PRINCIPLE PARADIGMS
REVISITING THE DUBLIN CORE *1:1 PRINCIPLE*

BY

RICHARD J. URBAN

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Library and Information Science
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2012

Urbana, Illinois

Doctoral Committee:

Professor Michael B. Twidale, Chair
Professor Allen H. Renear, Director of Research
Professor Carole L. Palmer
Associate Professor Jonathan Furner, University of California at Los Angeles

Abstract

The Dublin Core *1:1 Principle* asserts that “related but conceptually different entities, for example a painting and a digital image of the painting, are described by separate metadata records” (Woodley et al., 2005). While this seems to be a simple requirement, studies of metadata quality have found that cultural heritage metadata frequently does not conform to the *Principle.* Instead, representations commonly appear to make statements about multiple related resources, such as a painting and a digital surrogate that depicts the painting. Although these “violations” of the *Principle* are assumed to reduce metadata quality, they are widespread in cultural heritage metadata, and metadata creators indicate “a great deal of confusion” about what the *Principle* means and what constitutes a violation (Park & Childress, 2009).

A conceptual analysis of the *1:1 Principle* reveals that it is the product of an encounter between two different paradigms, with distinct approaches to how descriptions function, that have dominated the development of Dublin Core. The *knowledge organization (KO) paradigm* draws from a century of practice developing bibliographic representations and rules for description in libraries, archives, and museums (LAMs). This paradigm is primarily concerned with the organization and classification of document surrogates. It does not provide a formal account of how bibliographic languages describe and reference the resources they represent. In contrast, the *knowledge representation (KR) paradigm* is influenced by recent computer science, linguistics, and philosophy. This paradigm is primarily concerned with supporting automatic inference and data integration, mobilizing formal semantic theories to provide descriptions with grammatical structures that can be computationally modeled. It provides an explicit and formal, although not untroubled, account of reference and description.

Further analysis of how discussion of the *1:1 Principle* has been shaped by these two different approaches to description shows that *1:1 Principle* problems are as much about the design and

function of representation languages as they are about errors made by metadata creators. The *Principle* is most directly derived from the formal theories of description and reference that use proper names (or identifiers like URIs) to refer directly to resources.

Bibliographic records with fixed syntaxes, but informal, colloquial semantics, can successfully communicate descriptions that are meaningful to human interpreters and enable syntactic interoperability between systems. However, when viewed through the lens of formal semantics, bibliographic representations may appear incoherent—as failing to unambiguously reference any resource, or provide a single, shared semantic interpretation of descriptions. These problems are exacerbated by the need to make subtle ontological choices about descriptions of resources involved in equivalent, derivative, or descriptive relationships.

Because knowledge organization representations rely on relatively informal semantics, often not more than a sentence or two of natural-language prose, the heuristics for identifying *1:1 Principle* violations depend heavily on the implicit and informal, shared conceptual framework of cultural heritage professionals. A formal interpretation of these heuristics, therefore, requires more than operational definitions of the *Principle*; it requires highly expressive ontologies that make that understanding logically explicit. Moreover, even formalizing traditional ontological distinctions will fail to identify violations when contingent historical facts play a role in interpretations of metadata records.

While these difficulties make fully reliable methods for automatic detection of *1:1 Principle* violations impossible in principle, they usefully reveal obstacles that will require attention as the cultural heritage community moves towards adopting more formal representation practices that are the basis of the Semantic Web and Linked Data. Recognition of the fundamental differences between the knowledge organization paradigm and the knowledge representation paradigm can lay the groundwork for reconceptualizing how we represent related resources.

In Memoriam

Antoinette (DelConte) Radomile

Geno A. Radomile

Albert Radomile

Andrew Urban

Helen (Grace) Urban

Acknowledgements

This dissertation represents the culmination of a long journey into cultural heritage metadata, one that began when I joined the Colorado Digitization Program in 2001. It was while working on the CDP's *Dublin Core Metadata Best Practices* that I first ran afoul of the Dublin Core *1:1 Principle* and its host of problems. I owe a great debt of thanks to Liz Bishoff, Nancy Allen, Brenda Bailey-Hainer, Betty Meagher, and the many members of the CDP's Metadata Working Group for introducing me to the fascinating world of metadata and bibliographic description.

