.nih.gov/mesh/MBrowser.html
NBII: http://www.nbii.gov/portal/server.pt/community/biocomplexity_thesaurus/578
ITIS: http://www.itis.gov/

Coding Key

Codes range from 1 to 5 and relate to the degree of match between the term searched and the retrieved term from the vocabulary. The guide/coding scheme is listed below

1: perfect match: preferred vocabulary term

Terms matches exactly in spelling and tense. The term in the vocabulary occurs as the entry term or preferred term.

2: match, alternate (in HIVE) non-preferred vocabulary term

Term has an exact match (in spelling and tense), but in the vocabulary is listed non-preferred or alternative term.

Example: Your search: Bushmeat.

Result: found preferred term is: Wildlife meat, but Bushmeat is an exact term in alternatives.

3: partial match, preferred vocabulary term

Term has a partial match (spelling difference, tense difference) and the vocabulary term is listed as the entry term or preferred term.

Example 1: Your search: "electronic commerce"

Result: "e-commerce" as the entry term/preferred term

Example 2: Your search: "Bird"

Result: "Birds" as the entry term/preferred term

4: partial match, alternate (in HIVE) non-preferred vocabulary term

Term has a partial match (spelling difference, tense difference) and the vocabulary term is listed as the alternative term or non-preferred term.

Example: Your search: "port"

Result: "Port (wine)" as the alternative term/non-preferred term

5: no match

Appendix H: Vocabulary Descriptions

ITIS: Integrated Taxonomic Information System
Taxonomy url: http://www.itis.gov/
Description: ITIS was created to improve the organization of and access to standardized nomenclature. Its purpose is to provide taxonomic data and a directory of taxonomic expertise. It provides a reference database for scientific and common names for species. As of January 2012, ITIS was no longer funded by the government and removed from online

LCSH : Library of Congress Subject Headings
Thesaurus url: http://classificationweb.net/
Description: The Library of Congress Subject Headings (LCSH) are a controlled vocabulary for use in subject cataloging and indexing. It was originally designed for the Library of Congress collection, but many other libraries have adopted the system as well, especially academic libraries.
It covers all subjects generally and is updated weekly.
Information directly from website: http://liswiki.org/wiki/LCSH

MeSH: Medical Subject Headings
Thesaurus url:
http://www.nlm.nih.gov/mesh/meshhome.html
Description: MeSH is the National Library of Medicine's controlled vocabulary thesaurus. It consists of sets of terms naming descriptors in a hierarchical structure that permits searching at various levels of specificity.

MeSH descriptors are arranged in both an alphabetic and a hierarchical structure. At the most general level of the hierarchical structure are very broad headings such as "Anatomy" or "Mental Disorders." More specific headings are found at more narrow levels of the twelve-level hierarchy, such as "Ankle" and "Conduct Disorder." There are 26,142 descriptors in 2011 MeSH. There are also over 177,000 entry terms that assist in finding the most appropriate MeSH Heading, for example, "Vitamin C" is an entry term to "Ascorbic Acid." In addition to these headings, there are more than 199,000 headings called Supplementary Concept Records (formerly Supplementary Chemical Records) within a separate thesaurus.

Information directly from website:
http://www.nlm.nih.gov/pubs/factsheets/mesh.html
NBII: CSA-NBII Biocomplexity Thesaurus
Thesaurus url:
http://www.nbii.gov/portal/community/Communities/Toolkit/Biocomplexity_Thesaurus/

Description: Development of the CSA-NBII Biocomplexity Thesaurus began in 2002-2003 through a partnership between the NBII and CSA, a worldwide information company with more than 30 years experience as a leading bibliographic database provider. The original Biocomplexity Thesaurus, first made available online in 2003, and was a merger of five individual thesauri:

- the CSA Aquatic Sciences and Fisheries Thesaurus
- the CSA Life Sciences Thesaurus
- the CSA Pollution Thesaurus
- the CSA Sociological Thesaurus
- the CERES/NBII Thesaurus

In 2004, the CSA Ecotourism Thesaurus was also merged into the Biocomplexity Thesaurus.

