FLORIDA STATE UNIVERSITY

COLLEGE OF COMMUNICATION AND INFORMATION

RESEARCH DATA CURATION PRACTICES IN INSTITUTIONAL REPOSITORIES AND DATA IDENTIFIERS

By

DONG JOON LEE

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

2015

Dong Joon Lee defended this dissertation on June 22, 2015.

The members of the supervisory committee were:

Besiki Stvilia
Professor Directing Dissertation

Anke Meyer-Baese
University Representative

Corinne Jörgensen
Committee Member

Richard J. Urban
Committee Member

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

ACKNOWLEDGMENTS

I would like to express the deepest appreciation to my major professor, Dr. Besiki Stvilia, who has provided great support. Without his support and guidance, there would be no dissertation. Thanks also go to my committee members, Dr. Corinne Jörgensen, Dr. Richard Urban, and Dr. Anke Meyer-Baese, for their support in bringing this dissertation to completion.

I would like to thank all of the interview participants. Without their input, this dissertation could not be completed so well. I would also like to thank faculty and colleagues in the School of Information. And last, thanks to Seungyeon, Haryn, parents, and my family. I have been able to successfully complete this long journey because you were always there for me and gave me your endless love and support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

The access and sharing of research data have been emphasized by the government, funding agencies, and scholarly communities. The increased access to research data increases the impact, as well as the efficiency and effectiveness, of scientific activities and funding. The access, however, is facilitated not just by appropriate policies but also by the employment of effective infrastructure mechanisms, including enhancing data with effective metadata (Simmhan, Plale, & Gannon, 2005). Identifiers are important metadata that traditionally have been used for entity identification, linking, and referencing in various domains (Altman & King, 2007). To enable effective metadata creation support for research data, it is essential to gain a better understanding of the current uses of identifier systems with research data.

As many research institutions plan to provide some types of research data services (Tenopir, Birch, & Allard, 2012), it is important to study the current practices of data curation in IRs. In particular to develop effective data management infrastructure configuration templates, it is essential to understand user needs and related activities for data curation in IRs, including different roles played by IR staff and role-specific differences in needs for skills and infrastructure support (Foster, Jennings, & Kesselman, 2004). Furthermore, it is important to investigate both the current practices of identifier use and the requirements for quality and functionalities for identifier schemas in order to design effective metadata support for research data curation in IRs.

Studying the practices of research data curation requires multifaceted contextual analysis (Borgman, Wallis, & Enyedy, 2007). Hence this study, too, required a research design that could help examine and capture various sociotechnical and cultural factors that may affect data curation, including the selection and uses of identifier schemas for data. The study used Activity

Theory (Engeström, 1987; Leontiev, 1978) and Information Quality Assessment Framework (Stvilia, Gasser, Twidale, & Smith, 2007) to guide the design of a protocol for semi-structured interviews.

This study reports on data collected from fifteen participants from thirteen different universities in the US. The selection of participants was guided by two criteria. To be eligible for participation in the study, participants had to work for an IR that stored and curated research data objects and housed by one of the 108 institutions classified as RU/VH (very high research activity) in the Carnegie Classification of Institutions of Higher Education.

The study identified data curation activities and contexts (i.e., tools, norms, rules, and division of labor), perceived roles played by IR staff (e.g., data curator, IR manager, and metadata specialist), role-specific sets of activities and skills, and perception of quality identifiers in IRs. The findings of this study can inform the development of best practices and effective infrastructure support for data curation in the context of IRs, as well as teaching data curation in LIS schools.

CHAPTER 1

INTRODUCTION

1.1 Problem Statement

1.1.1 Background

Many research universities have operational institutional repositories (IRs) that provide open access to the digital content produced by the universities' communities. However, if the repositories are filtered by the inclusion of research data objects, the number of universities that have such IRs is dramatically decreased. According to Lee and Stvilia (2012), in 2012 only half of Association of American Universities (AAU) member universities, which are the leading 62 public and private research universities located in the United States and Canada, had IRs that contained research data objects. In addition, the Association of College and Research Libraries (ACRL) published a report about current practices and plans for the future of research data services in academic libraries. According to the report, only a small number of academic libraries in the United States and Canada currently offer research data services, but about 25 to 30 percent of the 351 ACRL-member libraries are planning to provide some research data-related services within the next two years (Tenopir et al., 2012).

The access and sharing of research data have been emphasized by the government (Office of Science and Technology Policy, 2013), funding agencies (IMLS, 2011; NIH, 2010; NSF, 2010a) and scholarly communities (Aalbersberg & Kähler, 2011; Thomson Reuters, 2012). The increased access to research data elevates the impact, as well as the efficiency and effectiveness, of scientific activities and funding opportunities. The access, however, is facilitated not just by appropriate policies but also by the employment of effective infrastructure mechanisms, including enhancing data with effective metadata (Simmhan et al., 2005). Identifiers are important metadata that traditionally have been used for entity identification, linking and referencing in various domains (Altman & King, 2007). To enable effective metadata creation support for research data, it is essential to a gain better understanding of the current uses of identifier systems with research data, as well as the needs for identifier system functionalities and the functionalities of currently available identifier systems.

Providing effective metadata support including identifier schemas is essential to achieve the objectives of IRs which include but are not limited to sharing, accessing, controlling and preserving knowledge and data (Markey, Rieh, St.Jean, Kim, & Yakel, 2007; Westell, 2006). As many research libraries plan to offer some types of research data services (Tenopir et al., 2012), it is important to investigate both the current practices of identifier use in IRs and the requirements for quality and functionalities for identifier schemas in order to support data curation in IRs.

Furthermore, as Linked Data technologies are increasingly used to expose, discover, link and integrate knowledge, metadata, and data curated by libraries and IRs (e.g., Latif, Borst, & Tochtermann, 2014; Park, 2015), understanding and coordinating identifier metadata, the essential component of any RDF based serialization and consequently any Linked Data implementation, ensuring the quality and reusability of those identifiers become increasingly important too. For example, the identifiers (e.g., ORCID) that can be assigned to and reused to link to various entities (e.g., person and event) and related identity profiles currently have significant commercial and community input (Warner, 2010). There is a significant body of research on metadata quality and reusabilty in general (e.g., Shreeves et al., 2005; Stvilia, Gasser, Twidale, Shreeves, & Cole, 2004). There is a dearth of research, however, on the quality and reusability of identifier metadata. Particularly, there is a lack of research on various socio-technical aspects that make identifiers reusable, or alternatively hinder identifier reuse.

1.1.2 IRs and Research Data

According to Witt and Cragin (2008), three types of repositories (i.e., domain, discipline and institutional) exist. The main difference between them is the granularity of the organizations that operate the repositories. For example, a chemistry community may develop a domain repository; a crystallography community may operate a discipline repository; and a university may run an institutional repository (IR). IRs can be defined as "a set of services that an institution provides to the members of its community for the management and dissemination of digital materials created by the institution and its community members" (Lynch, 2003, p.2). IRs offer various benefits to support curatorial activities, including preserving, discovering, controlling, reusing and repurposing institutional intellectual content (Markey et al., 2007; Rieger, 2007). In particular, IRs as alternative channels in support of content dissemination and communication may increase institutional name value through access to intellectual work produced by the communities (Lynch, 2003; Rieger, 2007; Witt & Cragin, 2008). This access, along with an emphasis of research data archiving and sharing from major funding agencies (IMLS, 2011; NIH, 2010; NSF, 2010b), inspires many institutions to put their efforts into the development of IRs and services for research data (Tenopir et al., 2012; Witt, 2012). According to many researchers, IRs storing and curating research data can help reuse and repurpose the data (Heidorn, 2008) and increase the value and credibility of the data (Witt & Cragin, 2008).

1.1.3 Curation for Research Data and Identifiers

Curation for research data is the process of managing research data through its lifecycle for long-term availability and reusability (Cragin, Heidorn, Palmer, & Smith, 2007; Curry, Freitas, & O'Riáin, 2010; Lord & Macdonald, 2003). The curation and its activities facilitate discovery, retrieval, quality and value management, and reuse of research data (Cragin et al., 2007). Research data curation activities proposed in the literature consist of discover, identify, select, obtain, verify, analyze, manage, archive, publish, and cite (Qin, Ball, & Greenberg, 2012; Stvilia et al, 2015). This study uses the set of data curation activities as an umbrella context to understand and reason about the practices, uses, and quality requirements of identifiers for research data. Each activity associated with data curation requires the use of different types of metadata to describe, administer, and package research data. Identifiers are essential elements for each of these types of metadata (i.e., descriptive, administrative and structural.). Specific data activities that involve identifiers include identification, citation, linking, and annotation (Lee & Stvilia, 2014). Gaining a better understanding of the metadata needs of those activities, including a better understanding of quality requirements for identifier systems can improve research data curation and scholarly communication practices and education.

The use of identifiers is context specific because different communities/organizations manage different data on different entities/uses. Identifiers are tailored to the community's data practices and curation needs (Stvilia et al., 2013). In biology, Life Science Identifiers (LSIDs), which identify and integrate data objects, are a major identification scheme (Wu, Stvilia, & Lee, 2012). In chemistry, various domain-specific identifiers (e.g., Chemical Abstracts Service (CAS) Registry Number, International Chemical Identifier, Chemical Entities of Biological Interest, etc.) and their associated metadata are used to discover chemical substances and compounds

(Akhondi, Kors, & Muresan, 2012). Also, different IRs use different identifier systems based on their sociotechnical context (e.g., policies, systems, practices) (Lee & Stvilia, 2012). In the context of large organizations, academic publishers have made important changes too. For example, Thomson Reuters announced its development of a data citation index and started indexing research data from repositories available to them across disciplines and around the world to supplement articles in the Web of Knowledge with associated research data. Appropriate identifier systems are essential for making these connections (Thomson Reuters, 2012). The current identifier practices from diverse communities reflect the increased importance of identifiers and their context-specific uses in the current research environment.

1.1.4 IRs and Identifiers

IRs are an essential infrastructure for digital scholarship and data curation, and identifiers are one of the key metadata needed for successful management and use of data stored in IRs (Lynch, 2003). Lee and Stvilia (2012) found that IRs use various identifiers such as Handle System, Uniform Resource Identifier (URI), Digital Object Identifier (DOI), Archival Resource Key (ARK) and Universally Unique Identifier (UUID) to support their data curation activities. However, there are few studies on the practices, functionalities, uses, and quality requirements of different existing identifier systems within the context of IRs, and therefore limited and unsystematic knowledge of identifier practices for research data.

1.2 Purpose and Significance of the Study, and Research Questions

1.2.1 Purpose

The purpose of this study is to examine data curation practices in IRs, with the emphases on the uses of identifier schemas and the needs and requirements for identifier quality and functionalities in data curation activities. With the increased push for research data services in many research libraries (Tenopir et al., 2012) and diverse uses of identifiers for research data (Lee & Stvilia, 2014), the current identifier practices within IRs storing research data is an essential research topic. Although many previous studies have been conducted on the theme of either identifiers or IRs, to the best of this researcher's knowledge no systematic investigation has yet been done of practical uses of identifiers for research data curation and use activities within IRs. By interviewing IR curators and analyzing their identifier practices for research data management, this study builds a knowledge base of identifier system uses for research data which can be used in data curation planning and education as well as in designing and planning IR data services.

1.2.2 Significance and Research Questions

Different communities use different identifiers to support their different data types. Various identifiers exist in support of a specific domain or across multiple domains. The push by major funding agencies and the government toward research data's long term availability, sharing and reuse (NSF, 2010a; Office of Science and Technology Policy, 2013) increases the importance of effective and efficient use of metadata schemas, including identifier schemas, in supporting those goals. According to several researchers (Duerr et al., 2011; Lee & Stvilia, 2014), examining the functional requirement of identifier schema's design for data, the issues of current practices of identifier selection and use for data, and identifier quality evaluation can provide valuable knowledge to data curation research and practice communities. The increasing need for providing metadata support for complex and diverse entity types of research data from different disciplines and different communities calls for the creation of such knowledge bases and best practices guides (Lee & Stvilia, 2014). To contribute towards that objective, this study will examine the following research questions:

RQ 1. What are the types of data activities in IRs and what are the structures and metadata requirements of those activities?

RQ 2. What are the major types of research data and their entity types within IRs for which identifiers are used?

RQ 3. What is the awareness of IR curators about different currently available identifier schemas?

RQ 4. How do IR curators perceive the quality of identifiers for research data? \

1.3 Brief Notes of Methodology

Studying the practices of research data curation with an emphasis on identifier uses requires multifaceted contextual analysis (Borgman et al., 2007; Lee & Stvilia, 2014; Stvilia et al., 2015). The perception of an identifier system's quality depends on the context of its use (Akhondi et al., 2012; Clark, Martin, & Liefeld, 2004; Lee & Stvilia, 2012). Hence this study requires a research design that can examine and capture various sociotechnical and cultural factors that may affect data curation, including the selection and uses of identifier schemas for data.

Activity Theory (Engeström, 1987; Leontiev, 1978) was used as a guiding theory for modeling the general context of data work in IRs. This theory helps identify and reason about activities generated by subject (e.g., IR curator) and object (e.g., data identification, citation, etc.) interaction and many different mediating factors (e.g., rules, norms, conventions, instruments, divisions of labor, and other work arrangements) within an organization or community. In other words, activities reflect the context of a specific organization and its culture, and the theory is able to provide a direction to explore that context(s) (Nardi, 1996; Wilson, 2008). In the case of IRs, the community comprises curators and managers of those IRs. Many researchers have utilized Activity Theory in their studies of data or metadata practices (e.g., Borgman et al., 2007; Stvilia, 2007; Stvilia et al., 2015). They use the theory as a knowledge tool and a predictive mechanism for activity structure (i.e., subject, object, instrument, rules, community, and division of labor) and relationships to guide the development of research questions and design. To conceptualize individual instances of activity and identify activity specific data requirements or issues, Activity Theory is often supplemented with scenario based task analysis (Go & Carroll, 2004a, 2004b) (Huang, Stvilia, Jörgensen, & Bass, 2012; Stvilia, 2007). In addition, prior studies also used semi-structured interviews guided with Activity Theory as a method of empirical data collection (Borgman et al., 2007; Stvilia et al., 2015).

The semi-structured interview is a type of research method often used in data practice- or metadata-related studies. Many researchers use the interviews to identify sociotechnical aspects (e.g., policies, system designs, practices) of community knowledge. This study will explore and understand the practices of research data curation in IRs. Hence the semi-structured interview is a suitable research method that can investigate the curation activities and practices within the IR community, probing the interviewees with both prepared interview questions and questions that evolve through the interview process. The interview data would allow the researcher to make sense of the community’s practices, language and data relationships. Thus, to understand and analyze a community’s data practice and its use of identifier system(s), the use of the semi-structured interview guided by Activity Theory is a relevant research design. Each concept and relationship from Activity Theory (i.e., subject, object, instrument, rules, community and division of labor) can be the basis of interview questions.

In addition, the Information Quality (IQ) Assessment Framework (Stvilia et al., 2007), which defines general relationships among information activity types, quality problem types, related quality dimensions, and metrics, guides this study. The framework can be used as a predictive mechanism for the types of identifier quality problems and activity relationships, which then could be probed for through interview questions during the data collection. The model conceptualizes data curation activities related to the use of identifiers by their activity types, quality problem types, and related quality dimensions.

CHAPTER 2

LITERATURE REVIEW

2.1 Identifiers for Research Data

2.1.1 Definition

Identifiers can be defined in many different ways depending on the purposes (e.g., identification, reference, annotation) for which they are applied. The Oxford English Dictionary (OED) defines identifier as “a thing used to identify someone or something” or “a sequence of characters arbitrarily devised to identify or refer to a set of data, a location in a store, or a point in a program.” This definition highlights the purposes of identifying and referencing objects. Altman and King (2007), who discussed a possible schema for data citations, characterized identifier as “a character string guaranteed to be unique among all such names, which permanently identifies the data set independent of its location.” Their definition points to the importance of identifier systems’ performance (e.g., persistent access) as well as the purposes of data entity disambiguation. Pepler and O’Neil (2008) in their definition of identifier, specified the resources (e.g., person, house, color, employee, journal paper, or file) referenced by the identifier. In a recent report from NISO/NFAIS (2013), a definition highlighting identifiers’ overall purpose was offered: Identifiers “provide discoverability of and linking to content.” Based on these definitions, a conclusion is that the definition of identifiers should mention identifier system features, assigned entity types and purposes of identifiers. The set of data entity types that need to be referenced by identifiers is contextual and varies from one discipline to another. Likewise, different identifier systems can be used for referencing different kinds of entities. However, the activities (purposes) of identifiers do not change much. Based on literature analysis, this research defines a data identifier as a sequence of symbols designed to identify, cite, annotate and/or link research data and their associated metadata.

2.1.2 Importance of Identifiers

Funding agencies now require applicants to submit plans for disseminating and providing access to research data (IMLS, 2011; NIH, 2010; NSF, 2010b). In addition, many journals and article databases now require the submission of data along with manuscripts, as well as the annotation and integration of the manuscript's content with the data (Aalbersberg & Kähler, 2011). All of these requirements were intended to increase the access and use of research data. Access to research data used in the production of outcomes has become essential for understanding the research (Brase & Farquhar, 2011). The need for greater access and sharing of research data to increase the impact and efficiency of scientific activities and funding has been emphasized by the government and various funding agencies (NSF, 2010b; Office of Science and Technology Policy, 2013). Greater access to research data, however, is enabled not just by appropriate policies but also by the deployment of effective infrastructure mechanisms, including augmenting data with effective metadata (Simmhan et al., 2005). Identifiers are important metadata that traditionally have been used for entity identification, linking and referencing in various domains (Altman & King, 2007). To enable effective metadata creation support for research data, it is essential to gain a better understanding of the current uses of identifiers with research data, as well as the needs for identifier system functionalities and the functionalities of currently available identifier systems.

With the increased push for data sharing and reuse by the government, funding agencies and scholarly communities (NSF, 2010b), there is increased attention on the design of metadata for data, including identifier schemas (Duerr et al., 2011; Lee & Stvilia, 2012; NISO, 2013). As different communities manage different data for different entities, identifier schemas are contextual and tailored to the community's data practices (Stvilia et al., 2013). In molecular biology, Life Science Identifiers (LSIDs) are used to identify and integrate data objects distributed in multiple databases (Wu et al., 2012). In chemistry, chemical identifiers (e.g., Chemical Abstracts Service (CAS) Registry Number, International Chemical Identifier) and their associated metadata assist discovery of chemical substances and compounds (Akhondi et al., 2012). Large academic publishers have made important changes too. For example, Thomson Reuters announced its development of a data citation index and started indexing research data from repositories available to them across disciplines and around the world to supplement articles in the Web of Knowledge with associated research data. Robust identifier systems are essential for making these connections (Thomson Reuters, 2012). Likewise, Elsevier decided to use Open Researcher and Contributor ID (ORCID) to create more robust links between scholarly works and their authors (Aalbersberg & Kähler, 2011; Guess, 2012). These changes from diverse disciplines and major publishers reflect the increased uses and importance of identifiers in the current research environment.

Furthermore, data identifier research enriches research data sharing and integration, particularly in the current research milieu (Costas, Meijer, Zahedi, & Wouters, 2013). Many researchers agree that identification, citation, linking and annotation activities for research data and their metadata elements require the appropriate use of identifier systems and facilitate access to research data from different communities (Altman & King, 2007; Duerr et al., 2011; Green, 2009; Qin et al., 2012; Wu et al., 2012). Identification can be defined as "confirming that the entity described corresponds to the entity sought [by the user], or distinguishing between two or more entities with similar characteristics" (IFLA, 2009, p.79). The identification of data is, in general, performed by referring to the metadata schemas' elements developed by various communities, and these elements including identifier(s). Many projects attempt to improve the practice of research data management (e.g., DataUp, DataONE, DataCite) by proposing their own set of metadata elements.

The goal of data citation is to make a connection between an identifier and its assigned data object at any point in the future (Duerr et al., 2011). The minimum component of the connection is a persistent identifier (Altman & King, 2007). Many institutional data repositories assign a permanent identifier connecting various types of entities to each data object (Lee & Stvilia, 2012). The assigned identifiers currently enable users to expand the boundaries of usable research data with persistent connections.

The effect of linking can be defined as the connection between data that was not previously linked, or the connection of data lowering the barriers to linking data currently linked using other methods (Heath, n.d.). It is also understood as a set of best practices for publishing and connecting structured data on the web (Bizer, Heath, & Berners-Lee, 2009). Linked data principles outlined by Berners-Lee (2006) emphasize the active use of Uniform Resource Identifiers (URIs) in Hypertext Transfer Protocol (HTTP) and in the standards such as Resource Description Framework (RDF) and SPARQL (RDF query language). Finding proper identifiers for data objects and entities is essential to fully construct fully linked data.

Annotation is the process of adding notes or commentary to informational sources. The value of scholarly work annotated with relevant data increases in such aspects as explanation, description and interpretation (Abbott, 2008), and the use of identifiers greatly improves annotation. If research data in a data repository are not associated with the relevant articles, the data are hidden, thereby limiting the data's use and reuse. The identifiers can serve as the solution to address such limited use.

Surprisingly, the practical use of identifier systems for research data and their activities has not yet been systematically studied in the literature. Studies examining the gap between identifier systems used by different communities and analyzing them along different facets of their design and use would be invaluable and could be used by data managers and curators as a knowledge tool in selecting an identifier system(s) for their data repositories. The studies can also inform policy development for institutional data repositories with regard to identifier systems selection and use (Davidson, 2006).

2.1.3 Current Identifier Systems

Different communities use different identifier systems for different data entities. The researcher selected fourteen identifier systems referenced by multiple articles in the literature that are used for research data entities and described them for this investigation.

2.1.3.1 Archival Resource Key (ARK). In 2001, Kunze and Rogers at the U.S. National Library of Medicine originally developed the ARK. It is currently maintained at the California Digital Library. The ARK, which is used to identify research data in institutional repositories, is a domain-independent identifier (Lee & Stvilia, 2012). It enables users to access the metadata of the assigned object (Paradigm, 2008). The identifier is able to identify digital objects, physical objects, living beings and groups and intangible objects (CDL, 2012). The ARK uses a Uniform Resource Locator (URL) scheme to support long-term or permanent access to information objects, and they are sequences of characters following a label "ark:/."

The syntax of ARK consists of four different segments, and when an ARK is embedded in a URL, it has six segments. The ARK embedded in a URL has the following format: [Protocol]/[Name Mapping Authority Hostport (NMAH)/]ark:/[Name Assigning Authority Number (NAAN)]/[Name]/[Qualifier] (Paradigm, 2008). ‘Protocol’ indicates the Internet protocol (e.g., http://) used to form a URL. ‘NMAH’ identifies the provider of services, and “ark:/” is a prefix that indicates the beginning of ARK. ‘NAAN’ is a number assigned to each Name Assigning Authority (NAA), which is a 5- or 9-digit decimal number. A name comprised of ASCII characters is assigned by the NAA. It is a unique element within the NAA. ‘Qualifier’ is an optional component of an ARK that specifies the subcomponents of a digital object and their hierarchies with a slash. The following is an example of ARK: “http://example.org/ark:/12025/654xz321/s3/f8.05v.tiff” (CDL, 2012).

2.1.3.2 Digital Object Identifier (DOI). DOI is a digital identifier of an object, rather than an identifier of a digital object (Paskin, 2010). The scope of the DOIs exceeds the range of digital objects, and they can be used to identify digital, physical and abstract objects. DOI is a typical, domain-independent, identifier system designed by the International DOI Foundation (IDF), which is a non-profit, member-funded organization. IDF created the DOI system for the persistent identification of content within a digital environment (Paskin, 2010). DOIs can be assigned to content-related objects, such as text documents, datasets, sound carriers, books, photographs, serials, audio, video, audiovisual recordings, software, abstract works and artwork. An assigned DOI resolves to the bibliographic metadata records of the objects. The metadata records contain current information of the object being assigned the DOI.

DOI includes a prefix and a suffix separated by a forward slash. The prefix is assigned to a particular DOI registration agency, which consists of a directory code and registrant code (e.g., 10.1006). The directory code is always 10 while the registrant code is a unique, four-digit, alphanumeric string. The registrant provides the suffix, which is a sequential number or a combination of ASCII characters (e.g., jmbi.1998.2354). The suffix can cooperate with other identifier schemas (e.g., ISSN and ISBN) (Paskin, 2010). The following are examples of DOIs: "10.1006/jmbi.1998.2354" and "10.1038/issn.0028-0836."

2.1.3.3 Handle system. The Handle System is a domain-independent identifier schema for Internet resources. It was first developed in 1994 by the Corporation for National Research Initiatives (CNRI), and it was used mainly to resolve DOIs (Tonkin, 2008). However, Handles also can be used separately. Many institutional repositories use the Handle System as a standalone identification system for research data (Lee & Stvilia, 2012). Handle System identifiers persist over changes of time, location, ownership and any other conditions (CNRI, 2012). Similar to the scope of DOI, the Handle System is assigned to digital, physical and abstract objects. They resolve to typed metadata records of the assigned objects (Lannom, 2000).

The syntax of the Handle System consists of two parts (i.e., a prefix and a suffix separated by the ASCII character, "/"). The prefix identifies the body of the object administration. Each prefix is globally unique, and the Handle System manages the prefix. The suffix of a Handle System is a unique, local name within the prefix. The suffix is a number or an alphanumeric number. Examples of the Handle System are "4263537/4000" and "10.1045/january2010-reilly" (CNRI, 2012).

2.1.3.4 Persistent Uniform Resource Locator (PURL). PURL was developed by the Online Computer Library Center (OCLC), and it is commonly used as a domain-independent identifier in many institutions (Shafer, Weibel, Jul, & Fausey, n.d.). PURL consists of a URL that is a web address that has the feature of persistency. Unlike URLs, which link directly to the locations of Internet resources, PURLs link to middle resolution systems. The PURL Resolution Service maintains the connections between PURLs and their actual URLs and returns the URLs (current locations of resources) to the users. PURLs are linked to metadata records of the assigned objects, such as documents, articles, datasets, web pages and cataloging systems (Shafer et al., n.d.).

There are three parts to the syntax of PURL (i.e., 1) the protocol, 2) the resolver address and 3) a name) (Shafer et al., n.d.). To make a connection to a resolver address (e.g., purl.oclc.org), an access protocol (e.g., HTTP, IMAP, or SMTP) is used, as well as a unique name (e.g., keith/home) of an object following the resolver address. An example of PURL is http://purl.oclc.org/keith/home.

2.1.3.5 Uniform Resource Identifier (URI). URIs are persistent domain-independent identifiers of various web resources. They function by sharing the syntax of the Uniform Resource Locator (URL), Name (URN), or Citation (URC). Many identifier systems that have been designed by using URL or URN syntax can be used as URIs (Paskin, 2008; W3C, 2001). URLs specify the location of a resource, URNs specify the name of a resource, which is independent of location, and URCs point to metadata rather than the resource itself (W3C, 2001). In the classic version (i.e., web of document), the URLs are sufficient as web addresses, although as the locations of web documents change, broken links often occur. However, in the contemporary version (i.e., web of data), which highlights the persistent and unique access of the resource, the condition of non-permanent URLs is no longer sufficient. The changes of the web from classic to contemporary require the use of persistent and unique URIs (Berners-Lee, 1998).

The syntax of URI consists of a URI scheme, a scheme-specific string (authority and path), an optional query and an optional fragment identifier (Masinter, Berners-Lee, & Fielding, 2005). The URI scheme has diverse types (e.g., http:, urn: and mailto:,), and a URI always begins with one of the types. Examples of URIs are: “http://en.wikipedia.org/wiki/Uniform_resource_identifier” as a URL and “mailto:username@example.com.” as a URN.

2.1.3.6 Universally Unique Identifier (UUID). Originally, UUID was a domain-independent identifier standard used in the computing environment or in software development. The importance of data uniqueness and persistency expanded the usage of UUID from software construction to data identification. UUID supports practical uniqueness guaranteed across space and time (Leach, Mealling, & Salz, 2005). UUID is also generated by its algorithm without a centralized authority, making it less costly. Most other identifiers offer a guaranteed uniqueness that is administrated via authorities. However, the uniqueness is not unique from a practical perspective if the administrations no longer operate. Conversely, UUIDs are likely to be unique identifiers with their own algorithm, regardless of any authority. Currently, UUIDs are being used within institutional repositories to identify a variety of research data objects and to link to metadata records of the assigned objects (Lee & Stvilia, 2012).

UUID begins with the prefix “uuid:” and consist of five fields separated by hyphens, “-.” The size of a UUID is 128 bits, which can be transcribed by a hexadecimal digit string, which is case insensitive. The five fields include data of time stamp, clock sequence and unique node identifier (i.e., host address) (Leach et al., 2005). The practical uniqueness of UUIDs is completed by the combination of three data. An example of a UUID is “uuid:f81d4fae-7dec-11d0-a765-00a0c91e6bf6.”

2.1.3.7 National Center for Biotechnology Information’s (NCBI’s) Accession Number. Since the publication of the human genome project in 2001, biology has entered into a new age within gene and protein sequences (Higgs & Attwood, 2005). With the advances of high-throughput sequencing techniques, data on large numbers of genes and proteins must be curated (MacMullen & Denn, 2005; Wu et al., 2012). NCBI's accession number is a unique, domain-dependent, identifier scheme assigned to sequence records when the records are submitted to GenBank, which is a comprehensive database that contains publicly available biological sequence data developed by the NCBI, or to Reference Sequence (RefSeq), which also is a public database for nucleotide and protein sequences synthesized from the sequence data available in GenBank (Pruitt, Tatusova, & Maglott, 2005). The accession numbers are unique numbers that can be embedded in LSID, which is a type of URN, and the embedded number resolves the metadata of the sequence records.

The syntax of the accession number consists of a combination of letters and numbers. The format of the combination is usually one letter followed by five digits of numbers or two letters followed by six digits of numbers. The prefix of letter(s) is allocated depending on a stored database and the type of submitted data (NCBI, 2012). Examples of unique accession numbers are "JN587088" and "AC_123456."

2.1.3.8 Chemical Abstracts Service (CAS) Registry Number. The number of chemical substances registered in the CAS Registry rapidly increases. According to their report (CAS, 2012a), about 15,000 substances are updated on a daily basis. The CAS Registry contains various types of unique organic and inorganic substances and sequences in their database systems. The substances, such as alloys, coordination compounds, minerals, mixtures, polymers and salts, have distinctive names and structures within the registry. The official titles of substances are used globally to identify the chemical substances. In addition to the CAS Registry, the CAS provides the CAS Registry Number, which is a numeric identifier designed for only one substance (CAS, 2012a). Similar to the NCBI Accession Number, the CAS Registry Number can also be embedded in a URL, and the numbers resolve metadata records of the chemical substances.

CAS Registry Numbers consist of three groups divided by hyphens. The first part can include from two to seven digits. The second part has two digits. The last part has only one digit, which is used to verify the validity and uniqueness of the registry number (CAS, 2012b). An example of a CAS Registry Number is that used for caffeine (i.e., “58-08-2”).

2.1.3.9 Life Science Identifier (LSID). LSID is a domain-dependent identifier to identify the entities of life science (Duerr et al., 2011). The Interoperable Informatics Infrastructure Consortium (I3C)began its development in 2003 . The entities of life science include both concrete and abstract types (e.g., individual proteins or genes, transcripts, experimental datasets, annotations, ontologies, publications and biological knowledge-bases). LSID is an interoperable identifier, so that a namespace, such as an NCBI Accession Number, can be embedded in a LSID, and the LSID can also be embedded in a URN. LSIDs were designed to identify and access biological data in a simple and common way. LSIDs enable their users to access data from various existing resources (e.g., relational databases, applications and public data sources) (Clark et al., 2004).

The syntax of LSID has the following format: URN:LSID:<AuthorityID>:<AuthorityNamespaceID>:<ObjectID>[:<RevisionID>] (Clark et al., 2004). Each part is separated by an ASCII character, a colon. URN:LSID is a mandatory preface of LSIDs embedded in a URN. Authority ID identifies an organization that assigns LSIDs to the entities. Authority namespace ID indicates a specific namespace within an authority organization. Object ID is a unique name of an entity. Revision ID represents the version number of an entity and is optionally used. The first parts of LSID (URN:LSID:<AuthorityID>:) are case-insensitive, but the second part of LSID is case sensitive (i.e., (<AuthorityNamespaceID>:<ObjectID>[:<RevisionID>]). An example of an LSID is "URN:LSID:ncbi.nlm.nih.gov:GenBank.accession:NT_001063:2."

2.1.3.10 International Standard Name Identifier (ISNI). International Organization for Standardization (ISO) developed ISNI and the specification of the valid ISNI standard was published in 2012. ISNI identifies public identities across multiple fields of creative activity. People's roles in creation, production, management and content distribution chains can be recognized accurately, and the content created from the public identities can be managed effectively (ISO, 2012). ISNI is allocated to any party that is or was a natural person, a legal person, a fictional character, or a group of such parties, whether or not incorporated (ISNI, 2012). An ISNI is 16 numerical digits displaying as four blocks of four digits. Each block is separated by a space (ISO, 2012). An example of an ISNI is the following: "ISNI 1234 6834 9573 0495." ISNI can also be used as a namespace of a URL.

2.1.3.11 Open Researcher and Contributor ID (ORCID). In 2012, the ORCID service was launched by the ORCID community and developed to disambiguate scholars with the same name and make connections between research (e.g., research articles and research data) and researchers (Bryant, 2013; ORCID, n.d.). The ORCID community maintains it as a registry service, and it has many participants, such as Elsevier and CrossRef (CrossRef, 2011; Guess, 2012). The main goals of ORCID are to provide a reliable identifier and to support its communication and authentication (ORCID, n.d.). The format of ORCID is compatible with the format of ISNI, i.e., 16 alphanumerical digits (e.g., "0000-0002-4510-0385").

2.1.3.12 ResearcherID. ResearcherID was designed by Thomson Reuters in 2008 to solve the ambiguity of authors' names in scholarly communications. Researchers registered with ResearcherID.com are given ResearcherID identifiers. ResearcherID enables researchers to manage their publication lists, check their number of citations, identify future collaborators and avoid author misidentification (ResearcherID, n.d.). Also, ResearcherID information integrates with the data citation index developed by Thomson Reuters, so that researchers can easily discover the publication and its related data from the repository (Thomson Reuters, n.d.). A ResearcherID consists of alphanumeric characters and contains the year the researchers registered within its string. If a ResearcherID is “A-3308-2013,” it means that the researcher was the 3308th person to sign up with ResearcherID in 2013 (Enserink, 2009).

2.1.3.13 OpenID. An open source community trying to solve the difficulty of identity metadata management developed OpenID in 2005. OpenID is not limited to the scholarly domain. OpenID is mainly designed for identity authentication for logging on to Web sites. However, it has a potential to be used in open systems as an identifier (Warner, 2010). People may easily create an OpenID with their preferred OpenID providers. Once they have OpenID, it can be used to facilitate the communication of the users' attributes, such as name and institution, between the provider and the OpenID acceptors (OpenID Foundation, 2013).

OpenID has a form of URL, and uses a target Extensible Resource Identifier (XRI). Through the connection of an URL and a XRI, the users can be protected from exposing their identities. A XRI includes i-names and i-numbers. The names and numbers are usually registered simultaneously, and the re-assignable i-names are connected to the fixed i-numbers. The linkage between i-names and i-numbers is communicated with the OpenID (a form of URL) via the

Extensible Resource Descriptor Sequence (XRDS) document (OpenID Foundation, 2007). The following is an example of OpenID: “johndoe.myopenid.com.”

2.1.3.14 GeoNameID. GeoNameID is an identifier system used by GeoNames.org. GeoNames is a worldwide database of public geographical data from various sources (GeoNames, n.d.). It contains more than 10 million geographical names in several layers. In addition, the names of places, latitudes, longitudes, elevations, population and postal codes are stored among its data. The data from GeoNames are freely accessible through various web services. GeoNameID is a 7-digit random number.

2.2 Issues, Characteristics, or Contexts of Identifier Systems

Various identifier systems exist to support the identification and linking of different types of data in different communities. With the increase of data-driven research and the push for data reuse and sharing by the government, the effectiveness and reuse of metadata schemas, including identifier schemas, gain new importance. Some of the issues, characteristics or contexts related to data identifier systems’ design, use and evaluation are presented in the following subsections as discussed in the literature.

2.2.1 Domains

Research data can generally be defined as “the recorded factual material commonly accepted in the scientific community as necessary to validate research findings” (Office of Management and Budget, 1999). However, the types, formats of and the expertise needed to interpret and curate research data are contextual and domain dependent (Huang et al., 2012; Stvilia et al., 2013). In addition, researchers in scientific disciplines are more inclined to use domain-specific repositories than institutional general data repositories for their research data (Erway, 2012).

Various research institutions and communities (e.g., National Center for Biotechnology Information and Chemical Abstracts Service) developed domain-specific identifier schemas to meet their specific needs for identifying and linking datasets, research concepts and entities (Akhondi et al., 2012; Pruitt, Tatusova, Klimke, & Maglott, 2009). At the same time, international or national standard organizations (e.g., International Organization for Standardization (ISO), National Information Standards Organization (NISO) and International DOI Foundation) developed general identifier schemas that are independent of particular domains.

General identifiers are not limited by disciplines. They have more availability and viability than domain-specific identifiers (Tonkin, 2008). Because of their flexible designs, limitations on their uses and assigned entities are lower than other identifiers.

Domain-specific identifiers are designed for particular needs and purposes. To identify the specialized entities of targeted domains, communities analyze data entities and develop their own domain-specific identifiers. Since these identifiers are tailored to the needs of the domain, they might be less interoperable than general identifiers (Wu et al., 2012).

2.2.2 Entity Types

Research data may include different types of entities determined by their targeted domains and community norms and policies. Many data repositories store data as application specific computer files (Lee & Stvilia, 2012). The types of data may include row tabular data, data analysis files, images and drawings, power point presentations and text data files (Stvilia et al., 2013). In addition, community data repositories may also store and maintain knowledge organization tools such as taxonomies, controlled vocabularies and ontologies, which define different concepts, entities and relationships of the community's knowledge.

A number of researchers (European Library Automation Group, 2010; LeBoeuf, 2005) have sought to build a map between the identifiers used for traditional library resources (e.g., books, audio-visuals, serials, images) and entity types (in most cases, the FRBR conceptual model's group 1 entities: *work*, *expression*, *manifestation* and *item*). The map linking identifiers with entity types can be a helpful resource in the construction of interoperable data management infrastructure, including data service interoperability, and effective uses of the identifiers (Baker & Dekkers, 2003; Wynholds, 2011).

Several conceptual data models from library, museum and data preservation communities have been proposed in the literature (Caplan, 2009; ICOM, 2012; IFLA, 2009). The models include entities these communities collect and organize data for. The Open Archival Information System (OAIS) is an ISO conceptual reference model designed to inform the development of systems for long-term digital data curation. The Preservation Metadata: Implementation Strategies (PREMIS) led by the OAIS is a preservation metadata vocabulary (Caplan, 2009). The PREMIS is being widely used in various disciplines (Library of Congress, 2011). The PREMIS data model consists of five high level entities: intellectual entities, objects, events, agents and rights.

In the 1990s, libraries were facing a changed information environment that included the variety of data media and new information and data technologies which created new opportunities for more sophisticated uses, aggregation, sharing, analysis and visualization of data in general and bibliographic data in particular. To support the new uses of bibliographic data, the community needed a more systematic model for bibliographic records. The International

Federation of Library Associations and Institutions (IFLA) developed and published such a model – the Functional Requirements for Bibliographic Records (FRBR) conceptual model in 1998. The model focuses on supporting four user tasks: to find, identify, select and obtain bibliographic entities using a library catalog (IFLA, 2009). FRBR is composed of ten different entities and several relationships among the entities. The ten entities are categorized into three groups. The entities in the first group represent the bibliographic resources in a library catalog. The group entities include work, expression, manifestation and item. The entities in the second group represent those responsible for the first group's entities. The group's entities are the person and corporate body. The entities in the third group represent the subject of the first group's entities and include entities of concept, object, event and place (IFLA, 2009). In the Bibliographic Framework (BIBFRAME) recently developed by the Library of Congress for linked data, bibliographic entities are divided by two classes: creative work and instance (Library of Congress, 2012a).

The International Committee for Documentation (CIDOC) Conceptual Reference Model (CRM) is a formal ontology supporting the museum community developed by the CIDOC of the International Council of Museums (ICOM). The CRM is designed to integrate, mediate and interchange cultural heritage information (ICOM, 2012). Due to the variety and complexity of information that needs to be organized by cultural heritage communities, version 5.1 of the CRM is composed of 90 entities and 152 properties. The following paragraphs review the different data entity types identifiers are used for, as referenced in the literature and conceptual models.

2.2.2.1 Intellectual Entity. The PREMIS defines the type of Intellectual Entity as "a set of content that is considered a single intellectual unit for purposes of management and description" (Library of Congress, 2012b, p.6). A book, map, photograph, database, or dataset is an example of an Intellectual Entity. In FRBR, this type can be mapped to the Group 1 entities (i.e., work, expression, manifestation and item). To articulate this type of entity, PREMIS used a book *Animal Antics* published in 1902 as an example (Library of Congress, 2012b). A library digitized the book that created one image file (i.e., TIFF type) for each of 189 pages, and the library also created an XML file to structure the image files. The library also used Optical Character Recognition (OCR) on the image files to create a single large text file. The text file was created as an SGML file. The library repository contains *Animal Antics* as an Intellectual Entity that includes two representations, one consisting of 189 image file objects and an XML file object, and the other consisting of one SGML file object (see Figure 2.1). Each representation of the Intellectual Entity is full version of *Animal Antics*.

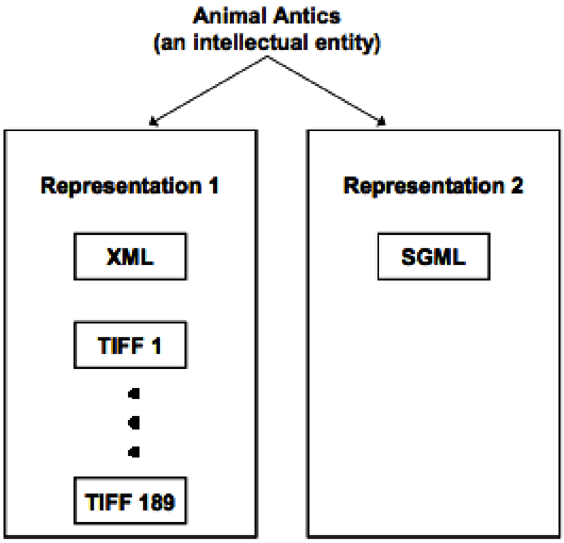


Figure 2.1. An example of Intellectual Entity, Animal Antics (Library of Congress, 2012b)

According to Carlyle (2004), abstract entities such as work and expression make FRBR difficult to understand for some because their existence is not observable. Discussions of the entities of expression and manifestation have also focused on their ambiguity related to XML documents (Renear, Phillippe, Lawton, & Dubin, 2003). Buckland (1998) and Floyd and Renear (2007) raised the lack of clarity in distinguishing what is a document within the digital environment, reflecting the difficulties of identifying item entity. All of the discussions about the ambiguities support the Intellectual Entity as being a single entity type for digital content, although it can be mapped to the multiple entities of the FRBR Group 1 (Caplan, 2009; Vitiello, 2004). In addition, Halpin (2008) mentioned that many different identifiers, such as URN and DOI, fail in accessing the entity that the identifiers are being assigned to, because the identifiers actually direct to the metadata descriptions of the information objects, rather than the entity itself. In this context, Intellectual Entity is a relevant entity type for digital data identifiers. PREMIS specified the URI, Library of Congress Control Number (LCCN) and Handle System as the identifier schemas for Intellectual Entities (Library of Congress, 2012b).

2.2.2.2 Object. Most research data in the digital environment exist as computer files or bitstreams. To store the data in digital repositories, the content of the data needs to be digitized. In the PREMIS, the Object is defined as “discrete units of information in digital form” (Caplan, 2009). The Object can be thought of as media/carriers of information, such as files, bitstreams, or representations. A dataset (i.e., an example of the Intellectual Entity) can be constructed by many computer files, and each file is an example of the Object entity type (Library of Congress, 2012b). As seen in Figure 2.1, each file is an Object. Many different data repositories and data management application tools (e.g., Dryad, DataUp, EZID, etc.) provide platforms for researchers managing and archiving research data. In most cases, the researchers upload their data file(s) in targeted repositories via the applications, and they get a unique identifier (e.g., DOI, ARK, etc.) associated with the data.

2.2.2.3 Symbolic Object. Scientific research data (i.e., an example of Intellectual Entities) in many cases have forms of symbolic representation (Huerta, 2013). Gene and protein sequences in biology and chemical compounds and structures are major examples. Every day scientists discover new DNA strands or chemical substances, and they store the discovery in data repositories and use the data to publish research articles. Alphabetic letters, specialized symbols/signs, etc. describe such scientific objects. The Concept entity in FRBR is defined as "an abstract notion or idea" (IFLA, 2009). Knowledge, theories, practices and techniques are examples of concept. The Symbolic Object in the CIDOC CRM is an entity type that can be matched with the Concept in FRBR. The Symbolic Object can be defined as identifiable concepts and any aggregation of concepts with an objectively recognizable structure (ICOM, 2012); the examples that the CRM provides are characters, texts, images, computer program codes, mathematical formulae, etc. The accession numbers from GenBank or Reference Sequence (RefSeq) databases, for example, are assigned to gene or protein sequences, and such sequences have the entity type of Symbolic Object.

2.2.2.4 Person. For the identification of any digital object, a Person entity is necessary to determine those who create and maintain the objects (Wynholds, 2011). Due to the malleable nature of digital data (i.e., easy to modify, aggregate, integrate) (Pollard & Wilkinson, 2010), metadata about who created, modified and/or accessed a particular data object is essential for discovering the data, and assessing its relevance and quality (Simmhan et al., 2005; Stvilia et al., 2007; W3C, 2013b). All of the three conceptual models used in this response include the entity representing human beings. Both FRBR and the CIDOC CRM have a Person entity in the models (ICOM, 2012; IFLA, 2009). In the PREMIS, an Agent entity exists that is defined as actors that affect the information. The Agent can include people, organizations and software applications.

The entities from the three models can help identify research data and control authority data of the assigned research data. As mentioned previously, Elsevier decided to use an author identifier (i.e., ORCID) to create links between scholarly works and their authors.

2.2.2.5 Organization. Organization is an entity type to identify organizations preserving, managing, or creating research data. The type has similar goals as the person entity, which helps access and retrieve correct research data with controlled authority metadata. This entity type exists in FRBR and the CIDOC CRM. The Corporate Body from FRBR is the entity corresponding to the Organization. It is defined as an organization or group of individuals (IFLA, 2009). The Legal Body entity corresponds to the organization in the CIDOC CRM. The CRM defines the entity as organizations or groups that have obtained legal recognition (ICOM, 2012). In the PREMIS, the organization is embedded in the agent entity. The International Standard Name Identifier (ISNI) designed by ISO is a type of author identifiers, which also identifies organizations as public identities (International Organization for Standardization, 2012). The ISNIs are mainly used in the Library of Congress to disambiguate the public parties involved in media content.