This dissertation would not have been possible without the ongoing financial support and leadership of the Institute of Museum and Library Services (IMLS). Their support and encouragement for the Collaborative Digitization Program, the IMLS Digital Collections and Content project (LG-06-07-0020), and other forms of collaboration between libraries, archives, and museums, has been an essential component of my development. Thanks to Joyce Ray, Chuck Thomas, and the many staff members and volunteer reviewers who stand behind their work.

Thanks to the GSLIS community for nurturing me through the process of MSLIS and PhD. Dean John Unsworth's guidance and support provided the opportunity to "see just how deep the rabbit hole goes." As an open and welcoming forum for discussion, the GSLIS Metadata Roundtable allowed me to explore my interests in metadata for cultural heritage materials. The Eclectic Design Research Group afforded opportunities to make creative leaps through its playful explorations. My writing benefited greatly from Dr. David Dubin's patient reviews in the Research Writers Group. But in these closing days of writing, perhaps the biggest contribution to this dissertation came from Karen Wickett, the Conceptual Foundations Group (CFG), and the Collection/Item Metadata Relationships (CIMR) working group. Here my eyes were opened to a different way of seeing my world.

Thanks to Dr. Jonathan Furner for offering his good sense on the philosophical problems

of cultural heritage materials. Dr. Carole L. Palmer, the Center for Informatics Research in Science and Scholarship (CIRSS), and the staff of the IMLS Digital Collections and Content project provided both the financial support and the raw materials for this inquiry. Dr. Michael B. Twidale has been a constant source of creative inspiration and encouragement to explore my interests in museum informatics. In addition, Dr. Allen H. Renear deserves special thanks for his many hours of dedicated work he devoted to helping me make sense of a very messy world.

Thanks to Rob “Eagle Eye” Lancefield, who patiently suffered my typos and grammatical errors for many years while we were colleagues in the Museum Computer Network. I could think of no one better to give this major project a final polish.

Thanks to Sunah Suh for opening your home to those of us who need a halfway house and to Mikki Smith for offering a harbor of refuge in stormy seas. Sarah Shreeves has been a constant colleague and friend who has seen me through many danger zones during my time at GSLIS. Thank you for your help in keeping my feet on the ground.

As with all of my accomplishments, I could not have done this work without the love and support of my family.

Table of Contents

List of Abbreviations

AACR: Anglo-American Cataloging Rules (1 or 2 indicates edition)

AAT: Art & Architecture Thesaurus

AHDS: Arts & Humanities Data Service

AI: Artificial Intelligence

CDWA: Categories for the Description of Works of Art

CIMI: Computer Interchange of Museum Information

DCMI: Dublin Core Metadata Initiative

DCAM: Dublin Core Abstract Model

DCAP: Dublin Core Application Profiles

DC(M)ES: Dublin Core (Metadata) Element Set

DLG: Directed Labeled Graphs

DTD: Document Type Definition

EDM: Europeana Data Model

FRBR: Functional Requirements for Bibliographic Records

HTML: HyperText Markup Language

IMLS DCC: IMLS Digital Collections and Content project

ISBD: International Standard for Bibliographic Description

KO: Knowledge organization

KR: Knowledge representation

LAM: Libraries, archives, and museums

LOD: Linked Open Data

MARC: MAchine Readable Cataloging

MCF: Meta Content Framework

OAI-PMH: Open Archives Initiative Protocol for Metadata Harvesting

OAI-ORE: Open Archives Initiative Object Reuse and Exchange

OCLC: Online Computer Library Center

OWL: Web Ontology Language

PICS: Platform for Internet Content Selection

RLG: Research Libraries Group

RDF: Resource Description Framework

RDFS: RDF Schema

RLIN: Research Libraries Information Network

SGML: Standard Generalized Markup Language

TGM: Thesaurus for Graphic Materials

URI: Uniform Resource Identifier

URL: Uniform Resource Locator

VRA: Visual Resources Association

W3C: World Wide Web Consortium

XML: eXtensible Markup Language

XMLWC: XML Web Collections

Conventions

Typographic Conventions

This thesis includes a discussion about several different metadata standards represented using a variety of syntaxes. Examples of full records are included in Appendix A, while in-line examples are provided using the `teletype` typeface.