Merging and reconciliation of the terms in these thesauri was performed by Jessica Milstead, a leading expert in the development of scientific thesauri. The NBII Thesaurus Working Group oversees expansion of the thesaurus and addition or modification of terms.

In 2006-2007, the Thesaurus will be expanded to include new terms to support the fire ecology and management communities. Thesauri and glossaries that will be evaluated for this effort include:

- E.V. Komarek Fire Ecology Thesaurus, Tall Timbers Research Station
- Fire Effects Information System (FEIS) Glossary
- Northwest and Alaska Fire Effects Clearinghouse Glossary
- National Wildfire Coordinating Group Glossary of Wildland Fire Terminology
- Encyclopedia of Southern Fire Science
- Wildland Fire Lessons Learned Center Topics

Information directly from website:
http://www.nbii.gov/portal/server.pt?open=512&objID=578&&PageID=1658&mode=2&in_hi_userid=2&cached=true

As of January 2012, NBII was no longer funded by the government and removed from online.

TGN: Getty Thesaurus of Geographic Names Online
Thesaurus url:
http://www.getty.edu/research/conducting_research/vocabularies/tgn/index.html
Description: The TGN is a structured vocabulary containing around 912,000 records, including 1.1 million names, place types, coordinates, and descriptive notes, focusing on places important for the study of art and architecture.
Its scope includes terminology needed to catalog and retrieve information about the visual arts and architecture; it is constructed using national and international standards for

thesaurus construction; it comprises a hierarchy with tree structures corresponding to the current and historical worlds; it is based on terminology that is current, warranted for use by authoritative literary sources, and validated by use in the scholarly art and architectural history community; and it is compiled and edited in response to the needs of the user community.
Information directly from website:
http://www.getty.edu/research/conducting_research/vocabularies/tgn/about.html

REFERENCES

Ackoff, R. (1989). From data to wisdom. *Journal of Applied Systems Analysis, 16,* 3-9.

Anderson, James D. & Perez-Carballo, J. (2005). *Information retrieval design : principles and options for information description, organization, display, and access in information retrieval databases, digital libraries, and indexes*. St. Petersburg, Fla. : Ometeca Institute.

Aristotle. (199x). *Metaphysics.* Internet Classics Archive. http://classics.mit.edu/Aristotle/metaphysics.html. (Original work published1st Century CE)

Babb, N. (2005). Cataloging spirits and the spirit of cataloging. *Cataloging and Classification Quarterly, 40,* 89-122.

Barreau, D. (1995) *Context* as a factor in personal information management systems. *Journal of the American Society for Information Science, 46*, 327-339.

Barreau, D. (2006). Personal information management in context. Presented at the 2006 *ACM SIGIR Workshop on Personal Information Management* in Seattle, Washington.

Barreau, D. (2008). The persistence of behavior and form in the organization of personal information. *Journal of the American Society for Information Science*, *59,* 310-317.

Barreau, D. & Nardi, B. (1995) Finding and reminding: File organization from the desktop. *ACM SigChi Bulletin*. *27,* 39-43.

Bedford, D., Greenberg, J., Hodge, G., White, H., & Hlava, M. (2010) Knowledge organization: Evaluating foundation and function in the information ecosystem. Panel presentation at *ASIS&T '10 Annual Meeting,* Oct. 25, 2010, Pittsburgh, PA.

Big Data. (2008). *Nature, 455,* 1–136.

Bland, R. & Stoffan, M. (2008). Returning classification to the catalog. *Information Technology & Libraries, 27,* 55-60.

Briet, S. (2006) *What is documentation?: English translation of the classic French text.* Lanham, Maryland: Scarecrow Press. (Original work published 1951).