2.2.2.6 Place. Along with the development of the Geographic Information System (GIS), the potential values and uses of geographic data have increased. The data are being actively used in various domains, such as business, economics, history, urban planning and oceanography (Data & GIS Lab, 2013). In addition to the GIS data, the importance of the accurate geographic location (i.e., latitude and longitude) as research data has also increased. Knowing the precise location is important to research in oceanology, glaciology, meteorology, etc. The Place entity in FRBR is defined as a geographical location (IFLA, 2009). The CIDOC CRM's definition of

Table 2.1. Definitions of Entity Types for Data Identifier Systems

Entities	*Definitions*	*Sources*
Intellectual Entity	A set of content that is considered a single intellectual unit for purposes of management and description	PREMIS
Object	Discrete units of information in digital form. Can be files, bitstreams or representations. Objects are what are actually stored and managed in the preservation repository	PREMIS
Symbolic Object	An identifiable symbol and any aggregation of symbols, such as characters, data sets, images, multimedia objects, or mathematical formulae that have an objectively recognizable structure and that are documented as single units	CIDOC CRM
Person	An individual; Real person	FRBR & CIDOC CRM
Organization	An organization or group of individuals and/or organizations acting as a unit; Institutions or groups of people that have obtained a legal recognition as a group and can act collectively as agents	FRBR & CIDOC CRM
Place	A geographical location; It comprises extents in space, in particular on the surface of the earth, in the pure sense of physics	FRBR & CIDOC CRM
Time	Specific forms of historical periods or dates; Abstract temporal extents, having a beginning and an end	CIDOC CRM
Event	An action or occurrence; Actions that involve an Object and an Agent known to the system; Changes of states in cultural, social or physical systems, regardless of scale	FRBR, PREMIS, & CIDOC CRM
Topic	A hierarchy of topics used to organize the content of the dataset.	Google DSPL

Place is more specific: spatial extents on the surface of the earth (ICOM, 2012). Both models support the geographic location data with this entity. GeoNames is a geographical database that freely provides over 10 million geographical names and locations to the general public. It uses GeoNameID to identify its location data. The identifier schema includes geographical names, latitude and longitude, elevation, timezone, population, etc. as its metadata elements (GeoNames, n.d.).

2.2.2.7 Time. It is critical that time information is collected as a part of research data. Recorded time helps find proper and accurate data to meet users' needs. If accurate time records do not exist, researchers might have difficulty identifying and classifying data. For example, if a high- performance camera, which takes tens or hundreds of photos per second, does not record the exact time each photo was taken, the classification and organization of the data may not be possible. In the real example of the camera in space science, researchers organize a number of images of the sun in chronological order to observe and record the rate of changes on the surface of the sun. Such observations have great value as documents and forecasts and as research data. In the CIDOC CRM, two different entities may convey time information. The entity of Date is defined as specific forms of historical periods or dates, and the entity of Time-Span is defined as abstract temporal extents, having a beginning and an end (ICOM, 2012). Time is essential information in research data, but identifiers are not yet assigned to the information.

2.2.2.8 Event. With the malleable nature of the digital environment, the issue of data reliability and the quality of provenance metadata become even more important. In this context, the importance of all the changes that affect the digital objects is emphasized. The PREMIS defines Event as actions that involve an object and an agent associated with intellectual entities (Library of Congress, 2012b). The Event from the CIDOC CRM effectively reflects the features of the community information (i.e., cultural heritage information) from the definition of the entity. It is defined as changes of states in cultural, social, or physical systems (ICOM, 2012). Finally, the Event from FRBR is defined as an action or occurrence (IFLA, 2009). The W3C

Provenance (PROV) Working Group recently published its model for provenance metadata (i.e., PROV Model). The model defined an Activity entity being compared with Event entity as something that occurs over a period of time and acts upon or with entities. The entity includes the actions of consuming, processing, transforming, modifying, relocating, using, or generating entities (W3C, 2013b). The PROV Model Primer also defines three kinds of provenance perspectives for its users: agent-, object- and process-centered provenance. PREMIS supports the provenance of information with the event entity, which focuses on agent- and object-involved information. In the research data, the position of provenance information has been particularly emphasized. In a scientific experiment, a small change to an experimental variable can bring about great changes to the experiment's outcome; thus, any change in variables, such as an event, must be accurately recorded. The identifiers designed specifically for the Event entity do not currently exist or have not been reported in the literature yet; however the California Digital Library developed an identifier schema with a wide scope and uses it, the Archival Resource Key (ARK) which is able to refer Event entities.

2.2.2.9 Topic. In many cases, topics of bibliographic resources are included within metadata schemas as their elements. Authors of the resources or the domain experts usually assign topics or keywords to resources, which are then used in discovering relevant data and information resources. To bring related data together by disambiguating and reducing vocabulary variance in metadata (i.e., referring to the same concept with different terms and using the same term to refer to different concepts), scholarly communities (e.g., National Center for Biotechnology Information (NCBI), the American Institute of Physics (AIP), and the Library of Congress (LC) use different thesauri, controlled vocabularies and ontologies. For example, in 1970 the AIP developed the Physics and Astronomy Classification Scheme for classifying scientific literature using a hierarchical set of alphanumeric codes. The scheme has been used internationally, including by major physics journals. Google developed the Dataset Publishing Language (DSPL) for visualizing public data and released the DSPL schema to the public. According to the DSPL schema, topics of data can be used as an entity type, which includes at least one unique identifier (Google Developers, 2012). The third group entities of FRBR (i.e., Concept, Object, Event and Place) correspond to the Topic entity in the DSPL. In FRBR, the entities in the third group have a bidirectional relationship, entitled “has a subject,” with the work entity in the first group (IFLA, 2009). The relationship indicates that the third group entities explain the subjects of creative work. In the DSPL schema, which is used to manage public web sources, the topics or subjects of sources can be expressed by the Topic entity. The Topic in DSPL schema can be identified and referenced by URIs.

2.2.3 Activities

Four types of activities that use data identifiers were identified from the literature. They are identification, citation, linking and annotation of research data.

2.2.3.1 Identification. The identification task in FRBR is defined as “confirming that the entity described corresponds to the entity sought [by the user], or distinguishing between two or more entities with similar characteristics” (IFLA, 2009). When identifiers point location information, the activity of obtainment can also be used by the identifier systems. The identification activity is performed by utilizing identity metadata elements, most importantly identifiers (NISO, 2013).

Qin et al. (2012) discussed identity metadata for scientific data. They defined the identity metadata as the properties of entities (e.g., agent, event) that when encoded as metadata can be used to verify the identity of data resources. These entities may also have assigned metadata such as identifiers. For example, author identifiers (e.g., ORCID, ResearcherID, etc.) identify agent/person entity, and resource identifiers (e.g., DOI, URI, Handle System, etc.) are assigned to publication, event and/or dataset entities.

The DataUp project ran by the California Digital Library developed an open-source add-in for Microsoft Excel software. The add-in targeting data management of earth, environmental, and ecological sciences helps users with documenting and depositing data into a data repository. DataUp add-in uses ARKs as a persistent identifier for deposited datasets (DataUp, n.d.).

The Organization for Economic Co-operation and Development (OECD) Publishing proposed a metadata standard for data publishing (Green, 2009). The OECD Publishing specified DOIs as a mandatory identity metadata element for dataset entity. Similarly, Altman and King (2007) suggested using unique global identifiers for research data identification and citation, and they recommended using URIs taking URN syntax, LSIDs, DOIs and Handle Systems.

2.2.3.2 Citation. The main goal of data citation is to build the connection between an identifier and its associated data object at any time in the future (Duerr et al., 2011), and the minimum component of the connection is a persistent identifier (Altman & King, 2007). Many institutional data repositories assign identifiers to data objects to connect them to various types of entities (Lee & Stvilia, 2012). Citation metadata also can serve as data itself in evaluating the productivity and impact of individual researchers, teams, laboratories and communities (Hinnant et al., 2012; Stvilia et al., 2011).

Major funding agencies, such as NSF and NIH, have changed their policies and now require applicants to submit plans for distributing and providing access to research data. This pressure from the funding agencies encourages libraries and data centers to found projects like

DataCite to help researchers find, access and reuse data. DataCite also provides services and tools for data publishers to generate associated metadata. DataCite uses DOIs as its only allowed value of identifiers (DataCite, n.d.).

Several other tools/instruments have been developed to help institutions publish, cite and discover research data. The Dataverse Network (DVN) developed by the Institute for Quantitative Social Science at Harvard University is an open-source application providing useful guidelines and tools for data citation (Crosas, 2011). The application intended to motivate researchers to share data through enabling persistent data citation using a global persistent identifier and universal numerical fingerprint. The DVN specifies Handle Systems and its Global Handle Registry as their persistent identifier system. Also, DOIs, which use Handle System infrastructure for their name resolution, can easily be used as the standard identifier system with the DVN application.

The Data Observation Network for Earth (DataONE) is a National Science Foundation (NSF) supported project, which intends to improve access to, and preserve data. The DataONE community developed a method for data citation in the areas of life and earth science. The Dryad repository – a member repository of the DataOne - asks its users who cite data in Dryad to use either DOIs or ARKs. DOIs used by the Dryad are registered at DataCite, and the DOI registration information contains data citation metadata elements required by the Dryad (i.e., author(s), date, title of the data package, repository name and data identifier) (Michener et al., 2011).

2.2.3.3 Linking. The activity of linking can be defined as the connection between data that was not previously linked, or the connection of data lowering the barriers to linking data currently linked using other methods (Heath, n.d.). W3C introduced the concept of linked data in

2006. It is understood as a set of best practices for publishing and connecting data on the web (Bizer et al., 2009). In brief, data is serialized and published on the Web using the RDF based format, which potentially allows connecting the data with other related datasets at a low cost. Linked data are not just about uploading data on the web, but also about generating links (Berners-Lee, 2006).

Linked data principles outlined by Berners-Lee (2006) emphasize the use of HTTP Uniform Resource Identifier (URI). A datum is represented by a URI, and the two related URIs are linked by another URI. The three URIs accordingly form a RDF triple (W3C, 2004). Many data identifiers designed by using URL or URN syntax (see Table 2.2), can be used as URIs (Paskin, 2008; W3C, 2001). If identifiers are used as HTTP URIs, it is possible to generate RDF links (Bizer, Cyganiak, & Heath, 2007).

If research data in a data repository are not associated with the relevant articles, the data are hidden, limiting its use and reuse. The frequency of data use can be closely related to the value of the data, and the value can be improved by connecting them to entities of the relevant articles (Stvilia et al., 2013). The entities, in this context, can be defined as discipline-specific concepts used in the research (Aalbersberg & Kähler, 2011).

Elsevier currently provides linking services that aim to add values to scientific articles. The data are connected to the articles (i.e., dataset linking) or to the entities of the articles (i.e., entity linking) by identifiers. The dataset linking service makes the linking based on the DOIs assigned to articles. The entity linking service accepts various accession numbers (from GenBank, Protein Data Bank, Cambridge Crystallographic Data Centre, Molecular Interactions Database and Universal Protein Resource Knowledgebase) with URI syntax as its identifiers

(Aalbersberg & Kähler, 2011). Elsevier also stores data as RDF documents in a Linked Data Repository.

Table 2.2. Data Identifiers as URIs by using URL/URN Syntax

Identifiers	*URI*		*Sources*
	URL	*URN*	
ARK	Yes		CDL, 2012
DOI	Yes	Yes	DOI, 2013
Handle System	Yes		CNRI, 2012
PURL	Yes		OCLC, n.d.
UUID		Yes	Leach et al., 2005
NCBI Accession Number		Yes, with LSID	Clark et al., 2004
CAS Registry Number	Yes		Common Chemistry, 2013
LSID		Yes	Clark et al., 2004
ORCID	Yes		ORCID, n.d.
ResearcherID	Yes		ResearcherID, n.d.
GeoNameID	Yes		Pabón, Gutiérrez, Fernández, & Martínez-Prieto, 2013

2.2.3.4 Annotation. Annotation is a process of adding notes on or commentary to informational sources. Annotations may enhance the value of data by connecting or supplementing it with relevant descriptions, explanations and interpretations (Abbott, 2008). Annotating and integrating research data with relevant scholarly works tend to rapidly increase with data-driven research in scientific disciplines (Wu et al., 2012).

The National Center for Biotechnology Information (NCBI) developed Reference Sequence (RefSeq) database, which has authority over biological sequences within the GenBank database. Biological scientists use the RefSeq as an authority file by having access to well annotated genomic DNA, transcripts and protein sequences (Pruitt et al., 2009). RefSeq uses its accession number as identifiers for scientific annotations.

W3C Open Annotation Community Group recently published Open Annotation Data Model, which provides a framework for annotation. The framework proposes the open annotation following the linked data principles (W3C, 2013a). The annotation is considered to be a set of linked resources including “body” and “target.” In most cases, the body explains the target. The model recommends using URIs to make connections between the resources.

2.2.4 Quality Dimensions

To support the activities of identifiers and evaluate their quality, many researchers have suggested or developed different quality requirements (Akhondi et al., 2012; Altman & King, 2007; Berners-Lee, 1998; Brand, Daly, & Meyers, 2003; Callaghan et al., 2012; Clark, 2006; Clark et al., 2004; Crosas, 2011; Duerr et al., 2011; Juty, Le Novère, & Laibe, 2011; Lagoze et al., 2006; Lee & Stvilia, 2012; Michener et al., 2011; NISO/NFAIS, 2013; Paskin, 2010; Pepler & O’Neil, 2008; Tonkin, 2008; Vitiello, 2004). Quality is usually defined as “fitness for use” (Wang & Strong, 1996). Quality is multidimensional and contextual and there could be tradeoffs among different quality dimensions (Eppler, 2003; Stvilia et al., 2007). Table 2.3 shows the definitions of the quality dimensions and sources referencing the dimensions. The following section discusses seven quality dimensions in more detail: simplicity, opacity, verifiability, contextuality, interoperability, actionability and granularity.

2.2.4.1 Simplicity/Transparency & Opacity. Identifiers within different contexts have different requirements on their strings. In the context of data aggregation, communities prefer transparent and simple strings (Berners-Lee, 1998). Information about the characteristics of data objects encoded in identifier strings in a transparent way can be helpful in the disambiguation, aggregation, or clustering of the data objects along those characteristics. On the other hand, when

Table 2.3. Definitions of the Quality Dimensions and their Sources

Dimensions	*Definitions*	*Sources*
Uniqueness	The requirement that one identifier string denotes one and only one data object	Altman & King, 2007; Michener et al., 2011; Lagoze et al., 2006; Paskin, 2010
Persistence/ Volatility/ Legacy support	The requirement that once assigned, an identifier string denotes the same referent indefinitely	Altman & King, 2007; Berners-Lee, 1998; Brand et al., 2003; Callaghan et al., 2012; Duerr et al., 2011; Michener et al., 2011; NISO/NFAIS, 2013; Lagoze et al., 2006; Paskin, 2010; Tonkin, 2008; Vitiello, 2004
Simplicity/Transparency	The degree of cognitive simplicity of an identifier string	Berners-Lee, 1998; Duerr et al., 2011; NISO/NFAIS, 2013; Tonkin, 2008
Opacity	The extent to which the meaning can be inferred from the content, structure or pattern of an identifier string	Brand et al., 2003; Clark, 2006; Duerr et al., 2011; Michener et al., 2011; NISO/NFAIS, 2013; Tonkin, 2008
Verifiability	The extent to which the correctness and validity of an identifier string is verifiable or provable	Akhondi el al., 2012; Duerr et al., 2011; Juty et al., 2012; Tonkin, 2008
Contextuality	The degree to which an identifier system and string meets the needs of a targeted community	Clark et al., 2004; Juty et al., 2012; Tonkin, 2008
Compatibility	The ability to use with the main internet naming schemes (i.e., URL or URN)	Duerr et al., 2011
Interoperability	The ability to use an identifier system and string in services outside of the direct control of the issuing assigner	Altman & King, 2007; Berners-Lee, 1998; Duerr et al., 2011; NISO/NFAIS, 2013; Paskin, 2010; Vitiello, 2004
Actionability	The ability of the identifier system to locate the object using an identifier string	Altman & King, 2007; Brand et al., 2003; Callaghan et al., 2012; Duerr et al., 2011; Juty et al., 2012; Michener et al., 2011; NISO/NFAIS, 2013; Lagoze et al., 2006; Paskin, 2010; Tonkin, 2008; Vitiello, 2004
Granularity/Flexibility	The extent to which the identifier system allows referencing data at a different granularity	Juty et al., 2012; Michener et al., 2011; Tonkin, 2008; Vitiello, 2004;
Authority	The degree of reputation of an identifier system in a given community	Altman & King, 2007; Duerr et al., 2011; NISO/NFAIS, 2013; Tonkin, 2008
Scalability	The ability of an identifier system to expand its level of performance or efficiency (e.g., support RDF)	Duerr et al., 2011; Juty et al., 2012; Lagoze et al., 2006
Security	The extent to which the resource of an identifier system is protected from unauthorized administrative access or modification	Duerr et al., 2011; Juty et al., 2012; Tonkin, 2008

the data is sensitive, opaque identifier strings are preferred (Clark, 2006; NISO/NFAIS, 2013). Opaque identifiers could be more robust as they are not sensitive to changes in the characteristics of data (e.g., entity name change) (Tonkin, 2008).

2.2.4.2 Verifiability. Identifier strings often have a complex syntax. The complexity causes various issues related to verification and validity of the strings. Often checksums or other error-correction mechanisms are used to ensure identifier string validity. Identifier string verification for digital resources can be relatively simpler than the one for physical resources. Network connection might provide a quick solution, checking the correctness or validity of the strings by returning the associated data objects.

2.2.4.3 Contextuality. Many identifier systems are developed to meet specific community's needs. Data-driven research trends also accelerate the use and development of community-driven identifiers and repositories (Erway, 2012). The large amount of and various types of research data require more sophisticated curation, including the development of identifiers schemas, which are tailored towards the community's data management needs (Clark et al., 2004; Juty et al., 2011). In addition, an identity tension exists on determining the type (i.e., domain and entity type) of a URI (Halpin, 2011). Berners-Lee (2003a, 2003b) takes a position that the type of a URI is whatever the owner intended. Hayes (2004), meanwhile, takes a different position that the type of a URI is determined by linked structured resources (i.e., RDF triples) within the Semantic Web. Halpin (2011) grafts community context onto the type of a URI, so the type could be defined as the use of URIs by a community. Therefore, Halpin's approach on the type of a URI accords closely with the developmental needs of contextual identifier schemas.

2.2.4.4 Interoperability. Interoperability aiming at a shared understanding of data can be defined as the exchange and use of information in an efficient and uniform manner across multiple organizations and systems (Abbott, 2009). In this context, Paskin (2008) identified three distinguishing identifier interoperabilities in the aspects of syntax, semantics and community. Syntactic interoperability is the ability of systems to read and recognize more than one identifier syntax string within an identifier string. For example, LSIDs use a form of URN and can include an identifier string, such as NCBI Accession Number, within their syntax strings (Clark et al., 2004). Semantic interoperability is the ability of systems to determine how two associated data objects are semantically related. It can be attained by using widely used structured metadata or ontologies. The CIDOC CRM, Online Information Exchange (ONIX) and Resource Description and Access (RDA) can be used as the standards for semantic integration (Dunsire, 2007; Paskin, 2008). Finally, community interoperability is the ability of systems to collaborate and communicate between different identifier systems without any restrictions on each system's use. The community interoperability first requires community policies that state willingness to share and compare their metadata management plans with other communities; otherwise, the interoperability can not be viable (Paskin, 2008). According to Paskin, these three aspects are dependent. Syntactic interoperability is a required condition of ensured semantic interoperability, which is necessary to ensure community interoperability. Pabón et al. (2013) mentioned that all of "legal compatibility, semantic interoperability, technical aspects of information systems, organisational cooperation and a favourable political climate" are necessary for interoperable services in reality (p. 1803).

2.2.4.5 Actionability. In general, resolution systems are bridge systems including both input and output. The input is an identifier string as a key, and the output is the current information associated with the identified objects (Paskin, 2010). IDF strongly recommended that identifier systems to have a resolver to track down dynamic locations of data objects. An identifier system with a resolver does not require any change of identifier strings, even when the physical location of the identified object is changed.

2.2.4.6 Granularity. Research can be driven by multiple research data. A dataset usually contains multiple data files. According to Lee and Stvilia (2012), many institutional repositories are storing various types of research data files (e.g., single data files, compressed data files and database files). The need for different granularity happens when a researcher wants to cite only one specific file from a dataset (Michener et al., 2011): for example, if a dataset contains one hundred files and the researcher wants to cite only one file in that dataset. To support this need, the identifier system needs to support data referencing at multiple granularity.

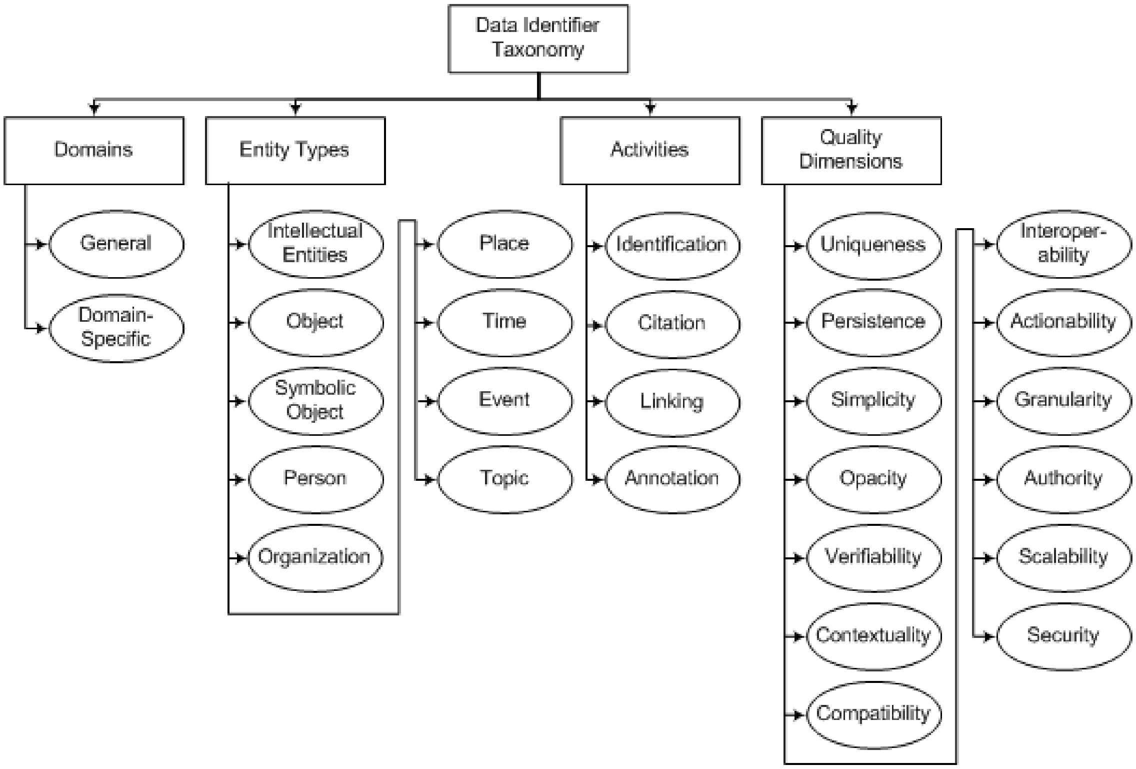


Figure 2.2. Data identifier taxonomy

Drawing from all of the discussions in the literature, the taxonomy above presents a summary of the concepts related to identifier schema design, use and evaluation (Figure 2.2). The taxonomy consists of four main categories and their sub-elements.

2.3 Practical Uses of the Identifier Systems

The 14 identifier systems reviewed in the previous section were selected for an analysis based on the characteristics defined by the taxonomy, and the conceptual analysis of those identifiers was conducted based on technical specifications, user documentation, and published journal articles. Table 2.4 briefly summarizes the results of the analysis. The empty cells within the table indicate the absence of a particular property or use, and the cells marked with "Yes" indicate the opposite. This analysis has limitations. In some cases, the literature used in this analysis provided clues rather than a direct answer for individual cells, and many results obtained from the literature do not include comparative elements among the identifier systems. The results provide a conceptual understanding based on the literature analysis and require further research using empirical data. In the following subsections, we discuss the results of this analysis.

2.3.1 Domains and Entity Types

Six identifiers were identified as domain-independent identifiers: ARKs, DOIs, Handles, PURLs, URIs and UUIDs. These are primarily assigned to the Intellectual Entities within many institutional data repositories (Lee & Stvilia, 2012), and some of them (i.e., ARK and URI) can be assigned to author- and subject-related entities.

Table 2.4. Summary of Practices of Data Identifier Systems

	ARK	*DOI*	*Handle System*	*PURL*	*URI*	*UUID*	*NCBI Accession Number*	*CAS Registry Number*	*LSID*	*ISNI*	*ORCID*	*Research-erID*	*Open ID*	*GeoName ID*
Domains														
General	Yes	Yes	Yes	Yes	Yes	Yes				N/A	N/A	N/A	N/A	N/A
Domain-specific							Yes	Yes	Yes	N/A	N/A	N/A	N/A	N/A
Entity Types														
Intellectual entities	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes					
Object	Yes	Yes	Yes	Yes	Yes	Yes			Yes					
Symbolic object	Yes				Yes		Yes	Yes	Yes					
Person	Yes				Yes					Yes	Yes	Yes	Yes	
Organization	Yes				Yes					Yes				
Place	Yes				Yes									Yes
Time														
Event	Yes				Yes									
Topic	Yes				Yes									
Activities														
Identification	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Citation	Yes	Yes	Yes	Possibly	Possibly	Possibly	Possibly	Possibly	Yes	Possibly	Possibly	Possibly	Possibly	Possibly
Linking	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Annotation	Possibly	Yes	Possibly	Possibly	Yes	Possibly	Yes	Yes	Yes	Possibly	Possibly	Possibly	Possibly	Yes
Quality Dimensions														
Uniqueness	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Persistence	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Simplicity	Yes	Some	Some	Yes	Yes				Yes				Yes	
Opacity	Yes	Yes	Yes	Yes	Yes	Very	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Verifiability	Very	Yes	Yes	Yes	Yes		Yes	Very	Yes	Very	Yes	Yes	Yes	Yes
Contextuality							Yes	Yes	Yes					
Compatibility	Yes	Yes	Yes	Yes	N/A	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Interoperability		Yes	Yes						Yes		Yes			
Actionability	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Granularity		Yes												
Authority	Some	Very	Very	Some	Very	Some	Very	Very	Very	Very	Very	Very	Some	Some
Scalability	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Security	Very	Very	Very	Some	Some	Very	Very	Very	Very	Very	Very	Very	Some	

Before the mapping of the identifiers to the Intellectual Entities, it is worth thinking about the distinctions and issues between individual FRBR Group 1 entities. The distinctions are not easily, unambiguously defined, as we previously mentioned for the abstract entities in the digital environment (Buckland, 1998; Carlyle, 2004; Floyd & Renear, 2007; Renear et al., 2003). Along with the uncertainties, mapping the domain-independent identifiers to Intellectual Entity (i.e., a broader entity including all the abstract entities) is effective and efficient. Previous similar studies have not mapped the identifiers to the individual FRBR Group 1 entities, or concluded that the mapping is meaningless (ELAG, 2010; Halpin, 2008; Vitiello, 2004). For example, many identifiers, in the digital environment, provide access to the metadata descriptions of the information objects, which include all different types of entities. These six identifiers can be considered as being mapped to Intellectual Entities within the bibliographic universe. The mapping could be developed with better accuracy (e.g., mapping to each of the FRBR Group 1 entities), but it might also cause unnecessary or incorrect results. The current mapping provides an efficient mapping method for data identifiers.

ARKs and URIs can also be used with entities in other groups—namely, author and subject. ARKs can be assigned to various types of objects (e.g., digital, physical and intangible objects and living beings and groups) with a flexible and wide range of scopes (CDL, 2012). URIs are compatible with all the identifiers (see Table 2.2).

In biology, alphabetic letters express gene or protein sequences. Accession numbers from the NCBI assigned to the expressed alphabetic records can be mapped to the Intellectual Entities and Symbolic Objects. The sequence records can be considered as intellectual concepts or the symbolic expression of intellectual concepts.

CAS Registry Numbers are associated with molecules of chemical substances, which are the smallest amount of a chemical substance. In most cases, the molecules are intangible and invisible to the naked human senses. An object assigned a CAS Registry Number is, therefore, a molecular expression of the substance written by chemical formulas and symbols. The registry numbers can be matched with Intellectual Entities and Symbolic Objects in the same manner as the NCBI Accession Numbers.

LSID is an identifier associated with data resources related to life sciences. The data include both concrete and abstract objects. LSID has a wide scope similar to domain-independent identifiers, but it is only applied to the resources in life sciences. It can be associated with protein or gene sequences by cooperating with various namespaces (e.g., GenBank, Protein Data Bank (PDB), GeneOntology) and data files in the field of life sciences (Clark et al., 2004). LSIDs can be mapped with the Intellectual Entities, Object and Symbolic Object.

Both ORCID and ResearcherID designed to associate with researchers can be mapped to the Person entity. Lastly, GeoNameID is an identifier that identifies accurate geographic location.

2.3.2 Activities

If the use of an identifier system in a particular activity is mentioned in the literature, the corresponding cell is marked with “Yes” in Table 2.4, and “Possibly” means that the identifiers seem to be applicable, but their use has not been reported in the literature. All the identifiers except UUID fully support the activity of identification. UUIDs are random numbers using the current time. They do not encode information about the properties of data objects they are assigned to. However, UUIDs can be used as a unique and persistent URI merging with a URN (Leach et al., 2005). All the identifiers can support the linking activity. All of them can be used as a URI (one of the requirements of linked data implementation) following the format of URIs, URLs, or URNs (see Table 2.2). ARKs, DOIs, Handle Systems and LSIDs are currently used as identifiers in different data citation models. The rest of the identifiers with URL- or URN-syntax also can be used as the identifiers in data citation (Altman & King, 2007; Crosas, 2011; Duerr et al., 2011; Michener et al., 2011) but their use has not been reported in the literature yet. For example, the American Psychological Association's (APA) publication manual only allows DOI and URL as citable identifiers (APA, 2010). This policy, however, is quite flexible and means that any identifier which has the URL syntax can be used in data citation. Three domain specific identifiers (i.e., NCBI Accession Numbers, CAS Registry Numbers and LSIDs), DOIs, URIs and GeoNameIDs are currently used within annotation activity (CAS, 2012a; Clark et al., 2004; GeoNames, n.d.; Paskin, 2005; Pruitt et al., 2009). The other identifiers do not seem to have any barriers for supporting annotation, but their use within the annotation activity has not been reported in the literature.

2.3.3 Quality Dimensions

As the survey lacks comparative analysis between the identifier systems, different levels indicating the conceptual extent of identifier systems' functions are used. The various levels include "Very," "Yes," and "Some," which represent the different values of quality dimensions of the data identifier systems. The ARKs, PURLs, URIs and LSIDs allow their schema users to generate identifier strings according to their own rules—a privilege that allows the identifiers to satisfy two conflicting dimensions: transparency and opacity. For instance, if the strings meet with the minimum requirements to be the schema strings, the remaining parts of the strings can be created for the users' convenience. In most cases, the strings generated by the users are transparent and simple. DOIs and Handle Systems also permit their schema users to create a part of their string—namely, the suffix of the string—so that they can also satisfy the dimension of simplicity/transparency. All the identifier strings except UUIDs are verifiable with an Internet connection, and ARKs and CAS Registry Numbers are additionally supported by checksum functions. The quality of interoperability is also important for the purpose of identifier synthesis (Paskin, 2008). DOIs are interoperable with the Handle system and many ISO identifiers (e.g., ISBN, ISSN, etc.), and the Handle systems share much of the technology (e.g., protocol) with DOIs (Altman & King, 2007). URIs are compatible with all the other identifiers that use URIs as their basis (see Table 2.2). LSIDs include a name authority within its syntax, such as the NCBI Accession Numbers embedded in LSIDs. ORCIDs share its syntax with ISNIs. Granularity is one of the more difficult quality dimensions. Most identifiers do not fully support multiple granularities. However, DOIs support identification at multiple granularities at Dryad, which is a repository for research data in biosciences (Michener et al., 2011). At Dryad, suffixes of assigned DOIs are generated by their own rules, displaying the relationship between data collection and a single data file within the collection. In addition, the identifiers (e.g., ARK, LSID, etc.) that support simplicity/transparency might potentially contain the granularity dimension using the same method as Dryad.

2.4 Activity Theory

Activity Theory (Leontiev, 1978; Vygotsky, 1962) has been used to analyze cultural practices of diverse areas, such as work, technology and education, within the developmental, historical and cultural contexts (Rogers, 2012). This theory provides a conceptual framework for analyzing the activity of informational artifacts (e.g., metadata, ontologies) and its context. Work initiated in the 1960s has evolved as a result of the efforts of several researchers and in different ways (Engeström, 1987; Wilson, 2006). The newer versions of the theory have been popular in several areas of study, such as analyzing the context of work, technology, or education (Rogers, 2012).

2.4.1 Proposition

The proposition of activity theory is that "consciousness is formed through activity" (Wilson, 2008 p. 120). According to Bedny, Seglin and Meister (2000), human consciousness is closely related with cultural and historical context (as cited in Wilson, 2008). In activity theory, the contexts can be considered as the mediators of activity as well as the activity itself (Nardi, 1996; Wilson, 2008).

2.4.2 Origin and Development

Activity theory originated from Soviet psychology as an alternative to behaviorism emphasizing the relationship between the stimulus and response of human consciousness. Initially, it focused on the theory of human consciousness; the theory evolved to highlight the activity influenced by the human consciousness (Wilson, 2008). Activity can generally be understood as interactions between a subject (e.g., an actor) and an object (e.g., an entity) in a community (Kaptelinin & Nardi, 2012; Leontiev, 1978). It can also be considered as the behavior of a human acting to transform something within a specific community (Kuutti, 1996; Wilson, 2008).

Although many different researchers have worked on developing activity theory, three key persons have significantly improved the theory: Lev Semyonovich Vygotsky (1896–1934), Alexei Nikolaevich Leontiev (1903–1979) and Yrjö Engeström.

2.4.2.1 Vygotsky's Activity Theory. Vygotsky's (1962) activity theory included a simple structure consisting of the mediating artifacts, subject and object (see Figure 2.3). In his theory, artifacts—including language, writing, mathematics, maps and other symbol structures—mediate activity (i.e., the relationship between subject and object) (Nardi, 1996; Wilson, 2008).

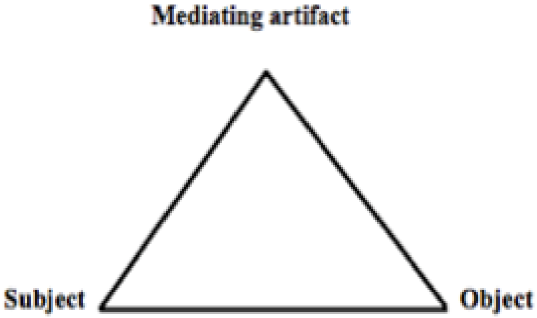


Figure 2.3. Vygotsky's Activity Theory (Wilson, 2008)

2.4.2.2 Leontiev's Activity Theory and its principles. In a later study, Leontiev (1978) developed the theory to incorporate more mediatory notions (e.g., cultural, historical and work related) around activity. However, the notions became concrete as mediators with Engeström's (1987) work. Leontiev also developed the hierarchical structure and relationships of activity constructed by activity, actions and operations. According to Leontiev (1978), an activity can have many actions, and an action can have many operations (see Figure 2.4). The activity structure can also be connected with human motivation, consisting of motive, goal and conditions. Wilson (2006) diagrammed the relationships (see Figure 2.5).

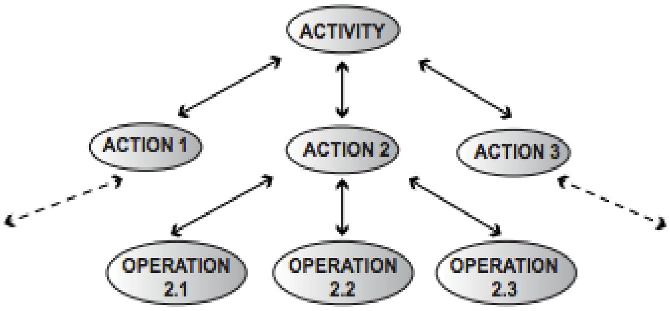


Figure 2.4. Hierarchical structure of activity (Kaptelinin & Nardi, 2012)

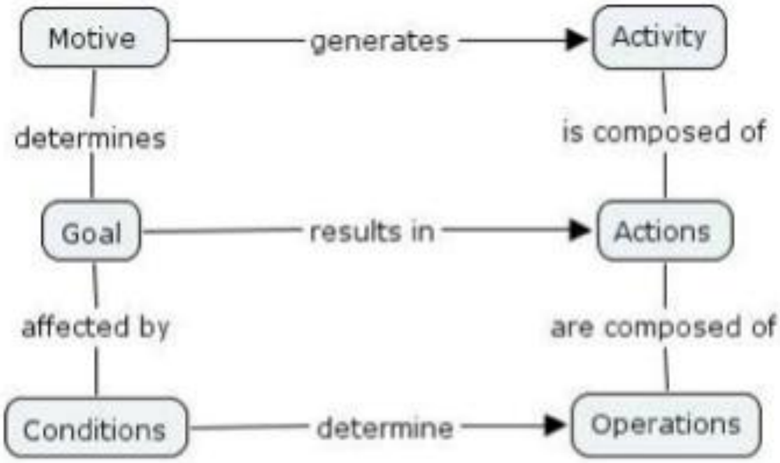


Figure 2.5. Activity and human motivation (Wilson, 2006)

The three levels of the activity are defined as follows (Leontiev, 1978):

1. Activities: Motives that provide a minimum meaningful context for understanding the individual actions.
2. Actions: Goals characterized by conscious planning.
3. Operations: Conditions or routinized behaviors that require little conscious attention.

Leontiev's activity theory can be explained using five basic principles identified by Kaptelinin and Nardi (2006). The principles effectively describe the cultural and historical aspects of activity theory. Object-orientedness, the first principle, is that human activity is directed toward objects. Activity can be broadly understood as interactions between a subject and an object. Therefore, subjects act for their objects. For example, when a person rings a doorbell, the person is the subject, the doorbell is the object and the ringing is the activity. The subject is acting to achieve the object, and the object is the ultimate reason for the activity.

The second principle is the hierarchical structure of activity. As previously mentioned, an activity consists of an activity, actions and operations whereas the structure is related with motives, goals and conditions (see Figure 2.5). Learning how to drive is a good example from

Kaptelinin and Nardi (2012) for explaining a multilayer activity system. A subject will want to learn how to drive when he reaches a certain age, which is in the cultural context a motive. Learning how to drive a car is an activity that requires multiple actions. For the activity, the subject needs to register for a driving school and buy instructional materials with conscious planning, which are the actions of the activity. When the subject is in a lecture, he will take notes, but he might not recognize that he is actually in the process of writing, which is an operation with little conscious attention. Instead, he is focusing on traffic rules.

The third principle is mediation. Tools mediate activity and contain various things from the socio-cultural and -technical context, such as policies, technologies, norms and practices. In Engeström's activity system model (1987), the tools became concrete as mediators.

The fourth principle is internalization and externalization, which states that human consciousness is formed by external activities being internalized or by internal concepts being exposed as external activities. To explain the principle, Kaptelinin and Nardi (2012) used an example of young children doing arithmetic. Children often use their fingers to count numbers, which is an external activity; the values calculated using their fingers are internalized to form consciousness. Kaptelinin and Nardi also offered another example: sketching a design idea. A designer sketches her idea on a whiteboard. The sketch comes from her internal concepts and is fixed through her external drawing activity.

The last principle is development. This implies the importance of understanding how activity theory operates. Activities undertaken by subjects are directed toward objects. The subjects can share the objects within a community to find a preferred outcome. Many different socio-cultural and -technical tools mediate the relationship among subjects, objects and the community (Wilson, 2008).

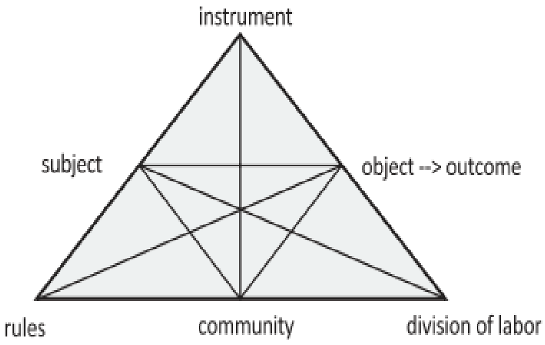


Figure 2.6. Engeström’s activity system model (1987)

2.4.2.3 Engeström’s Activity System Model and its Contradictions. Engeström (1987) extended the theory along Leontiev’s (1978) mediatory notions from Vygotsky’s (1962) original theory, diagramming it as activity system model (see Figure 2.6).

The activity system model contains four additional concepts related to an activity: instrument, community, rules and division of labor. The individual concepts are defined as follows (Engeström, 1987):

1. Instrument: Artifacts, signs and means that mediate the subject and object.
2. Community: Those who share the same object.
3. Rules: A set of agreed-upon conventions and policies in a community.
4. Division of labor: The primary means of classifying the labor.

Engeström’s model (1987) reflects individual activity as well as collective activity. As the first step of the extension, Engeström added an element of community at the interaction between a subject and an object, creating a three-way interaction as shown in Figure 2.7. The interaction allows for subjects to share objects within the community, which produces a collective activity.

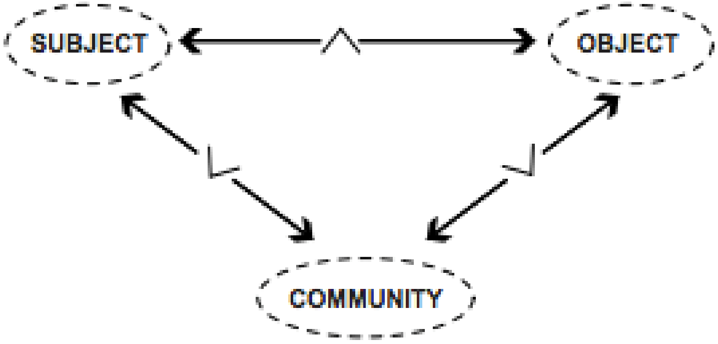


Figure 2.7. The first step of extension from Leontiev's theory (1978)

In the second step of the extension, Engeström suggested using mediators (i.e., instrument, rules and division of labor) for three interactions from the first extension (see Figure 2.6). Instrument is a mediator for the interaction between subject and object, rules are a mediator for the interaction between subject and community, and the division of labor is a mediator for the interaction between community and object. In addition, the model includes an element of outcome, which is produced by the activity system. The outcome resulting in a system can be used by other activity systems. It indicates the continuity of activity systems and networks of interrelated activities (Engeström, 1999; Kaptelinin & Nardi, 2012). An activity system is usually not sufficient to complete real-life projects, such as the user interface design of a computer application. To finish the project, many different instruments, tasks, rules and teams must be harmonized. Designing the user interface would require several partial outcomes, which would need to be integrated to build the user interface.

In addition, Engeström (1987, 1999) argued that activity systems are constantly developing. The idea becomes the core of the networks in activity systems. He further suggested three generations of activity theory (1999). The first generation was affected by the concept of mediation (Vygotsky, 1962). Figure 2.3 summarizes the first generation of activity theory. The second generation was influenced by Leontiev's work (1978) and diagrammed by Engeström (1987). Figure 2.6 represents the second generation of activity theory. Figure 2.8 describes the third generation activity theory developed by the concepts of collective activity and structure of the social world. The third generation of activity theory requires a minimum of two interacting activity systems and operates with contradictions that exist in and between activity factors. The contradictions play a role in forcing the evolution of the activity systems.

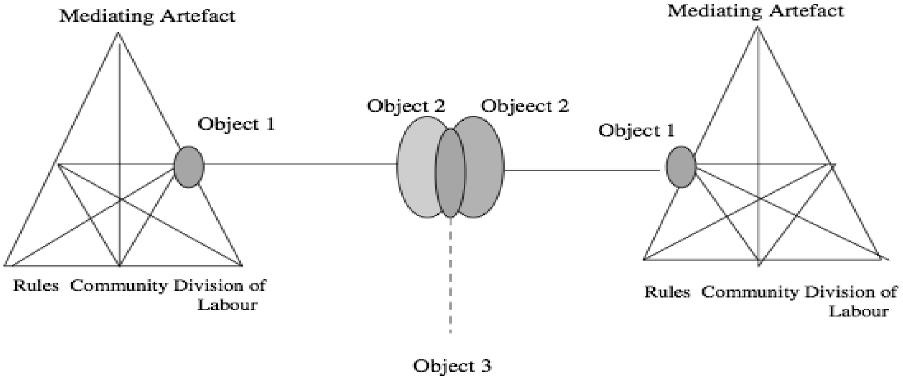


Figure 2.8. Third generation activity theory model (Engeström, 1999)

Engeström (1987) discussed four types of contradictions in activity systems. The first is the inner contradiction that exists within each component (e.g., subject, instrument, rules) of an activity. A subject can use an instrument, but the instrument is affected by various other mediating means, such as costs and legal regulations. For example, a graphic designer can use the best possible software for his project, but the software is expensive. Thus, the designer uses more affordable software. This is not an aspect that activity theory recognizes.

The second contradiction occurs among the components of an activity system. Kaptelinin and Nardi (2012) used the example of medicine: A certain type of medicine might cause an allergic reaction in certain patients. The instrument (in this case, medicine) does not function with the object (in this case, certain patients). The example shows the contradiction between an instrument and an object. Wilson (2008) used a different example: Adopting new technology can be a slow process because of inflexible bureaucratic labor structures. This might be a contradiction between an object and a division of labor. These examples describe specific situations that can exist between the components and can prevent the components from functioning.

The third contradiction explains the tension existing between old systems and new systems. In real life, some people resist changing current systems to new systems because they are dominantly using the old systems and adopting the new systems requires expensive cost (e.g., time, labor).

Finally, the fourth contradiction discusses the conflict arising among multiple activity systems. When an outcome is produced by multiple activity systems, none of the systems should fail to produce their own outcome. Kaptelinin and Nardi (2012) mentioned the example of a patient who had surgery and is influenced by two different activity systems: the actual surgery and the follow-up rehabilitation. Although the activity of the surgery had a positive outcome on the patient, if the activity of rehabilitation did not come with a positive result, the outcome of the surgery activity is no longer positive.

Activity system model considers the contradictions as tensions among the components (Wilson, 2008). The tensions can be used as dynamics for activity systems that are constantly evolving. Attempting to reduce the tensions can be meant to develop the activity system by incorporating other related activity systems (Engeström, 1987, 1999). To solve the conflicts, an activity system tends to be more complex and collective (Engeström, 1999). As mentioned previously, in complex real-life projects, a single activity system is not sufficient. Such projects need to be applied using the networks of activity systems, which are motivated by the contradictions of activity systems (Engeström, 1999; Kaptelinin & Nardi, 2012).