Language Conventions

The *1:1 Principle*

References to the Dublin Core *1:1 Principle* take many forms as they appear in various Dublin Core Metadata Initiative documents, listservs, and independent publications. These include "One-to-one Principle,""one to one principle," "1:1 Principle," etc. For consistency I have adopted *1:1 Principle,* except when a different form appears in direct quotations.

Chapter 1

Introduction

1.1 Introduction

As Operations Coordinator for the Colorado Digitization Program (CDP),[1] I was introduced to the complexities of creating metadata for cultural heritage resources. When I became involved in large-scale digitization efforts, Dublin Core was emerging as an essential piece of infrastructure for recording, storing, and exchanging information from libraries, archives, and museums (LAMs). As one of the later editors of the CDP *Dublin Core Metadata Best Practices (DCMBP)* (Collaborative Digitization Program, 2006), I had a front-row seat in the deliberations of CDP's Metadata Working Group. This group of experienced professionals sought to provide guidance for creating good descriptions of cultural heritage beyond the simple definitions provided by the Dublin Core Metadata Initiative (DCMI). Ultimately, CDP's recommendations closely conformed to existing library and museum rules of description, especially rules for describing *reproductions.* In order to make it clear that some information was about *original* resources and some information was about *digital* reproductions, CDP extended the basic Dublin Core element set to include properties such as `date.original` or `format.digital` (Bishoff & Garrison, 2000; Cronin, 2008). The distinction between information about "digital" vs. "original" things was also preserved in our local data model, which stored this information in separate relational tables. Members of the working group saw these recommendations as a pragmatic and efficient solution to the problem at hand. As an early set of guidelines, the *DCMBP* served as a model for other regional and local practices and was widely promoted by the Institute of Museum and Library Services and in other high-level guidelines.[2] The impact of the CDP *DCMBP* continues through various adaptations and current

[1] Later the Collaborative Digitization Program.

[2] Such as the *NINCH Guide to Good Practice* (NINCH, 2002) and *A Framework of Guidance for Building Good Digital Collections* (National Information Standards Organization, 2007).

practices in many cultural heritage organizations in the United States.

While these recommendations worked well within the CDP environment, subsequent studies of cultural heritage metadata quality raised concerns that this kind of approach results in records that violate the Dublin Core *1:1 Principle:*

> In general, Dublin Core metadata describes one manifestation or version of a resource, rather than assuming that manifestations stand in for one another. For instance, a jpeg image of the *Mona Lisa* has much in common with the original painting, but it is not the same as the painting. As such the digital image should be described as itself, most likely with the creator of the digital image included as a Creator or Contributor, rather than just the painter of the original *Mona Lisa.* The relationship between the metadata for the original and the reproduction is part of the metadata description, and assists the user in determining whether he or she needs to go to the Louvre for the original, or whether his/her need can be met by a reproduction. (Hillmann, 2003)

These studies (Stvilia et al., 2004; Hutt & Riley, 2005; Shreeves et al., 2005, 2006; Han et al., 2009) have found that cultural heritage metadata records often lack coherence because they describe features of both physical and digital manifestations of resources. This lack of coherence creates unwanted barriers for integrating records from multiple institutions into large-scale aggregations and inhibits the kinds of search and retrieval services that aggregators can provide. Overall, such records are considered to be lower in quality than is desirable for robust digital library services and broad interoperability.

This body of research accepts the Dublin Core *1:1 Principle* as a necessary requirement for quality Dublin Core metadata. These analyses are based on fundamental assumptions about what the *1:1 Principle* applies to—for example, the assumption that it is *records* which must be about one, and only one, thing. Rarely do these studies attempt to provide a systematic analysis of the *Principle* itself, nor do they provide a common interpretation of what counts as a "violation." Park & Childress (2009) also demonstrate that there is "a great deal of confusion" about what the *Principle* means and how it should be implemented by practitioners tasked with conforming to it.