Bruce, R. (2008) Descriptor and folksonomy concurrence in education related scholarly research. *Webology 5*. Available at: http://www.webology.ir/2008/v5n3a59.html
Bruce, T., & Hillman, D. (2004). The continuum of metadata quality: Defining, expressing, exploiting. In D. Hillman & E. Westbrooks (Eds.), *Metadata in practice* (pp. 238–256). Chicago, IL: ALA Editions.

Buckland, M. (1991) Information as thing. *Journal of the American Society for Information Science*, *42*, 351-360.

Busha, C. H., & Harter, S. P. (1980). *Research methods in librarianship: Techniques and interpretation.* New York: Academic Press.

Chamez, K. (2006) *Constructing grounded theory: A practical guide through qualitative analysis.* Thousand Oaks: Sage Publications.

Cleverdon, C. W. (1970). *The effect of variations in relevance assessments in comparative experimental tests of index languages*. Cranfield, UK: Cranfield Institute of Technology. (Cranfield Library Report No. 3)

Cole, I. (1982). Human aspects of office filing: Implications for the electronic office. *Proceedings Human Factors Society,* Seattle, Washington.

Coyle, K. (2008) Machine Indexing. *Journal of Academic Librarianship 34*.

Creswell, J.W., & Clark, V.L.P. (2006) *Designing and conducting mixed methods research*. Thousand Oaks: Sage Publications.

Currier, S. & Barton, J. (2003). *Quality Assurance for Digital Learning Object Repositories: How Should Metadata be Created?* Available at: http://assessment.cetis.ac.uk/content2/20040402013222

Cutter, C.A. (1904). Rules for a dictionary catalog: selections. In M. Carpenter and E. Svenonius (Ed.), *Foundations of Cataloguing* (pp. 62-71). Englewood, CO: Libraries Unlimited, 1985.

Dervin, B. (1992). From the mind's eye of the user: the sense-making qualitative-quantitative methodology. In J. D.Glazier, & R. R. Powell, *Qualitative research in information management* (pp.61-84) Englewood, CO: Libraries Unlimited.

Dupre, J. (1993). *The disorder of things: metaphysical foundations of the disunity of science*. Cambridge, MA: Harvard UP.

Dupre, J. (2006). Scientific classification, *Theory, culture & society*, *23*, 30-32.

El-Sherbini, M. (2008) Cataloging and classification: Review of the literature 2005-06. *Library Resources & Technical Services, 52,* 148-163.

Elsweiler, D, Ruthven, I. & Jones, C. (2005) Dealing with fragmented recollection of context in information management, *Context-Based Information Retrieval (CIR-05) Workshop* in CONTEXT-05, 2005.

Erickson, T. (2006). From PIM to GIM: personal information management in group

contexts. *Communications of the ACM, 49,* 74-75.

Fidel, R. (1992). Who needs a controlled vocabulary? *Special Libraries*, *83,* 1-9.

Glaser, B.G & Strauss, A. (1967) *Discovery of Grounded Theory. Strategies for Qualitative Research.* Sociology Press.

Gordon-Murdane, L. (2006) Social bookmarking, folksonomies, and web 2.0 tools. *Searcher, 14*, 26-38.

Greenberg, J. (2009). Theoretical considerations of lifecycle modeling: An analysis of the Dryad Repository demonstrating automatic metadata propagation, inheritance, and value system adoption. *Cataloging & Classification Quarterly*, *47*, 380-402.

Greenberg, J. (2010). Metadata and digital information. In *Encyclopedia of Library and Information Science, Third Edition*, (3610-3623). New York: Marcel Dekker, Inc

Greenberg, J., Pattuelli, M. C., Parsia, B., & Robertson, W. D. (2001). Author-generated Dublin Core metadata for web resources: A baseline study in an organization. *Journal of Digital Information*. Available at: http://jodi.ecs.soton.ac.uk/Articles/v02/i02/Greenberg/.