2.4.3 Application

Activity theory has been used in various studies examining sociotechnical context (e.g., policies, systems, practices). The theory has particularly been used in studies investigating metadata or research data practices.

Stvilia (2007) developed a model to evaluate ontology quality by using a theoretical framework consisting of activity theory (Engeström, 1987; Leontiev, 1978) and an information quality assessment framework (Stvilia, 2006). To conceptualize the quality, he first conceptualized an activity system reflecting the Morphbank biodiversity research data repository's system. The system has contained all of the individual mediators of the activity system model, and allowed for an understanding of the cultural and community context, contextual activities and entity relations. Based on the activity system, Morphbank's user activities include determining the specimen's taxon; marking, tagging and aggregating a specimen within a taxon; identifying and finding errors; and evaluating the quality of the taxonomy. Stvilia has also developed scenarios (Go & Carroll, 2004a, 2004b) for each user activity to support the entity relations within the Morpbank's cultural and community context. The identified Morphbank activities were mapped using Stvilia's information quality assessment framework, and Stvilia then categorized the activities within activity types describing information quality problems, dimensions and metrics.

Huang et al. (2012) proposed scientists' perceptions of and priorities for data quality dimensions and skills needed in genome annotation. The study also used a methodology consisting of activity theory (Leontiev, 1978; Nardi, 1996), scenario-based design (Go & Carroll, 2004a, 2004b), and information quality assessment framework (Stvilia et al., 2007). Huang et al. used a method of interviewing to conceptualize genome annotation processes and activities based on activity theory. The activity theory helped identify the tools (e.g., rules, instruments, etc.) that affect the genome annotation and its related skills. The high level understanding from the interviews was crystallized with scenario-based task analyses. They then used the contextualized understandings to develop a survey instrument for prioritizing the quality dimensions and skills in genome annotation work.

Borgman et al. (2007) studied the data practices of a habitat ecology community, which started using embedded sensor networks. They used a combined method consisting of interviews and field observations, which allows constructing a description of data practices based on community context. The researchers developed interview questions guided by the principles and the conceptualizations of the activity structure of activity theory (Engeström, 1990). As activity theory analyzes human activities in a target community, the interview questions could contain various topics (community's motives, cultures, shared tools, rules, divisions of labor, power relations, etc.). The interviewees were habitat ecology scientists and their partners in computer science and engineering.

Stvilia et al. (2013) also studied the data practices of the Condensed Matter Physics (CMP) community. They included activity theory (Engeström, 1990) within their theoretical framework as well as Stvilia's information quality assessment framework (Stvilia et al., 2007) and value-based quality model (Stvilia & Gasser, 2008). The framework helped identify the context of the CMP community as the interrelation between cultural and community structures and community-specific activity structures (Engeström, 1990) and conceptualized the typified activities of data practice, including quality problems (Stvilia et al., 2007; Stvilia & Gasser, 2008). To conduct the study, Stvilia et al. (2013) used multiple data collection methods consisting of semi-structured interview and survey. The semi-structured interview could be well combined with activity theory to conceptualize CMP community activity systems. The interview questions contain various components identifying actions, tools, rules, roles, strategies, division of labor, etc. The details about semi-structured interviews are discussed in the following methods response section.

Many studies have used activity theory as a guiding knowledge tool and the predictive mechanism as an explanatory framework. To achieve the goal of the studies with activity theory, the researchers have normally utilized a data collection method or multiple methods (e.g., scenario-based analysis, semi-structured interview, etc.) supplementing the theory with identifying experiential narratives (Borgman et al., 2007; Huang et al., 2012; Stvilia, 2007; Stvilia et al., 2013). The narrative data help arrange a complex activity structure into detailed lists of elements of activity theory (Carroll, 1997). The three levels of activity and various mediators around the activity help researchers recognize how an activity is defined and evaluated. The activity also consists of context that enables the researchers to study cultural, historical and practical factors in communities (Nardi, 1996; Wilson, 2008).

2.4.4 Strengths and Limitations

As previously mentioned, activity theory can be considered as a guiding knowledge tool to study contexts. The importance of studying contexts has been proven in various fields of study, such as psychology, anthropology and computer science (Nardi, 1996). For instance, in real-life projects, an individual cannot accomplish her or his tasks without support from other individuals, tools and social groups. In this sense, various fields of study motivate the study of contexts. To study the complex contexts of real-life projects, various approaches—activity theory, situated action models and distributed cognition—have been developed with different descriptions of context. A comparison between the different approaches can provide an understanding of the strengths and/or limitations of activity theory.

2.4.4.1 Situated action models. Situated action models have different approaches to interpreting an activity than activity theory. The models mainly focus on the improvisatory nature of human activity, which emphasizes an emergent situation given for an activity and the spontaneous human acts in the particular setting. The relationships between individuals and specific environments describe the activity proposed by the models (Lave, 1988; Suchman, 1987). For example, people go to supermarkets to shop for certain items and find the items in certain aisles. The activities are spontaneous acts occurring from the relationships between the shoppers' needs and the supermarket's arrangement of merchandise. Furthermore, situated action models emphasize the objective representation of human activity in a setting rather than a subjective representation including various cultural and historical contexts.

There are two major differences between activity theory and situated action models. The first is their activity structures. The activity from activity theory is formed by the relationships between the subject and object, including many different mediating factors; on the other hand, the activity from situated action models is spontaneous and responsive to human behaviors based on the given environmental setting (Nardi, 1996). The second difference relates to the continuous structure of the activity, which explains how the structure shapes an activity. Activity theory contains various activity-mediating elements for understanding the activity structure and roles. The mediators, such as artifacts, institutions and cultural values, are utilized within the activity systems as the means that have continuous properties to shape activities. Unlike the activity theory, situated action models consider continuous and durable structures, which conflict with the concept of the emergent situation, as independent factors of particular situations.

The differences between two approaches explain the benefits of using each approach in studying a context and simultaneously describe the strengths and limitations of using activity theory. Activity theory, compared to situated action models, is a relevant theory for studies of a larger scope and a longer-term analysis (Holland & Reeves, n.d.; Nardi, 1996). On the other hand, the theory would have limitations in investigating responsive and spontaneous situations with a shorter-term analysis than situated action models (Nardi, 1996). Furthermore, Carroll (1997) suggested using activity theory along with different facets of scenario-based design (Go & Carroll, 2004a, 2004b) for studies of interaction between human activities and technologies. According to Carroll (1997), such contextual studies require full understanding of the human and the technology along with the cultural and community context. Such studies can be well guided with comprehensive narratives of use situations or experiences, representing human activities with systems.

2.4.4.2 Distributed cognition. The distributed cognition approach stems from sociology, cognitive science and psychology and emphasizes a cognitive system defined as the relationships between individuals and the artifacts they use (Hutchins, 1991). The cognitive system can consider the activity from the activity theory. Hutchins (1991) used the activity of flying a plane, which corresponds with the cockpit system of the plane, as an example. The system completes its goal of flying, with the cockpit and pilots as a combined system. The unity of individuals and the artifacts (in this case, pilots and cockpit) toward its goals produces a cognitive system. In addition, the distributed cognition approach highlights the coordination among agents, which are individuals, and the artifacts they use. The coordination can be compared with mediators of activity theory. However, in this approach, the “collaborative manipulation” only includes the concepts of sharing goals and plans between individuals and particular characteristics of the artifacts in the system (Hutchins, 1987).

Overall, distributed cognition has a similar approach to defining activity structure using activity theory. First, a cognitive system is formed by the system goal, which is similar to the concept of object in activity theory (Nardi, 1996). In activity theory, an object forms an activity in combination with the subject. The only difference between the two approaches is that the system goal of distributed cognition does not contain subjects’ consciousness that activity theory has. However, the activity structures from the two originated from motive and goals. Second, the continuous structure of activity is also similar in the two approaches. Activity theory uses various mediating factors to shape an activity while distributed cognition uses the concept of collaborative manipulation.

Although the two approaches have similar structures in terms of activity, a distinct conceptual difference also exists between them. The distributed cognition views humans and artifacts as conceptually identical. The agents in this approach include both individuals and artifacts, and the agents are based on the concept of unity to develop a cognitive system. On the other hand, activity theory includes conceptual differences among human activity, humans and things. The activity theory, which focuses on motive and human consciousness, sees things and artifacts as mediating factors that affect human activities. According to Nardi (1996), the position of activity theory has more potential for guiding the studies of the human–computer interaction field. In activity theory, humans can be seen as an active entity that controls the artifacts they use. Thus, the purpose that humans have for their artifacts can be reflected in activities; meanwhile, the position of distributed cognition does not allow for a creative purpose from the human on its activity system. The uneven structure between individuals and things in activity theory could be strength for exploring more details about human behaviors whereas distributed cognition has other strengths compared to activity theory. Because of its emphasis on systems, the distributed cognition approach is appropriate for detailed analyses of particular artifacts and able to provide broadly applicable and stable system design principles, rather than using human-focused analysis. In other words, activity theory would have limitations for studying system design without focusing on human behaviors.

Defining a quality (Wang & Strong, 1996) identifier schema for research data requires contextual analyses of communities using identifier schemas (Lee & Stvilia, 2012; Stvilia et al., 2007). Different communities use different identifier schemas for their different uses and needs. For example, a research institution wants to build an institutional data repository, and data curators from the repository then need to adopt identifier systems for their needs. The curators would, before some adoptions, want to understand the uses of various identifier schema based on their functionalities (e.g., domains, entity types, activities and quality dimensions). Sociotechnical aspects (i.e., policy, systems, practices, division of labor, etc.) of the institutional repository would also be important factors for adopting identifier systems. To conduct a study of a community's data practice and its use of identifier system(s), activity theory is an appropriate knowledge tool. The theory helps guide studies of cultural and community context integrated with humans' activities and their uses of technology.

2.5 Information Quality Assessment Framework

The IQ Assessment Framework has been used as a knowledge resource and as a guide to manage information quality. This framework defines general relationships among information activity types, quality problem types, related quality dimensions, and metrics. To understand and use the framework, reviewing concepts, structure, previous applications, and strengths and limitations of this framework is essential.

2.5.1 Concepts

2.5.1.1 Information. Along with analyses of many different definitions of information and based on research purposes, Stvilia et al. (2007) defined *information* as “data plus the context of its interpretation and/or use” (p. 1721). In order to help understand the definition, they also described the hierarchy of information and defined *data* as “a raw sequence of symbols” and *knowledge* as “a stock of information internally consistent and relatively stable for a given community” (p. 1721). According to the definition of information, information cannot be interpreted without the understanding of surrounding context of information such as culture, community, and technology. In brief, information is comprised of data and its context, and a compendium of information within a community forms knowledge.

2.5.1.2 Information quality. To understand and manage information within a given community, one needs to recognize the context of community. Context consists of the relationship between a subject and an object and its mediating factors, such as tools, norms, policies, community, and division of labor (Leontiev, 1978; Vygotsky, 1962). Therefore, quality information should meet the needs of subjects and the requirements of objects. Stvilia and his colleagues (2007) adopted Juran’s (1992) definition of quality: “fitness for use,” in their study. For their purposes, a general definition of quality was appropriate to encompass the needs of both subjects and objects.

2.5.1.3 IQ dimensions or criteria. Stvilia et al. (2007) defined an *IQ dimension* as “any component of the IQ concept,” and they considered the dimensions as entity attributes to measure the IQ (p. 1722). The central part of Stvilia's framework is a taxonomy of IQ dimensions. The taxonomy comprises 22 IQ dimensions organized into three categories: *intrinsic IQ, relational or contextual IQ,* and *reputational IQ*. In a number of related literatures, the concepts of IQ dimensions and IQ criteria are interchangeably used.

2.5.1.4 *Intrinsic IQ.* The IQ dimensions in this category can be assessed with relatively less contextual understanding and with objective attributes or characteristics of information (e.g., spelling mistake, HTML validation, etc.). The dimensions in this category include accuracy/validity, cohesiveness, complexity, semantic consistency, structural consistency, currency, informativeness/redundancy, naturalness, and precision/completeness.

2.5.1.5 *Relational or contextual IQ.* The dimensions in this category can be measured with contextual understanding in a given community. The measurement requires analyzed information entities reflecting some external condition (Stvilia et al., 2007). The dimensions include accuracy, precision/completeness, complexity, naturalness, informativeness/redundancy, relevance, semantic consistency, structural consistency, volatility, accessibility, security, and verifiability. Furthermore, all these dimensions except the last three can be subcategorized as representational IQ, measuring the extent of mapping between an information entity and the external condition in a given context.

2.5.1.6 *Reputational IQ.* The dimension in this category, authority, measures the degree of reputation of an informant object in a given community or culture.

2.5.1.7 IQ metrics. IQ metrics are used to measure quality directly or indirectly along a particular quality dimension. The metrics are developed by analyzing the attributes and characteristics of an information entity and by connecting them to the IQ problems (Stvilia, 2006). Stvilia's Framework includes 41 general IQ metrics, and the general metrics can be reused to estimate contextual IQ problems (Stvilia et al., 2007).

2.5.1.8 Reference bases. *Reference bases* are the sources that affect IQ dimensions. According to Stvilia et al. (2007), there are two types of reference bases. The first includes culture, language, norms, and conventions, and the second is the context of a given community (e.g., actions, goals, roles, and best practices).

2.5.1.9 IQ problems. Stvilia et al. (2007) defined an *IQ problem* as "occurring when the IQ of an information entity does not meet the IQ requirement of an activity on one or more IQ dimensions" (p. 1722). According to Stvilia et al. (2007), major sources of the IQ problems consist of both static and dynamic sources. *Mapping*-related IQ problems come from static sources and the IQ problems caused by changes to the information entity, such as changes to the underlying entity or condition, and context changes come from dynamic sources.

2.5.1.10 Static IQ problems. Mapping-related IQ problems occur "when there is incomplete, ambiguous, inaccurate, inconsistent, or redundant mapping between some state, event, or entity and an information entity" (Stvilia et al., 2007, p. 1722).

2.5.1.11 Dynamic IQ problems. Dynamic IQ problems are dependent on a given community context that includes culture and sociotechnical structures (Engeström, 1987; Stvilia et al., 2007; Vygotsky, 1962). Any change (temporal or spatial) in context influences the understanding and evaluation of the IQ. The IQ change can be direct to an information entity or indirect stimulating underlying entity or condition. Furthermore, the change can be positive to the IQ, eliminating or diminishing the problems, or negative to the IQ, reducing an IQ level on an IQ dimension.

2.5.1.12 Information activity types. *Information activity* is context of a given community (Engeström, 1987; Leontiev, 1978). The context is comprised of subjects, objects, actions, instruments, rules, and division of labor in a given community. Changes to any element within the context can change the IQ and its problems. Therefore, understanding and organizing categories of information activities based on IQ problem sources is important. Stvilia et al. (2007) identified the following four activity types in their IQ Assessment Framework:

2.5.1.12.1 Representation dependent activities. They depend on "how well one information entity represents another entity or some condition" (Stvilia et al., 2007, p. 1724).

2.5.1.12.2 Decontextualizing activities. They use "information outside its original context of creation" (Stvilia et al., 2007, p. 1724).

2.5.1.12.3 Stability dependent activities. They depend on "how stable the information or its underlying entity is" (Stvilia et al., 2007, p. 1724).

2.5.1.12.4 Provenance dependent activities. They depend on "the quality of metadata of the information's provenance, mediation, and upkeep" (Stvilia et al., 2007, p. 1724).

2.5.2 Structure and Employment of the Framework

2.5.2.1 Structure. Stvilia's IQ Assessment Framework is comprised of activity types, sources of IQ problems, a taxonomy of IQ dimensions, and reference bases, and they have complex relationships among themselves (see Figure 2.9). The taxonomy of IQ dimensions is the central part of this framework. The 22 IQ dimensions in the taxonomy are organized within three categories (i.e., intrinsic, relational, and reputational), and each dimension has a proposed set of generic IQ metrics. The component of reference bases has relationships with each category of IQ dimensions. Culture, language, norms, and conventions are the reference bases of intrinsic and reputational IQ dimensions, and activity system context includes actions, goals, roles, genres, community, etc. as the reference bases of relational and reputational IQ dimensions.

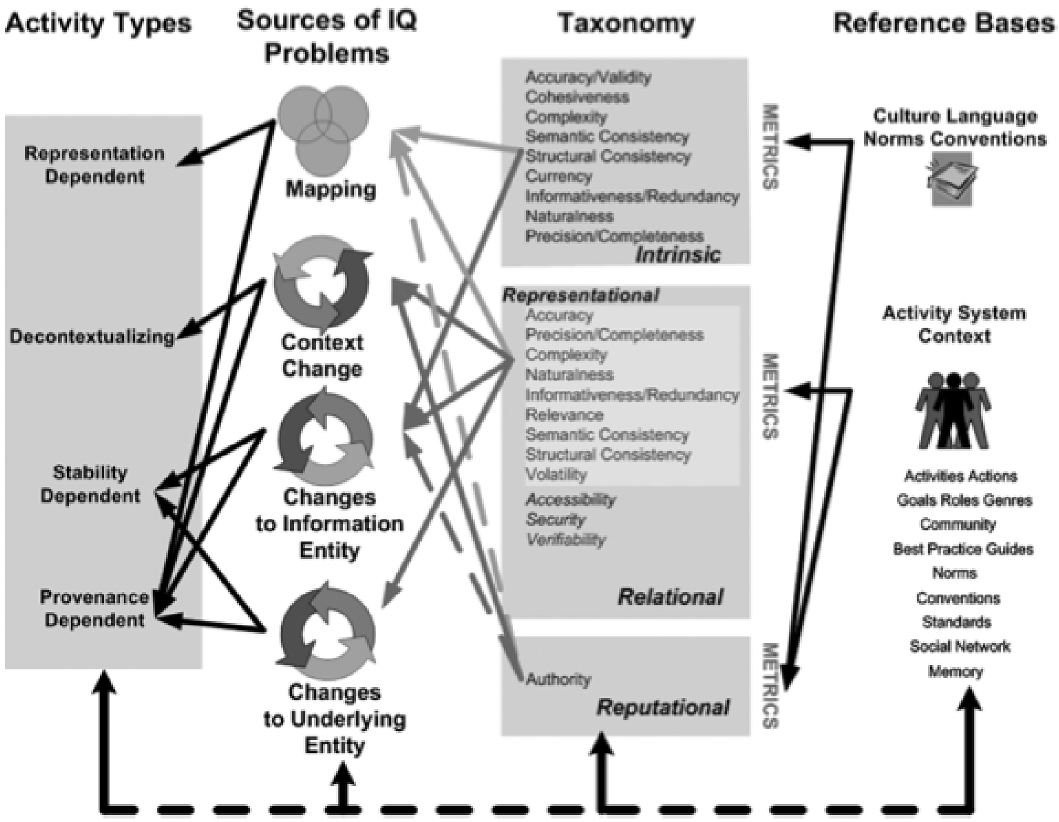


Figure 2.9. Conceptual model of IQ measurement (Stvilia et al., 2007, p. 1723)

The taxonomy part also has relationships with sources of IQ problems and activity types that are identified by Stvilia and his colleagues. Mapping-related IQ problems influence representation-dependent activities and provenance-dependent activities and tend to be IQ problems in intrinsic and relational dimensions. IQ problems with context change affect decontextualizing activities and provenance-dependent activities and can be prone to IQ problems in relational and reputational dimensions. IQ problems caused by changes to an information entity inform stability-dependent activities and provenance-dependent activities and have relation with intrinsic and relational dimensions. Lastly, IQ problems with changes to underlying entity influence stability-dependent activities and provenance-dependent activities, and only relational dimensions can be developed for this type of IQ problems.

2.5.2.2 Employment. To develop a context-specific IQ assessment model, the analysis of activity system (e.g., subjects, objects, instruments, rules, norms, and practices) within a specific community is the first step (Engeström, 1987; Stvilia et al., 2007; Vygotsky, 1962). After the analysis, the activities are organized by the types of activity in the framework. The next step is to identify and map the source of IQ problems to the activity types. Mapping with taxonomy and reference bases is the next step. IQ dimensions in the taxonomy, general IQ metrics, and reference bases are identified for the sources of IQ problems and activity types.

Once the activity system is identified with Stvilia's framework, the activities can be decomposed into the sub-elements (i.e., actions, operations, roles, and tools) of Activity Theory (Engeström, 1987; Leontiev, 1978). By doing that, one can analyze the relationships among the elements through information use scenarios (Go & Carroll, 2004a; Stvilia et al., 2007). Also, analyses of the information entities and their attributes are required to develop activity-specific IQ metrics. The final step of employment of this framework is IQ measurement aggregation. Since IQ measurements are context-specific, the measurements need to be aggregated and presented in "a tractable and actionable way" to help the stakeholders' information selection and reasoning (Stvilia et al., 2007, p. 1725).

2.5.3 Application

Stvilia’s framework has been used in various studies and in various domains examining information quality. The framework has particularly been used in studies investigating metadata or scientific data practices.

Stvilia (2007) developed a model to evaluate ontology quality by using a theoretical framework consisting of activity theory (Engeström, 1990; Leontiev, 1978) and his IQ Assessment Framework (Stvilia, 2006). To conceptualize the quality, he first conceptualized an activity system reflecting the Morphbank biodiversity research data repository’s system. Stvilia has also developed scenarios (Go & Carroll, 2004a, 2004b) for each user activity to support the entity relations within the Morpbank’s cultural and community context. The identified Morphbank activities were mapped using Stvilia’s Framework, and Stvilia then categorized the activities within activity types describing information quality problems, dimensions and metrics. The proposed IQ model contained specific IQ dimensions, metrics, and measurement costs.

Huang et al. (2012) proposed scientists’ perceptions of and priorities for data quality dimensions and skills needed in genome annotation. The study used a methodology consisting of Activity Theory (Leontiev, 1978; Nardi, 1996), scenario-based design (Go & Carroll, 2004a, 2004b), and IQ Assessment Framework (Stvilia et al., 2007). By conducting interviews, Huang et al. conceptualized genome annotation processes and activities and developed use scenarios. They then used the contextualized understanding to develop a survey instrument for prioritizing the quality dimensions and skills in genome annotation work. In this study, Huang et al. suggested a different value of the Framework to conceptualize the activities’, not the measurements’, related data quality.

Conway (2011) established a conceptual understanding for the linking of archival quality and information quality research. He specifically developed an error model for digitized books by using content preserved in HathiTrust, a large-scale preservation repository. To conduct this study, Conway adopted Stvilia's Framework for assessing IQ (Stvilia et al., 2007) and scenario-based task analysis (Go & Carroll, 2004a). He first identified different error sources and potential error types for digitized books, and then analyzed with scenarios to validate and test the findings. The error model may directly influence the trustworthiness of long-term preservation repositories.

Stvilia et al. (2015) also studied the data practices of the Condensed Matter Physics (CMP) community. They included Activity Theory (Engeström, 1990) within their theoretical framework as well as Stvilia's IQ Assessment Framework (Stvilia et al., 2007) and value-based quality model (Stvilia & Gasser, 2008). The framework helped identify the context of the CMP community as the interrelation between cultural and community structures and community-specific activity structures (Engeström, 1990) and helped conceptualize the typified activities of data practice, including quality problems (Stvilia et al., 2007; Stvilia & Gasser, 2008). To conduct the study, Stvilia et al. mainly used semi-structured interviews and a survey. The interviews helped develop a survey instrument, which asks survey participants the perception of data quality. By analyzing the survey results, Stvilia et al. developed a model of data quality perceptions in CMP community.

2.5.4 Strengths and Limitations

As Stvilia's framework encompasses aspects of sociotechnical and cognitive contexts, the framework can be used for general assessment purposes of IQ and as a guiding framework to develop community-specific quality models (Stvilia et al., 2007). Since the framework is context-dependent, Activity Theory (Engeström, 1987; Leontiev, 1978) and scenario-based task analysis (Go & Carroll, 2004a) are often used as supplements of the framework to effectively conceptualize the specific context (Stvilia et al., 2007). The framework contains 22 IQ dimensions and 41 general IQ metrics that can be reused for context-specific IQ assessment. In practice, the framework has been used in different fields to evaluate the quality of information objects (e.g., scientific data, metadata, ontologies, health information, Web pages, Wikipedia articles, and digitized books) (Conway, 2011; Huang et al., 2012; Stvilia, 2007; Stvilia et al., 2007, 2015; Stvilia, Mon, & Yi, 2009). In addition, the variety of successful uses of Stvilia's framework indicates its generalizability, validity, and extendibility. The framework was/is mainly used to assess IQ measurements, but also is now used to conceptualize IQ activities. Stvilia's framework based on contextual research is evolving its domains and is proving its generalizability and validity through ongoing studies.

Although validity of the framework is demonstrated by many studies, Stvilia's framework is relatively young, and has primarily been used by Stvilia and his colleagues and students at Florida State University. However, the numbers of scholars who use the framework in their own research, and citation-counts on published paper in the Journal of the Association for Information Science and Technology (JASIST) in 2007 are continuously increasing.

CHAPTER 3

METHOD

3.1 Research Questions

The purpose of this study is to examine the practices of research data curation within IRs and to build a knowledge base for identifier system uses, functionalities, and perception of quality in the curation activities. The knowledge base can inform not only the policy-related identifier use and quality requirements of identifier systems but also can be used by librarians, data managers, curators, scholarly communities and publishers as a guiding tool in selecting an identifier system for their IRs. To achieve that objective the study will examine data curation activities, the curation activities for which identifiers are used, data types and their entities, and perception of identifier quality in institutional repositories. In particular, the study will answer the following research questions:

RQ 1. What are the types of data activities in IRs and what are the structures and metadata requirements of those activities?

RQ 2. What are the major types of research data and their entity types within IRs for which identifiers are used?

RQ 3. What is the awareness of IR curators about different currently available identifier schemas?

RQ 4. How do IR curators perceive the quality of identifiers for research data?

The first question examines the types of data activities in IRs and the roles and requirements of metadata, including identifiers used to discover and link research data. To answer the question, it is important to be aware what data curation activities occur and what kinds of tools, policies, rules, norms, or best practices exist. Also, examining different types of metadata, including identifiers for research data, is important in identifying connections between the metadata and research data and discovering the roles and requirements of the metadata (Stvilia et al., 2013; Willis, Greenberg, & White, 2012). The second question seeks the major types of research data and their entity types within IRs, which can inform IR managers about the needs for various kinds of identifiers to support search, discovery, linking, and disambiguation of those entities in data curation and use activities (Lee & Stvilia, 2014). The third question examines IR curators' identifier literacy, and the fourth question investigates IR curators' perception of identifier quality – the properties of identifier systems that make them useful and usable in the context of IR.

3.2 Research Design

The study was guided by Activity Theory (Engeström, 1987; Leontiev, 1978) and the IQ Assessment Framework (Stvilia et al., 2007). Activity Theory (Engeström, 1987; Leontiev, 1978) can be used for modeling the general context of data curation work in IRs. This context comprises a system of different activities of the work and their structures including different roles (e.g., providers, users, curators), types of data, tools and skills needed, rules and policies used, and mediation relationships among those structures. As the second guiding framework, IQ Assessment Framework (Stvilia et al., 2007) was used as a predictive mechanism for the relationships among information activity types, quality problem types, related quality dimensions, and metrics. In this study, the types of identifier quality problems, identifier quality dimensions, and identifier activity types are, within the IR context, the main subjects of the inquiry guided by Stvilia's framework. Ultimately, the theoretical frameworks guided the development of questions for an interview protocol, which will be used in data collection. The relationships between the frameworks and the interview questions are specified in Section 3.2.2 Data Collection and Analysis.

To collect data effectively, the study used semi-structured interviews. The semi-structured interview is a relevant method for a multifaceted and contextual study (Galletta, 2013; Mason, 2002). They can be used to collect qualitative information and are worthwhile for learning about specific situations or for supplementing and validating information derived from other sources (Creswell, 2007). The semi-structured interview method allows the interviewer to structure the interview using an interview protocol containing the questions, topics, themes, or areas that need to be covered during the interview; at the same time, it is flexible and allows the interview to digress and be expanded with unexpected themes (Blee & Taylor, 2002). In detail, the interview begins with a structured interview protocol that allows systemic approach during the interview questioning process, but also facilitates following the flow of the interaction between the interviewer and the interviewee (O'Leary, 2005). The characteristics of the semi-structured interview include the ability to stay with the intended questions as well as ask follow-up and/or new questions as data emerge through the interview process (Galletta, 2013). The hybrid nature of semi-structured interviews effectively suits the exploratory and perceptional studies of communities (Barriball & While, 1994). Interviewers are able to clarify complex and sensitive issues from the interviewees and/or communities. Using semi-structured interviews also helps discover contextual and cultural differences of interviewees from professional, educational and historical backgrounds. In addition, the natural flow of the questioning helps add depth and breadth of information as well as probe interviewees' perspectives and experiences (Hardon, Hodgkin, & Fresle, 2004). Furthermore, personal interviews help to overcome poor response rates compared to the survey method (i.e., a structured research method). Interviews also facilitate the comparative analysis among all respondents with fully answered questions. Finally, interviews ensure that the respondents answer all the questions without any assistance from others (Barriball & While, 1994).

Many data or metadata practice studies are conducted using a structured or semi-structured interview method. As an example of the studies, Palmer and Knutson (2004) designed the Digital Collections and Content project to provide integrated access to IMLS National Leadership Grant (NLG) digital collections. They presented a paper on how metadata and collection items can best be represented in the project repository. In the study, they used an interview method along with survey and content analysis. To develop scenarios of the repository use, they interviewed participants to identify their experiences with metadata application and collection building.

Stvilia et al. (2013) conducted a data practice study in the Condensed Matter Physics (CMP) community. The researchers utilized semi-structured interviews with an aim to identify the structures of and relationships among the community’s data practices (e.g., data, activities, data quality and data quality value). They interviewed twelve different individuals to understand the community’s data practices, issues and problems.

In order to develop a context-sensitive aggregation strategy for digital cultural heritage metadata (i.e., Opening History), Palmer, Zavalina and Fenlon (2010) used a research design which included semi-structured interviews, among other methods. In particular, through the semi-structured interview and participant observation sessions conducted with academic historians, they examined the scholarly perspectives on collection-level metadata used for scholarly access and use.

Thus, to explore the IR community’s data curation practices, using semi-structured interviews guided by Activity Theory (Engeström,, 1987; Leontiev, 1978) and IQ Assessment Framework (Stvilia et al., 2007) is a relevant research design. Each concept and relationship from the theoretical framework (i.e., object, instruments, rules, division of labor, quality issues and problem, and quality dimensions) can be a basis for interview questions. The interviews provided this researcher with the understanding of the community's data curation practices, issues and problems.

3.2.1 Sampling Method

3.2.1.1 Identifying sample. The goal of this study was to explore the curation practices for research data in the context of IRs and develop a knowledge base of the quality use of identifiers in data curation. Therefore, the target population of this research was data curators who work for the IRs storing and curating research data objects. Data curators are, in this study, defined as the person who manages and promotes the use of research data objects from their point of storage to ensure they are fit for contemporary purpose and available for discovery and re-use (Higgins, 2008; Lord, Macdonald, Lyon, & Giaretta, 2004). The use of metadata including identifiers in the IRs aids the responsibilities of data curators in discovering and preserving research data and promotes the benefits of data curation (Lynch, 2003; Westell, 2006). Since IR data curators use metadata schemas in their daily work, they are the target population for this study.

A sample of data curators who work for IRs storing research data was sought. Subjects had to meet three conditions for this research. First, subjects had to be involved in curation in their IRs. In order to qualify, the job title of the subjects does not necessarily need to be "data curator." Different institutions use different job titles, such as metadata librarian, digital repository manager, digital repository architect, digital curator, digital service librarian, scholarly communication librarian, data management librarian, etc. Second, subjects had to work for IRs that store and curate research data. Third, subjects had to work for IRs being maintained by one of the 108 institutions classified as RU/VH (very high research activity) in the Carnegie Classification of Institutions of Higher Education, a leading framework for recognizing and describing institutional diversity in United States higher education. The institutions classified as RU/VH are comparable based on their research activities and sociotechnical context and are appropriate for examination. The sampling process of identifying whether or not IRs store research data objects built on a previous research. According to the previous research from this researcher (Lee & Stvilia, 2012), in 2012 only half of the AAU member universities, which are the leading 62 research universities, had IRs that contained research data objects. Thus, because of the limited numbers of the IRs storing research data objects, this researcher uses the Carnegie Classification. The 108 institutions also overlap with all AAU member universities except for two institutions located in Canada. Table 3.1 summarizes the target population and the criteria for sampling.

Table 3.1. Criteria for Sampling

Target population	*Criteria for sampling*
Data curators working for IRs storing and curating research data objects	1) Staff whose job responsibilities include curating or managing an IR 2) Staff who work for IRs that store and curate research data as their objects 3) Staff who work for IRs being operated by an institution classified as "RU/VH" in the Carnegie Classification

3.2.1.2 Recruitment of the sample. Data curators working for IRs storing research data objects were targeted population of this study. A sample for this study was sought based on the sampling criteria in Table 3.1. The use of nonprobability sampling methods was appropriate for this study because an intensive investigation of a small population was required (Schutt, 2009). Purposive and snowball sampling techniques was used due to the difficulty of reaching or identifying the members of the population (Schutt, 2009). Identifying the right interviewees by searching institutions' library staff directories, usually including employees' names and departments they belong to, job titles, email addresses, etc., is not an easy task and may produce an inappropriate list of interviewees. Therefore, key informants (e.g., heads of scholarly communication departments, IR software trainers, etc.) who have many connections with IR data curators are good sources who can help with the recruitment of the sample.

This researcher sent total 33 email invitations to subjects identified as potential interviewees. The email explained this study, the process of participation, the benefits and risks of this study, and the participant's rights, and included the study's consent form as an attachment. The first invitation was sent in March 3, 2014, and the last was sent in August 19, 2014. Total 20 people responded the email invitations, but only 13 respondents were valid with the sampling criteria. Some of the invalid respondents contributed to identify potential interviewees by providing their network. In addition, while the sampling process, the researcher attended a research data focused symposium (i.e. ASIS&T Research Data Access & Preservation [RDAP] Summit), held in San Diego, CA in March 2014, and met 8 potential interviewees who were already contacted by email invitations. Three of the eight participated in this study, and the rest helped to recruit the study participants. The first interview was conducted on March 5, 2014; and the last interview was conducted on September 4, 2014.

The 13 respondents agreed to participate in this study and sent electronically signed consent form back to this researcher. Thirteen interviews with 15 participants were conducted to understand data and data curation practices in IRs with an emphasis on identifier schemas. The participants were from 13 different institutions; however, two of the participants each requested that an additional person be interviewed with them, which brought the total number of participants to 15.

3.2.2 Data Collection and Analysis

This section discusses the key steps of a semi-structured interview and the issues that face during each step. The interview began with prior knowledge gleaned from a literature analysis; that knowledge became the basis of the interview protocol. The researcher conducted interviews using the protocol and analyzed the results using a coding schema.

3.2.2.1 Preparing the interview. The value of prior knowledge when preparing for a semi-structured interview is highly emphasized. Such prior knowledge helps the interviewer develop research questions, an interview protocol/guide, and an analytical framework. During the review of literature and development of the protocol, researchers should broadly understand the topic to avoid bias based on the researchers' specific interest in the topic.

This researcher investigated the practices of research data curation and data identifier systems within IRs of AAU member universities (Lee & Stvilia, 2012) and conducted an extensive literature analysis in the topic of data curation activities and identifier schemas for research data (Lee & Stvilia, 2014). These prior studies enabled this researcher to increase his understanding of IRs and its identifier systems. A collaborative study (Stvilia et al., 2015) in which this researcher participated used the same theoretical framework guided by Activity

Theory and IQ Assessment Framework to explore data practice in a Condensed Matter Physics community and helped prepare this researcher to conduct the proposed dissertation project.

3.2.2.2 Developing and testing the protocol. The protocol of a semi-structured interview can be designed in flexible ways to allow various questions, including narrative questions (e.g., open-ended), questions suggested by a theoretical framework, and finally narrative questions for important theoretical connections (Galletta, 2013; Mason, 2004; Schutt, 2009). According to Galletta (2013), the interview protocol should represent the empirical and theoretical axes of the research topic. At the beginning of the protocol, narrative questions can serve to engage interviewees to speak about their experiences on the research topic, which can represent the empirical and individual work context of the topic. In the middle of the protocol, specific questions allow the interviewers to ask the necessary questions supported by the theoretical framework. Interviewers will then form individual characteristics from the interviewees' responses, leading to a flexible extension of the interview. Finally, questions at the end of the protocol will include unexpected open-ended questions newly generated by an interaction between the interviewer and the interviewee.

This researcher developed an interview protocol based on both a theoretical framework, guided by Activity Theory (Engeström, 1987; Leontiev, 1978) and IQ Assessment Framework (Stvilia et al., 2007). The protocol consists of four different sections, including Demographic Information, Data Activities, Data Types, and Perception of Identifier Quality. The first two sections mainly focus on answering the first research question, representing data curation activities, issues, and requirements. The third section, Data Types, answers research question two, focusing on identifying different entity types of research data. The last section, Perception of Identifier Quality, contains questions designed to address the third and fourth research questions, examining the knowledge or awareness of identifier schemas and the perception of identifier quality by IR data curators. A map that illustrates the connections between the interview questions and research questions and that displays the relationship between the interview questions and the theoretical frameworks is shown in Table 3.2. The interview protocol is also available to see in Appendix A.

Readability and understandability of the protocol had been tested with five doctoral students in the School of Information at Florida State University (FSU) and an IR staff, scholarly communication librarian. Feedbacks from the pilot test participants were used to improve the protocol's readability and understandability.

3.2.2.3 Conducting the interviews. When conducting the interviews, this researcher as an interviewer considered two issues: (1) when and when not to engage with participants and (2) how to avoid the disjuncture in meaning and intent (Galletta, 2013). Engaging participants to clarify a concept, generate a definition of the concept, and create space for critical reflection is an important role for the interviewers who work on semi-structured protocol. However, the interviewers should carefully decide when to engage with interviewees. The decisions affect the extension of the interviews with expected or unexpected themes (Galletta, 2013; Mason, 2004). Therefore, an accurate understanding of the interview participants' narrative and intercommunication between interviewers and interviewees is essential for effective interviewing.

The hybrid nature (i.e., structured and unstructured) of semi-structured interviews allows for changes in both types of questions (e.g., closed, open, exploratory, explanatory, etc.) and data collection tools (e.g., surveys, observation, case studies, etc.) based on unexpected themes and interviewers' needs. Such variation has the potential to produce a huge gap between research purposes and actual interviews. Therefore, this interviewer constantly attempted to be aware of

Table 3.2. Design of the Interview Protocol

Interview Questions (Appendix A)	***Research Questions***	***Theoretical Framework***
Demographic Information		
1. Tell me a little about your position in your institution in regard to managing IR.	RQ 1	AT
a. What are the other positions existing in your institution to manage IR?		
2. What was your highest degree? What was the formal discipline of your degree and what are your specific areas?	RQ 1	AT
Data Activities		
3. What is the main objective of your IR?	RQ 1	AT, IQAF
4. How long has your IR allowed submission and searching of research data?	RQ 1	AT
5. What are some of the activities you perform managing and curating data in your IR?	RQ 1	AT, IQAF
6. What user activities (e.g., identifying, searching, browsing, social networking, annotating, citing, linking, etc.) does your IR currently support?	RQ 1	AT, IQAF
7. What is the division of labor in your IR - what are some of the roles related to those activities (e.g., curator, data provider, user, etc.)?	RQ 1	AT
8. What are some of the tools (i.e., software, instruments, etc.) you use to manage IR objects in your IR? Are they different from tools for research data?	RQ 1	AT
9. What are some of the tools (i.e., software, services, etc.) you provide for your IR user community to store, organize, analyze, visualize, share, communicate about, and/or interact with data?	RQ 1	AT
10. Does your institution manage its IR database by itself or does it use an outside company?	RQ 1	AT
11. What is the repository software (e.g., DSpace, EPrints, Fedora, etc.) that your IR uses?	RQ 1	AT
12. What are the metadata schemas (e.g., DC, PREMIS, MODS, TEI, etc.) used in your IR?	RQ 1	AT
13. What are the metadata schemas (e.g., DDI, DwC, EML, etc.) used for research data?	RQ 1	AT
14. What are the identifiers (e.g., DOI, Handle, ARK, UUID, etc.) used in your IR?	RQ 1	AT
15. Does your IR use different identifier(s) for research data? If so, what is it?	RQ 1	AT
16. Are there any policies, rules, norms, or best practices that guide data management and use in your IR? If yes, please name them. Do these policies, rules, or norms come from the government, funding agencies, community, or are developed locally?	RQ 1	AT

Table 3.2. Continued

Interview Questions (Appendix A)	***Research Questions***	***Theoretical Framework***
17. Are there any policies, rules, or norms that govern or guide identifier system selection and use in your IR? If so, please name them. Do these policies, rules, or norms come from the government, funding agencies, community, or are developed locally?	RQ 1	AT. IQAF
Data Types		
18. What major types of research data (e.g., raw data, slides, text documents, spreadsheets, laboratory notes, etc.) does your IR accept?	RQ 2	AT
19. What types of research data entities does your IR control metadata for (e.g., author, subject, geographic location, etc.)? The following page contains a list of research data entities found in the literature. Provide the attached list. After reviewing this list, are there any types of data that do not make sense or are not applicable in your work context? Or do any other entity types come to mind? What controlled vocabularies (e.g., SKOS vocabulary, etc.) does your IR use to control that entity metadata?	RQ 2	AT
Perception of Identifier Quality		
20. What are some identifiers or identifier systems you are familiar with? What do you know about them?	RQ 3	IQAF
21. What identifiers do you use at the data collection/set level, file/object level, or entity level?	RQ 3	AT, IQAF
22. Are you familiar with identifier quality assessment criteria (or models)? If so, have you used those criteria in practice and/or research?	RQ 3, RQ 4	IQAF
23. Can you recall a case when an identifier quality problem (e.g., access failure, incorrect access, etc.) led to disruption in IR activity? If yes, please describe it, and explain how you overcame the problem.	RQ 3, RQ 4	IQAF
24. The following page contains a list of identifier quality criteria for research data found in the literature. Provide the attached list. After reviewing this list, are there any other criteria that do not make sense or are not applicable in your work context? Or do any other criteria come to mind? How do you evaluate the quality of identifier systems for research data? On a scale where 1 indicates “extremely unimportant” and 7 indicates “extremely important,” please indicate the level of importance of each of the following data identifier quality criteria within the context of your IR. Can you briefly explain how you came up with the evaluation of the highest and lowest ranks?	RQ 4	IQAF

Note. AT = Activity Theory; IQAF = Information Quality Assessment Framework.

the intent of the questions and the responses while conducting the interviews. The researcher's reflexivity allowed the interviews to parallel the research goals.

Most interviews except two were conducted by using Skype, telecommunication application software, since the participants are spread all over the nation. However, one interviewee preferred to use phone instead of Skype, and then the researcher used phone as a mode for that interview. Also, one interviewee who met the researcher at RDAP Summit wanted to participate in the interview at the venue of the symposium, so that one interview was conducted in person. In addition, to establish a close rapport with interviewees and provide a sense of comfort and freedom, this researcher provided options to schedule preferred interview time and to select the style of the interviews between on and off of video camera, when online interviews are conducted. Interviewing with a comfortable feeling and at a familiar place is a critical element to collect a better quality data (Barriball & While, 1994).

The interviews were recorded and transcribed. Each interview took between 55-80 minutes. The transcribed interview data were imported into "QSR NVivo for Mac," a qualitative data analysis computer software, to conduct data analysis using the initial coding scheme developed by the researcher (see Appendix B).

3.2.2.4 Analyzing the results. A data analysis of qualitative research is not as simple as cause and effect. Multiple factors affect the analyses and results. Researchers must identify all the factors and their relationships using various and iterative ways that can explain the discussion between interviewers and interviewees (Galletta, 2013). In order to complete such analyses, this researcher transcribed, read, and organized the collected interview data. The processes allowed the data being prepared to be analyzed, and this researcher coded the data based on the coding schema developed with the top-level concepts of Activity Theory and IQ

Assessment Framework. QSR NVivo for Mac was used to classify, manage, and analyze the interview data.

3.2.3 Quality Control

A discussion of the verification of knowledge commonly includes the concepts of reliability, validity and generalizability (Kvale, 1996). Reliability refers to the consistency of interview results (Kvale, 1996). Validity refers to the extent to which the research explores the areas the researcher intends it to (Kvale, 1996; Schutt, 2009). Generalizability exists when a conclusion holds true for the population (Schutt, 2009). However, this exploratory research does not include a purpose of generalization.

When conducting semi-structured interviews, the interviewers put effort into enhancing the reliability of their interviews. The quality of the responses obtained during the interviews is largely dependent on how the interviewers conduct the interviews (Patton, 1990). Probing can be used to improve the reliability of the interview data (Barriball & While, 1994). However, novice interviewers might struggle with the semi-structured protocol when conducting their interviews, especially those including unexpected themes. To probe without being directive or judgmental, interviewers trained with relevant skills (e.g., the skills of questioning, the ability to think of questions during the interview, and knowledge about the community culture and context) are needed (Hardon et al., 2004). In addition, the validity of interview data is closely related to the reliability of the interviewers, as well as to respondents' willingness to provide value information (Barriball & While, 1994).

To collect quality interview data, this researcher/interviewer studied the qualitative interviewing method and has practiced interviewing with IR staff at FSU and also via a previous research project involving semi-structured interviews. This researcher probed the data to clarify relevant issues based on this research purposes, to explore sensitive issues, to elicit valuable and complete information and finally to find inconsistencies. This researcher also contacted interviewees with comfortable and unstrained interactions. The friendly approach improved the rapport between this researcher and interviewees.

When coding the transcribed data, researchers need a coding schema to code and analyze the interview content consistently, and intercoder reliability can be an issue. Intercoder reliability refers to the extent to which two or more independent coders agree on the coding of the content when applying the same coding schema (Cho, 2008). When the reliability is not established, the interview data and its interpretations cannot be considered valid (Lombard, Snyder-Duch, & Bracken, 2002). In this research, the researcher performed an initial coding of the whole dataset. A second coder, who has experience in qualitative research and is familiar with the theoretical framework, applied the same coding scheme to 10% of the transcribed interview data (i.e., 1.3 interviews) for quality control. The two coders had some disagreements in how they applied the coding scheme, but they discussed the differences and were able to achieve consensus. After the discussion, the first coder recoded the whole dataset.

3.2.4 Ethical Consideration

Four ethical issues are discussed in this research including voluntary participation, harm, identity disclosure, and confidentiality (Schutt, 2009).

3.2.4.1 Voluntary participation. Participation in this study was completely voluntary. Participants, who were interested in this study, were required to send the consent form with their signature back to this researcher. The signed form meant that the participants understood the interviews were recorded and transcribed, and the recordings were stored in a secured hard drive for 1 year. Also, the participants could ask to turn off the recorder at any time.