Greenberg, J., Spurgin, K. & Crystal, A. (2006). Functionalities for Automatic-Metadata Generation Applications: A survey of metadata experts' opinions. *International Journal of Metadata, Semantics, and Ontologies*, *1,* 3-20.

Greenberg, J., White, H., C, Carrier, C., & Scherle, R. (2009). A metadata best practice for a scientific data repository. *Journal of Library Metadata*, *9*, 194-212.

Hank, C. & Wildemuth, B. (2009) Quasi-experimental studies. In B. Wildamuth (ED), *Applications of social research methods to questions in information and library science.* (pp.93-104). Westport, Conn: Libraries Unlimited, 2009.

Hara, N., Solomon, P., Kim, S., & Sonnenwald, D. (2003) An emerging view of scientific collaboration: Scientists perspectives on collaboration and factors that impact collaboration. *Journal of the American Society for Information Science and Technology. 54*, 952-965.

Heidorn, P.B. (2008) Shedding light on the dark data in the long tail of science. *Library Trends 57*, 280-299.

Hicks, B.J., Dong, A., Palmer, R, & McAlpine, H.C. (2008) Organizing and managing personal electronic files: a mechanical engineer's perspective. *ACM Transactions on Information Systems, 26*, 4-23.

Hjørland, B. (2007). Semantics and knowledge organization. *Annual Review of information Science and Technology, 41*, 367-405.

Hjørland, B. (2008) What is knowledge organization (KO)?. *Knowledge Organization, 35*, 86-101.

Hjorland, B. & Nissen Pedersen, K. (2005) A substantive theory of classification for information retrieval. *Journal of Documentation, 61,* 582-597.

Hlava, M. M. (2002), Automatic Indexing: A Matter of Degree. *Bulletin of the American Society of Information Science and Technology*, *29*: 12–15. doi: 10.1002/bult.261

Hodge, G. (2000) Systems of knowledge organization for digital libraries: beyond traditional authority files" Available at: http://www.clir.org/pubs/reports/pub91/contents.html

Hooper, R.S. (1965). *Indexer consistency tests—Origin, measurements, results and utilization.* IBM, Bethesda.

Huang, L. (2010) Usability testing of the HIVE: A system for dynamic access multiple controlled vocabularies for automatic metadata generation. *UNC Masters Thesis.*

Hubenthal, U. (1998). Interdisciplinary thought. In Newell (ed.) *Interdisciplinarity: essays from the literature.* New York: College Entrance Examination Board.

Hughes, A.V. & Rafferty,P. (2011) Inter-indexer consistency in graphic materials indexing at the national library of Wales. *Journal of Documentation*, *67,* 9–32.

Hunter, E.J. (2002) *Classification made simple.* (2nd ed). London: Ashgate.

Jahoda, G.E., Hutchens, R.D., and Galford, R.R. (1966). Characteristics and use of personal indexes determined by scientists and engineers in one university. *American Documentation*, *17*, 71-75

Jones, J. (2009) [On not] taming the information wilderness. *Legal Information Management*. *9*, 53-56.

Jones, W. (2007a). How people keep and organize personal information. In W. Jones and J. Teevan (Eds.), *Personal information management.* Seattle, Washington: University of Washington Press.

Jones, W. (2007b). Personal information management. *Annual Review of Information Science and Technology. 41*, 453-504.
Jones, W. & Teevan, J. (2007). Introduction. In W. Jones and J. Teevan (EDs.), *Personal Information Management.* London: University of Washington Press.

Kant, E. (2003) *The critique of pure reason*. Champaign, IL: Project Gutenburg. http://www.gutenberg.org/etext/4280. (Original work published 1781).

Kelly, D (2006). Evaluating personal information management behaviors and tools. *Communications of the ACM. 49,* 84-6.

Kennedy, P. (2009) How to combine multiple research methods: Practical triangulation. *Johnny Holland Magazine* 2009-08-20. http://johnnyholland.org/2009/08/20/practical-triangulation/

Kipp, M. E.I. and Joo, S. (2010). Application of structural equation modeling in exploring tag patterns: A pilot study. *Annual Meeting of the American Society for Information Science and Technology*, Pittsburgh, Pennsylvania, USA.

Kipp, M.E.I. (2011a) Tagging of biomedical articles on CiteULike: A comparison of user, author and professional indexing. *Knowledge Organization, 38*, 245-261.

Kipp, M. E.I. (2011b) User, author and professional indexing in context: An exploration of tagging practices on CiteULike. *Canadian Journal of Library and Information Science*, *35*, 17-48.

Klein, J. (1999) *Mapping interdisciplinary studies.* Washington, D.C: Association of American Colleges and Universities.

Kwasnik, B.H. (1989) *The influence of context of classificatory behavior*. Doctoral Thesis. Rutgers University.

Kwasnik, B. H. (1991) The importance of factors that are not document attributes in the organization of personal documents. *Journal of Documentation*, *47*, 389-98.

Lakoff, G. (1987) *Women, fire, and dangerous things.* Chicago: University of Chicago Press.

Lancaster, F. W. (1986). *Vocabulary control for information retrieval. 2nd. ed*. Arlington, Va.: Information Resources Press.

Langridge, D. (1992). *Classification: its kinds, elements, systems, and applications.* London: Bowker Saur.

Lansdale, M. (1988). The psychology of personal information management. *Applied Ergonomics*, *19*, 55-66.

Leonard, L. (1977) Inter-indexer consistency studies, 1954-1975: a review of the literature and summary of study results. *Occasional papers* (University of Illinois at Urbana-Champaign. Graduate School of Library Science).

Locke, J. (2004) *An essay concerning humane understanding.* Champaign, IL: Project Gutenberg. http://www.gutenberg.org/etext/10615. (Original work published 1690)

Lopez-Huertas, M.J. (2008) Some current research questions in the field of knowledge organization. *Knowledge Organization*, *35*, 113-136.

Malone, T.W. (1983) How do people organize their desks? Implications for the design of office information systems. *ACM Trans Office Info Systems, 1*, 99-112.

Mathes, A. (2004) Folksonomies - cooperative classification and communication through shared metadata. *Computer Mediated Communication, LIS590CMC (Doctoral Seminar*), Graduate School of Library and Information Science, University of Illinois Urbana-Champaign.

Mayernik, M. S. (2010a). Metadata Realities for Cyberinfrastructure: Data Authors as Metadata Creators. *iConference 2010 Proceedings*,148-153.

Mayernik, M. S. (2010b) Metadata Tensions: A Case Study of Library Principles vs. Everyday Scientific Data Practices. *ASIS&T '10 Annual Meeting,* Oct. 25, 2010, Pittsburgh, PA.

Meadow, C.T., Boyce, B., Kraft, D. H., & Barry, C. (2007). *Text information retrieval systems.* Orlando: Academic Press.

Medelyan, O. (2009) Human-competitive automatic topic indexing. *The University of Waikato Thesis.*

Medelyan, O. & Witten, I. (2006). Thesaurus based automatic keyphrase indexing. *JCDL'06 Proceedings of the 6th ACM/IEEE-CS joint conference on Digital Libraries*, 1-10.

Medelyan, O. & Witten, I. (2008) Topic indexing with Wikipedia. *Proceedings of the AAAI WikiAI Workshop,* 19-24.

Miksa, F. (1998). *The DDC, the universe of knowledge and the post-modern library.* Albany, New York: Forest Press.

Miller, G.A. (1968) Psychology and information. *American Documentation*. July, 286-289.

Mostafa ,J. Quiroga , L.M. & Palakal, M. (1998) Filtering medical documents using automated and human classification methods. *Journal of the American Society for Information Science*, *49*, 1304-1318.