3.2.4.2 Harm. This researcher considered carefully how to avoid harm to subjects when he developed the interview protocol. However, it is not possible to avoid every theoretical possibility of harm while interviews are being conducted (Schutt, 2009). The Institutional Review Board (IRB) of the FSU Human Subjects Committee has approved this research as safe to protect subjects' welfare, and this researcher maintained the confidentiality of research subjects.

3.2.4.3 Identity disclosure. This researcher disclosed the goals and purposes of this study and his own information as a researcher to interview participants. Also, after the completion of this study, the findings will be shared with the participants if they so request.

3.2.4.4 Confidentiality. This researcher maintained the identity information of all subjects as confidential. All data from interviews were stored in secure place and will be kept for 1 year, and after that period the data will be destroyed. When quotations from the interviews are included in publications, the identity information will be converted into unique identification codes.

3.3 Limitation

The semi-structured interview has various limitations, and since it includes unstructured interview styles, interpreting interviews and creating and asking unexpected questions while conducting interviews requires trained interviewers with relevant skills (Hardon et al., 2004). Even with trained interviewers, possibilities of communication errors exist in the process of interviewing. The difficulty of reliable coding is also a limitation through a qualitative interview method. Lastly, the interviews and their analyses are time-consuming. As a result, a only limited number of the qualitative interviews can be conducted and their results cannot be generalized the whole population.

CHAPTER 4

FINDINGS

4.1 Demographics of the Interviewees

A total of 15 participants from 13 institutions participated in this study. Eight (53%) of them were female, and seven (47%) of them were male. In terms of their education level, five interviewees (33%) had a doctoral degree, and 10 interviewees (67%) had a master's degree. Eighty percent (12 of 15) had a master's degree in Library and Information Science (LIS), and 20% (3 of 15) who did not have a LIS degree had a degree in Ecology or English Literature with a concentration in digital literature. Interviewees with LIS degrees had specializations in digital libraries, data curation, and reference services. In terms of position, the interviewees primarily worked as data curators (6), heads of IR departments (4), or IR managers (4). One interviewee was an IR software developer. Table 4.1 illustrates a summary of the demographics.

Table 4.1. Demographics of the interviewees

Gender				
Female 53%		Male 47%		
Education Level				
Master 65%		PhD 33%		
Specific Areas of Education				
Digital Literature 7%	Ecology 13%	Library and Information Science 80%		
		Digital Library 53%	Data Curation 20%	Reference Services 7%
Position				
Department Head 27%	Data Curator 40%	IR Manager 27%	IR Developer 6%	

4.2 Objectives of the IRs

The main objective of the IRs is to collect, access, store, preserve, and share research scholarship as well as other materials that reflect the intellectual life of the university. All of the IR staff who were interviewed mentioned similar obejctives. An interviewee explained that collecting research output in the university is one of the main objectives of his IR:

> Like a lot of institutional repositories, part of what we want to do is to capture the scholarship and research outputs in the university, whether that is publications or datasets, so that one of the primary objectives is to have pre-prints or post-prints, and wherever possible the actual published version of the faculty member's work, within the institutional repository. (s15)

The preservation and accessibility of scholarly output through IRs were also presented by one of the interviewees as goals: “I see the main objective of our repository service as making accessible and preserving the scholarly output of our research communities” (s13). Lastly, an interviewee described IRs as providing data storage and long-term preservation, while also enhancing the circulation of data among researchers:

> It is housing research datasets but also part of digital collection managed by library. So, this IR is incorporated with the existing library collections, and so, as far as research dataset side, main objective is basically to provide both data storage and preservation to those who produce research datasets on campus, and also to increase data sharing by researchers. (s4)

As an optional function, a few interviewees introduced the idea that one of their main objectives is to publish open access journals through their IRs:

> Main objective of IR is two folds. The first is to host and make available publications from scholars at the institution, and the second, which is a growing element of the IR, is to actually publish open access materials, primarily journals. Those are kind of edited and hosted by scholars in the institution. Those are two primary goals. (s1)

In addition to the general objectives of the IRs, some interesting objectives regarding research data curation were identified in relatively new IRs. These newer IRs were built after January 2011 when NSF required its grant applicants to submit a research data management plan (NSF, 2010a). The institutions had mainly developed new repositories in response to major funding agencies' policy change on research data management. Five out of 13 IRs (38%) were developed after January 2011. The IRs' services started with a research data focus:

> We developed [the IR] in 2012 after never having had a repository service, which made us a little bit of an outlier for a big research university. You might be wondering why we did it in 2012, and a big part of the answer was, it was in response to the NSF data management plan requirement. (s12)

Two of the five IRs created after 2011 were designed to support the entire research data lifecycle (e.g., Digital Curation Centre [DCC] Curation Lifecycle Model, OAIS), from planning and creating to publishing and disseminating their research data on the Web:

> We want to support the entire research data lifecycle from data management planning, grant application, initial research project, data staging, virtual research environment for collaboration through publication with DOI, dissemination of data openly on the Web, and finally, this is a really important piece, preservation. (s3)

Additionally, the departments supporting those IRs provide consulting services, instruction in writing grant proposals, and data management training for their content providers in order to support a holistic approach to research data curation: “Our department grew services for archiving through the [IR] software, and then also having additional services around consulting and preparing data management plans for grant proposals or other data management training in general” (s14).

4.3 Research Data Activities in IRs

The interview data identified a variety of research data activities in IRs, including different types of curation activities as well as other related activities. Data curation activities mainly support the research data lifecycle (e.g., conceptualizing, planning, creating, uploading, and publishing); other related activities help facilitate or motivate data management through IRs (e.g., data analysis, policy development, and education). Table 4.2 shows each activity type and its corresponding actions within the two major categories of research data activities.

4.3.1 Understanding Data Curation Needs

Most interviewees described spending significant time determining the extent of a data provider’s data curation needs. One of the interviewees explicitly emphasized the importance of the first meeting with a data provider: “I think how much assistance each researcher needs really depends on the first interview. You are getting to know how well they have organized their data” (s4). First meetings with data providers to assess their data and curation needs notably affects later activities, such as receiving data and creating metadata. The meetings determine what types of help the providers need, who is the right person to help them create metadata, and how the data could be organized and stored:

> I specifically meet with researchers who are interested in actually depositing their datasets into our repository. So, I am kind of that first staff to talk about what we have

> and get to know their data needs, and then bring in other specialists in our library, such as metadata librarians, who would be involved in the project, the digital library project. We get them into the conversation once it is appropriate. (s4)

Table 4.2. Research data activities and their actions in IRs

	Activities	Actions
Curation Activities	Understanding data curation needs	Interviewing researchers
		Communicating with IR or library staff
		Consulting with researchers
	Managing and sharing data	Receiving or transferring data files
		Cleaning data
		Converting data to a different file format
		Developing and adding metadata
		Validating data
		Packaging data
		Uploading and publishing data into IR
	Ensuring that data is accessible and reusable	Annotating data for relevant entities
		Optimizing data to search engine
		Keeping data up to date into mirror repository
	Re-evaluating data for long term preservation	Selecting dataset for long term preservation
Other Related Activities	Analyzing data usage	Managing descriptive statistics of data usage
		Providing researchers with data tracking results
	Creating policy and administrating infrastructure	Understanding local needs and creating local policies and rules
		Building infrastructure component
	Educating people about data management	Training librarians
		Educating researchers
		Providing workshops for data analysis tools
		Providing outreach for data curation
	Continuing education	Learning the best practices for research data management
		Learning future technologies

In addition to communicating with researchers, dialogue with other IR or library staff also helps data curators identify and connect the right person to each specific researcher, in order to address the issues detected in the first meeting: “Going and meeting with people, getting people to coordinate work, and talking to people about understanding each project" (s1). Interviewees also consult with researchers to provide support for the researchers' data activities (e.g., depositing, documenting, and organizing). When IR staff are asked for help with those activities, they actively work with researchers to enable better access to datasets stored in IRs:

> We have some items that just start coming in. The researchers are depositing them [with] no interaction between myself and another librarian and the users. But, in other cases, we are actively working with researchers to do the deposit, and this is happening more around research data. We are actually doing work consultation on how to describe their research data, as well as how to organize it for download purposes, and also for segmenting, especially for larger datasets, segmenting it into particular files. I guess essentially to enable better access to those datasets. (s5)

Interestingly, one interviewee indicated that his IR has a very strong emphasis on communicating with researchers about their projects. The IR staff endeavor to keep in touch with the researchers periodically in order to stay current on each project's status (e.g., grant proposal, reward, creating data, depositing data, publishing data):

> If we were to go from the beginning we would be the first people to meet with researchers when they were interested in archiving data. It might be part of when they're working on a proposal or at a later date. We try to ideally meet with them in person and discuss what they'd want to go in and inform them of our protocols and such. We keep in touch with them. If it's something like a grant proposal, we keep in touch of whether or not they get their reward, and when they get their reward we try to contact them right at the beginning of a project and just kind of reiterate what their schedule may be for when they would be depositing data. Then periodically we would keep in touch with

> them. In other cases we’ll get someone at the end of a project where they just have data for us so we can meet with them and just kind of go over the process and give them some tips on how to organize their data. When they are ready to transfer data, in most cases we like to meet with them directly and have them transfer data right to a portable hard drive. (s14)

4.3.2 Managing and Sharing Data

This activity typically begins with receiving or transferring data files. Data providers generally deliver their files through file sharing services or portable hard drives:

> I do file transfer, so in some cases, if you have a large file set, too many files, or it's too big size-wise, our work history is to transfer either through a file sharing service or by going to somebody's office with a hard drive. (s15)

Once the curators receive data from researchers, they either help researchers clean up their data or the curators clean it themselves. One of the interviewees mentioned that they specifically work on discrepancies in how the researchers’ names are spelled: “We do some metadata clean up. People tend to submit things with their name spelled different ways. We do some metadata clean up for that” (s6). An interviewee stated that they also clean data from a disciplinary perspective (e.g., cleaning variable headings, organizing file directories): “I would probably help earlier on with getting their data cleaned up from the sort of discipline side of things [like cleaning] basic variable headings, file organizations, all that” (s4). Converting proprietary file format into non-proprietary format is a prevalent activity in the interviewees’ IRs. Some of the data files need their formats converted because they use fairly expensive software, and since it is important to be able to reuse, share, and preserve data, employing inaccessible software is not an ideal choice:

> We do recommend that they try to convert proprietary formats into nonproprietary when they can. We have some cases of that. There are a few where they used plotting software. It was called sigma plot or something. To actually use the file, someone would need fairly expensive software to get at it. In that case, we took the time to kind of export things to just standard tables. . . . It ended up taking a lot of time because it didn't transfer the role headings and metadata; the process of copying all that took more time than what we can really do in a lot of cases. (s14)

A different but related issue is the need to convert files into flattened file formats. Some research data files have more than one type of data. For example, a spreadsheet may also contain screenshots of image data generated by proprietary software. In such cases, creating the metadata and characterizing the file formats are not simple tasks. However, in order to curate research data, these activities are happening in practice:

> One of the things I have seen that really gets complicated is what people do to record their research. They do what they have to do. Sort of make their research more efficient. I met with someone who was doing medical research, and he had an Excel file. He did complicated micro-species research. [The file contained] thousands of images he was creating, and he would then ingest those from proprietary analysis software to create figures that he could use to analyze data, and he also could use them for publication. Right? So, he actually took a screenshot of the proprietary software and then embedded a JPEG of that screenshot into an Excel file. He created extremely complicated digital objects that if you are trying to flatten them it would be very hard. But so we actually haven't solved the problems. But I think this is a kind of problem. It's kind of come up. I

> mean, it goes beyond the normal equations that are built into Excel—other files embedded in files, how do you explore that in digital objects? (s1)

IR staff also help researchers develop metadata for their research data. They try to generate not only descriptive metadata, but also administrative and preservation metadata:

> I do the initial ingest process. What we do is a virus check; we do integrated scanning format checking. I help develop metadata for the data that's both descriptive metadata and also administrative and preservation metadata in terms of what we've done with the data. I put the data into a consistent package and upload it to the repository and again do things like integrity checking in virus scanning before the data goes into the repository. (s15)

Another interviewee also mentioned that he helps researchers document their metadata. Even though the researchers have their own understanding of metadata, they need further assistance to fully describe the data in the ways that they want it described:

> In my position, a lot of the work I do with researchers contributes to that metadata piece. Because even though they are describing, they are also creating a record for digital objects very similar to what you see in a published online journal article. But a lot of the time, they want someone working through the process and [then we] help them with the stages of the data [until] they get there. To describe their data in the ways that they want it described. (s3)

Researchers' different interpretations of metadata schemas requires intensive conversation between IR staff and researchers in order to create appropriate metadata:

> You can contribute, upload, select, and describe your dataset submitted to us without having talk to us. So, I guess it's kind of a human piece. Actually, we do a lot of back

> and forth with them. They want to know that things are going to be exactly what they want them to be. Kind of learning the system; learning what they can or cannot do with presentation of metadata or metadata record. That's why I have a lot of interactions and conversations, sometimes in month-long conversations at the stages of dataset. It is more than the human service side of things, in that respect, because the metadata form, DC term, is very straightforward. However, when people are presented with the form, they interpret in many ways. Even though it is straightforward. (s3)

Most interviewees talked about helping researchers create a Read Me file within a dataset. Because of the limited set of descriptive metadata elements IR provide, researchers often want to add supplementary information about the data by using a Read Me file, a sort of ‘workaround’ (Gasser, 1986). The file typically contains more specific information or disciplinary information about the data, along with different metadata schemas or vocabularies. The files tend to be only relevant to the domain specialists of that data due to the specificity of the information:

> Whenever we acquire a new dataset, typically from a faculty member, we work with them to describe the dataset using a qualified Dublin Core set of metadata. That gives us a generic bibliographic record for the content. We then work with them to create a Read Me file for the dataset, which gives more specific information about the dataset that may use different vocabulary that's more focused on the discipline and include other kinds of information that are really applicable only to the users of that data. (s10)

In addition, a few interviewees directly stated that they write metadata for the data researchers provide. While creating metadata, IR staff spend a significant portion of their time collecting documents that may help them understand the research data:

> When the researcher has time, we do try to sit down and interview them to talk about what they did, and we write the narratives, not a full transcript, but we do produce a narrative of that conversation. If they had NSF funding, I try to get the proposal that they have in IRB. I get the IRB and we will pass them or look up for ourselves a kind of representative sample of publications based on their data, so that we know something about it; we know the context of it, and so, based on all those things, we will write a lot of metadata ourselves. And so basically, I would say 70-80% of the time, it's just accumulating documentation. (s15)

Before a research dataset is uploaded into a repository, there are two more steps to do. The first is data validation, which checks whether the data contains any errors in its content. An interviewee provided an example of this:

> We talk with them and make sure the data is represented in the way they want it to be. We help take a look at [their dataset] like a tabulator dataset and ask questions like, “Are there supposed to be no values within those or should they all be zeroed out?” and they may have a very good reason for what they are doing. (s10)

The second is preparing the data package. Some datasets contain multiple files, and the files are different types (e.g., image, text, audio-visual). Most IRs require the depositors to submit a zip file package instead of submitting every single file: “Usually we just package the dataset with supplementary information like a data dictionary. Like in a zip package with the dataset and any other documentation” (s1). After the previously mentioned activities, the last activity for depositing data is uploading and publishing the data into an IR. As some of the interviewees mentioned that their curation team had the researchers give data to them, and then they would ingest data into their system. In addition, an interview participant described a specific service his institution supports that can be used just before publishing the uploaded data. The service allows researchers to look at their metadata in a test view and enables them to have one last opportunity to edit the metadata and its supplementary information:

> We upload everything onto a test instance that's not online in our IR platform so that we can share a link with just them [researchers] that's not public yet. They get to look at the metadata and can make final edits. Then, if they approve it, we move it to our regular online instance. (s14)

4.3.3 Ensuring that Data is Accessible and Reusable

IR staff aim to ensure that data is accessible and reusable online after the data is deposited and published. Some of the interviewees introduced different efforts to improve the accessibility and reusability of the data (e.g., managing metadata for search engine optimization, managing and maintaining the links between a central repository and a mirror repository). Similarly, an interviewee presented an effort to create maps connecting researchers' names with their affiliation information. The mapping may not only improve reusability of the data, but also reduce ambiguity regarding researchers' names:

> We do some mapping from collection, because as you are aware not all researchers are housed in one department or research lab. So, someone is in one research lab, but they are also in another department. I want to map things from place to place, so they can be found. (s6)

4.3.4 Re-Evaluating Data for Long-Term Preservation

A few of the interviewees shared that they conduct re-evaluation activity based on their preservation policy, but the activity has never actually been started and completed before. Their

IRs have not had a long enough period of operation to do the re-evaluation activity. Re-evaluating data is for long-term preservation; select IR staff, such as archival specialists, subject specialists, and IR managers evaluate the data to see whether they should be preserved or deselected. In order to be re-evaluated, the data must have been stored for the duration specified in the IR's data retention policy. Five or ten years were the examples described in the interview data. An interviewee presented his institution's policy: "We have a collection policy [to] dictate what happens to [data] 10 years after the deposit" (s3).

4.3.5 Analyzing Data Usage

Many interviewees described analyzing data usage as one of their data activities. They manage data statistics for different purposes: (1) to facilitate their own administrative work: "We are doing things like managing statistics to try to understand the number of downloads, that sort of thing" (s5). And, (2) to provide tracking results to the content providers: "Over time we will be giving the researchers tracking results on how often their data has been downloaded and that sort of thing" (s14).

4.3.6 Creating Policy and Designing and Administering System Infrastructure

Policy development and system improvement take time to be built up to a satisfactory level that follows current state of the art practices and harmonizes with other local policies. IR staff specifically needs to understand the recommended policies, rules, and norms around data management, and be able to select the policies that they want to adopt. In addition, the policies from external sources must align with current local policies or rules to produce successful results. One of the interviewees talked about the process of policy construction:

> We are trying to figure out things like data retention policies, how to interact with collection development policies, and institutional policies for research data management that are being communicated now. Even though we are accepting stuff [research data] into the system, it takes a long time to create the policies. So, we are kind of working in parallel. (s1)

A similar experience an interviewee described is the process of designing and administering system infrastructure. Building and managing the system is a highly time-consuming activity for IR staff: “I've been responsible for a lot of the infrastructure components and a lot of the thinking about how the pieces of this come together” (s10). Another interviewee provided a prominent example of this kind of activity:

> To some extent, there are policies or rules, but identifiers and identifiers on the web has been an area of interest of mine for a good 15 years, so to some extent, I was vocal about what I thought we should do and I think a lot about what we should do. . . . so that maybe the only rules or norms were the ones in my head that I kept repeating at people. (s12)

4.3.7 Educating People about Data Management

Managing research data is a new area of study in information organization. The community practices are still developing and academic libraries are accumulating this growing knowledge. However, there are currently a variety of teaching activities associated with the IRs in this study. First of all, IR staff teach the librarians about IR resources so they can communicate that information to library patrons, including both students and faculty members:

> We are training all the librarians and people that are in reference desks and everyone in the library who deals with any of our patrons, any of our faculty. To make sure that

> everyone does really know, “Hey, we have an IR. Your data can be open accessed,” and “Hey, if you have bigger data or sensitive data, whether it's HIPPA sensitive or privacy sensitive, here is our website that can explain it to you, and here is the different contexts you can have.” So you can always come through us. (s11)

Second, IR staff design events to provide research data management practice for their campus communities. The events aim to educate researchers in data management, as well as answer any of their data-related questions:

> I’ve done some campus-wide events. We’ve had ones on big data, little data, and having all of that. We do a research-computing day once a semester. The next one should be coming up in October. We’re also looking at doing a “bring out your data” event where we have a bunch of data experts . . . A whole bunch of people in the same room where people can just come and ask everyone at once about their data needs. I set up a lot of events like that. We need to make sure the right people are in the room. (s11)

The third type of data management education focuses on how to use IR platforms and data analysis tools:

> I suppose a lot of work that we do is educating users how to use the [IR] platform itself. We, like many IRs, don’t have a lot of self-in [systems] to do that kind of content ingestion on behalf of users. So, there is a lot of trying to help people manage data by themselves. (s2)

One of the interviewees also mentioned teaching users how to implement data analysis tools: “We do provide workshops on data analysis tools for users” (s15). The last education type focuses on IR staff provides outreach to promote the use of research data curation services within their IRs:

> We have done a multi-media campaign starting last January. We did a postcard mail out too, we distributed postcards across campus, provided multiple workshops internally for librarians, and multiple workshops externally for data management planning IR use. I've done 60 plus consultations with researchers, and 40 plus presentations to various faculty groups. (s3)

4.3.8 Continuing Education

IR staff learns the best practices, policies, rules, norms, and technologies of research data management services from the data, data curation, and archival communities, which they can then use to develop their own IR systems. Because data curation is an emerging topic, IR staff accumulates knowledge from different communities, in order to ensure both the effective processing and long-term preservation of research data. More specifically, they learn by attending conferences or data curation-specialized trainings, taking coursework, reading articles, and benchmarking the practices of peer institutions:

> The best practices come from the data curation community and the archival community, and policies and rules are being created now locally for our system. So, we are trying to figure out things like data retention policies, how to interact with collection development policies and institutional policies for research data management that are being communicated now. Even though we are accepting stuff [research data] into the system, it takes a long time to create the policies. So, we are kind of working in parallel. (s1)

> Some of the best practices are sort of what we learned in conferences or in training or coursework we are taking. We look at universities that have been at the research data management game for a little bit longer than we have. So . . . looking to see what they

> are suggesting for best practices. Kind of combined those things into our local setting. (s1)

In addition, IR staff collaborate with research and technology experts on their campuses to prepare for future changes in IR systems: "Liaising with researchers, computing people, and our research computing advisory committee to look at how we build up the technologies that are needed for the future" (s11).

4.4 The Activity Structure

The structure or context of research data activities in IRs consists of communities, division of labor, tools/instruments, and policies, rules, or norms (Engeström, 1987; Leontiev, 1978). This section describes the activity structure identified by the interview data.

4.4.1 Communities and Division of Labor

People's roles around research data activities in IRs based on their tasks could broadly be divided into three communities: (1) IR staff, (2) data providers, and (3) users. Each community is a group of people who share the same object (Engeström, 1987).

4.4.1.1 IR staff. IR staff are the employees who work for IRs and their job responsibilities include curating or managing an IR. IR staff, based on the interview data, can be divided into seven different roles: (1) head, (2) data curator, (3) IR manager, (4) metadata specialist, (5) developer, (6) subject specialist, and (7) graduate assistant. Table 4.3 is a map of the relationships between the roles and the position titles that exist in the interviewees' IR staff. The positions of data curator, IR manager, and developer exist in almost all IRs. On the other hand, the positions corresponding to head, metadata specialist, subject specialist, and graduate assistant only appeared in less than half of the 13 IRs. Some of the IRs only had one person to perform multiple roles of IR staff. For example, three different cases were identified: (1) head, data curator, and IR manager, (2) head and data curator, and (3) data curator and IR manager.

Table 4.3. IR position titles mapped into identified IR staff's roles

Roles	Job titles that include a particular role	# of IRs that have the roles
Head	Head of IR, Director of Scholarly Communication, Head of Digital Publishing, Assistant Dean for Digital Libraries, Head of Publishing and Curation Services	6
Data Curator	Data Service Librarian, Science Data Management Librarian, Repository Specialist, Technical Analyst, Repository Coordinator, Data Curation Specialist, IR Coordinator, Data Librarian, Curation Librarian, Digital Scholarship Librarian, Digital Collections Curator, Digital Content Strategist, Data Management Consultant, Data Curation Librarian, Digital Projects Designer	13
IR Manager	IR Manager, Repository Specialist, Repository Coordinator, IR Coordinator, IR Production Manager, Data Management Consultant, System Administrator, Digital Collections Curator	12
Metadata Specialist	Metadata Specialist, Head of Digital Project Unit, Digital Metadata Head	3
Developer	Developer, IR Administrator, System Administrator, Technology Architect, Software Developer, Senior Computer Specialist	13
Subject Specialist	Collection Administrator, Subject Librarian, Community Administrator, Subject Specialist	5
Graduate Assistant	Graduate Assistant	3

The role-related activities of head positions included three different tasks. First, they plan and build their research data IR services and further design the services to fit within their library systems. As the lead person of an IR group, they are responsible for planning and building their infrastructure:

> [This] is an evolving position. I think what we're trying to achieve by having that position is to develop a holistic view on the way that we are developing our digital collections, including what goes into our IR, but also thinking holistically about

> collections, so not making the digital collections like a silo, away from the other kinds of collections that exist in the university or that the libraries collect. (s13)

Table 4.4. IR staff's role-related activities

Roles	Role-Related Activities
Head	- Build or plan data governance structure in their IR - Communicate with researchers - Provide outreach for their IRs
Data Curator	- Consult with data providers and connect them to metadata specialists or IR managers - Facilitate communication across different entities - Evaluate or view research data to see whether the dataset would be continue to be maintained or whether it would be deselected - Build or plan data governance structure in their IR - Outreach and educate campus community
IR Manager	- Manage IRs on a daily basis - Work with data providers to help add metadata and upload data into IRs - Answer questions about IR use and data management - Outreach and educate campus community
Metadata Specialist	- Help data providers to create appropriate metadata for their dataset - Design metadata schema for their IR
Developer	- Maintain and update IR software
Subject Specialist	- Evaluate or view research data to see whether the dataset should be continued to be maintained or whether it should be deselected - Manage and approve incoming submissions to their own collections - Provide support and help to the management of the IR from their subject/user community specific perspectives
Graduate Assistant	- Assist data curators or IR managers

Another interviewee described a similar idea: "I am responsible for a lot of the infrastructure components and a lot of the thinking about how the pieces of this [IR] come together" (s10) In addition, the heads of IR groups communicate with researchers to address their concerns and answer their questions regarding IR use. They also provide outreach services to encourage use of their IRs (see Table 4.4):

> The other piece of that is providing user services, so being the person that works directly with researchers and hearing from them about what their requests are, what the challenges are in using our IR, and doing a lot of outreach for the IR and that kind of thing. (s13)

Unlike the position of the heads, which was found at only six of the institutions, all of the interviewees mentioned that they have one or more than one data curators in their IRs. As it implies, the responsibilities of data curators include numerous activities, some of which partially overlap with the responsibilities of the heads. One of their primary responsibilities is consulting with data providers to understand and address their needs, which is often accomplished by connecting them to the right person:

> We have a subject specialist-centered service model. So, each time I am in contact with a researcher, each time they created a project in our IR, each time they submitted a publication, or when they have a successful grant application or where they used our IR in their data management plan, I would like their subject specialist librarian to know, and, the three of us, in a perfect world, work together to get them to use our IR. (s3)

Another interviewee also described a similar work model in her institution. That work model motivated the interviewee to teach librarians about research data management:

> I can direct them to others or they contact their librarians, any of the subject specialist librarians. So, we have to make sure that all of them are ready for answering those basic questions on the IR about research data support on campus. (s11)

Data curators also play the role of facilitator between different work units. Specifically, the main area of responsibility is facilitating technical communication among stakeholders:

> I think of myself mainly as a translator because I have the depth and breadth of the technical knowledge as well as the academic research practices, knowledge and means, as well as the library practices, needs, and technologies. There are a couple of people on campus like me. . . . Sometimes other people call us “glue people.” We help things connect and stick together. (s11)

The curators also collaborate with librarians to select datasets for long-term preservation. Not all of the datasets are appropriate for permanent storage. They select datasets for preservation through the lens of archivist, subject specialist, and data curator: “Datasets would be viewed [by digital archivist, subject specialist, and data curator] to see whether the datasets should be continued to be maintained or whether they should be deselected” (s3). More than half of the institutions enabled data curators to perform typical responsibilities of the head position. Like the heads, the curators plan data governance structure for their IRs, and follow the current and future trends of the research data curation community:

> I do the next thing, which is considering how we leverage the repository information structure in order to support changes in library publishing, changes in data curation, changes in new forms of digital scholarship, alternative scholarly work, and so on. (s11)

Outreach for the IR services was a widespread activity throughout the IR staff, including head, data curator, and IR manager:

> We have done a multi-media campaign starting last January. We did a postcard mail out too, we distributed postcards across campus, provided multiple workshops internally for librarians, and multiple workshops externally for data management planning IR use. I've done 60 plus consultations with researchers, and 40 plus presentations to various faculty groups. (s3)

A primary role of IR managers is managing IRs. Simply, they manage IRs on a daily basis: “She [IR manager] does a majority of day to day managing [of the IR]” (s2). They ensure the ongoing production flow of adding new materials to the IRs. Along with managing IRs, they answer any questions regarding IR use: “My primary job is to be the frontline person for our IR. I handle any incoming questions about the IR, requests for consultations, and demonstrations about the IR” (s3). IR managers also communicate with data providers to create and add metadata, and to upload the data into IRs: “Major role . . . is helping users provide and add their metadata for their research project and curating. But, I also communicate with them to provide interaction between what they want and what I want” (s3). Another responsibility of IR managers is to provide outreach for their IRs and educate the campus community. These activities are a shared responsibility between the head, data curator, and IR manager roles: “From the user side, [I do] any internal and external outreach for education, demos, and trouble shooting questions in terms of when someone actually runs into a bug or issue using the IR” (s3). A few interviewees mentioned a separated metadata specialist position within their IR staff. Those interviewees indicated that there is a high demand for metadata specialists:

> Our metadata staff is currently influx. We have one digital metadata head, who is very much involved [in the data curation team]. She is kind of key player for sure. We also, in the process, put out the position and hiring process. . . . We are going to have one full metadata specialist working on our research data curation team and also interact with the researchers to get their data documented for and deposited into our IR. (s4)

The complex and diverse types of research data also required the continuous development of metadata schema modeling for research data: “[The metadata specialist] helps a lot with the metadata schemas and modeling as we add new content into the repository” (s10). All of the 13

IRs included one or more than one person working as the IR software developer. They generally maintain their IRs and update the system as needed: “We also have part time on some of our systems people or library system staff who help manage the backend [systems], make sure all the backups are there, do upgrades to the software, etc.” (s9). In addition to that, one of the interviewees mentioned that they purchase a developer service from an outside company: “Technical support comes from an external company; we are purchasing services from them” (s1).

Almost half of the interviewees stated that subject specialists connected to different departments. They can be subject librarians working closely with different departments or people who have been designated by the department to administer their own departmental collections. One of the responsibilities of subject specialists is evaluating research data for long-term preservation. As mentioned in the description of data curator roles, the evaluation is co-conducted by an archival specialist, subject specialist, and data curator: “Dataset would be viewed by digital archivists, subject specialist librarians, and digital data repository specialists to see whether the dataset should be maintained further or whether it should be deselected” (s3). The primary tasks of the subject specialists draw on their domain knowledge. They manage departmental collections, approve incoming submissions to their own collections, and help manage the IR from their specific disciplinary perspectives: “These are the people who have been designated by the department or by the research group to manage and approve incoming submissions to their repositories. And also oftentimes make deposit themselves” (s2). Another interviewee also spoke about the responsibilities of subject specialists: “All of our subject specialist librarians connected to different departments inform and provide support and help to the management of the IR from their specific lens and their constitutive perspectives” (s11).

One of the interviewees assigned a wide-ranging set of tasks the subject specialist role in her IR. In that IR, the subject specialists can establish disciplinary workflows, submission processes, and setup a new collection:

> We also have, within the institution, community members [in a department, institute, or research center, etc.] who are actually able to manage a collection within our IR. We give them appropriate permission so that they can establish the workflow and submission processes they want, or they can setup new collections, they can setup groups of submitters, and they have some level of control over the entire repository. So, we call those community administrators. We have documents that outline their roles and responsibilities, which have been used as guidelines for them. (s5)

A few IRs had graduate assistants who assist data curators or IR managers. Their responsibilities are very flexible based on the tasks set by their supervisors: "I have a limited amount of graduate hourly support. There is a graduate student who works for me on an hourly basis to complete tasks" (s5).

In addition to the IR staff's role-related activities, the power structure among the IR staff implied by the interview data can be described with some flexibility (see Figure 4.1). For IRs without a head position, data curators tended to be higher-level positions based on their responsibilities to design and organize research data service in their IRs. Data curators' responsibilities, in most of the interview data, overlapped with the responsibilities of head position. The positions of metadata specialist and developer also had some flexibility. The interviewees indicated that those positions could belong to not only the IR staff, but also to other groups within the institutions.

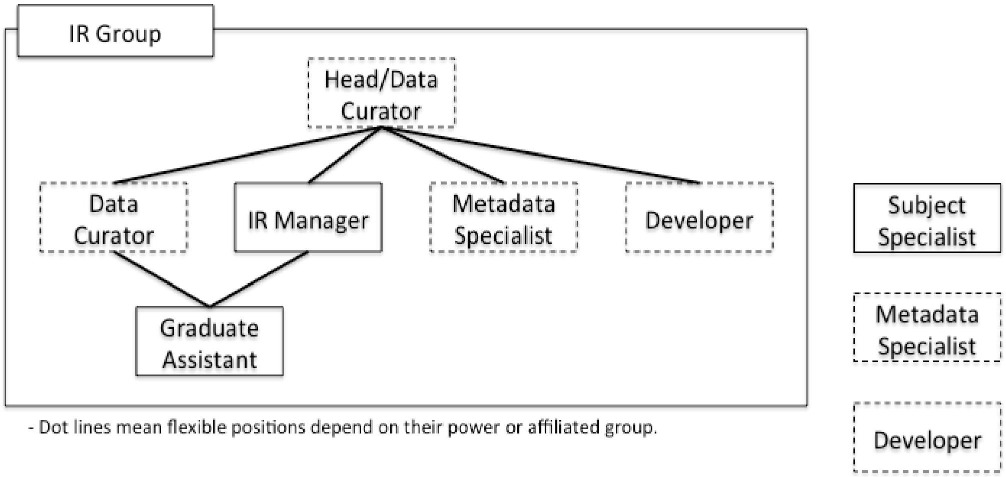


Figure 4.1. Hierarchical structure of IR staff

4.4.1.2. Data providers. Data providers can be integrated into the category of researcher. Researchers are mainly faculty members, postdoc researchers, and graduate students in each institution. A few institutions also allowed undergraduate students to submit their data with faculty advisors' permissions. A main activity of data providers in relation to IR is the submission of data files and their metadata (see Table 4.5). There are two different types of submission processes. First, researchers provide their data and its metadata directly to IR staff: "The users provide us with the data files and spreadsheet that contains all the metadata that they want to import to the repository" (s2). Second, researchers upload their data and the metadata into the IR systems: "You upload your files for publication phase. And then using a wizard . . . you can contribute metadata so you can describe the dataset" (s3).

The second main activity identified is the conversion of file formats. In most cases, data providers consult with IR staff, and the staff recommend nonproprietary file formats for long-term accessibility and preservation: "We do recommend that they try to convert proprietary formats into nonproprietary when they can" (s14). In addition to those two main activities, some of the interviewees described activities associated with services provided by IR systems. Using

IR services, data providers are able to track their research data's usage and number of downloads:

> The researcher can create their own profile they call selective works I guess. So, there is a sort of dashboard you get and when you sign up for that it allows you to look at downloads and track what is happening with the publication. (s1)

Table 4.5. Data providers and their activities

Data Providers	
Faculty members, postdoc researchers, graduate students, and undergraduate students	
Activities	
- Provide data files and its metadata	- Share data through social networking sites
- Convert file formats to nonproprietary formats	- Share recommended citation and contribute citation data
- Track data usage (download, publication uses, etc.)	- Collaborate with researchers in IR project space
	- Transfer data ownership from student to faculty advisor

Sharing activity through social networking sites is also supported by IR systems: "We have, through the IR, a social networking widget that allows you to tweak [share] the paper or data" (s1). Data providers can share a recommended citation and build citation information for their data:

> We have made a big push to use RDA to share machine-readable metadata with various tools that can scrape that data. It's really easy to link to items. For every item in the repository, we have a suggested citation if you want to quickly share a citation. You can just copy and paste a generic format citation if you don't have very specific format you want. (s10)

Another interviewee mentioned IR services associated with citations: "We provide recommended citation and allow contribution of citation. Or you can link to other digital objects or just provide citation to non-digital resource" (s3). There are some activities only provided by a single IR. The first one is collaboration between a project team as an activity supported by IR systems. Project team collaborations are facilitated by using wiki capability: "You can upload your files [in project space]. It has a wiki capability. You can assign tasks, which is a project management functionality" (s3). The second activity available for data providers at one IR is the ability to transfer ownership of their datasets. This activity is important because it enables graduate students to transfer data ownership to their advisors:

> Users can also allow another user to deposit on their behalf. . . . That user can also transfer ownership back to the owner of the file. There is an ability to let somebody else, like a graduate assistant for a faculty member, let's say, deposit files on his or her behalf and then transfer the ownership of those files once that has been done. (s13)

4.4.1.3 Users. Data users of the IRs can be anyone who has Internet access and is interested in the stored research data. The activities of the users are dependent on the IR systems (see Table 4.6). There are two different types of IRs: (1) IRs that only provide typical repository services (e.g., identifying, searching, browsing, and downloading): "We support searching, browsing, download of materials" (s5); We don't actually offer any direct tools that do that [analysis]. We make our data available to the end users and we pretty much expect them to take that data and move it into their tool of choice" (s10) and (2) IRs that provide not only repository services but also diverse user services. Many interviewees mentioned sharing and social networking ability of their IRs:

> Our dataset is searchable. You can browse as you see. There is a social networking ability, where you can disseminate a dataset, in the upper right corner of every dataset

> record. You can select how you want to distribute the dataset using Facebook or Twitter, for example. (s3)

Table 4.6. Users' data activities in IRs

Users' Activities through IR User Services	
Typical Repository Services	Additional Services by the IRs
Identifying	Sharing
Searching	Social Networking
Browsing	Full-text Searching
Downloading	Bookmarking

Some other interviewees also talked about full-text searching and bookmarking services:

> We have mechanisms to search for data and to browse through items in our repository. They can search the full text of the items. They can share items that they found with various social networking infrastructures. (s10)
>
> We also have the bookshelves. . . . I think people normally save their favorites and save their items elsewhere whatever their other external system is, but people have really used those [bookshelves] a lot, especially in teaching. We've gotten a number of responses from people doing graduate research saying, "This is so easy. I can save my searches. I can save my browsers. I can save my favorite collections. I have my collections page so then I can really use this as my work space within the collections." (s11)

Some of the interviewees from the IRs with only typical repository services explained why they decided to not provide additional services. They assumed that the users would have the tools to work with the data:

> Honestly, we don't really provide much of anything. I mean our data is available for download. So the assumption is that users would have the tools necessarily to interact with it. I am actually situated with a unit that does a lot of work around numeric data

> analysis, GIS data, as well as qualitative data analysis. But you know that's a separate unit, not explicitly connected with the IR. (s5)

Some of the interviewees also thought that it is a repository, not a user interactive site: “Repository services. I actually advise users to use alternate systems whether it’s Google or something else just to get access to the file. It’s a repository, not a user interactive site” (s9).

4.4.2 Tools

The interview data displays some of the current practices regarding tools for data curation in IRs. The tools include IR software, metadata schemas, ontologies, identifier schemas, controlled vocabularies, and tools for data curation (see Appendix C).

4.4.2.1 IR software. IRs use different software for their platforms (see Table 4.7). Many of the IRs employed external and widely used solutions for repository services (e.g., Bepress Digital Commons, DSpace, Hydra, Dataverse, and HUBzero). Digital Commons is a software managed by an outside company called Bepress. IRs that use the software usually have fewer personnel available to work on software development. Some of the IRs, however, used local solutions (e.g., Aubrey and SobekCM) for their specific goals and needs. IRs select their software based on available resources and needs. The performance and available personnel of an IR software development team affects the selection of IR software. Some institutions try to use the software that can be controlled and developed by them. Other institutions purchase services from an outside company to outsource the software management of their IRs. In addition, political factors as well as the software development company’s perceived reliability may affect decisions about IR software:

> I think that there were really only three players in the field in terms of platform, back when they made the decision to go with DSpace. E-print was really only used by folks in UK, so that was kind of off the table. DSpace has a lot of support and a big user community here in the US. Also, it [DSpace] was something that we could manage locally, and I think that's why we chose that over Bepress Digital Commons. Because our team, in the library technology unit, would always like to control the development and use of the software. (s2)

Some of the institutions have very specific purposes and needs for their IR platform. The system has to suit the other repository infrastructure within the institution as well as all of the other local policies and norms for data curation:

> This is one that we have built locally. Locally, we call it Aubrey and it's a Python-based repository infrastructure we built. . . . then we have a sister piece to it called Coda which is the archival storage repository that we use to build archival storage systems within our institution. Both of them are locally developed. We are in the process of sharing components of them, but we're probably not going to be making them generally accessible because they're very, very tied to how we've wanted to do things. (s10)

4.4.2.2 Metadata schemas. The interview data indicate that most of the IRs provide simple descriptive metadata. Even though some of the IRs provide diverse metadata elements for specific disciplines, the elements that are regularly filled out by researchers are limited by required elements. Metadata from the all of the IRs based on Dublin Core (DC) metadata schema and some of the IRs offer additional metadata schemas such as DataCite metadata. One interviewee mentioned that they use DC metadata because it is simple for all researchers: "We have been using DC, because we want all researchers to be able to use it, so we keep it really simple” (s3). Most of the IRs locally modified the DC metadata for their needs. Another interview participant stated that they use qualified DC metadata for everything in her IR as well as DataCite metadata for data objects:

> We allow people to put in qualified Dublin Core or DataCite, which they could do for either data objects or regular objects. If people have more specific metadata schemas, they need to add that as an attached file. The qualified DC is for everything and then DataCite is optional. (s9)

The interview data also identified some metadata elements particularly added for research data. They include citation information, abstract, location, temporal coverage, methodology, note, collection, related URL, and partner institution. However, the main elements provided by most of the IRs are close to simple DC metadata elements (DCMI, 2012): title, creator, subject, description, publisher, contributor, date, type, format, identifier, source, language, relation, coverage, and rights:

> For the descriptive metadata, we use a locally qualified [modified] Dublin Core Element Set. We’ve added a few additional fields that we feel are important, including a field to hold citation information. We’ve added in a note field. We’ve added fields for defining collections and partner institutions, but mainly it's the 15 Dublin Core Elements that are then locally qualified, specifying what kind of title, what kind of creator, what kind of contributor, what kind of identifier, what kind of date is being used? As items are added to the preservation repository for every file, a PREMIS record gets created, a PREMIS object record. (s10)

Table 4.7. IR software, metadata, and tools for data curation

IR Software						
Bepress Digital Commons	DSpace	Hydra	Dataverse	HUBzero	Aubrey	SobekCM
Metadata Schemas						
Modified Simple Dublin Core (DC), Qualified DC, DataCite Metadata, MODS, METS, PREMIS, MIX, EAD						
Metadata Schemas used in Supplementary Space						
Darwin Core, EML, DDI, TEI, FGDC, ISO 19115 Geographical Metadata						
Identifier Schemas						
DOI	Handle, DOI	ARK, DOI, HTTP URI	DOI	DOI	ARK	Permanent local URL
Controlled Vocabularies						
DC Contolled Vocabularies, Library of Congress Subject Headings (LCSH), Medical Subject Headings (MeSH), Faceted Application of Subject Terminology (FAST), Only with Hydra: DC RDF Ontology, FOAF, RDF Schema						
Applications for Data Curation						
Creating and Editing Metadata: Microsoft Word, Microsoft Excel, Text Editor (WordPad, Notepad++), Oxygen XML Editor, Morpho (Ecology Metadata Editor), Nesstar Editing Images or Videos: SnagIt Photoshop for images, Handbreak for audiovisual Cleaning Data: Open Refine Storing Data: Dropbox, Google Drive Identifying and Validating Data Files: DROID, PRONOM, Git for version control, FITS for file characterization Transferring Data: BagIt Indexing Data for Search: Apache Solar Tracking and Measuring Data: Altmetric						

In spite of the simple metadata elements from the IRs, research data is complex and diverse. Unlike most research articles, which are usually submitted in PDF format, data are submitted in different formats and types. The complexity and diversity increase issues with creating and adding metadata, and then a lack of metadata is also connected to issues of reusing, sharing, and searching the data:

> However, we sometimes add extra files. You can give us a survey; you may also give us a codebook; and you can also give us DDI file, or all three of them together. But it doesn't display out as metadata and it's not really searchable in a structured way within DSpace. (s5)

Current practice for creating and adding metadata consists of filling out the form of metadata elements and creating and uploading separate files that include supplementary information about the data. Most of the IRs adopt the practice of metadata creation. As mentioned earlier, the main metadata elements of the IRs are based on simple DC metadata, which is not sufficient to describe research data. The result of the insufficient metadata elements results in abuse of supplementary information space:

> You have to give it [the data] a title, synopsis, and abstract. You could fairly abuse the abstract field. Near the end of the publication workflow, there is also a place you can add notes. That's a kind of space to catch all for the data. (s3)

The supplementary information cannot be searched in a structured way, but the information can be found by using full-text searches. Interview participants also mentioned some of the disciplinary metadata schemas currently used by researchers within their IRs for supplementary information. These included Text Encoding Initiative (TEI), Federal Geographic Data Committee (FGDC), Data Documentation Initiative (DDI), Darwin Core (DWC), Ecological

Metadata Language (EML), and ISO 19115 Geospatial Metadata. In addition to the descriptive metadata, most of the IRs also support administrative, structural, and technical metadata (i.e., METS, PREMIS).

4.4.2.3 Identifier schemas. According to the interview data, IR software has its own identifier schema embedded into the established infrastructure. Many IRs employ the underlying identifier schemas in the software without a strong discussion about selecting an identifier schema that fits their specific goals. However, recent active movements toward the improvement of research data curation and semantic web technologies have changed the perceived importance of identifier schema selection for IRs:

> We use Handle because DSpace uses it. I think, in terms of the decision to move to DOI, it was made because DOI is becoming more and more important. And, there wasn't necessarily any sort of formalized process to choose that [DOI]. It was just kind of a recommendation from referred inspiration. (s2)

Another interviewee also discussed her own ideas about her IR situation moving to or adopting a new identifier schema. The changes might not only satisfy the IR content providers but also expand the IR services:

> We do have researchers who are explicitly asking for DOI. We use Handle. Especially for publishers, I don't think they have the same knowledge of other identifier systems. They appear to be saying that you need DOI, so we are getting asked for DOI. Often when we probe a little bit about that [their need for DOI], we find out Handle is sufficient, but not preferred. So, researchers, at least in my experience, so far have been a little nervous about using something that is not DOI. That's not what their familiar with in their publication process. I also think we are interested in potential citation tracking.