Noruzi, A. (2007) Editorial: Folksonomies: Why do we need controlled vocabularies? *Webology 4*. Available at: http://www.webology.org/2007/v4n2/editorial12.html

Olsen, H. (1998) Mapping beyond Dewey's boundaries: Constructing classificatory space for marginalized knowledge domains. *Library Trends*, *47*, 233-254.

Otlet, E. (1990) The science of bibliography and documentation. In Rayward (ed. and trans.) *International organization and dissemination of knowledge*: *Selected essays of Paul Otlet.* New York: Elsevier. https://www.ideals.illinois.edu/handle/2142/4004 (Original work published 1903).

Papadakis, I., Kyprianos, K., Mavropodi, R., & Stefanidakis, M., (2009). Subject-based information retrieval within digital libraries employing LCSHs. *D-Lib Magazine 15*. Dlib.org.dlib/09papadakis.html.

Pickard, A.J. (2007). *Research Methods in Information*. London: Facet.

Plato (1999). *Theaetetus.* Champaign, IL: Project Gutenburg. http://www.gutenberg.org/etext/1726 *(Original work published 369BC)*

Powell, R.R. (1991) *Basic research methods for librarians*. 2nd ed. London: Ablex Pub Corp.

Price SA and Gittleman JL (2007) Data from: Hunting to extinction: biology and regional economy influence extinction risk and the impact of hunting in artiodactyls. Dryad Digital Repository. doi:10.5061/dryad.82

Price SA, & Gittleman, JL (2007) Hunting to extinction: biology and regional economy influence extinction risk and the impact of hunting in artiodactyls. *Proceedings of the Royal Society B* 274: 1845-1851. doi:10.1098/rspb.2007.0505

Qin, J and D'Ignazio, J. (2010a) Lessons learned from a two-year experience in science data literacy education*". International Association of Scientific and Technological University Libraries, 31st Annual Conference*. Paper 5. http://docs.lib.purdue.edu/iatul2010/conf/day2/

Qin, J and D'Ignazio, J (2010b) The Central Role of Metadata in a Science Data Literacy Course, *Journal of Library Metadata, 10*, 188-204.
Rafferty, P. and Hidderley, R.(2007). " Flickr and Democratic Indexing: Dailogic Approaches to Indexing." *Aslib Proceedings: New Information Perspective*. *59,* 397-41.
Rayward, W. (1975) *The universe of information: The work of Paul Otlet for documentation and international organization.* Moscow: International Federation of Documentation.

Richardson, E. (1930). *Classification: Theoretical and practical.* New York: H.W. Wilson.

Robertson, G., Czerwinski, M,. Larson, K., Robbins, D.C., Thiel, D., & van Dantzich, M. (1998) Data mountain: using spatial memory for document management. In the Proceedings of *UIST '98: User Interface Software and Technology*.

Reich, P, & Biever, E. J. (1991). Indexing consistency: The input/output function of thesauri. *College & Research Libraries 52*, 336-342.

Rolling, L. (1981). Indexing consistency, quality and efficiency. *Information Processing and Management*, *17*, 69–76.

Rowley, J. (1994) The controlled versus natural indexing language debate revisited: a perspective on information retrieval practice and research. *Journal of Information Science, 20*,108-119.

Salo, D. (2010) Retooling libraries for the data challenge. *Ariadne 64*. Available at: http://www.ariadne.ac.uk/issue64/salo/

Salton, G. (1975) *A Theory of Indexing*. Society for Industrial and Applied Mathematics.

Schwartz, C. (2008) Thesauri and Facets and Tags, Oh My! A Look at three decades in subject analysis. *Library Trends, 56*, 830-842.

Shiri, A. and Chase-Kruszewski, S. (2009) Knowledge organization systems in North American digital library collections. *Program: electronic library and information systems,. 43,* 121-139.

Shirky, C. (2005) Ontology is Overrated; Categories, Links and Tags. Available at: http://www.shirky.com/writings/ontology_overrated.html

Sievert, M.C., & Andrews, M.J. (1991). Indexing consistency in Information Science Abstracts. *Journal of the American Society for Information Science, 42*, 1-6.

Sparck Jones, K.(1972). 'A statistical interpretation of term specificity and its application in retrieval', *Journal of Documentation*, 28, 11-21.

Spiteri, L. F (2007) The Structure and Form of Folksonomy Tags: The Road to the Public Library Catalog. *Information Technology and Libraries, 26*, 12-25.

Spurgin, K. (2006) The sense-making approach and the study of personal information management. Personal information management- a *SIGIR 2006* workshop.

Strauss, A. and Corbin, J. (1990) *Basics of qualitative research: Grounded theory procedures and techniques.* Thousand Oaks, CA: Sage.

Strout, R. (1969) The development of the catalog and cataloging codes in *The Catalog and Cataloging (pp. 3-33).* New York: Shoe String Press.

Svenonius, E. (1986) Unanswered questions in the design of controlled vocabularies. *Journal of the American Society for Information Science. 37*, 331-340.

Svenonius, E. (2001). *The intellectual foundation of information organization*. Cambridge, MA: The MIT Press.

Symonds MRE, Tattersall GJ (2010) Data from: Geographical variation in bill size across bird species provides evidence for Allen's rule. Dryad Digital Repository. doi:10.5061/dryad.1421

Symonds MRE, Tattersall GJ (2010) Geographical variation in bill size across bird species provides evidence for Allen's rule. American Naturalist 176: 188-197. doi:10.1086/653666

Szostak, R. (2004). *Classifying science: phenomena, data, theory, method, practice*. Dordrecht, Netherlands: Springer.

Szostak, R. (2007). Interdisciplinarity and the classification of scholarly documents by phenomena, theories, and methods. In *ISKO Italia Proceedings 2007.*

Tait, J. (1970). *Authors and titles; an analytical study of the author concept in codes of cataloguing rules in the English language, from that of the British Museum in 1841, to the Anglo-American cataloguing rules*. Hamden, Conn.: Archon Books.

Vleduts-Stokolov, N. (1982) On automatic support to indexing a life sciences data base. *Information Processing and Management, 18*, 313-21.

Wallis, J.C, Mayernik, M.S., Borgman, C.L., & Pepe, A. (2010) Digital Libraries for Scientific Data Discovery and Reuse: From Vision to Practical Reality. *Joint Conference on Digital Libraries*'10, June 21-25, 2010, Goald Coast, Queensland, Austrialia.

Weinberger,D. (2007) *Everything is miscellaneous.* New York: Times Books.
White, H.C. (2008, September). Exploring evolutionary biologists' use and perceptions of semantic metadata for data curation. Poster session presented at the *International Conference on Dublin Core and Metadata Applications* 2008. Berlin, Germany, September 22-26, 2008.

White, H. C.(2010a) 'Considering Personal Organization: Metadata Practices of Scientists', *Journal of Library Metadata*, 10: 2, 156-172.

White, H. C. (2010b) Organization as a means for information repository design: The convergence of knowledge organization and personal information management in scientific data. *ASIS&T '10 Annual Meeting,* Oct. 25, 2010, Pittsburgh, PA.

White, H., Carrier, S., Thompson, H., Greenberg, J., and Scherle, R. (2008). The Dryad Data Repository: A Singapore Framework Metadata architecture in a DSpace environment. In *DC-2008: Metadata for Semantic and Social Applications*. International Conference on Dublin Core and Metadata Applications, 22-26 September, 2008, Berlin Germany, 157-162.

Whittaker, S., Bellotti,V. & Gwizda, J.(2006) "Email in Personal Information Management. *Communications of the ACM,* 9, 68-73.

Xu, H. and Lancaster, F. W. (1998) Redundancy and uniqueness of subject access points in online catalogs. *Library Resources and Technical Services*, *42*, 61-6.

Zerubavel, E. (1991). *The fine line: Making distinctions in everyday life.* Chicago: University of Chicago Press.