> We may possibly use DOI for that. In that way, I think there are two main reasons [to use DOI]: comfort level because of familiarity with DOI, and working with some of the services DataCite is trying to provide, as well as I understand, DataCite is working with CrossRef to provide more services for DOI. (s5)

For potential use in changing circumstances, the identifier schemas that the interview participants mentioned included DOI, ARK, Handle, HTTP URI, and permanent local URL (see Table 4.7). Many of the IRs use the identifiers that came with their IR software, and if the IR staff identified a need to adopt or move to a new identifier schema, they were planning or testing the adoption of a new identifier. Most of the efforts came with the DSpace and Dataverse software that use Handle as a default. In addition to the IRs planning a change, some of the other IRs had already made changes. One interviewee discussed the identifier schema use in her IR. Her IR uses both Handle and DOI with DSpace software, but the use of DOI was optional. The researchers who want to be assigned a DOI for their data have to request one: "We automatically have Handles with the DSpace software and then we also give an optional DOI but it's an opt-in. We don't automatically assign them" (s9). Another case was with Dataverse software. The IR also used both Handle and DOI, but only DOI was visible to the users of the IR. The interviewee strongly indicated that DOI was sufficient for his IR without Handle:

> Right now we just use DOI. Dataverse uses Handle as a default. . . . We had to wait. Right now it [DOI] is appearing on the screen. We were trying to get them [Handle] to not show, but apparently it's hardcoded in. We're hoping they [Dataverse] are going to be coming up with a new version where I think we can remove the Handle and just have DOI. (s14)

A similar change also occurred with Bepress Digital Commons software, which does not use a currently existing persistent identifier. One of the interviewees said that his institution hires Bepress to manage his institution's IR and operate DOI systems separately, in order to assign DOI to data.

Hydra solution under Fedora Commons software not only provides a flexible identifier system environment but also supports a linked data integration platform. The interview data showed that the IRs using Hydra can employ various identifier schemas. They used DOI, ARK, or HTTP URI. One interviewee shared that her institution's IR uses both DOI and ARK, and the two have different granularity levels for their assigned objects: "What I am seeing so far is that we assign DOIs at collection level, and ARKs get assigned to every single digital object within that collection" (s4). Another IR that uses Hydra also uses HTTP URI. In order to generate an identifier, the IR used a software called Nice Opaque Identifier (NOID). NOID can generate two different types of identifiers for short- or long-term uses. The short-term identifiers are more like random namespace-less numbers, but the long-term identifiers are persistent object names like DOI, ARK, and Handle (CDL, 2013). In the IR, researchers self-generate a NOID, and then the NOID is tacked on the IR's URL, which develops a HTTP URI.

> The software is called NOID, which I think is like Nice Opaque Identifier or something like that. It's been a number of years. Anyway . . . every file that's uploaded gets one of these NOIDs, which is in effect a namespace-less verifiable unique identifier, and then what we do is we take that bared namespace-less identifier and we tack on our IR URL, so our approach to a persistent unique identifier is to use HTTP URIs that we mint in the IR. . . . As I just said, we do provide an identifier yield for every deposit. I don't know

> how our users are using that field, so they could be self-populating DOIs or handles or other, maybe PubMed IDs, things like that. (s12)

4.4.2.4 Controlled vocabularies. Most of the IRs have not been using controlled vocabularies. However, two interviewees mentioned that they do have controlled subject lists, which contain Medical Subject Headings (MeSH), Library of Congress Subject Headings (LCSH), and Faceted Application of Subject Terminology (FAST). Two other interviewees also mentioned that they do not have any controlled vocabularies but use DC metadata type and certain controlled format list to organize the format or type of submitted contents.

One of the IRs is exceptionally different in its use of metadata. The IR uses Hydra solution under Fedora software. All of its metadata are modeled using RDF triples. The data is integrated with the concept of linking activity. Linked data principles (Berners-Lee, 2006) emphasize the use of HTTP URI. A datum is represented by a URI, and the two related URIs are linked by another URI. The three URIs accordingly form an RDF triple. The IR's metadata for research data is a series of RDF triples. The IR uses locally modified DC metadata schema and sources of the elements from RDF ontology, Friend of a Friend (FOAF), and RDF Schema. One other interview participant indicated that her institution is planning to change IR software from DSpace to Hydra in order to fully set up an integrated linked data web.

4.4.2.5 Applications for data curation. The interview participants presented diverse applications or tools used within their data curation workflow (see Appendix C). The applications are tied to their purposes: creating and editing metadata, editing images or videos, cleaning data, storing data, identifying and validating data files, transferring data, indexing data for search, tracking and measuring data (see Table 4.7). In the process of editing metadata, different types of text, XML, or disciplinary metadata editors are used. Particularly, simple text editors to create ReadMe files that contain information about data files seem to be widely using by IR staff. The files are normally stored as supplementary metadata for the data objects:

> Honestly, in many cases, we were actually creating [metadata on] just plain text files, so we do a lot of work with ReadMe files, we create just basic ReadMe files that have some amount of structured data. But often it is more of a big sort of place where researchers can talk in unstructured ways about the data. We get a huge variation in datasets. So, often there isn't a standard used. So, yes, for metadata, we tend to rely a lot on plain, readme.txt. (s5)

Another interviewee explained that the main tasks for him and his colleague are to make sure the data is understandable through the metadata documentation. In order to do that, the interviewee not only uses simple text file editors but also employs, in extreme cases, data and metadata conversion and editing tools (i.e., Nesstar). Nesstar is an advanced data management tool that helps the curators easily get DDI-formatted documentations.

> Whatever the researcher uses is what we do. This is a discussed idea, and I do such discussions with my colleague sometimes, because they want to talk to the researchers about using like DDI [metadata]. “Your metadata should be DDI like. Oh, we [researchers] don't know what that means.” Researchers don't really need to worry about it. If the researchers are using spreadsheets, all we want for them is to go through and make sure that they documented their variables and if they have codes, you would see this all the time. The field heading is a three letter code. If you don't have a schema for that code, then you don't know what it means. So, we make sure that we have the schema and then we map it on our side to DDI. And sometimes that's as simple as a

> ReadMe text file, but in extreme cases when we are doing a high-level curation, I would use the Nesstar, a data publisher, where you can put in variable information and tabular dataset information and it will spit out DDI complying xml that can be converted to a PDF or html or whatever. So we try to work with the researchers. We try to be as agnostic as possible, and meet them where they are at and let them use the tools that they prefer to use. (s15)

The interviewees also described conducting data cleaning for the researchers, but they only do it a little bit. The tools identified for the task was Open Refine, which helps clean data and transform data from one format into another.

Software tools or guidelines for data file identification and validation were another interesting finding identified from the interview data. IR staff use a file format identification tool (i.e., DROID) developed by The National Archives to perform automated batch identification of file formats as well as a file format registry (i.e., PRONOM) to support digital preservation. They also used GitHub repository for version control, due to the frequent events (e.g., updates) on data. File Information Tool Set (FITS) developed by Harvard University Library is also a useful tool for identifying, validating, and extracting technical metadata in file formats. IRs could use the file format metadata collected by the FITS for long-term preservation:

> F-I-T-S. File Information Tool Set. What it does is it wraps a bunch of other tools. And what they're responsible for is basically pulling out as much file format metadata as they can. It's a PDF conforming to the PDF 1.7 spec. It's got a width of this; it's got a height of this; it's got a color palette of this; it's . . . many frames per second, that kind of stuff. What we do is we throw all of that away in the repository, so that we have it there for

> the long term and we can search against it, and then, if at some point, when we have a staffing model for format migration, we discover that, for instance, Adobe Version 1.6 is going to be obsolete, we can do a really quick search, find all the content that matches that, and then migrate it to a format that works better. (s12)

To validate, transfer, and package data easily, IRs use BagIt, which is a tool developed by the Library of Congress and their partners in the National Digital Information Infrastructure and Preservation Program. The tool helped IR staff make OAIS Archival Information Packages in order to transmit the archival essence of data and its metadata into their IRs.

4.4.3 Policies, Rules, and Norms

The interview data analysis not only identified some current policies, rules, and norms that the IR staff use for data curation, but also the rationales for their actions around the practice. Some of the IRs currently are developing and improving their practices around policies, rules, and norms. Even though they have some established practices, they keep trying to figure out the best practices from various sources:

> Policies and rules are being created now locally for our system. So, we are trying to figure out things like data retention policies, how to interact with collection development policies, and institutional policies for research data management that are being communicated now. Even though we are accepting stuff [data] into the system, it takes a long time to create the policies. So, we work on that in parallel. (s1)

One of the interview participants indicated that his institution provides data curation services through the IR, but the IR does not have many policies in place:

> We don’t have a whole lot of policies in place right now about what kinds of items get out, how they could be used, and how they're supposed to be done. Over the next year,

> we’re going to be looking at cleaning up some of these areas where we're having gaps in policies and documenting them fully in the hopes of doing a TRAC, Trustworthy Repositories Audit & Certification. (s10)

Some of the IRs tend to start their data service with minimal policies and rules, and then they develop and improve their policies and rules while they operate the services in a bottom-up approach. The sources of the policies they develop were also identified from the interview data. In many cases, the IR staff learn best practices from participating in various academic conferences and curation community training, taking coursework, and benchmarking peer institutions:

> Some of the best practices are sort of what we learned in conferences or in training or coursework we are taking. We look at universities that have been at the research data management game for a little bit longer than we have. So . . . we look to see what they are suggesting for best practices. Kind of combine those things into our local setting. (s1)

The policies identified from the interview data pertain to data management workflow, scope of data, deposit, copyright infringement, accessibility, collection, and preservation. In comparison to other policies, preservation policies have distinct differences between the institutions. One of the institutions has a fairly elaborate preservation policy. It has three different preservation levels, depending on the format of the material. If a material has a proprietary file format, or a file format that is not widely adopted, etc., the material would be categorized as low confidence level, which only provides basic preservation. If a material satisfies the criteria for moderate confidence level or highest confidence level, the material would be preserved by the corresponding level’s preservation actions: “We have some standard preservation activities that are running. I will say that we have different preservation levels depending on the format of the material" (s5). A few IRs also have preservation contracts with the data providers storing datasets. When a dataset is submitted into the IRs, it automatically has a set-year contract (e.g., 5 years or 10 years). When the initial contracts are terminated, the IR staff and the data providers evaluate the datasets for long-term preservation:

> We upload data, write the catalogue in metadata and then over time we will give the researchers tracking results on how often their data has been downloaded and that sort of thing. For these projects that are on a 5 year contract, after the end of the 5 years, which is the term of their contracts, we will meet with them [the data providers] and have them assess what they want to do with the dataset, and they'll have choices. If they don't think it's relevant anymore we can deaccession it or we can see if there's other repositories out there that have come up in the meantime that might be more appropriate for it. (s14)

The duration of the contracts are different between the IRs. One institution uses a 5-year model and the other uses a 10-year model. There is not an agreed-upon model for the contract; they seemed to select the period based on their experiences or other policies existing in their institution:

> Our institution has a policy. It's one of the few schools that ever got their act together and has a data retention policy, where they require 5 years retention of data for any publication; that's where we got our 5 years. I think NFS and other funders vary as far as how long they ask for data to be shared; it could be 3 years or unspecified. I think we mostly got 5 years from our institution policy. It seems like actually we don't really know how long data stays useful, so we figured 5 years is a good starting point, and then

> we'll see how that goes. Some researchers have said, "We think it should be forever." (s14)

There are some recommendations for data curation services, including metadata fields for research data and recommended file formats. The IRs could make them requirements. However, they tend to make them be recommended guidelines in order to keep a higher volume of data in their IRs. They try to keep fewer policies, but then have recommended guidelines: "We do have format guidelines. I am sure you know that people aren't running to put stuff in most institutional repositories, so you really don't want to constraint them" (s15). One other interviewee also mentioned loose guidelines rather than strict policies:

> We don't have very strict policies on file formats. There are file formats we think we can curate longer into the future than others, typically text-based file formats, file formats that have open standards to them, and file formats that are really common and have open source tools that have been created to read and write them. If there's a proprietary format or proprietary binary format that they want to use because it's meaningful within their discipline, or it's meaningful to them, we try to accommodate that whenever possible. If there is an open alternative, we encourage them to deposit both. . . . For example, if a faculty member brings an Excel file and through talking with them we find they don't use any of the complicated features of Excel and their dataset would naturally work well as a CSV file or tab-separated file, we work with them on possibly either converting it or including both the original Excel and a text-based alternative. (s10)

The IRs also use data curation service guidelines (e.g., Data Curation Profiles Toolkit, Data Management Plan (DMP) Tool) developed by well-known institutions. However, many of the

IRs used the guidelines with local modifications. They tended to tailor the guidelines for their needs:

> We've experimented with both the Data Curation Profiles Toolkit and DMP Tool but we don't use either. In most of our data management plans, we just have a good base documentation that we provide on our website. We find that it is more effective than the DMP tool. (s9)

Another interviewee also mentioned a similar idea:

> Data Curation Profiles Toolkit, you know, we have modified that when we meet with the researcher, because a data curation profile isn't essentially a data-oriented reference interview. It is just extremely detailed. . . . So, we do stick within those concepts, but for the most part, it's really slimmed down. (s15)

There is a norm identified by the interview data: some of the IRs apply a subject specialist-centered service model in their data curation services. Because of the complex and diverse types of research data, the roles of subject specialists who can provide support based in disciplinary knowledge and practice are emphasized within the data curation processes in IRs:

> We have a subject specialist-centered service model. So, each time I am in contact with a researcher, each time they create a project in our IR, each time they submit a publication, or when they have a successful grant application or where they use our IR in their data management plan, I would like their subject specialist librarian to know, and they, the three of us, in a perfect world, work together to get them to use our IR. (s3)

A different interviewee also described her institution's service model for data curation:

> We always recommend that people use norms from their disciplinary practice. We don't have a full list of those written, but we refer to them. That's one of the reasons that we

> work through this specialist or liaison model to make sure that we do know those norms from the different communities for any of the technical data support that we're doing. (s11)

4.5 IR Staff Role-Related Skillset

The interviewees discussed skillsets that they think they need for research data curation. The identified skillsets can help develop a team of data curators' professional expertise. Most of the skillsets discussed are interchangeable between the roles of IR staff, but the researcher mapped the skillsets based on the staff's role-related activities (see Table 4.8). Dotted lines within the table mean that the skillset is not limited to the specified role; solid lines mean that the skillset is more or less limited to the specified role. The main skillsets contain eight distinct types of knowledge or skills, including metadata, domain knowledge, research practice, curation lifecycle, software, library technology skill, data description and documentation, and communication. Almost all of the interviewees mentioned knowing how to create metadata as a necessary skill. One of the interviewees clearly expressed that hiring someone who already knows metadata and diverse disciplinary practices will be more efficient in completing their work:

> . . . . and then as far as the training for the folks working with the creation of metadata . . . That's usually where the most work has to be done when bringing in someone new. Because it just takes awhile to see all the different variants that you have. You can have documentation, but it takes a lot of time. You just need to see a variety of different kinds of content work with a variety of different disciplines to understand how they differ. (s10)

Table 4.8. IR staff’s role-related skillset

Head	Data Curator	IR Manager	Metadata Specialist	Developer	Subject Specialist	Graduate Assistant
Understanding of data curation lifecycle						
	Long term preservation knowledge					
	Familiarity with research data (e.g., Ability to handle data complexity and diversity)	Collection management skill	Metadata knowledge particularly for research data	Technical details of repository software, server, and its architecture	Understanding disciplinary metadata, workflows, and knowledge	
	Academic research practice	Software skill			Collection management skill	
• Library practices, needs, and technologies • Data management practice • Soft skill (i.e., communication)				• Ability to communicate and work within a team • Data description/documentation skill • Time management		

Another interviewee considered knowledge of metadata and plenty of experience of creating metadata as an important skill for IR staff:

> My theoretical conception of what metadata is and what it should be doesn't align very well with what the researcher’s conception is. There has to be some kind of compromise there, and we try to understand how we take this. There are some highly detailed and highly formalized schemas, and then take these things that are really amorphous, and how do you put those things together, and still have metadata that is useful and serves its purposes. I think that is going to be one of the technical skills that we are going to have to have more of. (s1)

An interviewee even mentioned metadata knowledge particularly for research data as the most important and ideal skill for staff who work on research data curation:

> The most difficult thing that we struggle with is the metadata on all three of those levels [data curators, data providers, and users]. That would be the ideal skill for someone to have, to be someone who can figure out the correct level of metadata and the range of data. (s9)

In order to support metadata creation, data curators need to understand disciplinary knowledge, research practices, and data curation lifecycle (workflow). One interviewee emphasized domain knowledge as a skill of data curators: “Domain knowledge, especially data complexity and diversity to bridge metadata librarians and researchers” (s4). Another interviewee highlighted the importance of understanding academic research practices: “Obviously, familiarity, deep familiarity with research practices in academic research and its research data needs, experience with and knowledge of library systems and practices, and how each of those operate with the other” (s11). However, this does not mean the IR staff need to be biologists, chemists, or physicists. One other interviewee stated that his institution does not hire librarians with degrees in fields other than library studies. Instead, his team wants people who know the data curation lifecycle and its workflow:

> Librarians who are willing to understand some of the various research methods that get used within these disciplines, and are able to communicate and ask questions to learn more about those datasets and different kinds of scholarship. I don’t need librarians with degrees in Biology to go talk to a biologist. I want people that know what we’re going to do with the data once we get it and to be able to ask questions of the biologist that get them to give more meaningful information. I don’t think it's possible for us to have subject experience in the vast quantity of subjects here in the library, but it's more a matter of finding people that can ask good questions, know the curation lifecycle, and can go through and communicate well with the faculty. (s10)

Managing software and understanding library information technology is another skill or knowledge needed for IR staff. However, IR staff have different levels of familiarity with certain technologies. One interviewee tried to explain what the right amount of technological knowledge might be. But it added more ambiguity:

> I would say it is substantially different in that it depends on what the data is. We had to use tools you know and understand. [For example,] Excel. There is a difference in how to use Excel and how to tease or parse data in Excel. Also, using an
> Excel spreadsheet for multiple sheets—what that means and the problems with that. . . . In order to have them (IR staff) ask the right questions . . . in my mind, it's little more than just having a certain level of comfort working with software. (s5)

An interviewee introduced a typical case illustrating the need for technological skill:

> We prefer to provide a text file because it is more accessible to people, and then we work with them to create a text file version of that [original data file]. Some researchers are comfortable with that; they say, “Okay, I'll just export it and give you another version.” Some people say, “I don't know how to do that.” So, we have to handle the process. (s1)

The interviewee also presented an extreme case requiring technological skill that entailed using software for a complex research data type:

> One of the things I have seen that really gets complicated is what people do to record their research. They do what they have to do. Sort of make their research more efficient. I met with someone who was doing medical research, and he had an Excel file. He did complicated micro-species research. [The file contained] thousands of images he was creating, and he would then ingest those from proprietary analysis software to create figures that he could use to analyze data, and he also could use them for publication. Right? So, he actually took a screenshot of the proprietary software and then embedded a JPEG of that screenshot into an Excel file. He created extremely complicated digital objects that if you are trying to flatten them it would be very hard. But so we actually haven't solved the problems. But I think this is a kind of problem. It's kind of come up. I mean, it goes beyond the normal equations that are built into Excel—other files embedded in files, how do you explore that in digital objects? (s1)

The final abilities identified for IR staff include interpersonal, communication, and documentation skills. Some of the interviewees stated that interpersonal and communication skills are important because research data curation is a relatively new field, which means there is a lot of uncertainty or ambiguity. Clear communication between IR staff and data providers enable smooth data curation processes:

> Communication, I would say it's even more important because it's a relatively new part of the field. There's a lot of uncertainty, there's a lot of ambiguity, the researchers are not always comfortable with talking to somebody who might not be a domain expert about the data and managing their data, and so just being able to have an intelligent and productive conversation with people is far more important than whether you can write the script of Python or whatever. (s15)

One other interviewee also presented a similar perspective on communication skills:

> Having someone with great interpersonal skills, having someone who is not only comfortable with changing unknown practices, but someone who's excited about [research data curation], that's hard for a lot of people, but that's really important. Being comfortable with ambiguity, being comfortable with confusion, and being really comfortable with failure, because if you want to succeed, double your failure rate. You have to be able to be comfortable and make other people feel comfortable and calm. (s11)

In addition to that, some interviewees explained that both communication and documentation skills are important for increasing researchers' trust in their services and encouraging them to share their data. One of them stated:

> The research data, that's really a big deal and you're asking them to give it to you? This is either more work for them, or you're asking them to trust you with walking their baby. They really need to feel comfortable and to feel supported. They need to know that you know the answers or that you're going to find out the answers. A huge amount of work

> goes into establishing and supporting trust. Some of that trust is with documentation. (s11)

Another group of the interviewees described IR staff as almost all soft-skilled people. In order to complete all of the research data curation processes, a significant amount of presentation, communication, documentation, and teamwork skills are required in IR staff. One of them described the tasks that need soft skills:

> Our data curators are almost all soft-skilled people. They go to promote our services, and they work with a lot of faculty members on their data management plans. They are using the right language and explaining things like the difference between backup and preservation. (s9)

4.6 Major Types of Research Data and Their Entity Types Within the IRs

In response to the interview question, “What major types of research data does your IR accept?” all of the interviewees stated that they accept any type of data. They tend to accept any file format that can be downloaded. One of the interviewees explained:

> We’re able to accept just about anything in the digital format that they want to include, as long as it can be downloaded. Something like a database file would be downloaded just as a database file; we don’t have any interface for them to use it like a database online. They just download the file and run it on their own software. (s14)

Another interviewee also mentioned that they accept all types of data: “We accept all types. We’re pretty non-admonitory. We currently have a lot of raw data, text documents, spreadsheets, and things like SPSS files and stuff like that” (s9). One interviewee even said that if data providers want to share it, they will put it in the IR. Table 4.9 shows the major types of research data deposited in the IRs, as well as some criteria for data-type limitation. The responses to the

Table 4.9. Major types of research data and their entity types

Major Types of Research Data

Any types of data (e.g., Raw data, Text documents (e.g., Word, PDF, LaTeX, TXT), Spreadsheets (e.g., Excel), Slides (e.g., PowerPoint), Audios, Audio-Visuals, Images, Laboratory Notes, Statistical data files, Databases (e.g., Access, MySQL, Oracle), Software codes, Tabular data files)

Criteria for Data File Properties

File Capacity, The Number of Files, Proprietary Files Extension (e.g., .exe)

Entity Types	**Metadata Elements**	**Identification Schemes**
Intellectual Entity (See page 25)	Title, Main Title, Other Title, Abbreviated Title, Subtitle, Abstract, Grant, Citation, Supplementary Information, Description, Material Type, Language, Target Audience, Reviews, Open Summary, Subject Summary, Identifier, Related URL, Right	DOI, ARK, Handle, HTTP URI, Permanent Local URL
Object	Title, Identifier, Related URL, File Format, Description, Supplementary Information, Note, Citation	DOI, ARK, Handle, HTTP URI, Permanent Local URL
Symbolic Object	Title, Identifier, Related URL, File Format, Description, Supplementary Information, Note, Citation	DOI, ARK, Handle, HTTP URI, Permanent Local URL
Person	Author, Creator, Contributor	Local Name Authority Records, ORCID
Organization	Larger body of work, Publisher, Source institution, Physical Container, Funder	Local Authority Control System
Place	Place of Publicaiton, Holding Location, Spatial Coverage, Coordinates, Physical Container	GeoName Database (GeoNameID)
Time	Date, Publication Date, Copyright Year, Temporal Coverage, Time	
Event	Process, Publication Status, Edition	
Topic	Subject Keyword, Methodology, Genre	LCSH, MESH, and FAST with HTTP URI

question on the major types of research data accepted into the IR led the researcher to make one additional interview question: “Is there any limitation regarding the file types that you accept in your IR?” Many interviewees answered this question with much more interesting responses. Some of the interviewees mentioned that the size and the number of files are restricted. One of them stated:

> We can't accept a terabyte of research data all at once. Also, because the only sort of interaction with the items is through downloading and uploading, you are restricted by the size of the items in that way as well. So, you can't add items that are just too big to download or too big to upload. So, size is a big restriction. (s5)

Another interviewee added that the number of files is also a criterion of file-type restriction:

> We have a mirror [repository] for some of our research dataset. If it's too big for our repository software and if there are too many files for our IR platform, we create a shelf record in the platform and then we have a link out to the files in the mirror. One of our datasets is almost a terabyte, something like 5000 files. (s15)

Proprietary file extension such as .exe could be a restriction, but only one IR actually restricts the file type. Most of the other IRs allow researchers to submit the files, although they recognize the problem that comes along with the file type and recommend not using the file format. The interviewee of the IR that restricts proprietary file types explained: “We don’t allow people to upload .exe files directly. They have to zip them because we don’t want anyone to have an .exe file on the server. Someone else could download it and have a problem on their machine” (s11).

The interview data identified some metadata elements that are offered by the IRs. The elements were analyzed and mapped into the research data entity types identified by the researcher’s previous study (Lee & Stvilia, 2014). Table 4.9 shows the mapping and identification schemes used for the entity types. All of the metadata elements could be mapped into one or more entity types, but many of the interviewees indicated that they frequently discuss the relationships between these abstract entities: intellectual entity, object, and symbolic object. The discussions primarily focus on selecting an appropriate entity level for a dataset when IR staff organize and deposit the dataset into the IRs. The IR staff could deposit a dataset into a collection level (i.e., intellectual entity), and they also could deposit each object of the dataset into an object level (i.e., object entity or symbolic object entity) along with different metadata. Currently the staff make the decision according to their communication with the data providers. However, in many cases, data providers just rely on the IR staff to decide. One of the interviewees mentioned that they deal with those kinds of discussions on a daily basis:

> One of the things that has been somewhat tricky and we've always tried to help researchers with is trying to find that right level of description for how things are grouped. Sometimes, we will actually create at the intellectual entity level a separate collection in DSpace, just because it's easier to group the different objects under that intellectual entity. Sometimes, the objects will have different metadata, but they're still one intellectual unit for the purpose of the dataset as a whole. . . . Those three actually, it's interesting. They're tricky concepts, but we deal with those on a daily basis. They're very much a reality when you're a researcher with a bunch of data in front of you, trying to figure out how you want to group the data. (s9)

One other interviewee also indicated that there is no systematic recommendation or guideline for this issue:

> Within the system, for us, any of those [datasets] can be modeled either as individual items, if you, as a researcher, wanted to deposit a number of research data samples as a

> single entity [intellectual entity], or as multiple items [object or symbolic object]; it could be done either way. It really just depends how much time you want to put into modeling a collection. Either way it could be done for us. The notion that it's symbolic object versus object, it really, once again, depends on how you want to formulate it as a researcher, and we try to allow for multiple representations, if that's how you [the researcher] want, or multiple avenues to represent it depending on what works the best for you. We typically have a conversation with the researcher to understand what they're trying to accomplish with it. Many times, they don't care. They just want us to tell them how to do it; however, sometimes they have a preference for it and so then we try to accommodate that if possible. (s10)

All of the entity types except time and event entities are at least identified by one of different identification schemes (see Table 4.9). Among the schemes, identifier schemas (i.e., DOI, Handle, ARK, HTTP URI, Permanent Local URL, and GeoNameID) identified the entities of intellectual entity, object, symbolic object, place, and topic. The identifiers that can be assigned to datasets could identify the first three entities. Place entity can be identified by GeoNames database (i.e., GeoNameID). Topic entity is specified by different subject lists that are supported by linked data technique (i.e., HTTP URI). One of the interviewees mentioned how they use identification schemes for different entity types:

> We have a system where we try to use the IR content with name authority records that we create here locally. We create authority records for all of the authors that contribute. We make use of GeoNames and its database for place names and geographic locations. We have a number of different subject lists we can use LCSH, MeSH, and FAST from OCLC as well as just keywords. . . . We make use of the extended date time format for

> time formats so you can go through and have a consistent machine-readable format for talking about dates. … Places, organizations, and people, we have a system for authority control on those as well. (s10)

In addition, person entity has a higher chance to be controlled by ORCID identifiers as some of the IRs are planning to adopt ORCID in their IRs:

> We were talking about ORCID. We were talking about pushing to get ORCID for every faculty member at the university. We are in the discussion right now about serving ourselves into that space. We are saying this is the identifier we are going to use at the university for people. (s1)

4.7 IR Staff Data Identifier Awareness

This section, based on the interview data analysis, describes IR staff's data identifier awareness, including their degree of familiarity with and reasons why they use a specific identifier within their IRs. Understanding their current awareness can provide insights into IR staff's current metadata literacy, particularly regarding identifier schema; it can also indicate directions toward future identifier systems and services used within IRs.

4.7.1 The Degree of Familiarity

The interview data indicated that most of the interviewees are familiar with at least one of the existing identifier schemas; however, the degree of familiarity differed between interviewees. More than half of the interviewees only knew general information about the identifier schemas widely used in the community. Less than half of the interviewees knew technical details about various identifier schemas. Figure 4.2 presents the number of IR staff who are familiar with specific identifiers. Even though the degree of familiarity is different, the interviewees who introduced themselves as familiar with specific identifiers were counted as such in the table.

The interview subjects who belong to the first category tend to know broad information about the identifiers or the limited information needed to use their IR software. One of the interviewees stated that he knows general information about the identifiers but not technical details: “I am familiar with DOIs. We have PURL, Handle, and ARK. My understanding of how any of those actually technically differs and what are the pros and cons of each are very limited” (s1). One other interviewee explained that he knows identifiers only well enough to use IR software:

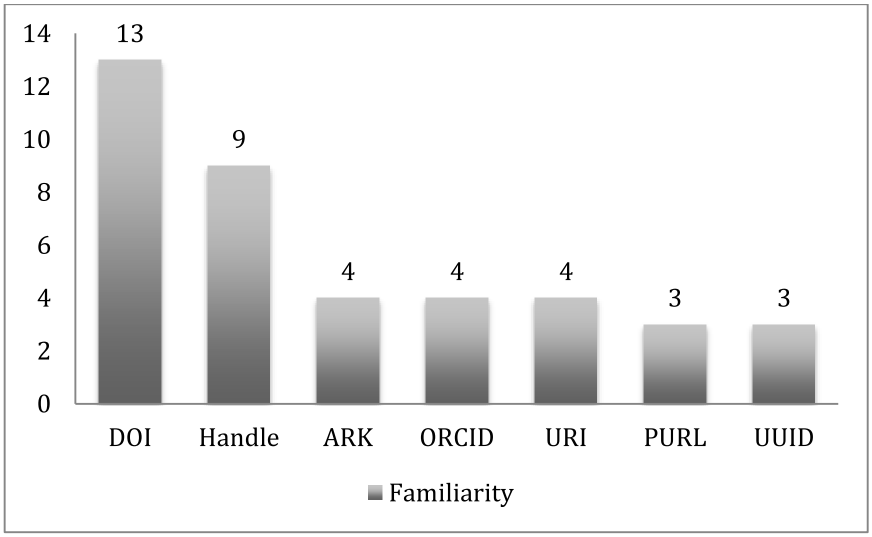


Figure 4.2. Number of IR staff who are familiar with specific identifiers

> This is where my ignorance comes out. [Laugh] I guess I am familiar [with identifiers] very broadly from using IR or IR platforms. For example, the persistent identifier in Fedora or Islandora. DOI, it is globally unique and resolvable. I am aware of URIs, and I am familiar with CrossRef, DataCite DOI, and their community focus. So, yes, very broad. I haven't had an opportunity to dig it deeper. (s3)

The interview participants in the second category know technical details about different identifiers. They learned about identifiers to meet institutional needs or to satisfy their own interests. One of the interviewees discussed that while his IR uses Handle as a main object identifier system, they have thought about adopting a new identifier schema (i.e., DOI). In that process, the interviewee learned details about DOI:

> We were investigating DOI implementation here. We thought briefly about minting our own. But we did some research and decided that it's just easier to join another service like EZID or whoever. So, by doing research, I learned a lot about policies behind DOI, the DOI founding body, and what one expects in making DOIs. (s15)

Another interviewee has an interest in identifier systems and plenty of experience managing the servers. He could explain the pros and cons for each identifier system that he is familiar with:

> There’re a lot of them. Handles, ARKs, DOIs, PURLs, URIs and URNs, ISBNs. I could probably name 20 of them. I just hinted at this in the last segment. I said this has been a focus of mine, so I’m pretty familiar with the different standards and what the tradeoffs are for using a DOI versus a Handle versus an ARK versus other. I’ve run the software before. I’ve run the PURL server before. I’ve run ARK servers before. (s12)

All of the IRs except one in this category use ARK, HTTP URI, or permanent local URL, although DOI and Handle are the most widely used identifier systems. They tend to be independent in selecting and adopting their own identifier systems in the context of their accumulated knowledge. On the other hand, the IRs in the first category (the ones with only broad knowledge) were somewhat dependent on the default identifiers of the IR software that they use.

4.7.2 Why Use the Identifier Schema

In order to select an appropriate identifier system for an IR, some of the IRs are independent of the IR software that they use. Instead, they use their own knowledge and library resources to select a system. On the other hand, some of the IRs are dependent on the IR software that they use. They use the default identifier system embedded into the IR software without any discussion. One interviewee supported this idea with her IR practice. In addition, the interview data show that there is a lack of policies, rules, or norms to govern or guide identifier system selection and use in the IRs. In response to a question about policies, rules, or norms, all of the interviewees mentioned that they are not familiar with them or that there are no such policies. One of the interviewees stated: “I am not really sure what made our decision [to use our current identifier system]” (s14). Another interviewee also explained:

> If we were talking about the built-in identifiers that we create, I wouldn’t say we have anything like rules or norms, anything like that. They’re just generated by the system. When we were originally building our repository, we were pretty free to do what we wanted to do, what matched, and what we really wanted out of an identifier system. I think I can say we’ve been really happy with our identifier selection since then. It’s never failed us. (s12)

There was another interviewee who talked about norms in the larger library community. She indicated that her IR looks at norms when they work on their identifier system: “The norms are the norms from the larger library community and research community. Any material has to have permanent identifiers. That’s what we have for our permanent URLs. Then that’s what we’re looking at [community norms]” (s11). According to the interview data, current unsystematic practices for selecting or adopting an identifier system may arise from the lack of specific policies, rules, or norms.

To find out practical reasons for unsystematic practices or to identify the current criteria used to select an identifier schema, the researcher asked interviewees to share the reasons why their IR uses one or more specific identifier(s). The interview data suggested five different criteria, which include (1) authorities, (2) local resources, (3) expert knowledge, (4) new technologies or services, and (5) community needs. The IR staff adopt an identifier system based on one or two of the five criteria. Figure 4.3 indicates the connections that may exist between the criteria for selecting an effective identifier system for local needs. The first criterion derived from the interview data is authority. IR staff tend to accept an identifier system based on its reputation in a given community:

> I think in terms of the decision to move to DOI, it was because DOI is becoming more and more important. So there wasn't necessarily any sort of formalized process to choose that. It was just kind of a recommendation from referred inspiration. (s2)

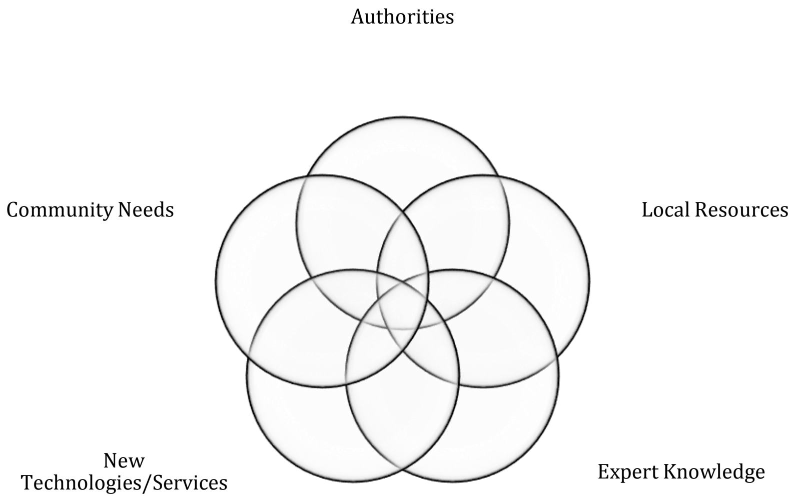


Figure 4.3. Current criteria for identifier schema selection

The availability of local resources also affects which identifier system IR staff decide to use. The second criterion, local resources, could encompass a variety of things, including financial, human, or sociotechnical resources. For example, one IR uses DOI because the institution is one of the DOI registrars and the identifier is a widely used identifier system. In this case, the IR uses the DOI system because it is already a local resource in that institution and the system has a reputation:

> DataCite is an international consortium. Our institution was interested in expanding its global outreach and the strategic goals of the library. Also, it is extremely important that our institution is a DataCite registrar. But, I think DataCite is the most internationally recognized digital object identifier (DOI) community. It contributes to our institutional goals in a number of ways as well as community goals. (s3)

One other interviewee also indicated that his IR adopted a specific identifier system because of local human resources. A group of people in his department were involved with the development of the specific identifier system that they use:

> We went with EZID because some of the people in our department were following that, and have been involved in that project since its beginning. . . . It's one of the reasons we're using that. I'm not sure about what the decision was not to use Handles. I think just EZID seemed to be more secure for preservation. I actually don't know why we didn't go with Handles. (s14)

There was another interviewee who explained that her institution uses a default identifier system that comes with its IR software without any argument. She also mentioned her IR selected an IR software based on institutional sociotechnical resources (e.g., division of labor, norms, practices, tools, community); her IR selected DSpace because when they made the decision, DSpace was the only software with large support and a big user community. Also her institution’s library technology team preferred to control the development and use of the software.

The third criterion is expert knowledge. One of the IRs developed a system based on expert knowledge. If an IR department has an employee with strong expertise in identifier systems, the department might rely on his or her knowledge to select an effective identifier system for the IR:

> To some extent, there are policies or rules, but identifiers and identifiers on the web has been an area of interest of mine for a good 15 years, so to some extent, I was vocal about what I thought we should do and I think a lot about what we should do. . . . so that maybe the only rules or norms were the ones in my head that I kept repeating at people. (s12)

One of the IRs plans to migrate to a different IR platform and its derived identifier system. The main reason for the migration is the fourth criterion identified, the adoption of new technologies or services:

> We are moving probably in a year or two to Hydra [an IR platform]. In that system, we already have a set up for us, a fully integrated controlled semantic web with link data framework. At that time, we’ll be converting all the data and we’ll be controlling a lot more, but in DSpace, we let it be pretty free flow. (s9)

The last criterion for identifier schema selection is community needs. One of the IRs is thinking about adopting a DOI system as a response to community requests. Along with the community needs, a new service that can support DOI adoption is another reason for the change:

> We do have researchers who are explicitly asking for DOI. We use Handle. Especially for publishers, I don't think they have the same knowledge of other identifier systems. They appear to be saying that you need DOI, so we are getting asked for DOI. Often when we probe a little bit about that [their need for DOI], we find out Handle is sufficient, but not preferred. So, researchers, at least in my experience, so far have been a little nervous about using something that is not DOI. That's not what their familiar with in their publication process. I also think we are interested in potential citation tracking. We may possibly use DOI for that. In that way, I think there are two main reasons [to use DOI]: comfort level because of familiarity with DOI, and working with some of the services. (s5)

According to the interview data, the lack of policies, rules, or norms could be one of many reasons for the existence of various and unstandardized criteria used to select identifier schemas for IRs. Also, the absence of recognized best practices could be an obstacle to increasing IR staff's data identifier awareness. As a result, diverse factors tend to influence decision making processes rather than a systematic understanding of identifier schemas.

4.8 IR Staff Perception of Data Identifier Quality

Data identifier quality can be defined as the degree to which the identifiers meet the requirements of the activities in which they are used (Stvilia et al., 2007; Wang & Strong, 1996). The concept of quality and its dimensions are usually perceived when there are quality problems, which happen when the existing identifiers do not meet the activity's needs. Data analysis on IR staff perceptions of data identifier quality not only identified some quality problems with the identifiers, but also described IR staff perceptions of identifier quality dimensions (Lee &

Stvilia, 2014). The analysis also presented different and interesting perspectives among IR staff on identifier services and its management within an IR setting.

4.8.1 Data Identifier Quality Problems

Many of the interview participants mentioned that they do not have any identifier quality problems, such as access failure or incorrect access. They simply responded, “No. We don’t have problems with our identifier (s4)” or “It’s never failed us” (s12). But some of the interviewees described identifier quality problems that are linked to identifier schemas, their servers, or IR software implementation. Table 4.10 summarizes the identifier quality problems and their types, sources, and corresponding assurance actions.

Table 4.10. Identifier quality problems and their types, sources, and assurance actions

	Quality Problems	**Problem Types**	**Problem Sources**	**Assurance Actions**
Mapping-Related Problems	Inconsistent identifier assignment	Inconsistency	Lack of related policies and best practices	
	Incomplete system update and maintenance	Incompleteness	IR domain name change	Maintain up-to-date with the IR
Context-Related Problems	Garbled identifier strings	Incompleteness	Software incompatibility	
	Inconsistent implementation of identifier systems	Spatial context change	Undocumented feature of identifiers	
	Unstable server condition	Temporal context change	Unstable server	

4.8.1.1 Inconsistent identifier assignment. An inconsistent level of identifier granularity means that an identifier can be assigned to any level of data objects (i.e., collection, file, or entity levels), depending on the type of data package. Although the inconsistency provides some flexibility with identifier assignment, it also delivers inconsistent levels of identifier granularity over the flexible assignments. In those cases, the same type of data package could have different levels of identifier granularity. The inconsistent level is a quality problem that happens due to lack of related policies or best practices:

> For some situations, files get DOI. . . . In some cases, those data files might be in multiple formats, so that, in that package, we give a DOI. . . . We just don't have a very good policy around defining what gets DOI and what doesn't. I think that’s because research data stuff is so new. So we have run into a lot of problems. (s1)

4.8.1.2 Incomplete system update and maintenance. This is a typical quality problem that can happen with identifier persistency. Incomplete or belated system updates and maintenance can cause identifier access failures or incorrect access when an institution upgrades or re-configures its IR platform. If an institution changes its domain name, it will also require a rapid system update or maintenance:

> I mean we had cases where our Handle server was disrupted, when we upgraded and misconfigured [the Handle server], because we had to and we also changed our domain name, we had to do a lot of work behind the scene to update the information. It didn't go smoothly. So, suddenly everything was disrupted. (s5)

4.8.1.3 Garbled identifier strings. A data identifier string is a sequence of symbols designed to identify, cite, annotate, and/or link research data. The sequences include different types of alphanumeric characters as well as punctuation characters including colon, slash, hyphen, parentheses, etc. Lately, in many mobile devices or instant messengers, the punctuation characters are being used to form various emoticons. As a result, when someone shares an identifier string in those systems, the system sometimes interprets the characters as an emoticon: “Depending on how you have your instant messenger set up, if you share an ARK identifier with another user, sometimes the system can interpret that colon slash as an emoticon” (s10).

4.8.1.4 Inconsistent implementation of identifier systems. There are different localized methods for implementing an identifier system in its underlying IR platform. In many cases, the methods and their detail functions are not documented as a specification to ensure that they are only used in specific IRs. The inconsistent implementation of an identifier system can produce confusion in its users:

> To fully support ARK identifiers, you are able to add a single question mark at the end of the identifier to get back a brief metadata record and double question marks at the end to bring back a service agreement. Implementing that function can be hard for some web frameworks because it’s an undocumented feature. It’s a weird behavior that’s not usually covered [by specification] and how the web servers are implemented. It may work for certain kinds of web servers and it may not in another. So, it’s hard to implement on the software side. It has nothing to do with the quality of the identifiers, but it’s more of just an implementation challenge you sometimes can run into. (s10)

4.8.1.5 Unstable server condition. Identifier systems depend on the Internet and server conditions. If the technology is not stable, some of the functions of identifiers cannot be conducted. However, a solution for the unstable conditions still remains in primitive ways: “Occasionally, our Handle server needs to be restarted, which is funny” (s9).

4.8.2 Data Identifier Quality Perception

The existing data identifier quality problems indicate the IR community’s perceptions of and priorities for data identifier quality (Huang et al., 2012; Stvilia, Twidale, Smith, &Gasser, 2008). To help understand the IR staff perceptions of data identifier quality, the interview participants were asked to rate the importance of 11 existing quality dimensions on a 7-point Likert scale from *extremely unimportant* to *extremely important*. The existing dimensions selected are based on the data identifier taxonomy (Lee & Stvilia, 2014). The researcher did not use the dimensions of contextuality and compatibility in this study. Contextuality can be defined as the degree to which an identifier system and string meets the needs of a targeted community, and compatibility can be defined as the ability to use the identifier scheme with the main internet naming schemes (Lee & Stvilia, 2014). The meanings of those two dimensions can be integrated with opacity, simplicity, and interoperability. The analysis of the interview data first of all identified the IR staff’s different perspectives in rating the quality dimensions, and provided some understanding of how the IR staff perceive and prioritize the quality dimensions.

4.8.2.1 Different perspectives in rating the quality dimensions. The interview data discovered IR staff have four different perspectives in rating data identifier quality dimensions. How they rate the dimensions depends on (1) the data provider’s activities, (2) repository domains, (3) identifier workflows and their core functions, and (4) the dimensions’ interconnectivity. One of the interviewees mentioned that researcher activities associated with sharing their research data are important criteria for rating the quality dimensions. She would rate the dimensions according to their value to researchers:

> My perception is a mix of what I think would be important to researchers. I guess I would sell ideas of different identifiers to someone wanting to share their research data. What would be valuable to the researcher? That's my perspective. (s8)

Considering researcher perspectives when rating the quality of identifier schemes is linked to one of IR staff's activities: outreach services to motivate and increase the usage of IRs.

The second perspective shaping how quality dimensions were rated is repository domains. An interviewee indicated that different domains of repositories could influence how repository staff rate the quality dimensions:

> I am not assigning something for a specific community; they get specific identifiers. When I think about DOI, I would assume that some publishers do care about different pieces of the DOI and how they are assigning that in working intention. It doesn't really matter to me. I don't think it matters much to my users either. (s5)

Repositories and their identifier systems have different purposes and users based on their domains or communities. Particularly, some data identifiers designed for hard science data have different identifier string requirements. For example, NCBI's accession numbers have specific rules for generating an identifier string for gene sequences, rather than assigning a random number (NCBI, 2012).

The third perspective identified by the interview data was that IR staff rate the quality dimensions based on identifier workflows and core functions. According to an interviewee, uniqueness and persistence are the most important quality dimensions for data identifiers. If those dimensions are satisfied, the next are the dimensions needed for verification of the identifiers, and then finally interoperability and actionability:

> Either interoperability or actionability might be something that comes later. It’s not necessarily inherent in the initial assignment. . . . Verifiability is with the same logic. Uniqueness and persistence are first, and then you need to verify those identification schemes and then think about and consider interoperability and actionability in your logic. (s9)

The last perspective on how to rate the quality dimensions that was discovered by analysis of the data is the dimensions’ interconnectivity. One interviewee rated all of the dimensions as *extremely important*. According to her perspective, all of the dimensions are interconnected in conducting their individual roles:

> Yes, they are all seven extremely important. You can’t have an identifier system that isn’t secure, that isn’t scalable, and that doesn’t have authority, granularity/flexibility, actionability, and accessibility. I mean those are core. Those are fundamental elements. . . . I would keep them all extremely important because they have to be unique. If it’s not unique, then it’s not resolvable and actionable. If it’s not persistent, then nothing else matters because it’s not there. If it’s not interoperable . . . If they don’t relate to each other and if they are not actionable and resolvable, then they are irrelevant. Why have an identifier that you can’t use. They can’t be usable unless they are persistent, unless they are unique. (s11)

4.8.2.2 Priorities for and perceptions of the quality dimensions. The interview participants responded to the rating question, and the results demonstrated that most of the interviewees think all of the dimensions are important for maintaining the quality of identifier

Table 4.11. Mean importance ratings for the identifier quality dimensions

Dimensions	Mean	Median	Std. deviation
Uniqueness	6.93	7	0.25
Persistence	6.92	7	0.26
Actionability/Resolvability	6.73	7	0.59
Scalability	5.86	6	0.83
Authority	5.73	6	1.36
Interoperability	5.73	6	1.75
Security	5.57	6	1.28
Verifiability	5.46	6	1.45
Granularity/Flexibility	5.26	5	1.16
Opacity	4.33	5	2.02
Simplicity	4.13	4	1.55

Table 4.12. Identifier quality dimensions and their definitions (Lee & Stvilia, 2014)

Identifier Quality Dimensions	**Definitions**
1. Uniqueness	The requirement that one identifier string denotes one and only one data object
2. Persistence	The requirement that once assigned, an identifier string denotes the same referent indefinitely
3. Actionability/Resolvability	The ability of the identifier system to locate the object using an identifier string
4. Scalability	The ability of an identifier system to expand its level of performance or efficiency (e.g., support RDF)
5. Authority	The degree of reputation of an identifier system in a given community
6. Interoperability	The ability to use an identifier system and string in services outside of the direct control of the issuing assigner
7. Security	The extent to which the resource of an identifier system is protected from unauthorized administrative access or modification
8. Verifiability	The extent to which the correctness and validity of an identifier string is verifiable or provable
9. Granularity/Flexibility	The extent to which the identifier system allows referencing data at a different granularity
10. Opacity	The extent to which the meaning can be inferred from the content, structure or pattern of an identifier string
11. Simplicity	The degree of cognitive simplicity of an identifier string

schemas. All of the mean values of each dimension were higher than four out of seven (see Table 4.11). Four indicated a neutral perspective on the dimension’s importance. The mean importance of the participants' rating is shown in Table 4.11. Since the number of the participants is not sufficient to generalize the results, the numbers within that table are only used to see their perceptions of the quality dimensions and to provide some insights for future studies. Table 4.12 presents the definitions of identifier quality dimensions.

4.8.2.2.1 ***Uniqueness.*** Uniqueness can be defined as the requirement that one identifier string denotes one and only one data object (Lee & Stvilia, 2014). One of the interviewees explained why he considers the dimension of uniqueness to be important. His concern for unique identifiers is motivated by the dynamic and fragile features of current URLs on the Web:

> I think the uniqueness is extremely important, and part of the reason is the expectation of using identifiers in today's Web world. If I put in a URL today, tomorrow it's changed. Something is gone and wrong. So, uniqueness, I want to make sure we keep the content around. If an identifier identifies a dataset about Ebola virus today, I want to make sure that it doesn't mistakenly identify a photograph of a fish skeleton in a week. It's important for us to maintain access, to make sure that those identifiers stay with it. (s12)

4.8.2.2.2 ***Persistence.*** Persistence refers to the requirement that once assigned, an identifier string denotes the same referent indefinitely (Lee & Stvilia, 2014). This dimension was perceived by the interviewees as a main characteristic of identifier schemas. One study participant stated: "Persistence differentiates the IR from a vendor-based repository. We intend for this thing to be around for as long as our other library collections are around. So, we really want everything to have a long persistence" (s12). One interviewee raised an interesting discussion question on the persistence dimension. According to him, persistence is institutions' efforts to make the data permanently accessible, rather than an identifier quality dimension.

Keeping persistent identifiers is a commitment of institutions managing identifier systems, not a quality criterion of identifier systems:

> That [persistence] is an institutional commitment to your identifiers. I think that one is commonly completely misguided in people's identification or selection of identifiers. They think that somehow Handles are more persistent than other formats and that's a completely incorrect statement in my opinion. . . . I don't think [persistence] is actually a valid characteristic of an identifier. . . . I know ARK identifiers that have been incredibly persistent for the past eight years. I've seen other people's ARK identifiers that are horribly persistent, and it has nothing to do with the identifiers. . . . It's just how you managed your system, which really doesn't have anything to do with identifiers. Persistence is great, but it's not a characteristic that's inherent to the identifier in my opinion. It's the implementation of the identifier. (s10)

4.8.2.2.3 ***Actionability/Resolvability*****.** Actionability/resolvability in locating the object using an identifier string was also one of the main properties of identifier schemas (Lee & Stvilia, 2014). One of the interviewees directly addressed the importance of actionability/resolvability: "If they are not actionable and resolvable, then they are irrelevant" (s11).

4.8.2.2.4 ***Scalability.*** Scalability can be defined as the ability of an identifier system to expand its level of performance or efficiency (Lee & Stvilia, 2014). Many of the interview participants consider an identifier system to be inseparable from a repository system. The relationship between repository and identifier system affected ratings of this identifier quality.

According to some perspectives, this dimension could be a quality dimension of repository systems, rather than a quality dimension of identifier systems:

> I mean we don't really have an identifier system that separates from our repository system. So, for authority, scalability, and security, those are little bit hard to answer because it's extremely important that our repository system be regarded as an authoritative and secure system. (s12)

4.8.2.2.5 ***Authority.*** Authority can be defined as the degree of reputation of an identifier system in a given community (Lee & Stvilia, 2014). As the previous quotation in the Scalability section, mentions about the authority, this dimension was also considered as very important and closely related to repository system.

Another interviewee raised a different thought on authority. According to him, the dimension would be rated differently, based on the degree of implementation of IRs and identifier systems. Authority would be less important if an IR already uses an identifier system. On the other hand, authority would be extremely important if an IR plans to adopt a new system:

> There is sense that it’s being widely used and more sense that we are internally widely using it. To switch to something else would be very difficult. So, I think at this point, [authority] has less to do with what is happening externally toward an organization and more to do what's happening internally toward an organization. We actually found internally there is some issue to switching to a different identifier system because it was solving a major problem we are having. We will probably to do that, but my sense from my … kind of management is this: unless there is major problem with something like identifiers, we are not going to move to a different system because of some external

> pressure to do so. I don't know. It's a complicated question. Your decision about the authority of an identifier system before you start using the identifier is very different from your decision related to the authority of the identifiers, some already widely used. (s1)

4.8.2.2.6 *Interoperability.* Interoperability can be defined as the ability to use an identifier system and string in services outside of the direct control of the issuing assigner (Lee & Stvilia, 2014). One interview participant gave a rating of six to interoperability, because she considered its potential to increase in importance in the coming days: "Interoperability is currently not very important, but I think it will grow to be very important" (s2). On the other hand, another interviewee gave a rating of two to the same dimension. In his perspective, an identifier system is a local thing, and the systems do not necessarily need to be interoperable:

> I've always thought an identifier was a local thing and I've never really thought that much about the interoperability of identifiers. . . . If I choose ARK, should it be able to interoperate with DOIs? No. They're different. If you chose Handle you wanted to work with other Handle systems, because they're the same. (s10)

4.8.2.2.7 *Security.* Security can be defined as the extent to which the resource of an identifier system is protected from unauthorized administrative access or modification (Lee & Stvilia, 2014). One interviewee specifically described his perspective on security at his IR. He emphasized a close relationship between repository system and identifier system and the importance of secure repository system:

> That one is kind of hard to answer because I am not really concerned at all about the security of our identifier system. But having our IR be secure is the most important

> because there is open access data in there, but there is also private data in there. It's not sensitive, but it's still private. I would hate for that to get out, and plus I would also hate for a situation like for instance, there was a researcher at our institution, a climate scientist who was getting into a lot of trouble because of claims that he colluded to lie about climate science. And if we had an insecure system, people would hack into it and attribute to that researcher things he never said. So, security is very important. A repository is a big part of research ECO system. We all want that to have high integrity. (s12)

4.8.2.2.8 *Verifiability.* Verifiability can be defined as the extent to which the correctness and validity of an identifier string is verifiable or provable (Lee & Stvilia, 2014). One of the interviewees perceived the importance of verifiability in identifier systems, but in the context of IRs, the verifiability was considered to be less significant:

> Verifiability, I've never found that to be as important as other people have found that to be. For example, the needs for check digits and all that. I think it's absolutely cool. I've just never considered it as being a problem. (s10)

4.8.2.2.9 *Granularity/Flexibility.* Granularity/flexibility can be defined as the extent to which the identifier system allows referencing data at a different granularity (Lee & Stvilia, 2014). The dimension is a concept that is currently being discussed, and only a few of the existing identifier systems support the function (Lee & Stvilia, 2014). Although most interviewees described the dimension as important, one of the interviewees indicated that granularity/flexibility is a relatively less significant dimension of identifier quality: "I think, in a lot of use cases for identifiers, having a collection level identifier is fine. So, granularity is not always needed. It will not be an issue with repository systems" (s15).

4.8.2.2.10 ***Opacity and Simplicity.*** Opacity and simplicity are the dimensions that are related to identifier strings. In the context of IRs, these dimensions did not seem to be as important as the other dimensions. Opacity can be defined as the extent to which the meaning can be inferred from the content, structure, or pattern of an identifier string; and simplicity can be defined as the degree of cognitive simplicity of an identifier string (Lee & Stvilia, 2014). According to the interview data, the dimensions of opacity and simplicity do not really matter for curating and managing research data. One of the interviewees described her perspective on opacity. She considered opacity a very important criterion for identifier schema, but her experience with it is somewhat limited to a different community:

> There is something that is really nice in the DOI about just having a random number. But at the same time, there is something nice about having an identifier that has some meaning. . . . That's the criteria that I have the hardest time trying to figure out how important it really is. For example, in a non-IR data project, I worked on newspaper projects. In that project, the identifiers actually have a meaning. We described the paper's LCCN, series, date, and page in the identifier. It's really easy to navigate and figure out what you want to find and even use the identifiers as a way of exploring the system in a way you can't when they're all opaque. (s9)

One other interviewee also presented her perception on opacity as well as simplicity. She considered the dimensions less important in her work context:

> For simplicity and opacity, in some ways, I don't really care. In fact, you have identifiers knowing prefix. They always refer to our IR. It is helpful, but I don't really pay attention

> at all. For the simplicity issue and for the most part, it [opacity] sounds like memorizing identifiers mostly . . . either doing sort of a machine reader or just copying and pasting. It doesn't really matter to me. (s5)

Another interview participant stated a similar idea: “I just think as a user. As long as the identifier string is unique and produces results correctly, the simplicity of the string doesn't matter” (s3).

CHAPTER 5

DISCUSSION

5.1 Data Activities in the IRs

According to Activity Theory (Engeström, 1987; Leontiev, 1978), activity can be generally understood as interactions between a subject and an object in a community, and the activity is mediated by contexts that include tools/instruments, policies, rules, norms, and division of labor. Various activities and their mediating factors exist in the context of IRs; identifying them and investigating their relationships can form a useful knowledge base that can be used in data curation planning and education, as well as in designing and implementing research data services in IRs. Data activities in the IRs could mainly be divided into two different categories: curation activities and other related activities (see Table 4.2). Curation activities include the activities that are directly related to data curation work; on the other hand, other related activities contain the activities associated with administration/management and user services of the IRs. Collecting knowledge about those identified data activities can guide for the institutions that currently provide or plan to provide institutional data repository services. In addition, understanding those data activities provides insights for further developing a general model of IR research data curation work.

There are many different general models of research data and related curation activities (DCC Curation Lifecycle Model, DataOne, OAIS). The curation activities identified in this study can be mapped to the DCC Curation Lifecycle Model. The mapping demonstrates the differences and similarities between a general model and the IR data activities identified in this study; understanding these similarities and differences can then generate further discussions among members of research data curation communities regarding the nature of curation activities within the IR context. The DCC model provides an overview of the stages required for curation of data from conceptualization and receipt through the publication and sharing of data (Higgins, 2008). The IR curation activities appeared throughout all of the sequential actions of the DCC model, and identified relations with the full lifecycle actions of the DCC model (see Table 5.1). The IR data curation activities were more detailed than the sequential actions of the DCC model. For example, *consulting with researchers, communicating with IR or library staff, developing and documenting metadata, validating data,* and *packaging dataset* can be mapped into the *preservation action* stage of the DCC model. In the *preservation action* stage, curators communicate with researchers to develop and document metadata for their specific research data. Within that communication process, IR or library staff who have domain knowledge in the research discipline may be involved in the process of developing metadata. After the process, curators and researchers collaborate on a validity check of the data and then package the data in order to store it within the IR. In addition, one particular IR action (i.e., *consulting with researchers*) coincided at least with one other action via all of the sequential actions. The *consulting with researchers* action as a full lifecycle action and its coincidence with different actions demonstrate the importance of communication and consultation between curators and researchers. In some cases of data curation services, curators are involved in the research projects from the planning stage, and the involvement is continued until the project ends. Some of the interview participants mentioned their wide spectrum of different points of involvement in data curators. The curation work can include meetings with researchers who are interested in storing and preserving data; providing assistance in developing grant proposals; regularly meeting with researchers to help them deposit their data throughout their project period; and helping transfer data from one place to another. In order to understand data practices (e.g., tools, rules, norms, policies, division of labor) of a specific research community and to curate the data in sociotechnically-integrated ways, communication between IR curators and data providers throughout the curation lifecycle is an essential activity to ensure high quality of data curation service.

The IR curation activities can also be mapped into the research lifecycle model developed by the Joint Information Systems Committee's (JISC). This is an informative resource identifying the specific research phases in which IR curators would become involved. JISC's conceptualization of research stages presents a typical funded-research process and has been frequently cited in the data curation literature (e.g., Tenopir et al., 2011). Table 5.2 shows widespread distribution of IR data curation activities throughout the research lifecycle of JISC model. Again, this implies that data curation service takes place during the whole research lifecycle and requires systematic investigation of the data and data curation practices.

The OAIS reference model (CCSDS, 2012) was designed to inform the development of systems for long-term preservation of digital information. It is a comprehensive conceptual model consisting of six different entities: administration, ingest, data management, archival storage, preservation planning, and access. The entities encompass all of the IR data activities except activities related to education (e.g., training librarians, educating researchers, learning the best practices for data management). Table 5.3 summarizes the mapping between OAIS model entities and the IR data activities. As the mapping shows, the IR data activities include not only data curation-related activities, but also the activities related to preparation for the future (e.g., education). In comparison with the OAIS model, the data activities from this study can be a practical guideline for developing research data curation services in the context of IRs.

Table 5.1. The comparison of the IR curation activities to the DCC Curation Lifecycle Model

		DCC Curation Lifecycle Model								
		Full Lifecycle Actions	Description and representation information Preservation planning Community watch & participation Curate and preserve							
		Sequential Actions	Conceptualize	Create or receive	Appraise and select	Ingest	Preservation action	Store	Access, use and reuse	Transform
Data Curation Activities	Understanding data curation needs		▨	▨	▨	▨	▨	▨	▨	▨
	Interviewing researchers		X							
	Consulting with researchers		X	X	X	X	X	X	X	X
	Communicating with IR or library staff				X		X			
	Managing and sharing data			▨	▨	▨	▨	▨		
	Receiving or transferring data files			X						
	Cleaning data				X					
	Converting data to a different file format					X				
	Developing and adding metadata						X			
	Validating data						X			
	Packaging data					X	X			
	Uploading and publishing data into IR							X		
	Ensuring that data is accessible and reusable							▨	▨	▨
	Annotating data for relevant entities								X	X
	Optimizing data to search engine								X	
	Keeping data up to date into mirror repository							X	X	X
	Re-evaluating data for long term preservation								▨	▨
	Selecting dataset for long term preservation								X	X

Table 5.2. The comparison of the IR curation activities to the JISC model of research lifecycle

		JISC model of research lifecycle							
					Research process				
		Ideas	Partners	Proposal writing	Simulate, experiment, observe	Manage data	Analyze data	Share data	Publishing
Data Curation Activities	Understanding data curation needs	▨		▨		▨	▨	▨	▨
	Interviewing researchers	X							
	Consulting with researchers	X		X		X	X	X	X
	Communicating with IR or library staff					X	X		
	Managing and sharing data					▨	▨	▨	▨
	Receiving or transferring data files					X			
	Cleaning data					X			
	Converting data to a different file format					X			
	Developing and adding metadata					X	X	X	
	Validating data					X	X		
	Packaging data							X	X
	Uploading and publishing data into IR							X	X
	Ensuring that data is accessible and reusable							▨	▨
	Annotating data for relevant entities								
	Optimizing data to search engine								
	Keeping data up to date into mirror repository							X	X
	Re-evaluating data for long term preservation								
	Selecting dataset for long term preservation								

Table 5.3. The comparison of the IR curation activities to the OAIS model entities

IR Data Curation Activities	OAIS Entities
Understanding data curation needs	Administration
Managing and sharing data	Ingest, Data management, Archival storage
Ensuring that data is accessible and reusable	Access
Re-evaluating data for long term preservation	Archival storage, Preservation planning
Analyzing data usage	Administration
Creating policy and administrating infrastructure	Administration
Educating people for data management	
Continuing education	

A survey conducted by Tenopir et al. (2012) identified research data activities currently offered by ACRL-member libraries or planned to be offered in the next year to two years. All of the member libraries are based on academic institutions in the United States and Canada. Interview data from the current study was collected in 2014, which is two years after Tenopir et al.'s survey data was collected. Therefore, comparing the activities identified by the two studies is appropriate to see whether there are any changes in practices. The comparison includes not only the curation activities, but also the other data-related activities in the context of IRs (see Table 4.2). Although many of the data activities from the current study can be mapped to the data activities identified in Tenopir et al.'s study, the activities in this study are of a finer granularity and contain additional activities (see Table 5.4 & 5.5). For instance, *consulting with researchers, cleaning data, converting file format, developing and documenting metadata, validating data*, *packaging dataset, uploading and publishing data into IR,* and *controlling authority data* can be mapped to the *preparing data/datasets for deposit into a repository* activity of Tenopir et al.'s study. In addition, the current study includes tasks that relate to *analyzing data usage* within IRs (e.g., managing descriptive statistics of data usage, providing researchers data tracking results). Although this study has a focus on data curation activities, the activities identified from the interview data include other data-related activities in the context of IRs. Since

Table 5.4. The comparison of the IR data curation activities to the data activities identified from Tenopir et al.'s project

Data Curation Activities	Research Data Services Currently Offered by the Library or Planned to Be Offered in the Future						
	Consulting with faculty, staff, or students on data management plans	Consulting with faculty, staff, or students on data and metadata standards	Directly participating with researchers on a project (as a team member)	Deaccessioning/deselection of data/datasets for removal from a repository	Preparing data/datasets for deposit into a repository	Creating or transforming metadata for data or datasets	Identifying data/datasets that could be candidates for repositories on or off campus
Understanding data curation needs							
Interviewing researchers	X	X					X
Consulting with researchers	X	X	X	X	X	X	X
Communicating with IR or library staff	X	X		X			
Managing and sharing data							
Receiving or transferring data files			X				
Cleaning data			X		X		
Converting data to a different file format			X		X		
Developing and adding metadata		X	X		X	X	
Validating data			X		X		
Packaging data					X		
Uploading and publishing data into IR					X		
Ensuring that data is accessible and reusable							
Annotating data for relevant entities					X	X	
Optimizing data to search engine						X	
Keeping data up to date into mirror repository				X			X
Re-evaluating data for long term preservation							
Selecting dataset for long term preservation				X			X

Table 5.5. The comparison of other data-related activities to the data activities identified from Tenopir et al.'s project

Other Data-Related Activities	Research Data Services Currenlty Offered by the Library or Planned to Be Offered in the Future					
	Outreach and collaboration with other research data services providers either on or off campus	Providing reference support for finding and citing data/datasets	Creating web guides and finding aids for data/datasets/data repositories	Discussing research data services with other librarians, or other people on campus, or RDS professionals, on a semi-regular frequency	Training co-workers in your library, or across campus, on research data services	Providing technical support for RDS systems (e.g., a repository, access and discovery systems)
Analyzing data usage						
Managing descriptive statistics of data usage						
Providing researchers data tracking results						
Creating policy and administrating infrastructure						
Understanding local needs and creating local policies and rules			X	X		
Building infrastructure component			X	X		X
Educating people about data management						
Training librarians	X				X	
Educating researchers		X				
Providing workshops for data analysis tools		X				
Providing outreach for data curation	X					
Learning the best practices						
Learning the best practices for research data management				X		
Learning future technologies				X		

data curation is still an emerging field and the work practices within IR settings are not yet matured, the IR curators spend a significant amount of time on activities related to IR system administration and education about data curation best practices.

5.2 Activity Structure

5.2.1 Communities and Division of Labor

The interview data identified seven different IR staff roles that provide research data curation services in IR and the roles of data providers and data users. The roles of IR staff are head, data curator, IR manager, metadata specialist, developer, subject specialist, and graduate assistant. In practice, role mapping to job title is not consistent. Institutions may use different job titles and there could be many to many relationships between the data curation roles and the job titles based on local needs (see Table 4.3). In addition, since IR staff collaborate, their tasks and required or preferred skills often overlap (see Table 4.8). However, each role's unique tasks and skillsets were also identified by the interview analysis (see Table 4.4 & 4.8). The knowledge of IR role specific curation tasks and skills needed can benefit the institutions planning to implement data curation services and to recruit new members for their IR data curation teams.

The current IRs' division of labor can be compared to Swan and Brown's (2008) study of the skills, roles, and career structures of data scientists and curators, commissioned by JISC. Swan and Brown used a mixed methodology (i.e., interviews, focus groups, and online surveys) to look at data scientists and curators' roles and skills. Their study participants included data scientists, librarians, library technologists, and library educators. Although their study focused on data curation in research institutions, their findings on division of labor were much broader than the current study. Furthermore, the role boundaries for research data curation work were fuzzy (Swan & Brown, 2008), and they used the actions of DCC Curation Lifecycle Model to specify the role-related activities. Mapping between the two studies is lossy because the current study is more detailed on research data curation's division of labor and tailored to the context of IRs (see Figure 5.1).

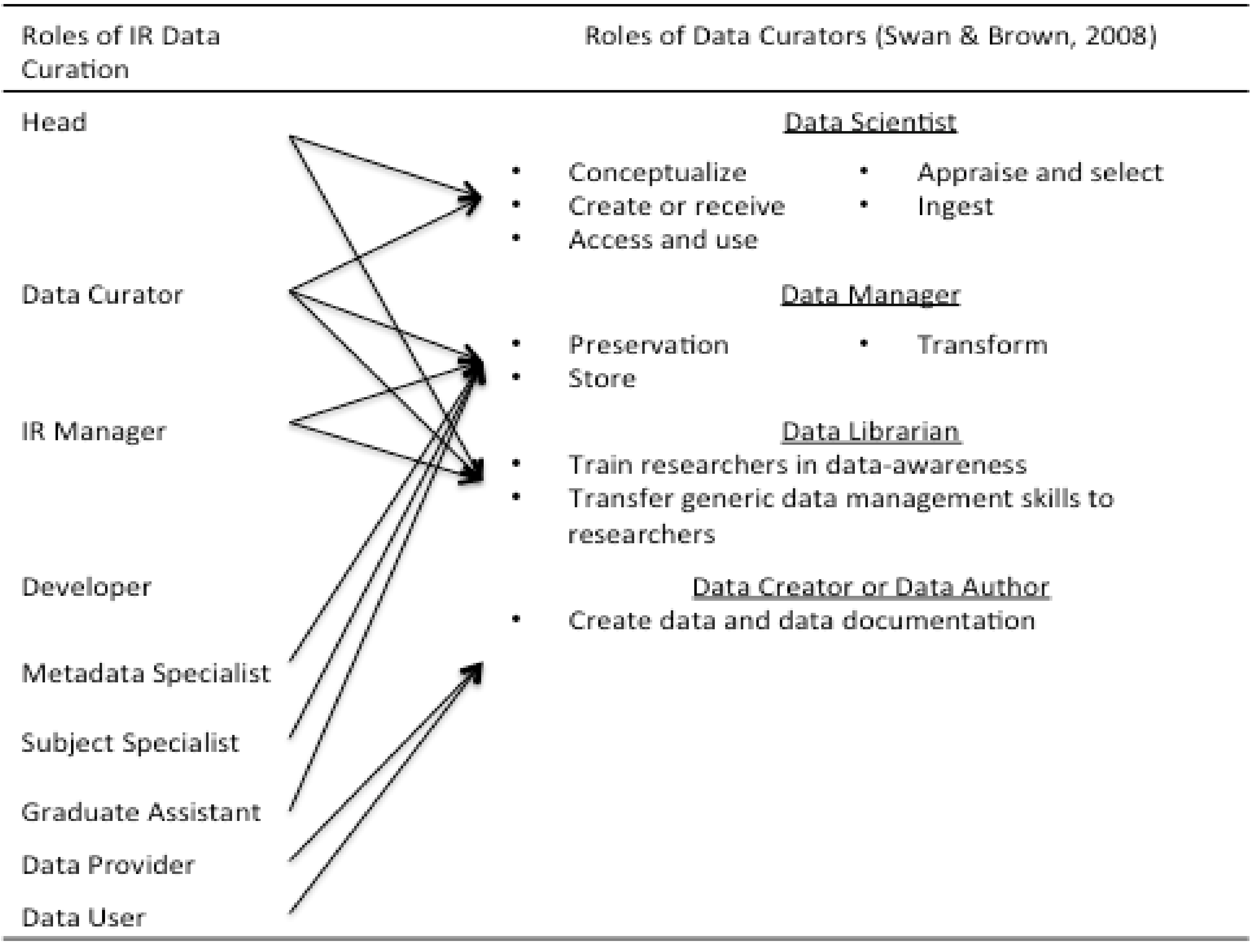


Figure 5.1. The comparison of division of labor between the current study and Swan and Brown's study (2008)

The current study also has fuzzy boundaries for role-related activities. However, in order to provide knowledge of data curation role-related activities to IR communities, a mapping between the current study and the DCC curation lifecycle will have significant value for assembling an effective data curation team for an IR context and work specialization structure (see Figure 5.2). Furthermore, a design chart of research data curation in IRs including different data activities and their contexts (i.e., tools, policies, skillsets, division of labor) can provide ideas for how to implement data services in IRs (see Figure 5.3).

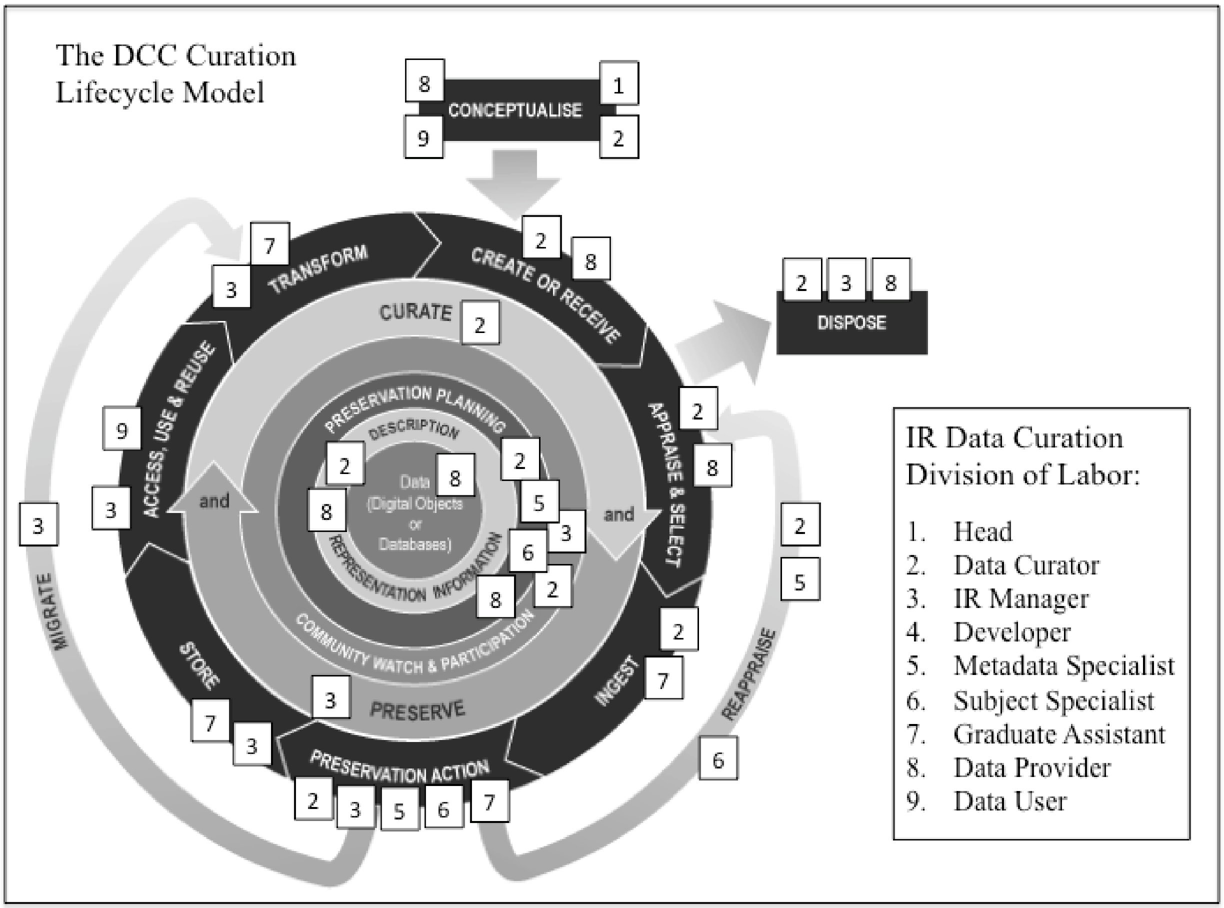


Figure 5.2. Division of labor of IR data curation over the DCC Curation Lifecycle Model (Higgins, 2008)

Comparing the IR staff skills identified by this study to the set of data skills needed for genome annotation curation (which is a type of data curation) identified by Huang et al. (2012) reveal some differences and similarities (see Figure 5.4). Interpersonal skills to communicate and collaborate with researchers do not have a clear match in the Huang et al. model. However, similar skills in his model fall under adaptive skills, which are the skills needed to determine and improve quality (e.g., value and relevancy) of research data. The current study’s metadata skills,

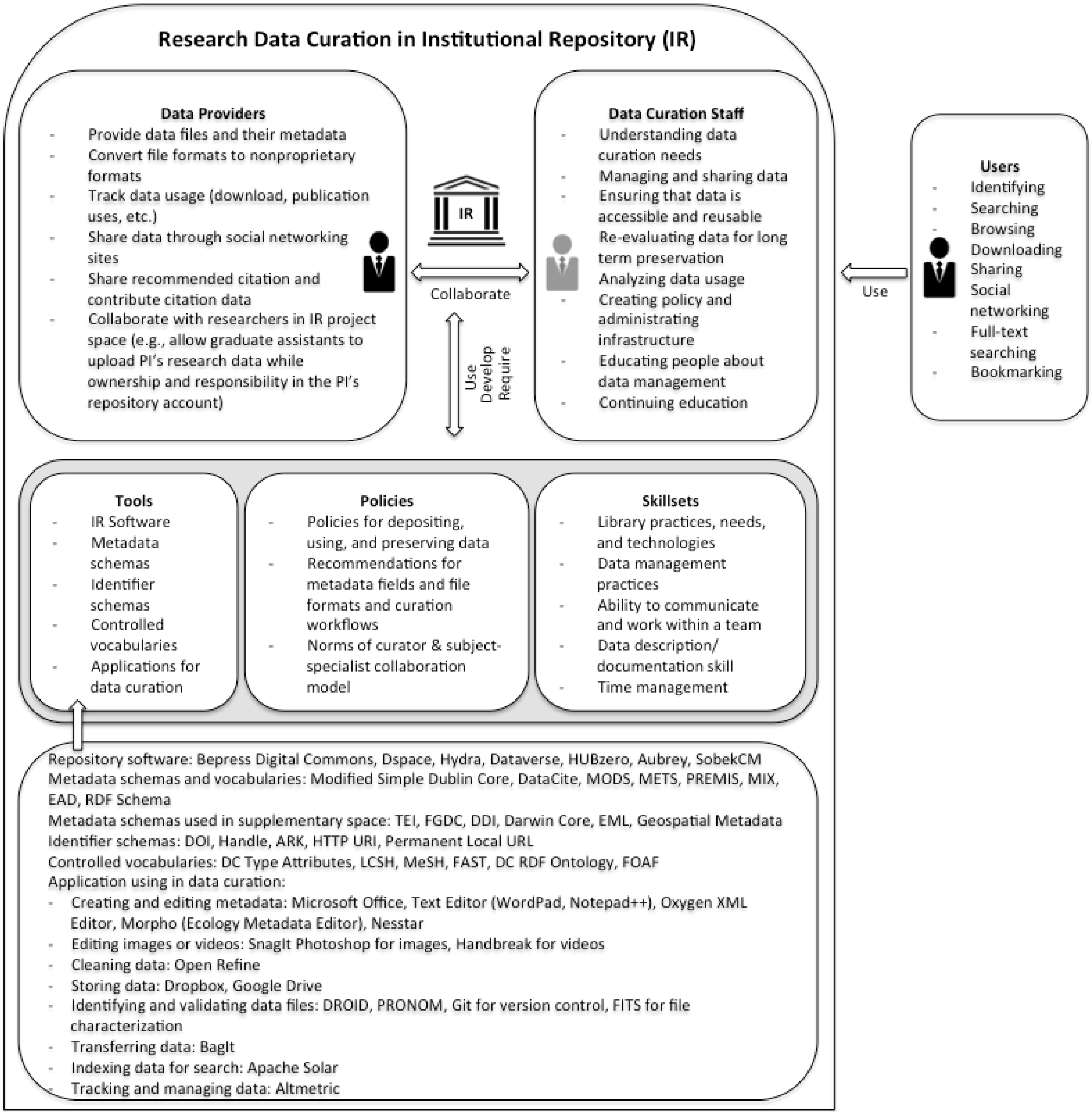

Research Data Curation in Institutional Repository (IR)
Data Providers
- Provide data files and their metadata
- Convert file formats to nonproprietary formats
- Track data usage (download, publication uses, etc.)
- Share data through social networking sites
- Share recommended citation and contribute citation data
- Collaborate with researchers in IR project space (e.g., allow graduate assistants to upload PI's research data while ownership and responsibility in the PI's repository account)
IR
Collaborate
Use
Develop
Require
Data Curation Staff
- Understanding data curation needs
- Managing and sharing data
- Ensuring that data is accessible and reusable
- Re-evaluating data for long term preservation
- Analyzing data usage
- Creating policy and administrating infrastructure
- Educating people about data management
- Continuing education
Use
Users
- Identifying
- Searching
- Browsing
- Downloading
- Sharing
- Social networking
- Full-text searching
- Bookmarking
Tools
- IR Software
- Metadata schemas
- Identifier schemas
- Controlled vocabularies
- Applications for data curation
Policies
- Policies for depositing, using, and preserving data
- Recommendations for metadata fields and file formats and curation workflows
- Norms of curator & subject-specialist collaboration model
Skillsets
- Library practices, needs, and technologies
- Data management practices
- Ability to communicate and work within a team
- Data description/ documentation skill
- Time management
Repository software: Bepress Digital Commons, Dspace, Hydra, Dataverse, HUBzero, Aubrey, SobekCM
Metadata schemas and vocabularies: Modified Simple Dublin Core, DataCite, MODS, METS, PREMIS, MIX, EAD, RDF Schema
Metadata schemas used in supplementary space: TEI, FGDC, DDI, Darwin Core, EML, Geospatial Metadata
Identifier schemas: DOI, Handle, ARK, HTTP URI, Permanent Local URL
Controlled vocabularies: DC Type Attributes, LCSH, MeSH, FAST, DC RDF Ontology, FOAF
Application using in data curation:
- Creating and editing metadata: Microsoft Office, Text Editor (WordPad, Notepad++), Oxygen XML Editor, Morpho (Ecology Metadata Editor), Nesstar
- Editing images or videos: SnagIt Photoshop for images, Handbreak for videos
- Cleaning data: Open Refine
- Storing data: Dropbox, Google Drive
- Identifying and validating data files: DROID, PRONOM, Git for version control, FITS for file characterization
- Transferring data: BagIt
- Indexing data for search: Apache Solar
- Tracking and managing data: Altmetric

Figure 5.3. A design chart of research data curation in IRs

which include disciplinary knowledge and its associated metadata knowledge, can be mapped to data quality literacy skills (which are the skills needed to understand and measure data quality) and adaptive skills. However both skills from Huang et al. contain many more detailed concepts than the metadata skills from the current study. For example, data-quality dimension, data-quality measurement, data-quality implication, data-quality cost/benefit, data-entry improvement, change process, organization policy, user requirement, and information overload are detailed concepts of the two skill constructs. Some of the differences in the number of detailed concepts between the current study and Huang et al.'s could be linked to the variations in the levels of data curation provided by IRs and subject-specific data repositories. Subject-specific data repositories and their staff are expected to provide deeper analysis of submitted data, including data annotation and quality assessment. Hence, curators of subject-specific data repositories are expected to be subject specialists with advanced degrees in those subject areas. On the other hand, curators in IRs collaborate with subject specialists to complete their tasks in research data curation.

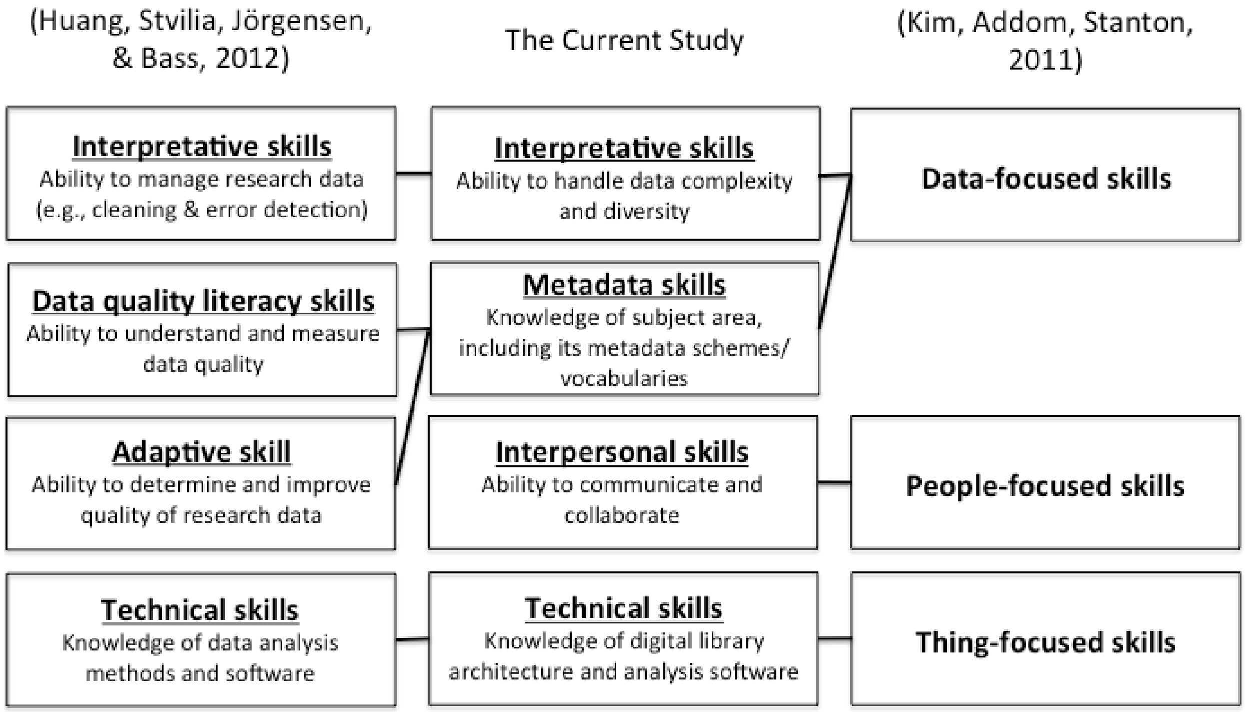

(Huang, Stvilia, Jörgensen, & Bass, 2012)
The Current Study
(Kim, Addom, Stanton, 2011)
Interpretative skills
Ability to manage research data (e.g., cleaning & error detection)
Interpretative skills
Ability to handle data complexity and diversity
Data-focused skills
Data quality literacy skills
Ability to understand and measure data quality
Metadata skills
Knowledge of subject area, including its metadata schemes/ vocabularies
Adaptive skill
Ability to determine and improve quality of research data
Interpersonal skills
Ability to communicate and collaborate
People-focused skills
Technical skills
Knowledge of data analysis methods and software
Technical skills
Knowledge of digital library architecture and analysis software
Thing-focused skills

Figure 5.4. Comparison of IR data curation staff's skills to data curation skills identified by Huang et al (2012) and Kim et al. (2011)

Similarly, Wu (2014) identified 16 different data curation skills in the context of biological ontology. Based on her interview data collected from biocurators and Gene Ontology users, domain knowledge (i.e., basic biological knowledge, domain-specific biological knowledge, staying current on developments in biological knowledge, reading scientific literature, bioinformatics) was the most frequently mentioned skill for data curation. Besides domain knowledge, interpersonal skills, interpretative skills, and technical skills were also identified as data curation skills in biology.

Kim, Addom, and Stanton (2011) conducted a study to understand the educational needs of eScience professionals. One of their findings was the relative importance and frequency of eScience professionals' tasks. People-, data-, and thing-focused skills were identified as three main skills and tasks (see Figure 5.4). People-focused skills, which are used to manage projects and analyze project/researcher needs, can be mapped to the interpersonal skills of the current study. Thing-focused skills, which are needed to work with content management tools and office productivity software, can be mapped to this study's technical skills. Data-focused skills, which are employed in everything from creating or receiving data to defining metadata, can be mapped to both interpretative skills and metadata skills. In the context of IRs, metadata skills are separated from the general data-focused skills (e.g., interpretative skills), as metadata specialists and subject specialists are independent roles in IR data curation services.

5.2.2 Tools

Knowledge organization tools (e.g., metadata, taxonomy, and ontology) for research data can be considered as essential in data discovery, use, and citation (Qin, Ball, & Greenberg, 2012). The interview data in the current study identified different types of tools/instruments (e.g., metadata schemas, identifier schemas, controlled vocabularies) that IR staff use for research data curation. The identified tools can be categorized by major types of knowledge organization tools in order to systematically understand the current practices that implement different kinds of knowledge organization tools and to form a comprehensive representation of data objects (see Figure 5.5).

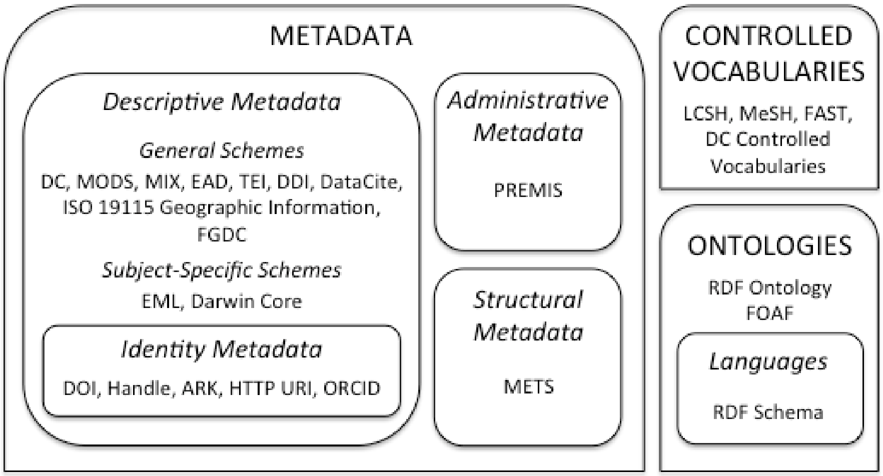


Figure 5.5. Current IR knowledge organization practice for research data curation

Metadata can be categorized as of descriptive-, administrative-, and structural-types (NISO, 2004) and descriptive metadata has a subtype of identity metadata. Descriptive metadata generally enables discovery and identification of a resource. In the current practices of the IRs, various metadata schemes (i.e., DC, MODS, MIX, EAD, TEI, DDI, DataCite, ISO 19115 Geographic Information, FGDC) are used for identification of research data. A few subject-specific schemes (i.e., EML, Darwin Core) are also used in IRs. In scientific context, these metadata can help verify, replicate, and reproduce research data (Qin et al., 2012). Most IR systems do not provide different types of disciplinary metadata. However, they do support an optional place to upload additional information about the data. Data providers who want to add discipline-specific metadata typically utilize the optional place. Data models of identity metadata includes the entities that have their own set of metadata elements for description purposes (Qin et al., 2012). For example, the entities of person, event, place, and object have various elements (e.g., name, role, location, time, and description) describing each entity. The current study identified various identifier schemas (i.e., DOI, Handle, ARK, HTTP URI, and ORCID) currently used and categorized as identity metadata. Many of the IRs also used PREMIS as their administrative metadata and METS as their structural metadata.

Controlled vocabularies (i.e., controlled vocabularies, ontologies, ontology languages) mainly includes subject-related vocabularies and their linking mechanism (Qin et al., 2012). The analysis of this study reveals that many of the IRs currently do not use a set of controlled vocabularies; only a few of the IRs mentioned using controlled vocabularies (i.e., LCSH, MeSH, FAST, DC Controlled Vocabularies) in their IRs. Also, only one of the IRs used ontologies in the form of linked data/RDF (Bizer et al., 2009), and one other interviewee indicated that her IR is planning to build a mechanism for the linked data web.

Current knowledge organization tools in IR research data curation (see Figure 5.5) indicates that the IRs' metadata models are aligned with the series of principles in modeling metadata for research data curation identified by Qin et al. (2012). Qin et al. (2012) presented three principles in modeling metadata for scientific data. The first is "The least effort principle." A number of databases exist to identify people, institutions, or funding agencies. Using the existing databases can reduce the effort expended to design new metadata schemas and possibly decrease the redundancies that can happen in data entry. Many IRs in this study used or planned to use existing descriptive metadata (e.g., DC, DataCite, DOI, ORCID) and controlled vocabularies (i.e., LCSH, MeSH, FAST). The second is "The infrastructure service principle." This principle requires building an architectural view of metadata for scientific data (see Figure 5.6). Qin et al. (2012) interpreted the architectural view of metadata requirements as metadata infrastructure. Based on the current study, IRs use at least one or more metadata schema(s) for each component of the architectural view (see Figure 5.6). The components of this architectural view capture the complexity of research data entities. In order to identify and link research data objects, efforts in identifying and annotating entities of those research data objects are essential. The third is "The portable principle." This principle means that metadata properties are modeled by using linked data/RDF technology (Bizer et al., 2009), which will support the linking and reuse of research data. According to the interview data, only one IR currently implemented their system in RDF structure (i.e., triple: subject, predicate, object), and another IR is currently planning to change its current IR software (i.e., DSpace) to different IR software (i.e., Hydra) in order to develop an RDF structured system.

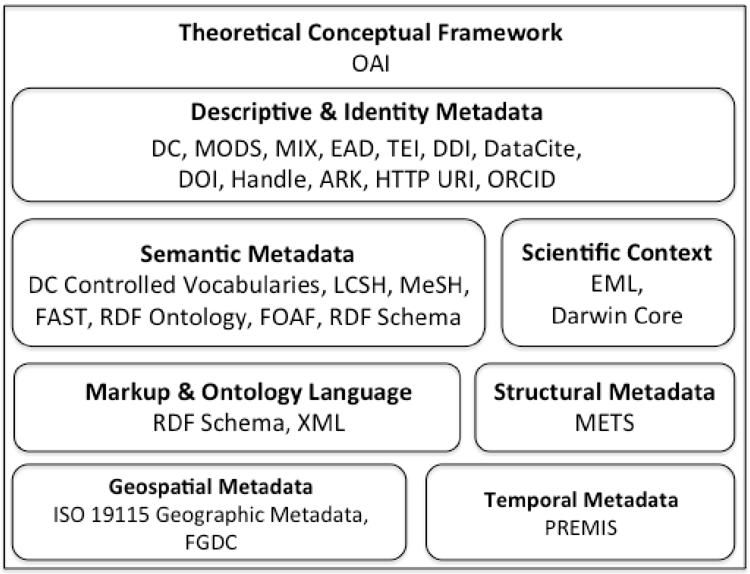

Theoretical Conceptual Framework
OAI
Descriptive & Identity Metadata
DC, MODS, MIX, EAD, TEI, DDI, DataCite,
DOI, Handle, ARK, HTTP URI, ORCID
Semantic Metadata
DC Controlled Vocabularies, LCSH, MeSH,
FAST, RDF Ontology, FOAF, RDF Schema
Scientific Context
EML,
Darwin Core
Markup & Ontology Language
RDF Schema, XML
Structural Metadata
METS
Geospatial Metadata
ISO 19115 Geographic Metadata,
FGDC
Temporal Metadata
PREMIS

Figure 5.6. Current IR metadata architectural view adapted from Qin et al.'s architectural view of metadata requirements.

5.2.3 Policies, Rules, and Norms

In 2007, Rieh, Markey, St. Jean, Yakel, and Kim conducted a national survey to understand current uses of IRs. At the time of the study, many colleges and universities were increasingly developing IRs to store, preserve, and reuse the intellectual products created by the institution members. Rieh et al. found that IRs are concerned with policies related to administration and access issues. For example, administrative and access policies determine who is entitled to submit material, what can be accepted, who is responsible, and who can access the material. The current study reflects those research data curation issues in its findings on policies, recommendations, and norms. The findings identified from the current study include the policies for depositing, using, and preserving objects and research data, the policies for collection and copyright, the recommendations for metadata, file formats and data curation workflow, and the norms of the curator and subject-specialist collaboration model. The policies from the current study are broader than the policies identified in Rieh et al.'s study, but at the same time more detailed about research data curation. In addition to policies, this study shows the recommendations and norms that support data curation tasks.

5.2.4 Contradictions

In an activity system, contradictions that can be understood as tensions, conflicts, or limitations among the components of that system (Engeström, 1987; Wilson, 2008). Identifying the contradictions that exist in and between activity components and seeking resolving those contradictions can lead to the evolution of and innovation in the activity system (Engeström, 1987). The following sections provide different examples of contradictions that occur between components of an activity structure and some solutions for resolving those contradictions are suggested. A better understanding of the types of contradictions found in IR curation work can benefit the institutions that currently plan to implement institutional data repositories. In addition, some of alternative solutions for the identified contradictions are suggested to help evolve the activity systems.

5.2.4.1 Contradictions within a single component. There is a tension closely related to the issues of tradeoffs. One institution only uses a Handle identifier system because of limited resources in its IR infrastructure. In order to adopt a new identification system, the institution had to consider the tradeoff between tool complexity and scalability. They seemed a little afraid of adopting a new system:

> Definitely, part of the reason we only use Handles is we don't have a generator to create a unique identifier. Because as soon as you start doing that, you will get into a lot of identifier criteria and quality problems. You will not be able to foresee it. If you come up with a simple system, then it's not going to be scalable, but if you come up with something that's really scalable, it's probably going to be not very simple. (s15)

5.2.4.2 Contradictions between objective and tool. One of the interviewees provided an example of a contradiction that occurred between a dataset and his IR software. The contradiction occurred during the process that satisfies the curation objective of data storage or preservation. Research data is very complex and diverse, and the number and the scale of data files and their types are also very different depending on the domain of the research. However, based on the interview data, the existing IR software and storage space did not yet sufficiently support the needs of research data curation services in IRs. An interviewee mentioned a workaround (Gasser, 1986) to avoid the contradiction. His IR uses a mirror repository as a backup system, a form of workaround, which is used for some of large research datasets. The mirror stores data files and then it has a link to the data description in his IR.

Another interviewee provided a different example of a contradiction between objective and tool. In response to the interviewer's question about the levels of identifier granularity for supporting identification of entity level data (i.e., What identifiers do you use at the data collection/set level, file/object level, or entity level?), she simply said: “That’s a system limitation” (s9). Granularity is the extent to which the identifier system allows data to be referenced at a different granularity (Lee & Stvilia, 2014). She explained that her IR does not assign data identifier strings to entity level objects, and it is because her IR software does not support the service. Throughout the interview data, many interviewees indicated that their data curation services frequently depend on what can be supported by their IR software. To resolve this contradiction, IRs can adopt an additional and external identifier system that can be assigned to an entity level’s objects. Also, IRs can replace their existing IR software with one that supports linked data/RDF technology. With such software, diverse existing controlled vocabularies that can be used to construct HTTP URIs can be assigned to the entity level’s objects.

5.2.4.3 Contradictions between tool and the best practice. Analysis of the data identified contradictions that exist between tools and best practices. One of the interviewees indicated that his institution could adopt a new identifier system for person entity to provide an effective authority control service, but the progress on it is very slow. One of the main reasons for the slow process is lack of established best practices. Knowledge about application of the new system was insufficient to actually adopt the system with low risk in his IR. His institution discussed the system adoption, but progress moved forward slowly without examples of best practices:

> The ORCID ID, I would love to see more uptakes. I think maybe one percent of people are really excited about it and using it. Ninety-nine percent of the people know that's out there and just think it's one more thing. . . . I think a lot of them are primarily concerned with getting it into the systems that they use. So, I don't know. We were watching it. We

> definitely want to implement it here. We've talked about adding either just dumb flat text ORCID field to DSpace or maybe one that was actually connected by the API to ORCID services. We just haven't made much progress on it. We are watching to see what's evolved. (s15)

To resolve this type of contradiction, IRs can conduct a pilot project to test the system performance. One of the interview participants mentioned her IR's pilot test for using DataCite DOI. She also explained that her institution would decide whether to adopt DataCite DOI or not based on the test result:

> We are just initiating a pilot to use the DataCite DOI for research data. We haven't started to do that yet. We have a handful of pilot participants we are going to contact. Essentially, we would be working with them to submit the item into our repository so we would get a handle, but we also find a DataCite DOI for the item, research data. (s5)

5.2.4.4 Contradictions among three or more components. In the context of IRs, many different resources affect the curation services or tools of IRs (Markey et al., 2007; Tenopir et al., 2012). The interview data also identified various resources intertwined with the services or tools. Some of the interviewees implied that the current availability of personnel affects their usage of tools. Some IRs control and maintain their IR server and systems with a strong development team; otherwise, if they don't have a strong development team, IRs may outsource the management of their IR server and systems. One of the interviewees stated that her IR has a strong development team:

> Our team in the library technology unit always likes to control the development and use of software. ... Mainly we issued the resources in terms of people we had around to do

> the data modeling and then think throughout about the issue and the programmer power we have available to enact a solution. (s2)

Financial funding is another resource affecting IR services. The annual budget for IRs is limited, and variations in budget size influences the kinds of services and tools that the IRs provide: "We need to consider the incremental cost for support services so we see how much they actually cost. We haven't even figured out what extended storage costs will be" (s14). As mentioned in the previous sections, the current tools and available best practices are also resources that impact IR data curation services.

Some of the contradictions recognized in the interview data involve components of data curation work in IRs. One interview participant presented a contradiction that occurs between norms, best practices, and division of labor. Her IR adopted a new model for identifier granularity, but the IR did not actually employ the new model. Her IR team did not have access to sufficient resources to use the model in their setting. For example, they did not have norms or best practices to help them adapt the new model to their existing IR setting. Also, her IR has insufficient human resources to devote to the software development that would have been necessary to implement the new model:

> We looked at and adopted Dryad model for granularity. But we ultimately decided not to really go in that direction just yet. Again, the question of resources. We didn't see a model, because Dryad is really—they [the developers of Dryad] were the only people who were doing anything with more granular DOI at that point. We didn't see a model for how we would make it work in DSpace. Mainly we had issues with resources in terms of the number of people we had around to do the data modeling and think through issues

> and the programmer power we have to enact. We didn't have a lot of either of those resources. So, we just decided on very cut and dry DOI issuing. (s2)

One last example of contradictions identified from the interview data occurred between the IR infrastructure and researcher needs imposed by publishers. Researchers want to follow the practices recommended by publishers, and therefore ask their IRs to meet those requirements. However, IR infrastructure is not always aligned with the publishers' practices:

> We do have researchers who are explicitly asking for DOI. We use Handle. Especially for publishers, I don't think they have the same knowledge of other identifier systems. They appear to be saying that you need DOI, so we are getting asked for DOI. Often when we probe a little bit about that [their need for DOI], we find out Handle is sufficient, but not preferred. So, researchers, at least in my experience, so far have been a little nervous about using something that is not DOI. That's not what their familiar with in their publication process. I also think we are interested in potential citation tracking. We may possibly use DOI for that. In that way, I think there are two main reasons [to use DOI]: comfort level because of familiarity with DOI, and working with some of the services. DataCite is trying to provide, as well as I understand, DataCite is working with CrossRef to provide more services for DOI. (s5)

One solution for this contradiction could be designing a pilot test to determine the value of adopting suggested tools or instruments. Based on the test result, IRs would be able to make an informed decision about the adoption.

5.3 Data and Data Entity Types

Investigating major types of research data and their entity types has significant value for understanding different characteristics of research data (Borgman et al., 2007) and for enabling semantic data linking envisioned by Berners-Lee (Berners-Lee, Hendler, & Lassila, 2001). Research data in the IR context includes any type of data (e.g., text documents, spreadsheets, slides, audio recordings, audio-visuals, images, laboratory notes, statistical data files, databases, software codes, executable files, and tabular data files). According to the interview analysis, all of the IRs do not have any restrictions on the types of research data they will receive, but do have minimum criteria for acceptable data characteristics (e.g., file capacity, the number of files, and proprietary file extension). The minimum criteria and IRs' discipline-independent nature enables the IRs to include all types of data from any discipline. For example, the IRs contain a greater variety of data types (i.e., raw data, text documents, slides, laboratory notes, spreadsheets, software codes, drawings, statistical data files, Website, and databases) than the data types of a Condensed Matter Physics community (Stvilia et al., 2015).

Identifying different types of research data curated in IRs can help in planning, deploying and using identifier systems, including for RDF encoded data (Berners-Lee et al., 2001; Qin et al., 2012). The identification systems are primarily used to identify, describe, locate, link, and group resources by assigning identifiers to appropriate entities (e.g., object, person, organization, place, time, event, topic). Therefore, identifying the current research data entity types curated in an IR context help not only with understanding how identification systems are being used, but also in selecting the appropriate identification systems for research data. According to many researchers, the minimum elements to access, cite, and link data are identifiers (Altman & King, 2007; Qin et al., 2012). From the data analysis, all nine different entity types (i.e., intellectual entity, object, symbolic object, person, organization, place, time, event, and topic) are documented for research data (Lee & Stvilia, 2014) in IRs. In addition, the findings demonstrated how identification systems are currently being used in the IRs (see Table 4.9).

Intellectual entity, object, symbolic object, place, and topic entities are currently being identified by some type of globally unique identification or subject schema (i.e., DOI, ARK, Handle, HTTP URI, permanent local URL, GeoNameID, LCSH, MeSH, and FAST). However, person and organization entities are still identified by local authority control systems, although a few of the IRs plan to adopt a global identification system (i.e., ORCID) for the person entity type. The lack of a global identification system for person and organization entities can hinder entity determination, disambiguation and linking with external sources (Wynholds, 2011). Person and organization entities are typical examples that can frequently change through various situations (e.g., marriage status, job changes, etc.). Particularly, family names (e.g., Zhang, Lee, Wang, Chen, etc.) from Asian countries are difficult to use for unambiguously identifying authors (Warner, 2010), because there are huge numbers of different people with the same family name. Although many organizations have put efforts into developing identification systems for person entity (e.g., ORCID, ResearcherID, OpenID), any of these systems are not used yet pervasively within scholarly communications communities. Among the different identification systems, ORCID is the fastest growing system with significant commercial and community participation (Warner, 2010). Lastly, none of the identification systems were mentioned as a system particularly used for time and event entities.

5.4 Identifier Awareness and Quality Perception of IR Data Curators

Metadata plays a dominant role in discovery, interpretation, selection and evaluation of research data (Qin & D’ignazio, 2010). Metadata is an essential knowledge organization tool widely used through all of the curation lifecycle phases. Identifier schemas are a key type of metadata needed for successful management and use of data stored in IRs (Lynch, 2003) and support identification, citation, linking, and annotation (Lee & Stvilia, 2014). However, there is a lack of systematic identifier-related studies in the LIS field (Lee & Stvilia, 2014). Moreover, although IR curators perceive the importance of identifier schema use, their level of awareness of the options available is low. In many cases, the IRs selected and adopted their identifier systems with a little dedicated research and quality comparison of different identifier systems to their needs (see Figure 4.3). In order to select an identifier system for an IR and/or a particular type of data, or reuse existing identifiers for data linking, understanding quality requirements of identifiers is essential.

To identify the IR curators' perceptions of identifier quality, the study asked the interviewees to rate 11 quality dimensions by their importance. The analysis discovered the IR curators' perceptions and understandings of identifier quality (see Table 4.11). According to this study's data analysis, many of the IRs already used existing identifier systems (e.g., DOI, ARK, Handle) in their IRs, and many of the interviewees revealed their interests in the use of another existing identifier system (i.e., ORCID). They also indicated that they would use the existing identifier systems because of their quality. One interviewee mentioned that authority, i.e. the reputation of an identifier system in a given community (Lee & Stvilia, 2014), is an important quality dimensions to consider when adopting an identifier system: "If nobody is using that identifier, nobody is going to use it" (s1). Another interviewee talked about the importance of actionability/resolvability, which is the ability of the identifier system to locate the object using an identifier string (Lee & Stvilia, 2014): "What I would love to be able to do is just use the ORCID identifier to be able to pull in that information automatically rather than just store that ORCID information, if we use that as an identifier for our author entities" (s5). Many of the IRs are also waiting to see what evolves in the use of the existing identifiers in their communities. According to Vrandečić and Krötzsch (2014), reusing existing identifiers provides great benefits in data exchange and integration across application boundaries. Identifiers (e.g., DOI, ARK, Handle, ORCID) include considerable data about the objects referenced, which helps connect objects to their related resources (e.g., authors, scholarly works, affiliation information, collaborators, grants).

The interview data also carried some discussion of specific quality dimensions (i.e., persistence, authority, scalability, and security). One of the interview participants discussed persistence of identifier systems. He raised a question: Is persistence a quality dimension, or an institution's commitment to support the persistence of an identifier system? He insisted that persistence is about how an institution manages its identification system. If an institution manages and supports its identifier system permanently, the system will be persistent. According to him, the persistence quality criterion is not a characteristic of an identifier schema. Another interviewee also discussed a system dependency between repository systems and identifier systems, and raised the issue of ambiguity in certain quality dimensions (i.e., authority, scalability, and security). He argued that the three dimensions are more closely related to the repository system quality than the identifier system quality. All of those discussions indicated and demonstrated the relationships between identifier system and its neighboring context (e.g., tools, policies, norms, rules, and division of labor). Therefore, an identifier system, its surrounding tools, and the efforts that support the system functionalities must all be considered to systematically evaluate the quality of an identification system.

Although many researchers have studied issues of design and the use of identifier schemas, to the best of the researcher's knowledge, Duerr et al.'s study (2011) is the only one to include a comprehensive and systematic review of the current identifier systems. Duerr et al. examined the utility of identifier schemas for digital earth science data. Their assessment is conducted by asking 14 different questions, categorized as either technical value, user value, or archive value. The interview participants' perception of identifier quality can be mapped to Duerr et al.'s identifier assessment categories (see Figure 5.7). The list of quality dimensions in Figure 5.7 indicates the mean importance ratings for the identifier quality dimensions (see Table 4.11). The mapping between the two studies presents the importance of technical value for identifier system evaluation. Also, according to the mapping, the end-user value of identifier systems seems to be more important than the archive value of identifier systems.

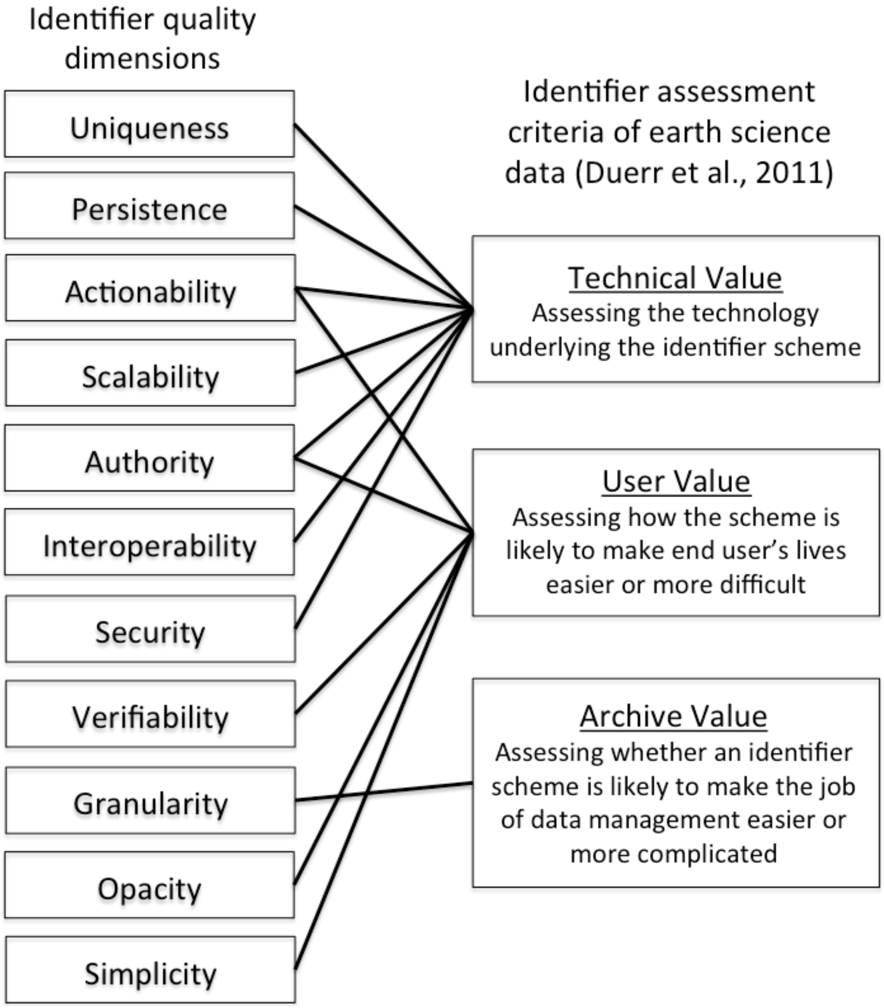

Identifier quality dimensions
Uniqueness
Persistence
Actionability
Scalability
Authority
Interoperability
Security
Verifiability
Granularity
Opacity
Simplicity
Identifier assessment criteria of earth science data (Duerr et al., 2011)
Technical Value
Assessing the technology underlying the identifier scheme
User Value
Assessing how the scheme is likely to make end user's lives easier or more difficult
Archive Value
Assessing whether an identifier scheme is likely to make the job of data management easier or more complicated

Figure 5.7. Comparison of identifier quality assessment criteria

CHAPTER 6

RECOMMENDATION, CONCLUSION AND FUTURE RESEARCH

This study examined research data curation practices in IRs based on Activity Theory (Engeström, 1987; Leontiev, 1978) and a theoretical IQ Assessment Framework (Stvilia et al., 2007). The study identified roles played by IR staff, role-specific sets of activities and skills, major data and entity types curated, as well as the perceptions of quality and the functionalities of identifiers in IRs.

Based on the analysis in the previous two chapters, the researcher provides the following recommendations or curation knowledge that can benefit institutions that currently manage or plan to implement institutional data repositories. The following sections discuss the practices or issues IR staff should be knowledgeable about before their implementation of IRs and conclude with directions for future research.

6.1 Data Activities in IRs and Designing Effective Teams for IR Data Curation

Data activities in IRs include curation-related activities (i.e., understanding data curation needs, managing and sharing data, ensuring that data is accessible and reusable, and re-evaluating data for long-term preservation) and other related activities (i.e., analyzing data usage, creating policy and administrating infrastructure, educating people about data management, and learning the best practices). For curation-related activities, data curators first communicate with data providers to understand the providers' research data. This helps the data curators coordinate the rest of the curation-related activities with skilled personnel. During the coordination process, communications among IR staff or between IR staff and subject librarians frequently occur to tailor the curation activities to the data provider's needs and the types of data submitted. Once the coordination is completed, each activity and their actions are conducted throughout the data and data curation lifecycles. In curation-related activities, communication between stakeholders (i.e., IR staff, librarians, data providers, and users), networking with subject specialists, and knowledge of data and data curation lifecycles are essential. For other related activities, IR staff analyze data usage for their IR users. IR staff also create curation-related policies. Lastly, they educate people (i.e., institutional researchers and librarians) on data management and in the best practices of data curation.

Although the activity of *understanding data curation needs* mainly requires communicative acts between stakeholders, the rest of the activities also require a fair amount of communication. Research data curation is a comprehensive service throughout a data lifecycle; data curators must continuously plan, collect, assure, describe, preserve, discover, integrate, and analyze (Cragin et al., 2007; Strasser, Cook, Michener, & Budden, 2012). Communication between the curation team and data providers throughout the data lifecycle is an essential task. One of the interview participants mentioned that he, as a data curator, spends about 80% of his time on communication. Because of the importance of communication for successful data activities, IR staff consider an individual's communicative skill when designing an effective team for research data curation (see Table 4.8). In addition, data curation teams prefer having IR staff with research experience who also know data and data curation lifecycles. The identified division of labor among curation staff indicated the importance of networks and relationships between IR staff and domain experts (e.g., subject specialists/librarians, departmental collection administrators) in order to create and add metadata for different research datasets. Having the power to develop software within a data curation team makes it easier to maintain the IR system. In the instance when a IR data curation team uses software development expertise present in a different team specialized on software development within the same library, keeping a good relationship between the IR data curation team and the other team can still render an effective solution for data curation.

6.2 IR Infrastructure for Research Data Curation

IR staff use diverse tools/instruments for different activities and purposes. Although many of the tools are designed for similar purposes, they have slightly different goals and uses. IR software is a typical example. All of them support general repository functions (e.g., storing, publishing, sharing), but they have different characteristics in their administration, system structure, and storing objects. Institutions should consider a variety of factors, such as data curation needs of target communities, policies, norms, rules, division of labor when they select tools for their institutional needs. In addition, institutions should think of the potential future directions available with the tools/instruments. Many of the tools do not have sufficient expandability for their functions.

Policies, rules, and norms enable the IR system and the curation teams to be systematic in their practices. But the research data curation field is still an emerging community. As a result, the current practices of the IRs that participated in the study are not yet unified or standardized among themselves. Policies, rules, norms, and current practices were identified in this study. The institutions that currently manage institutional data repositories can refer to the current data curation practices, and they can also share the knowledge about their own practices to the community. The institutions and their IR staff need to keep a close relationship to stay current with best practices.

When an institution develops its IR, the institution needs to consider the user services that can be provided through the IR. Without active IR content-contributors and end-users, development of an IR is meaningless; thus, providing appealing user services is important. Designing IR user services that can increase institutional members' IR use and enhance IR staff's outreach services is an activity to consider before IR implementation. All of the IR staff that participated in the study mentioned the outreach services they provide in order to increase the use of their IRs. The interview data identified different IR user services that are currently supported by the IRs. The IRs currently provide typical repository services (i.e., identifying, searching, browsing, and downloading) and several optional services (i.e., sharing, social networking, full-text searching, and bookmarking).

IRs operate in the context of the specific sociotechnical factors (i.e., tools/instruments, policies, rules, norm, culture, division of labor) of their community. But there are numerous conflicts between the factors. For example, some tools cannot be used without using another tool. A tool cannot be used because of limitations in other resources (e.g., system development power, budget, policy). Institutions must have a distinct goal for their IR systems and set a plan for future directions. All of the goals and plans should be attend various sociotechnical factors that may affect IR infrastructure.

6.3 Data Types

Research data tends to be complex and diverse. All of the IRs that participated in the study do not have restrictions on the types of research data they will receive. Only some technical limitations, such as file scale and the number of files, exist for acceptable data characteristics. In contrast with the diversity of accepted research data types, the controlled entity types used to identify complex research data do not reflect their complexity. The major controlled entity types used by the IRs are not different from the entity types of general library objects (e.g., books, journal articles). Instead, the IRs use supplementary ReadMe files to add domain specific metadata for the datasets. Although the supplementary information is indexed by search engine(s) used in IRs, the information is not as effectively and efficiently searchable as the controlled entity metadata. This informs directions for future studies including scientific metadata and data identifiers. More specifically, what would be the different metadata elements for research data (e.g., software code files, lab notebooks, databases) and what are the current data identification schemes that can be used to reference identify different data entities?

6.4 Identifier Selection

Based on the data analysis of IR staff's perceptions regarding identifier quality and quality problems, IR staff perceive that identifiers currently available for research data curation do not directly derive quality problems from the schemas themselves. However, a few interview participants mentioned some quality problems existing around the context of identifiers, including human-error, mapping, and changes to the underlying entity or condition (see Table 4.10). For example, an identifier cannot locate a research dataset due to incomplete system update and maintenance (i.e., mapping-related issue). As another example, an identifier system does not provide access to its user because of unstable server condition or software incompatibility (i.e., context-related issue). In order to avoid such problems, IR staff must not only maintain their systems and servers so that those are always up to date and operable, but also recognize and document the implementation details of identifier system features and functions for future uses.

Quality system selection and maintenance require a comprehensive understanding of current quality problems, institutional goals, and user needs of the system (Stvilia et al., 2004). Understanding current quality problems enables the system administrator to control and prevent the problems. The identification of institutional goals and user needs also helps create a better alignment between the system and its purpose. Since most of the identifiers do not have direct identifier system-related quality problems, the important factors that affect the decision to select an identifier system are the institutional goals for the IR services. Based on their goals, different institutions should select different identifier systems. For example, if an institution plans to set up an IR to support linked data sharing on the web (Berners-Lee et al., 2001), the institution might want to select Hydra software with the use of HTTP URIs. Also, if an IR wants to provide citation services to its institutional members, they might decide to use DOI.

In addition, the interview data identified the IR staff's perceptions of identifier quality requirements based on currently existing identifier quality dimensions (Lee & Stvilia, 2014). In the context of IRs, technical value of identifier systems (i.e., uniqueness, persistence, actionability, scalability, authority, interoperability, and security) seemed to be more important than user value (i.e., actionability, authority, verifiability, opacity, and simplicity) and archive value (i.e., granularity) of the systems. These findings provide insight not only for the development of identifier schemas, but also for the process of selecting an identifier schema. The findings can also inform IR staff, librarians, scholarly communities, and publishers about the needs and requirements for an identifier schema to help discover, aggregate, and cite data, and reveals some of the issues and problems related to current uses of identifier schemas for data in IRs.

6.5 Future Research

This study suggests two potential areas for future research: (1) to investigate IR infrastructure while considering different potential conflicts and their context, and (2) to examine complex data types, their entities, and data identifiers to improve the function of identification.

Developing an effective IR infrastructure that takes into consideration different potential conflicts is a significant challenge. However, the findings in the sections related to data activities can inform the development of best practices, infrastructure configuration templates, as well as the teaching of data curation in LIS schools. The findings also indicate other future research directions, such as: exploring data curation practices of IRs that use a specific tool; investigating IR user services to find ways to better motivate researchers' IR use, and to design and manage IR user community(s); and surveying different conflicts that can exist within IR infrastructure. In addition, similar studies that examine goals, perceptions, and uses of IRs from the perspectives of data providers, end-users, or university administrators can help the IR community overcome various challenges associated with the operations of IRs. Also, exploring existing institutional barriers to establishing IRs at universities that do not have them yet would be a great resource for the community. Conducting future research can expand the current study and also provide guidelines for designing and developing IR infrastructure.

With the emphasis on data sharing and reuse (NSF, 2010b), research data communities put their efforts in the design of metadata for data, including identifier schemas (Duerr et al., 2011; Lee & Stvilia, 2012; NISO, 2013). In many cases, metadata designs are tailored to the specific research disciplines' data practices, and it requires examining the data types and entities as well as current metadata schemes. This study was conducted with limited numbers of interview participants and therefore the findings on major research data types and their entities has limited generalizability. This limitation provides a clear direction for future research to expand the current study in order to conduct a survey and quantitative analysis based on the current findings. More specifically, there are a variety of potential quantitative studies that could be conducted to build on the current qualitative study: to investigate metadata elements for various and complex research data; to explore the current identifier schemas that can identify entities of the metadata elements; and to study data identifier quality priorities and metadata priorities in general.

Identifiers are fundamental metadata that traditionally have been used for entity identification, linking, and referencing cross domains (Altman & King, 2007). According to the study findings, many of the interview participants have plans to or interest in adopting new identifier systems for different entity types (e.g., person). The ORCID identifier system is one example of an identifier system that could be used for different entity types. But, some of the interviewees also explained their concerns about the lack of known practices for ORCID identifiers. Thus, a future research direction could be to understand the ORCID system thoroughly and to compare its metadata elements with other person entity-type identifiers. ISNI, which can be assigned to a natural person, a legal person, a fictional character, or a group, is a potential identification schema for a comparison study.

APPENDIX A

INTERVIEW PROTOCOL

Identifier Practices for the Curation of Research Data in Institutional Repositories:

Thank you for participating in my research study. The purpose of this study is to gain an understanding of the identifier practices for research data in institutional repositories (IRs). In particular, I am interested in how IR staff use, manage identifiers and perceive identifier quality. I have several questions to ask, and I hope you will feel free to talk about any experiences or ideas that come to mind.

I would like to record our conversation in order to facilitate note taking. The recording will be transcribed and the recording will be destroyed in 1 year. Please remember that you may ask to turn off the recorder at any time.

Demographic Information

1. Tell me a little about your position in your institution in regard to managing IR.

What are the other positions existing in your institution to manage IR?

2. What was your highest degree? What was the formal discipline of your degree and what are your specific areas?

Data Activities

3. What is the main objective of your IR?

4. How long has your IR allowed submission and searching of research data?

5. What are some of the activities you perform in managing and curating data in your IR?

6. What user activities (e.g., identifying, searching, browsing, social networking, annotating, citing, linking, etc.) does your IR currently support?

7. What is the division of labor in your IR - what are some of the roles related to those activities (e.g., curator, data provider, user, etc.)?

8. What are some of the tools (i.e., software, instruments, etc.) you use to manage IR objects in your IR? Are they different from tools for research data?

9. What are some of the tools (i.e., software, services, etc.) you provide for your IR user community to store, organize, analyze, visualize, share, communicate about, and/or interact with data?

10. Does your institution manage its IR database by itself or does it uses an outside company?

11. What is the repository software (e.g., DSpace, EPrints, Fedora, etc.) that your IR uses?

12. What are the metadata schemas (e.g., DC, PREMIS, MODS, TEI, etc.) used in your IR?

13. What are the metadata schemas (e.g., DDI, DwC, EML, etc.) used for research data?

14. What are the identifiers (e.g., DOI, Handle, ARK, UUID, etc.) used in your IR?

15. Does your IR use different identifier(s) for research data? If so, what is it?

16. Are there any policies, rules, norms, or best practices that guide data management and use in your IR? If yes, please name them. Do these policies, rules, or norms come from the government, funding agencies, community, or are developed locally?

17. Are there any policies, rules, or norms that govern or guide identifier system selection and use in your IR? If so, please name them. Do these policies, rules, or norms come from the government, funding agencies, community, or are developed locally?

Data Types

18. What major types of research data (e.g., raw data, slides, text documents, spreadsheets, laboratory notes, etc.) does your IR accept?

19. What types of research data entities does your IR control metadata for (e.g., author, subject, geographic location, etc.)? The following page contains a list of research data entities found in the literature. Provide the attached list.

> After reviewing this list, are there any types of data that do not make sense or are not applicable in your work context? Or do any other entity types come to mind?
>
> What controlled vocabularies (e.g., SKOS vocabulary, etc.) does your IR use to control that entity metadata?

Perception of Identifier Quality

20. What are some identifiers or identifier systems you are familiar with? What do you know about them?

21. What identifiers do you use at the data collection/set level, file/object level, or entity level?

22. Are you familiar with identifier quality assessment criteria (or models)? If so, have you used those criteria in practice and/or research?

23. Can you recall a case when an identifier quality problem (e.g., access failure, incorrect access, etc.) led to disruption in IR activity? If yes, please describe it, and explain how you overcame the problem.

24. The following page contains a list of identifier quality criteria for research data found in the literature. Provide the attached list.

> After reviewing this list, are there any criteria that do not make sense or are not applicable in your work context? Or do any other criteria come to mind?
>
> How do you evaluate the quality of identifier systems for research data? On a scale where 1 indicates “extremely unimportant” and 7 indicates “extremely important,” please indicate the level of importance of each of the following data identifier quality criteria within the context of your IR. Can you briefly explain how you came up with the evaluation of the highest and lowest ranks?

Definitions of Research Data Entity Types

Entities	Definitions
Intellectual Entity	A set of content that is considered a single intellectual unit for purposes of management and description
Object	Discrete units of information in digital form. Can be files, bitstreams or representations. Objects are what are actually stored and managed in the preservation repository
Symbolic Object	An identifiable symbol and any aggregation of symbols, such as characters, data sets, images, multimedia objects, or mathematical formulae that have an objectively recognizable structure and that are documented as single units
Person	An individual; Real person
Organization	An organization or group of individuals and/or organizations acting as a unit; Institutions or groups of people that have obtained a legal recognition as a group and can act collectively as agents
Place	A geographical location; it comprises extents in space, in particular on the surface of the earth, in the pure sense of physics
Time	Specific forms of historical periods or dates; abstract temporal extents, having a beginning and an end
Event	An action or occurrence; actions that involve an Object and an Agent known to the system; changes of states in cultural, social or physical systems, regardless of scale
Topic	A hierarchy of topics used to organize the content of the dataset

Data Identifier Quality Dimensions and Definitions

Dimensions	Definitions
Uniqueness	The requirement that one identifier string denotes one and only one data object
Persistence	The requirement that once assigned, an identifier string denotes the same referent indefinitely
Simplicity	The degree of cognitive simplicity of an identifier string
Opacity	The extent to which the meaning cannot be inferred from the content, structure or pattern of an identifier string
Verifiability	The extent to which the correctness and validity of an identifier string is verifiable or provable
Interoperability	The ability to use an identifier system and string in services outside of the direct control of the issuing assigner
Actionability/Resolvability	The ability of the identifier system to locate the object using an identifier string
Granularity/ Flexibility	The extent to which the identifier system allows the reference of data at different granularity
Authority	The degree of reputation of an identifier system in a given community
Scalability	The ability of an identifier system to expand its level of performance or efficiency (e.g., support RDF)
Security	The extent to which the resource of an identifier system is protected from unauthorized administrative access or modification

Please evaluate the importance of the following quality criteria of identifier systems in your IR context.

Criteria	**1 (extremely unimportant)**	**2**	**3**	**4**	**5**	**6**	**7 (extremely important)**
Uniqueness							
Persistence							
Simplicity							
Opacity							
Verifiability							
Interoperability							
Actionability/Resolvability							
Granularity/Flexibility							
Authority							
Scalability							
Security							

APPENDIX B

INITIAL CODING SCHEME

RQ 1. What are the types of data activities in IRs and what are the structures and metadata requirements of those activities?

Demographic Information

- Degree
- Discipline

Data Curation Activities in IR

- IR objectives
- Activity
    - Curator activity
    - Data provider activity
    - User activity
- Division of labor
- Tools
    - Software
    - Web apps
    - Metadata schemas
    - Identifier schemas
    - Repository software
- Norms, policies, rules
- Communities
- Contradictions
- Roles
- Skills
    - Curator skills
    - Data provider skills
    - User skills
- User services

RQ 2. What are the major types of research data and their entity types within IRs for which identifiers are used?

Data Types

- Data types
- Data entity types

RQ 3. What is the awareness of IR curators about different currently available identifier schemas?

Perception of Identifier Quality

- Data identifiers
- Familiarity of identifiers

• Identifier granularity • Identifier quality criteria • Identifier quality problems • Identifier quality problem source • Identifier quality assurance action
RQ 4. How do IR curators perceive the quality of identifiers for research data?
• Identifier quality criteria

DEFINITIONS

Degree
A course of study that you take at a university or college, or the qualification that you get when you have passed the course

Discipline
Particular area of study

IR objectives
What IRs are trying to achieve

Activity
A complex system of related elemetns, including roles (i.e., subject, objects), actions, rules, and tools

Division of labor
Both the horizontal division of tasks between members of the community and the vertical division of power and status

Tools
Artifacts, abstract or physical, used by the subject of an activity

Norms, policies, rules
Explicit or impolicit norms, conventions, regulations that enable or limit the actions, operations, and interactions within an activity system

Communities
A group of people who share the same object

Contradictions
Tensions, conflicts, or limitations among the components of an activity system

Roles
Particular task or function in a position

Skills
Type of work or activity, which requires special training and knowledge

User services
The services that IRs provide for their users

Data types
Different kinds of data (e.g., raw data, slides, text, spreadsheet, etc.)

Data entity types
That which constitutes the being of data; essence, essential nature

Data identifiers
A sequence of symbols designed to identify, cite, annotate, and/or link research data and their associated metadata.

Familiarity of identifiers
Knowledge of identifiers through long or close association or frequent perception by any of the senses

Identifier granularity
The extent to which the identifier system allows to reference data at different granularity

Identifier quality criteria
A set of attributes that represents a single aspect or construct of identifier quality (any component of the identifier quality concept)

Identifier quality problems
Any problem that occur to the identifiers when the identifiers cannot meet the needs and requirements of the activities in which they are used

Identifier quality problem source
Any source that leads to identifier quality variance and may suggest both the types of identifier quality problems and the types of identifier quality assurance actions

Identifier quality assurance action
Any action taken by the subject to improve the identifier quality to meet the needs and requirements of the activities in which the data are used

APPENDIX C

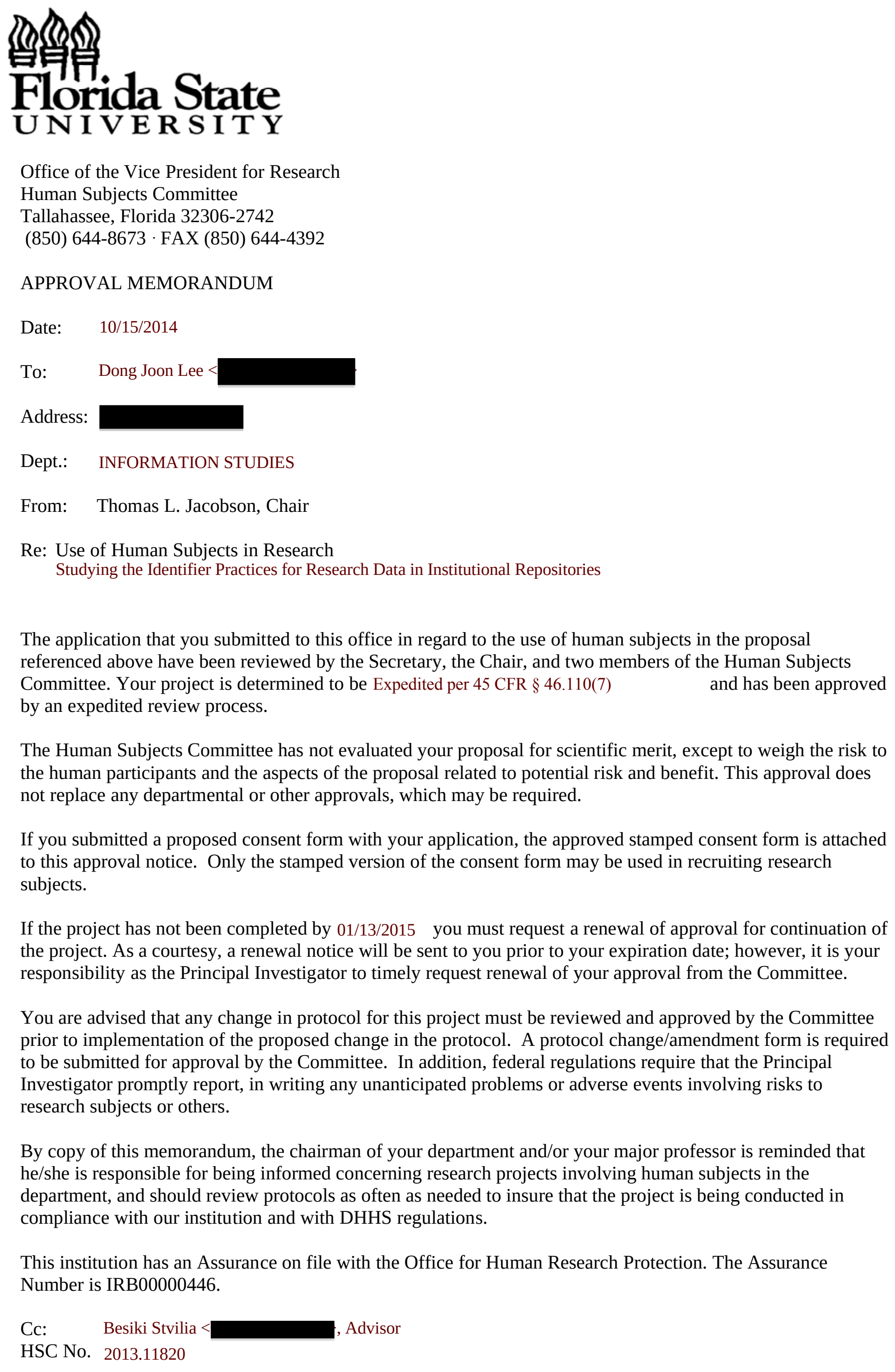
TOOLS

Categories	Tools	URLs
IR Software	Bepress Digital Commons	http://digitalcommons.bepress.com/
	DSpace	http://www.dspace.org/
	Hydra	http://projecthydra.org/
	Dataverse	http://dataverse.org/
	HUBzero	https://hubzero.org/
	Aubrey	Locally developed software
	SobekCM	http://ufdc.ufl.edu/sobekcm
Metadata Schemas	Dublin Core	http://dublincore.org/
	DataCite	https://schema.datacite.org/
	MODS	http://www.loc.gov/standards/mods/
	METS	http://www.loc.gov/standards/mets/
	PREMIS	http://www.loc.gov/standards/premis/
	MIX	http://www.loc.gov/standards/mix/
	EAD	http://www.loc.gov/ead/
	TEI	http://www.tei-c.org/index.xml
	FGDC	https://www.fgdc.gov/metadata/geospatial-metadata-standards
	DDI	http://www.ddialliance.org/
	Darwin Core	http://rs.tdwg.org/dwc/
	EML	https://knb.ecoinformatics.org/#external//emlparser/docs/index.html
	ISO 19115 Geographical Metadata	http://www.iso.org/iso/catalogue_detail.htm?csnumber=26020
Identifier Schemas	DOI	http://www.doi.org/
	Handle	http://www.handle.net/
	ARK	https://wiki.ucop.edu/display/Curation/ARK
	HTTP URI	http://www.w3.org/DesignIssues/HTTP-URI.html
Controlled Vocabularies	LCSH	http://id.loc.gov/authorities/subjects.html
	MeSH	https://www.nlm.nih.gov/pubs/factsheets/mesh.html
	FAST	http://www.oclc.org/research/themes/data-science/fast.html?urlm=168918
	RDF Ontology	http://semanticweb.org/wiki/Ontology
	FOAF	http://xmlns.com/foaf/spec/
	RDF Schema	http://www.w3.org/TR/rdf-schema/
Applications for Data Curation	Microsoft Office	https://products.office.com/en-US/
	WordPad	http://windows.microsoft.com/en-us/windows7/products/features/wordpad

	Notepad++	https://notepad-plus-plus.org/
	Oxygen XML Editor	http://www.oxygenxml.com/
	Morpho	http://knb.ecoinformatics.org/morphoportal.jsp
	Nesstar	http://www.nesstar.com/
	SnagIt	https://www.techsmith.com/snagit.html
	Handbreak	https://handbrake.fr/
	Open Refine	http://openrefine.org/
	Dropbox	https://www.dropbox.com/
	Google Drive	https://www.google.com/drive/
	DROID	http://www.nationalarchives.gov.uk/information-management/manage-information/policy-process/digital-continuity/file-profiling-tool-droid/
	PRONOM	http://apps.nationalarchives.gov.uk/PRONOM/Default.aspx
	Git	https://git-scm.com/
	FITS	http://projects.iq.harvard.edu/files/fits/files/fits_poster_final.pdf
	BagIt	https://wiki.ucop.edu/display/Curation/BagIt
	Apache Solar	http://lucene.apache.org/solr/
	Altmetric	http://altmetrics.org/manifesto/

APPENDIX D

APPROVALS FROM HUMAN SUBJECTS COMMITTEE



Florida State UNIVERSITY

Office of the Vice President for Research
Human Subjects Committee
Tallahassee, Florida 32306-2742
(850) 644-8673 · FAX (850) 644-4392

APPROVAL MEMORANDUM

Date: 10/15/2014

To: Dong Joon Lee <[redacted]>

Address: [redacted]

Dept.: INFORMATION STUDIES

From: Thomas L. Jacobson, Chair

Re: Use of Human Subjects in Research
Studying the Identifier Practices for Research Data in Institutional Repositories

The application that you submitted to this office in regard to the use of human subjects in the proposal referenced above have been reviewed by the Secretary, the Chair, and two members of the Human Subjects Committee. Your project is determined to be Expedited per 45 CFR § 46.110(7) and has been approved by an expedited review process.

The Human Subjects Committee has not evaluated your proposal for scientific merit, except to weigh the risk to the human participants and the aspects of the proposal related to potential risk and benefit. This approval does not replace any departmental or other approvals, which may be required.

If you submitted a proposed consent form with your application, the approved stamped consent form is attached to this approval notice. Only the stamped version of the consent form may be used in recruiting research subjects.

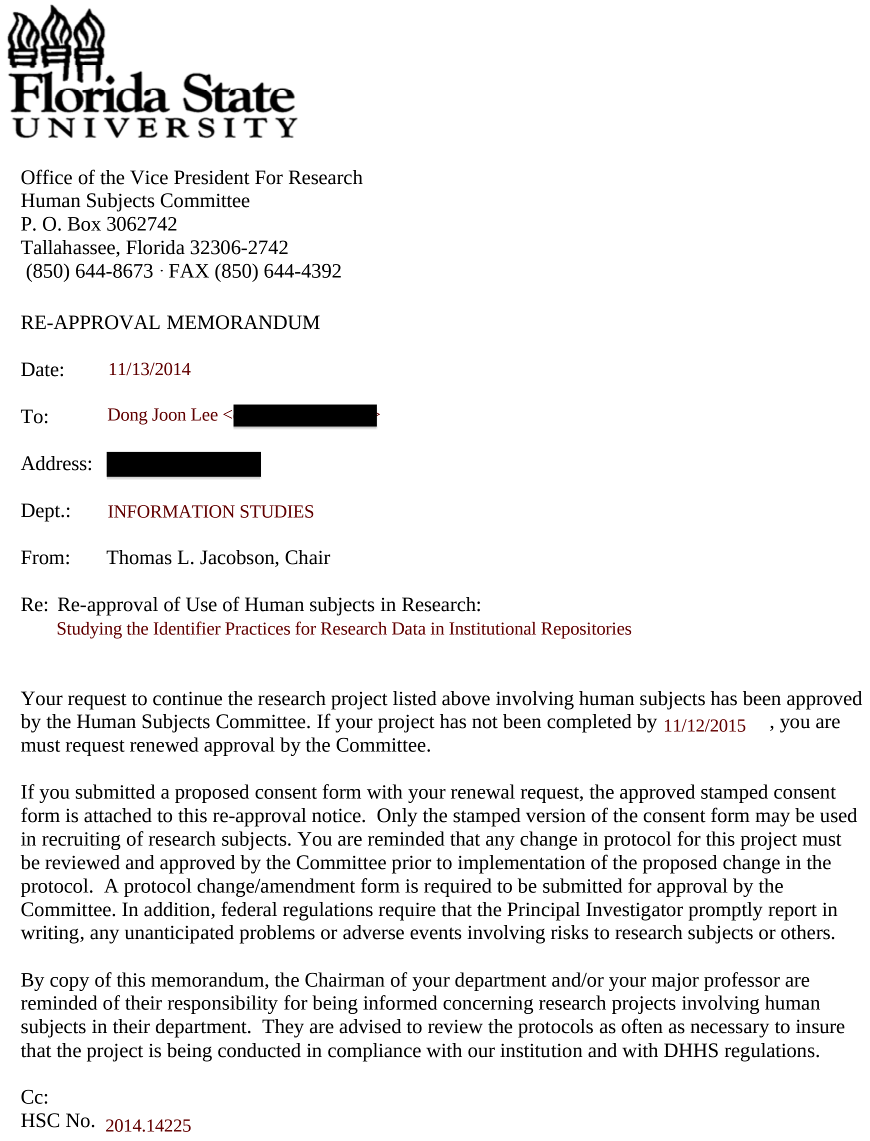
If the project has not been completed by 01/13/2015 you must request a renewal of approval for continuation of the project. As a courtesy, a renewal notice will be sent to you prior to your expiration date; however, it is your responsibility as the Principal Investigator to timely request renewal of your approval from the Committee.

You are advised that any change in protocol for this project must be reviewed and approved by the Committee prior to implementation of the proposed change in the protocol. A protocol change/amendment form is required to be submitted for approval by the Committee. In addition, federal regulations require that the Principal Investigator promptly report, in writing any unanticipated problems or adverse events involving risks to research subjects or others.

By copy of this memorandum, the chairman of your department and/or your major professor is reminded that he/she is responsible for being informed concerning research projects involving human subjects in the department, and should review protocols as often as needed to insure that the project is being conducted in compliance with our institution and with DHHS regulations.

This institution has an Assurance on file with the Office for Human Research Protection. The Assurance Number is IRB00000446.

Cc: Besiki Stvilia <[redacted]>, Advisor
HSC No. 2013.11820



Florida State UNIVERSITY

Office of the Vice President For Research
Human Subjects Committee
P. O. Box 3062742
Tallahassee, Florida 32306-2742
(850) 644-8673 · FAX (850) 644-4392

RE-APPROVAL MEMORANDUM

Date: 11/13/2014

To: Dong Joon Lee <[redacted]

Address: [redacted]

Dept.: INFORMATION STUDIES

From: Thomas L. Jacobson, Chair

Re: Re-approval of Use of Human subjects in Research:
Studying the Identifier Practices for Research Data in Institutional Repositories

Your request to continue the research project listed above involving human subjects has been approved by the Human Subjects Committee. If your project has not been completed by 11/12/2015 , you are must request renewed approval by the Committee.

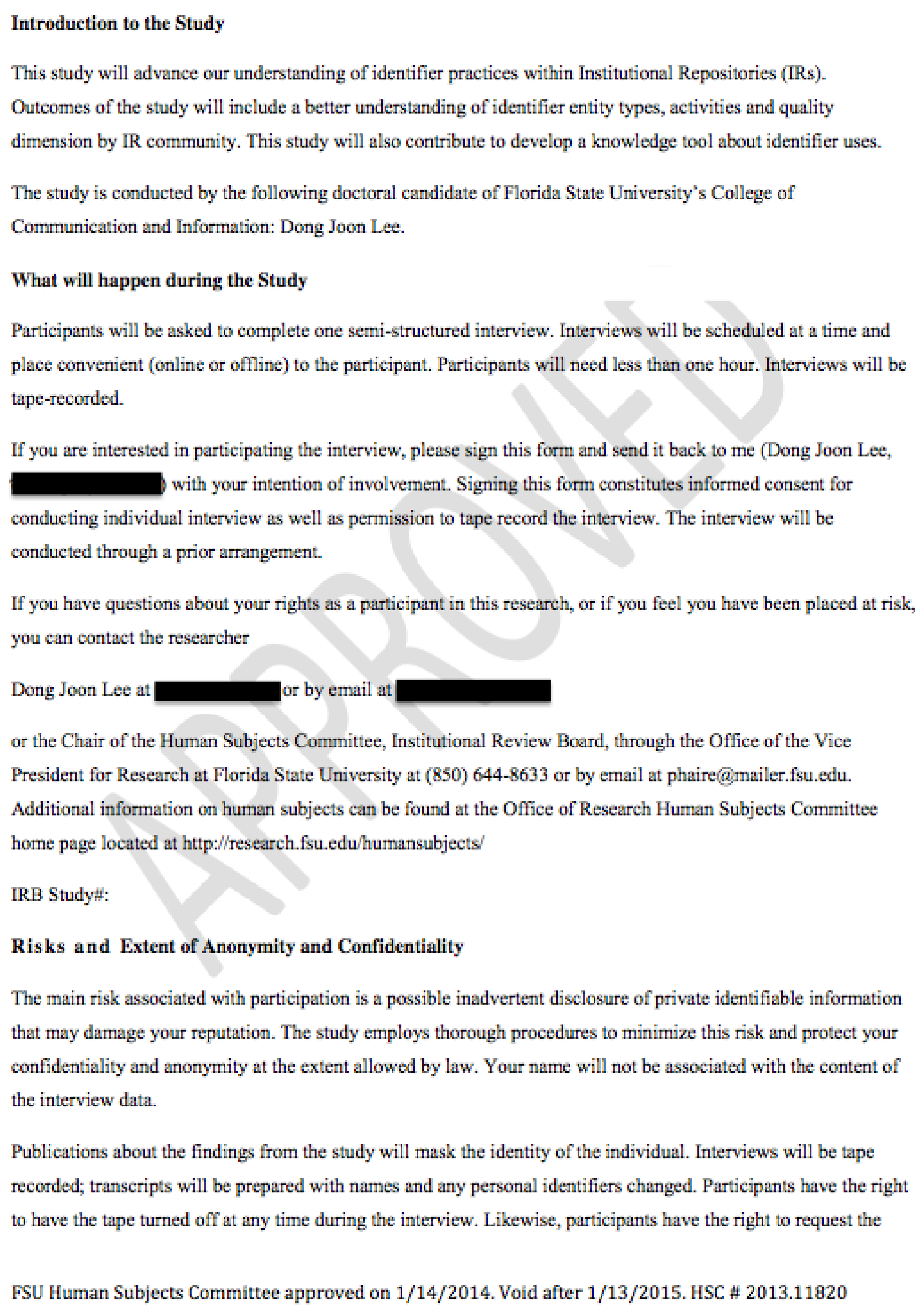
If you submitted a proposed consent form with your renewal request, the approved stamped consent form is attached to this re-approval notice. Only the stamped version of the consent form may be used in recruiting of research subjects. You are reminded that any change in protocol for this project must be reviewed and approved by the Committee prior to implementation of the proposed change in the protocol. A protocol change/amendment form is required to be submitted for approval by the Committee. In addition, federal regulations require that the Principal Investigator promptly report in writing, any unanticipated problems or adverse events involving risks to research subjects or others.

By copy of this memorandum, the Chairman of your department and/or your major professor are reminded of their responsibility for being informed concerning research projects involving human subjects in their department. They are advised to review the protocols as often as necessary to insure that the project is being conducted in compliance with our institution and with DHHS regulations.

Cc:
HSC No. 2014.14225

APPENDIX E

INFORMED CONSENT FORM



Introduction to the Study

This study will advance our understanding of identifier practices within Institutional Repositories (IRs). Outcomes of the study will include a better understanding of identifier entity types, activities and quality dimension by IR community. This study will also contribute to develop a knowledge tool about identifier uses.

The study is conducted by the following doctoral candidate of Florida State University's College of Communication and Information: Dong Joon Lee.

What will happen during the Study

Participants will be asked to complete one semi-structured interview. Interviews will be scheduled at a time and place convenient (online or offline) to the participant. Participants will need less than one hour. Interviews will be tape-recorded.

If you are interested in participating the interview, please sign this form and send it back to me (Dong Joon Lee, [redacted]) with your intention of involvement. Signing this form constitutes informed consent for conducting individual interview as well as permission to tape record the interview. The interview will be conducted through a prior arrangement.

If you have questions about your rights as a participant in this research, or if you feel you have been placed at risk, you can contact the researcher

Dong Joon Lee at [redacted] or by email at [redacted]

or the Chair of the Human Subjects Committee, Institutional Review Board, through the Office of the Vice President for Research at Florida State University at (850) 644-8633 or by email at phaire@mailer.fsu.edu. Additional information on human subjects can be found at the Office of Research Human Subjects Committee home page located at http://research.fsu.edu/humansubjects/

IRB Study#:

Risks and Extent of Anonymity and Confidentiality

The main risk associated with participation is a possible inadvertent disclosure of private identifiable information that may damage your reputation. The study employs thorough procedures to minimize this risk and protect your confidentiality and anonymity at the extent allowed by law. Your name will not be associated with the content of the interview data.

Publications about the findings from the study will mask the identity of the individual. Interviews will be tape recorded; transcripts will be prepared with names and any personal identifiers changed. Participants have the right to have the tape turned off at any time during the interview. Likewise, participants have the right to request the

FSU Human Subjects Committee approved on 1/14/2014. Void after 1/13/2015. HSC # 2013.11820 interview be stopped and notes destroyed. Tapes, transcripts and notes will remain in the possession of the primary researcher and stored on a password protected server.

Benefits of this Project

The results of this exploratory inquiry are expected to contribute to the advancement of both practical and theoretical knowledge of identifier uses in IRs.

Compensation

Participants will not receive compensation for this interview.

Participant's Rights

In accordance with Florida State University (FSU) policy, and as the researcher, I would like to assure you that:

- Participation in this study is entirely voluntary
- If you decide to participate, you have the right to withdraw your consent at any time. You are free not to answer any questions that you choose or to request that the tape recorder be turned off at any time during the interview.

All research on human volunteers is reviewed by a committee that works to protect your rights and welfare. If you have any questions or concerns regarding the study and would like to talk to someone other than the researcher(s), you are encouraged to contact the FSU IRB at telephone number 850-644-7900. You may also contact this office by email at humansubjects@magnet.fsu.edu, or by writing or in person at 2010 Levy Street, Research Building B, Suite 276, FSU Human Subjects Committee, Tallahassee, FL 32306-2742.

Participant's Permission

By signing this form below, you acknowledge that you've read and understood the above statement and consent to participate in this study.

If I participate, I may withdraw at any time. I agree to abide by the rules of this project.

Signature: ______________________________ Date: ____________________

REFERENCES

Aalbersberg, Ij. J., & Kähler, O. (2011). Supporting Science through the Interoperability of Data and Articles. *D-Lib Magazine*, *17*(1/2). doi:10.1045/january2011-aalbersberg

Abbott, D. (2008). Annotation. *DCC Briefing Papers: Introduction to Curation*. Retrieved from http://www.dcc.ac.uk/resources/briefing-papers/introduction-curation/annotation

Abbott, D. (2009). Interoperability. *DCC Briefing Papers: Introduction to Curation*. Retrieved from http://www.dcc.ac.uk/resources/briefing-papers/introduction-curation/interoperability

Akhondi, S. A., Kors, J. A., & Muresan, S. (2012). Consistency of systematic chemical identifiers within and between small-molecule databases. *Journal of Cheminformatics*, *4*(1), 35. doi:10.1186/1758-2946-4-35

Altman, M., & King, G. (2007). A Proposed Standard for the Scholarly Citation of Quantitative Data. *D-Lib Magazine*, *13*(3/4).

American Psychological Association (APA). (2010). *Publication manual of the APA* (6th ed.). Washington, D.C.

Baker, T., & Dekkers, M. (2003). Identifying Metadata Elements with URIs. *D-Lib Magazine*, *9*(7/8). doi:10.1045/july2003-baker

Barriball, L. K., & While, A. (1994). Collecting data using a semi-structured interview: a discussion paper. *Journal of Advanced Nursing*, *19*(2), 328–335. doi:10.1111/j.1365-2648.1994.tb01088.x

Berners-Lee, T. (1998). Cool URIs don't change. W3C. Retrieved from http://www.w3.org/Provider/Style/URI.html

Berners-Lee, T. (2003a). Message on www-tag@w3.org list. Retrieved fromhttp://lists.w3.org/Archives/Public/www-tag/2003Jul/0158.html

Berners-Lee, T. (2003b). Message to www-tag@w3.org list. Retrieved from http://lists.w3.org/Archives/Public/www-tag/2003Jul/0022.html

Berners-Lee, T. (2006). Linked Data. W3C. Retrieved from http://www.w3.org/DesignIssues/LinkedData.html

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic web. *Scientific American*, *284*(5), 34.

Bizer, C., Cyganiak, R., & Heath, T. (2007). How to publish linked data on the Web. Retrieved from http://www4.wiwiss.fu-berlin.de/bizer/pub/linkeddatatutorial/

Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked data - The story so far. *International Journal on Semantic Web and Information Systems*, *5*(3), 1–22. doi:10.4018/jswis.2009081901

Blee, K., M., & Taylor, V. (2002). Semi-Structured Interviewing in Social Movement Research. In B. Klandermans & S. Staggenborg (Eds.), *Methods of Social Movement Research* (pp. 92–117). University of Minnesota Press.

Borgman, C. L., Wallis, J. C., & Enyedy, N. (2007). Little science confronts the data deluge: habitat ecology, embedded sensor networks, and digital libraries. *International Journal on Digital Libraries*, *7*(1-2), 17–30. doi:10.1007/s00799-007-0022-9

Brand, A., Daly, F., & Meyers, B. (2003). *Metadata demystified*. Sheridan and NISO Press. Retrieved from http://www.niso.org/standards/resources/Metadata_Demystified.pdf

Brase, J., & Farquhar, A. (2011). Access to Research Data. *D-Lib Magazine*, *17*(1/2). doi:10.1045/january2011-brase

Bryant, R. (2013). ORCID supports the interoperable exchange of datasets. Retrieved from https://orcid.org/blog/2013/09/24/orcid-supports-interoperable-exchange-datasets

Buckland, M. (1998). What is a digital document? *Document Numerique (Paris)*, *2*(2), 221–230.

California Digital Library (CDL). (2012). ARK (Archival resource key) identifiers. Retrieved from https://wiki.ucop.edu/display/Curation/ARK

California Digital Library (CDL). (2013). NOID: Nice Opaque Identifier (Minter and Name Resolver). Retrieved from https://wiki.ucop.edu/display/Curation/NOID

Callaghan, S., Donegan, S., Pepler, S., Thorley, M., Cunningham, N., Kirsch, P., … Wright, D. (2012). Making data a first class scientific output: Data citation and publication by NERC's environmental data centres. *International Journal of Digital Curation*, *7*(1), 107–113. doi:10.2218/ijdc.v7i1.218

Caplan, P. (2009). *Understanding PREMIS*. Library of Congress.

Carlyle, A. (2004). *FRBR and the bibliographic universe, or, how to read FRBR as a model*. Presented at the ALA Annual Conference, Orlando, FL. Retrieved from http://www.ala.org/alcts/sites/ala.org.alcts/files/content/events/pastala/annual/04/Carlyle.pdf

Carroll, J. (1997). Human-computer interaction: Psychology as a science of design. *International Journal of Human-Computer Studies*, *46*, 501–522.

Chemical Abstracts Service (CAS). (2012a). CAS information use policies. Retrieved from http://www.cas.org/legal/infopolicy

Chemical Abstracts Service (CAS). (2012b). CAS registry - The gold standard for chemical substance information. Retrieved from http://www.cas.org/content/chemical-substances

Cho, Y. (2008). Intercoder reliability. In P. Lacrakas (Ed.), *Encyclopedia of survey research methods*. Thousand Oaks, California: SAGE Publications, Inc.

Clark, A. (2006). Anonymising research data. ESRC National Centre for Research Methods. Retrieved from http://eprints.ncrm.ac.uk/480/1/0706_anonymising_research_data.pdf

Clark, T., Martin, S., & Liefeld, T. (2004). Globally distributed object identification for biological knowledgebases. *Briefings in Bioinformatics*, *5*(1), 59–70.

Common Chemistry. (2013). About common chemistry. Retrieved from http://tinyurl.com/nnh9uj7

Consultative Committee for Space Data Systems (CCSDS). (2012). *Reference model for an open archival information system (OAIS)* (Recommended Practice No. CCSDS 650.0-M-2). Retrieved from http://public.ccsds.org/publications/archive/650x0m2.pdf

Conway, P. (2011). Archival quality and long-term preservation: a research framework for validating the usefulness of digital surrogates. *Archival Science*, *11*(3-4), 293–309. doi:10.1007/s10502-011-9155-0

Corporation for National Research Initiatives (CNRI). (2012, October). System fundamentals. Retrieved from http://www.handle.net/overviews/system_fundamentals.html

Costas, R., Meijer, I., Zahedi, Z., & Wouters, P. (2013). *The value of research data - Metrics for datasets from a cultural and technical point of view*. Retrieved from www.knowledge-exchange.info/datametrics

Cragin, M. H., Heidorn, P. B., Palmer, C. L., & Smith, L. C. (2007, June 25). An educational program on data curation. Retrieved from http://hdl.handle.net/2142/3493

Creswell, J. W. (2007). *Educational research: planning, conducting, and evaluating quantitative and qualitative research*. Upper Saddle River, N.J.: Pearson/Merrill Prentice Hall.

Crosas, M. (2011). The Dataverse Network®: An Open-Source Application for Sharing, Discovering and Preserving Data. *D-Lib Magazine*, *17*(1/2). doi:10.1045/january2011-crosas

CrossRef. (2011). CrossRef & ORCID. Retrieved from http://www.crossref.org/01company/orcid.html

Curry, E., Freitas, A., & O’Riáin, S. (2010). The role of community-driven data curation for enterprises. In D. Wood (Ed.), *Linking Enterprise Data* (pp. 25–47). Springer US. Retrieved from http://link.springer.com/chapter/10.1007/978-1-4419-7665-9_2

Data & GIS Lab. (2013). GIS across the disciplines. Retrieved from http://libguides.ucsd.edu/content.php?pid=42741&sid=1825758

DataCite. (n.d.). DataCite. Retrieved November 6, 2012, from http://datacite.org/

DataUp. (n.d.). DataUp. Retrieved from http://dataup.cdlib.org/

Davidson, J. (2006). Persistent Identifiers. *DCC Briefing Papers: Introduction to Curation*.

DOI. (2013). The DOI Systems. Retrieved from http://www.doi.org/factsheets.html

Dublin Core Metadata Initiative, (2012). Dublin core metadata element set, Version 1.1. Retrieved from http://dublincore.org/documents/dces/

Duerr, R., Downs, R., Tilmes, C., Barkstrom, B., Lenhardt, W., Glassy, J., … Slaughter, P. (2011). On the utility of identification schemes for digital earth science data: an assessment and recommendations. *Earth Science Informatics*, *4*(3), 139–160. doi:10.1007/s12145-011-0083-6

Dunsire, G. (2007). Distinguishing Content from Carrier. *D-Lib Magazine*, *13*(1/2). doi:10.1045/january2007-dunsire

Engeström, Y. (1987). *Learning by expanding: An activity-theoretical approach to developmental research*. Helsinki, Finland: Orienta-Konsultit Oy.

Engeström, Y. (1990). Learning, working and imagining: Twelve studies in Activity Theory. Orienta-Konsultit, Helsinki.

Engeström, Y. (1999). Learning by expanding: Ten years after. Retrieved from http://lchc.ucsd.edu/mca/Paper/Engestrom/expanding/intro.htm

Enserink, M. (2009). Are You Ready to Become a Number? *Science*, *323*(5922), 1662–1664.

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

Erway, R. (2012). *Lasting Impact: Sustainability of Diciplinary Repositories*. Dublin, Ohio: OCLC Research. Retrieved from http://www.oclc.org/research/publications/library/2012/2012-03r.html

European Library Automation Group (ELAG). (2010, June). *Workshop on FRBR and Identifiers*. Presented at the European Library Automation Group 2010, Helsinki, Finland. Retrieved from http://elag2010.nationallibrary.fi/files/2010/06/ELAG-2010-workshop-on-FRBR-and-identifiers.pdf

Floyd, I. R., & Renear, A. H. (2007). What exactly is an item in the digital world? *Proceedings of the American Society for Information Science and Technology*, *44*(1), 1–7. doi:10.1002/meet.1450440374

Foster, I., Jennings, N. R., & Kesselman, C. (2004). Brain meets brawn:why grid and agents need each other. In *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems, 2004. AAMAS 2004* (pp. 8–15).

Galletta, A. (2013). *Mastering the semi-structured interview and beyond: From research design to analysis and publication*. New York, NY, USA: New York University Press.

Gasser, L. (1986). The integration of computing and routine work. *ACM Transactions on Information Systems 4*(3), 205-225.

GeoNames. (n.d.). About GeoNames. Retrieved from http://www.geonames.org/about.html

Go, K., & Carroll, J. (2004a). Scenario-based task analysis. In D. Diaper & N. Stanton (Eds.), *The handbook of task analysis for human-computer interaction* (pp. 117–133). Mahwah, NJ: Lawrence Erlbaum Associates.

Go, K., & Carroll, J. (2004b). The blind men and the elephant: Views of scenario-based system design. *Interactions*, *11*(6), 44–53.

Google Developers. (2012). Google Schema. Retrieved from https://developers.google.com/public-data/docs/schema/dspl9

Green, T. (2009). We need publishing standards for datasets and data tables. OECD Publishing. Retrieved from http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.40/2010/wp.8.e.pdf

Guess, A. (2012, October). Elsevier joins ORCID in launch of ORCID registry. *Semanticweb.com*. Retrieved from http://semanticweb.com/elsevier-joins-orcid-in-launch-of-orcid-registry_b32762

Halpin, H. (2008). The principle of self-description: Identity through linking. *Proceedings of the 1st IRSW 2008*. Retrieved from http://ceur-ws.org/Vol-422/irsw2008-submission-13.pdf

Halpin, H. (2011). Sense and Reference on the Web. *Minds and Machines*, *21*(2), 153–178.

Hardon, A., Hodgkin, C., & Fresle, D. A. (2004). *How to investigate the use of medicines by consumers*. World health organization (WHO).

Hayes, P. (2004). RDF semantics. W3C Recommendation, Retrieved from http://www.w3.org/TR/rdf-mt/

Heath, T. (n.d.). Linked Data. Retrieved from http://linkeddata.org/home

Heidorn, P. B. (2008). Shedding Light on the Dark Data in the Long Tail of Science. *Library Trends*, *57*(2), 280–299. doi:10.1353/lib.0.0036

Higgins, S. (2008). The DCC Curation Lifecycle Model. *International Journal of Digital Curation*, *3*(1), 134–140. doi:10.2218/ijdc.v3i1.48

Higgs, P., & Attwood, T. (2005). *Bioinformatics and molecular evolution*. Malden, MA: Blackwell Publishing Company.

Hinnant, C. C., Stvilia, B., Wu, S., Worrall, A., Burnett, G., Burnett, K., … Marty, P. F. (2012). Author-team diversity and the impact of scientific publications: Evidence from physics research at a national science lab. *Library & Information Science Research*, *34*(4), 249–257. doi:10.1016/j.lisr.2012.03.001

Holland, D., & Reeves, J. (n.d.). Activity theory and the view from somewhere: Team perspectives on the intellectual work of programming.

Huang, H., Stvilia, B., Jörgensen, C., & Bass, H. W. (2012). Prioritization of data quality dimensions and skills requirements in genome annotation work. *Journal of the American Society for Information Science and Technology*, *63*(1), 195–207. doi:10.1002/asi.21652

Huerta, M. (2013). *Data, data everywhere, but not a byte to eat*. Presented at the A Symposium of the Board on Research Data and Information, Washington, D.C. Retrieved from http://sites.nationalacademies.org/PGA/brdi/PGA_081268

Hutchins, E. (1987). *Metaphors for interface design* (No. ICS Report 8703). La Jolla: University of California, Department of Cognitive Science.

Hutchins, E. (1991). *How a cockpit remembers its speed*. Ms. La Jolla: University of California, Department of Cognitive Science.

ICOM/CIDOC CRM SIG. (2012). *Definition of the CIDOC conceptual reference model*. Retrieved from http://www.cidoc-crm.org/docs/cidoc_crm_version_5.1.pdf

Institute for Museum and Library Services. (2011). Specifications for projects that develop digital products. Retrieved from http://www.imls.gov/applicants/projects_that_develop_digital_products.aspx

International Federation of Library Associations and Institutions. (2009). *Functional requirements for bibliographic records*. Retrieved from http://www.ifla.org/files/assets/cataloguing/frbr/frbr_2008.pdf

International Organization for Standardization. (2012). *Information and documentation - International standard name identifier (ISNI)* (No. ISO 27729). Retrieved from http://www.iso.org/iso/catalogue_detail?csnumber=44292

ISNI. (2012). ISNI. Retrieved from http://www.isni.org/

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

Juty, N., Le Novère, N., & Laibe, C. (2011). Identifiers.org and MIRIAM Registry: community resources to provide persistent identification. *Nucleic Acids Research*, *40*(D1), D580–D586. doi:10.1093/nar/gkr1097

Kaptelinin, V., & Nardi, B. (2012). Activity Theory in HCI: Fundamentals and Reflections. *Synthesis Lectures on Human-Centered Informatics*, *5*(1), 1–105. doi:10.2200/S00413ED1V01Y201203HCI013

Kim, Y., Addom, B. K., & Stanton, J. M. (2011). Education for eScience Professionals: Integrating Data Curation and Cyberinfrastructure. *International Journal of Digital Curation*, *6*(1), 125–138. http://doi.org/10.2218/ijdc.v6i1.177

Kuutti, K. (1996). Activity theory as a potential framework for human-computer interaction research. In B. Nardi (Ed.), *Context and consciousness: Activity theory and human-computer interaction*. Cambridge,MA: MIT Press.

Kvale, S. (1996). *InterViews: An introduction to qualitative research interviewing*. London, UK: Sage.

Lagoze, C., Krafft, D., Cornwell, T., Dushay, N., Eckstrom, D., & Saylor, J. (2006). Metadata aggregation and "automated digital libraries": a retrospective on the NSDL experience. In *Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries* (pp. 230–239). New York, NY, USA: ACM. doi:10.1145/1141753.1141804

Lannom, L. (2000). Handle system overview. In *IFLA Conference Proceedings*. Jerusalem. Retrieved from http://archive.ifla.org/IV/ifla66/papers/032-82e.htm

Latif, A., Borst, T., & Tochtermann, K. (2014). Exposing data from an Open Access repository for Economics as linked data. *D-Lib Magazine, 20*(9/10), doi:10.1045/september2014-latif

Lave, J. (1988). *Cognition in Practice*. Cambridge, UK: Cambridge University Press.

Leach, P., Mealling, M., & Salz, R. (2005). A Universally Unique IDentifier (UUID) URN namespace. Internet Engineering Task Force (IETF). Retrieved from http://www.ietf.org/rfc/rfc4122.txt

LeBoeuf, P. (2005, May). *Identifying "textual works": ISTC: controversy and potential*. Presented at the FRBR in 21st Century Catalogues: An Invitational Workshop, Dublin, Ohio. Retrieved from http://www.oclc.org/research/activities/frbr/frbr-workshop/program.html

Lee, D. J., & Stvilia, B. (2012). Identifier schemas and research data. *Proceedings of the American Society for Information Science and Technology*, *49*(1), 1–4. doi:10.1002/meet.14504901311

Lee, D. J., & Stvilia, B. (2014). Developing a data identifier taxonomy. *Cataloging & Classification Quarterly*, *52*(3), 303–336. doi:10.1080/01639374.2014.880166

Leontiev, A. (1978). *Activity, consciousness, personality*. Englewood Cliffs, NJ: Prentice Hall.

Library of Congress. (2011). Premis Implementation Registry. Retrieved from http://www.loc.gov/standards/premis/registry/

Library of Congress. (2012a). Bibliographic Framework as a Web of data: Linked data model and supporting services. Retrieved from http://www.loc.gov/marc/transition/pdf/marcld-report-11-21-2012.pdf

Library of Congress. (2012b). PREMIS data dictionary for preservation metadata, version 2.2. Retrieved from http://www.loc.gov/standards/premis/v2/premis-2-2.pdf

Lombard, M., Snyder-Duch, J., & Bracken, C. C. (2002). Content Analysis in Mass Communication: Assessment and Reporting of Intercoder Reliability. *Human Communication Research*, *28*(4), 587–604. doi:10.1111/j.1468-2958.2002.tb00826.x

Lord, P., & Macdonald, A. (2003). *E-Scicence curation report: Data curation for e-Science in the UK: An audit to establish requirements for future curation and provision*. Bristol, UK: The JISC Committee for the Support of Research. Retrieved from http://www.jisc.ac.uk/uploaded_documents/e-ScienceReportFinal.pdf

Lord, P., Macdonald, A., Lyon, L., & Giaretta, D. (2004). From data deluge to data curation. *Proceedings of the UK e-Science All Hands Meeting*. Retrieved from http://www.ukoln.ac.uk/ukoln/staff/e.j.lyon/150.pdf

Lynch, C. (2003). *Institutional repositories: Essential infrastructure for scholarship in the digital age* (No. No. 226). Association of Research Libraries. Retrieved from http://www.arl.org/storage/documents/publications/arl-br-226.pdf

MacMullen, W. J., & Denn, S. O. (2005). Information problems in molecular biology and bioinformatics. *Journal of the American Society for Information Science and Technology*, *56*(5), 447–456. doi:10.1002/asi.20134

Markey, K., Rieh, S. Y., St.Jean, B., Kim, J., & Yakel, E. (2007). *Census of institutional repositories in the United States: MIRACLE project research findings* (No. CLIR pub 140). Washington, D.C.: Council on Library and Information Resources. Retrieved from http://www.clir.org/pubs/reports/pub140/reports/pub140/pub140.pdf

Masinter, L., Berners-Lee, T., & Fielding, R. T. (2005). Uniform Resource Identifier (URI): Generic Syntax. Retrieved June 24, 2013, from http://tools.ietf.org/html/rfc3986

Mason, J. (2002). *Qualitative researching* (2nd ed.). London, UK: Sage.

Mason, J. (2004). Semistructured interview. In M. Lewis-Beck & T. Liao (Eds.), *Encyclopedia of Social Science Research Methods Encyclopedia of social science research methods*. Thousand Oaks, CA: SAGE Publications, Inc.

Michener, W., Vieglais, D., Vision, T., Kunze, J., Cruse, P., & Janée, G. (2011). DataONE: Data Observation Network for Earth — Preserving Data and Enabling Innovation in the Biological and Environmental Sciences. *D-Lib Magazine*, *17*(1/2). doi:10.1045/january2011-michener

Nardi, B. A. (1996). *Studying context: A comparison of activity theory, situated action models, and distributed cognition.. Context and Consciousness : Activity Theory and Human-Computer Interaction.*

National Institutes of Health. (2010). NIH data sharing policy and implementation guidance (NIH Publication No. 03-05-2003). Retrieved from http://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm

National Science Foundation. (2010a). Grant proposal guide (gpg 11001). Retrieved from http://www.nsf.gov/publications/pub_summ.jsp?ods_key=gpg

National Science Foundation. (2010b, May). Scientists Seeking NSF Funding Will Soon Be Required to Submit Data Management Plans (NSF 10-077). Retrieved from http://www.nsf.gov/news/news_summ.jsp?cntn_id=116928

NCBI. (2012). Accession number prefixes: Where are the sequences from? Retrieved from http://www.ncbi.nlm.nih.gov/Sequin/acc.html

NISO. (2004). *Understanding metadata*. Bethesda, MD: NISO Press.

NISO. (2013). *Improving OpenURLs Through Analytics (IOTA): Recommendations for Link Resolver Providers* (No. NISO RP-21-2013). Retrieved from http://www.niso.org/apps/group_public/download.php/10811/RP-21-2013_IOTA.pdf

NISO/NFAIS. (2013). *Recommended practices for online supplemental journal article materials* (No. NISO RP-15-2013). Retrieved from http://www.niso.org/workrooms/supplemental

OCLC. (n.d.). PURL. Retrieved from http://tinyurl.com/yc6kbon

O’Leary, Z. (2005). *Researching real-world problems*. Sage Publications.

Office of Management and Budget. (1999). Uniform Administrative Requirements for Grants and Agreements With Institutions of Higher Education, Hospitals, and Other Non-Profit Organizations (OMB Circular 110). The White House. Retrieved from http://www.whitehouse.gov/omb/circulars_a110#36

Office of Science and Technology Policy. (2013). Expanding public access to the results of federally funded research | The White House. Retrieved June 25, 2013, from http://www.whitehouse.gov/blog/2013/02/22/expanding-public-access-results-federally-funded-research

OpenID Foundation. (2007). *OpenID authentication 2.0 - Final*. Retrieved from http://openid.net/specs/openid-authentication-2_0.html

OpenID Foundation. (2013). OpenID. Retrieved from http://openid.net/

ORCID. (n.d.). What is ORCID? Retrieved from http://about.orcid.org/about/what-is-orcid

Pabón, G., Gutiérrez, C., Fernández, J. D., & Martínez-Prieto, M. A. (2013). Linked Open Data technologies for publication of census microdata. *Journal of the American Society for Information Science and Technology*, *64*(9), 1802–1814. doi:10.1002/asi.22876

Palmer, C., & Knutson, E. (2004). Metadata practices and implications for federated collections. *In Proceedings of the 67th ASIS&T Annual Meeting*.

Palmer, C., Zavalina, O. L., & Fenlon, K. (2010). Beyond size and search: Building contextual mass in digital aggregations for scholarly use. Presented at the ASIST, Pittsburgh, PA.

Park, O. N. (2015). Development of linked data for archives in Korea. *D-Lib Magazine, 21*(3/4), doi:10.1045/march2015-park

Paskin, N. (2005). Digital Object Identifiers for Scientific Data. *Data Science Journal, 4*. Retrieved from http://www.doi.org/topics/041110CODATAarticleDOI.pdf

Paskin, N. (2008). Identifier interoperability. *Briefing Papers - Digital Preservation Europe*. Retrieved from http://www.digitalpreservationeurope.eu/publications/briefs/

Paskin, N. (2010). Digital Object Identifier (DOI®) System. In *Encyclopedia of Library and Information Sciences, Third Edition* (pp. 1586–1592). Taylor & Francis. Retrieved from http://www.tandfonline.com/doi/abs/10.1081/E-ELIS3-120044418

Patton, M. (1990). *Qualitative evaluation and research methods*. Newbury Park, CA: Sage.

Pepler, S., & O'Neil, K. (2008). *Preservation intent and collection identifiers* (No. CLADDIER Project Report II). Retrieved from http://epubs.cclrc.ac.uk/bitstream/2359/Report_II_PreservationIntentAndCompoundObjectIdentifiers-1.pdf

Personal Archives Accessible in Digital Media (Paradigm). (2008). Persistent identifiers - Archival resource key (ARK). Retrieved from http://www.paradigm.ac.uk/workbook/metadata/pids-ark.html

Pollard, T., & Wilkinson, J. (2010). Making Datasets Visible and Accessible: DataCite's first summer meeting. *Ariadne*, *64*. Retrieved from http://www.ariadne.ac.uk/issue64/datacite-2010-rpt

Pruitt, K. D., Tatusova, T., Klimke, W., & Maglott, D. R. (2009). NCBI Reference Sequences: current status, policy and new initiatives. *Nucleic Acids Research*, *37*(Database), D32–D36. doi:10.1093/nar/gkn721

Pruitt, Kim D., Tatusova, T., & Maglott, D. R. (2005). NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research*, *33*(Database Issue), D501–D504. doi:10.1093/nar/gki025

Qin, J., Ball, A., & Greenberg, J. (2012). Functional and architectural requirements for metadata: Supporting discovery and management of scientific data. In *Proceedings of International Conference on Dublin Core and Metadata Applications*. Kuching, Sarawak, Malaysis.

Qin, J., & D'ignazio, J. (2010). The Central Role of Metadata in a Science Data Literacy Course. *Journal of Library Metadata*, *10*(2-3), 188–204. http://doi.org/10.1080/19386389.2010.506379

Renear, A., Phillippe, C., Lawton, P., & Dubin, D. (2003). An XML document corresponds to which FRBR group 1 entity? In *Proceedings of Extreme Markup Languages 2003*. Montreal, Quebec. Retrieved from https://www.ideals.illinois.edu/handle/2142/11885

ResearcherID. (n.d.). What is researcherID? Retrieved from http://www.researcherid.com/Home.action?returnCode=ROUTER.Unauthorized&SrcApp=CR&Init=Yes

Rieger, O. (2007). Select for success: Key principles in assessing repository models. *D-Lib Magazine, 13*(7/8). Retrieved from http://www.dlib.org/dlib/july07/rieger/07rieger.html

Rieh, S. Y., Markey, K., St. Jean, B., Yakel, E., & Kim, J. (2007). Census of institutional repositories in the U.S.: A comparison across institutions at different stages of IR development.*D-Lib Magazine, 13*(11/12).

Rogers, Y. (2012). HCI Theory: Classical, Modern, and Contemporary. *Synthesis Lectures on Human-Centered Informatics, 5*(2), 1–129. doi:10.2200/S00418ED1V01Y201205HCI014

Schutt, R. (2009). *Investigating the social world*. Thousand Oaks, California: Sage.

Shafer, K., Weibel, S., Jul, E., & Fausey, J. (n.d.). Introduction to persistent uniform resource locators. Retrieved from http://purl.oclc.org/docs/long_intro.html

Shreeves, S., Knutson, E., Stvilia, B., Palmer, C., Twidale, M., Cole, T. (2005). Is quality metadata 'Shareable' metadata? The implications of local metadata practices for federated collections. In H.A. Thompson (Ed.) *Proceedings of the 12th National Conference of the Association of College and Research Libraries.* (pp. 223-237). Minneapolis, MN. Chicago, IL: Association of College and Research Libraries.

Simmhan, Y. L., Plale, B., & Gannon, D. (2005). A survey of data provenance in e-science. *SIGMOD Record*, *34*(3), 31–36. doi:10.1145/1084805.1084812

Strasser, C., Cook, R., Michener, W., & Budden, A. (2012). Primer on data management: What you always wanted to know. DataONE. Retrieved from https://www.dataone.org/sites/all/documents/DataONE_BP_Primer_020212.pdf

Stvilia, B. (2006). *Measuring information quality* (Ph.D.). University of Illinois at Urbana-Champaign, United States -- Illinois. Retrieved from http://search.proquest.com/docview/305328745/abstract?accountid=4840

Stvilia, B. (2007). A model for ontology quality evaluation. *First Monday, 12*(12). Retrieved from http://firstmonday.org/ojs/index.php/fm/article/view/2043

Stvilia, B., & Gasser, L. (2008). Value based metadata quality assessment. *Library & Information Science Research, 30*(1), 67-74.

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

Stvilia, B., Gasser, L., Twidale, M. B., & Smith, L. C. (2007). A framework for information quality assessment. *JASIST*, *58*, 1720–1733.

Stvilia, B., Hinnant, C. C., Schindler, K., Worrall, A., Burnett, G., Burnett, K., … Marty, P. F. (2011). Composition of scientific teams and publication productivity at a national science lab. *J. Am. Soc. Inf. Sci. Technol.*, *62*(2), 270–283. doi:10.1002/asi.21464

Stvilia, B., Hinnant, C. C., Wu, S., Worrall, A., Lee, D. J., Burnett, K., … Marty, P. F. (2015). Research Project Tasks, Data, and Perceptions of Data Quality in a Condensed Matter Physics Community. *Journal of the American Society for Information Science and Technology*. 66(2), 246-263.

Stvilia, B., Hinnant, C. C., Wu, S., Worrall, A., Lee, D. J., Burnett, K., … Marty, P. F. (2013). Studying the data practices of a scientific community. In *Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries* (pp. 425–426). New York, NY, USA: ACM. doi:10.1145/2467696.2467781

Stvilia, B., Mon, L., & Yi, Y. J. (2009). A model for online consumer health information quality. *J. Am. Soc. Inf. Sci. Technol.*, *60*(9), 1781–1791. doi:10.1002/asi.v60:9

Stvilia, B., Twidale, M., Smith, L.C., & Gasser, L. (2008). Information quality work organization in Wikipedia. *Journal of the American Society for Information Science and Technology, 59*(6), 983–1001.

Suchman, L. (1987). *Plans and Situated Actions*. Cambridge, UK: Cambridge University Press.

Swan, A., & Brown, S. (2008). *The skills, role and career structure of data scientists and curators: An assessment of current practice and future needs*. UK: JISC. Retrieved from http://www.jisc.ac.uk/media/documents/programmes/digitalrepositories/dataskillscareersfinalreport.pdf

Tenopir, C., Birch, B., & Allard, S. (2012). *Academic libraries and research data services* (ACRL White Paper). Association of College and Research Libraries. Retrieved from http://www.ala.org/acrl/sites/ala.org.acrl/files/content/publications/whitepapers/Tenopir_Birch_Allard.pdf

Thomson Reuters. (2012, June). Thomson Reuters unveils data citation index for discovering global data sets. Retrieved from http://thomsonreuters.com/content/press_room/science/686112

Thomson Reuters. (n.d.). Data citation index. Retrieved from http://wokinfo.com/media/pdf/dci_fs_en.pdf

Tonkin, E. (2008). Persistent identifiers: Considering the options. *ARIADNE*, *56*. Retrieved from http://www.ariadne.ac.uk/issue56/tonkin

Vitiello, G. (2004). Identifiers and Identification Systems. *D-Lib Magazine*, *10*(1). doi:10.1045/january2004-vitiello

Vrandečić, D., & Krötzsch, M. (2014). Wikidata: A free collaborative knowledgebase. *Communications of the ACM*, *57*(10), 78–85. http://doi.org/10.1145/2629489

Vygotsky, L. (1962). *Thought and Language*. Cambridge, MA, USA: MIT Press.

Wang, R., & Strong, D. (1996). Beyond accuracy: What data quality means to data consumers, *12*(4), 5–35.

Warner, S. (2010). Author identifiers in scholarly repositories. *arXiv:1003.1345*. Retrieved from http://arxiv.org/abs/1003.1345

Westell, M. (2006). Institutional repositories: proposed indicators of success. *Library Hi Tech*, *24*(2), 211–226. doi:10.1108/07378830610669583

Willis, C., Greenberg, J., & White, H. (2012). Analysis and synthesis of metadata goals for scientific data. *Journal of the American Society for Information Science and Technology*, *63*(8), 1505–1520. doi:10.1002/asi.22683

Wilson, T. D. (2006). A re-examination of information seeking behavior in the context of activity theory. *Information Research*, *11*(4).

Wilson, T. D. (2008). Activity theory and information seeking. *Annual Review of Information Science and Technology*, *42*(1), 119–161. doi:10.1002/aris.2008.1440420111

Witt, M. (2012). Co-designing, Co-developing, and Co-implementing an Institutional Data Repository Service. *Journal of Library Administration*, *52*(2), 172–188. doi:10.1080/01930826.2012.655607

Witt, M., & Cragin, M. (2008). Introduction to Institutional Data Repositories Workshop. *Libraries Research Publications*. Retrieved from http://docs.lib.purdue.edu/lib_research/83

World Wide Web Consortium (W3C). (2001). URIs, URLs, and URNs: Clarifications and recommendations 1.0. Retrieved from http://www.w3.org/TR/uri-clarification/

World Wide Web Consortium (W3C). (2004). Resource Description Framework (RDF): Concepts and abstract syntax. Retrieved from http://www.w3.org/TR/rdf-concepts/

World Wide Web Consortium (W3C). (2013a). Open Annotation Data Model. Retrieved from http://www.w3.org/ns/oa

World Wide Web Consortium (W3C). (2013b). PROV model primer. Retrieved from http://www.w3.org/TR/prov-primer/

Wu, S. (2014). *Exploring the Data Work Organization of the Gene Ontology* (Doctoral dissertation). Retrieved from http://diginole.lib.fsu.edu/etd/9267/

Wu, S., Stvilia, B., & Lee, D. J. (2012). Authority Control for Scientific Data: The Case of Molecular Biology. *Journal of Library Metadata*, *12*(2-3), 61–82. doi:10.1080/19386389.2012.699822

Wynholds, L. (2011). Linking to Scientific Data: Identity Problems of Unruly and Poorly Bounded Digital Objects. *International Journal of Digital Curation*, *6*(1). doi:10.2218/ijdc.v6i1.183

BIOGRAPHICAL SKETCH

Dong Joon Lee is a doctoral candidate in the School of Information at the Florida State University (FSU). I have a MS in Information Management (MSIM) from University of Washington, and a BS in Computer Information Systems from Grove City College, Grove City, Pennsylvania.

Dong Joon's areas of research interests are data and information management, and his current research focus is on data curation practices in institutional repositories, development of a model for data identifier evaluation, metadata for entity-type identifiers.