The objective of this dissertation is to provide the clarity that is missing from current research and guidance for metadata practitioners about the Dublin Core *1:1 Principle.* In what follows, we will explore the fundamental concepts that Hillmann (2003) tries to communicate and how these concepts do or do not align with accepted cataloging practices within the cultural heritage domain. Unpacking the *1:1 Principle* reveals that the problems in this area arise not simply from the behavior of metadata creators, but more fundamentally from the nature of how *descriptions* function as meaningful linguistic constructs. Probing our confusions about the *1:1 Principle* suggests two distinct paradigms for solving the problem of description. By outlining the contours of each paradigm, we can improve our understanding of the apparently rampant violations observed by studies of metadata quality. While the *1:1 Principle* may seem like a minor and insignificant problem, this study demonstrates that it is extremely important to understand the underlying paradigms as the cultural heritage community moves towards adopting Semantic Web and Linked Data approaches to representation and description.

1.2 Approach

The study that follows is carried out as a combination of historical and conceptual analysis. Viewed abstractly, it involves first identifying publications and documents that represent the evolving attitudes, ideas, and practices of key communities during the relevant periods of time. It closely analyzes portions of documents that use or discuss the concepts of interest. This analysis focuses not solely on essential logical features, but also on the objectives and influences that have driven the adoption of particular descriptive frameworks. This is partly to conjecture likely patterns of influence and intention, but also to present the historical context needed for insight into, and interpretation of, the Dublin Core *1:1 Principle* in the face of apparent incoherence.

This study is primarily weighted towards conceptual analysis, although it does rely on historical reference materials. Among these are evolving versions of Dublin Core standards documents (Hillmann, 2003; Powell & Johnston, 2003; Woodley et al., 2005; Powell et al., 2007; Nilsson et al., 2009), DCMI community/listserv archives (Bearman, 1997; Dublin Core Metadata Initiative, 2000a,c,b, 2012a,b), and contemporary publications. The objective is not an account of historical events, but rather an analysis of how some terminology and concepts have been understood, what the liabilities of those understandings are, and how a clearer, more consistent understanding might be evolved in order to better support information and access in the 21st century.

Conceptual analysis itself is not at all uncommon as a research method in information science, even though it is not always identified as such. It appears as the dominant method, at least in its information version, in many of the great classics of library and information science (LIS) literature. These include works by Patrick Wilson (1968), Michael Buckland (1991, 1997), Biger Hjørland (1998, 2003, 2005, 2007), Richard Smiraglia (2001), and Elaine Svenonius (2000, 2004). In an article intended to demonstrate the value of this approach for information studies, Jonathan Furner (2006) characterizes conceptual analysis in this way:

> Conceptual analysis is a technique that treats concepts as classes of objects, events, properties, or relationships. The technique involves precisely defining the meaning of a given concept by identifying and specifying the *conditions* under which any entity or phenomenon is (or could be) classified under the concept in question. The goal in using conceptual analysis as a method of inquiry into a given field of interest is to improve our understanding of the ways in which particular concepts are (or could be) used for communicating ideas about that field.

My use of conceptual analysis is contextually situated. We are concerned to give an account of the *1:1 Principle* as a development influenced by historical trajectories of the LAM/LIS community and of computer scientists interested in knowledge-representation problems. Understanding the connection between these two paradigms provides insight into the various understandings of the *1:1 Principle* and an explanation of the role that they play in the development and evolution of metadata practice and theory. The Dublin Core *1:1 Principle*'s conceptualization of relationships between resources and metadata is at the center of this analysis. However, understanding this concept quickly leads us to explore related questions about what metadata *records* are, what it means for them to *describe* resources, and how *violations* of the *Principle* are conceptualized.

Somewhat surprisingly, there is little prior analytical or historical work on the *1:1 Principle* itself. There are, of course, many references to the *Principle* in the metadata literature; however, these are not sustained analytical treatments, but rather informal explanations that are part of recommending practice (for example, (Caplan, 2003; Hillmann & Westbrooks, 2004; Zeng & Qin, 2008)), or relatively brief accounts that reveal difficulties, confusion, or irritation (Shreeves et al., 2005; Hutt & Riley, 2005; Park & Childress, 2009; Park, 2009; Miller, 2010). Here, this literature functions as a source of evidence, rather than as prior work. As such, it is taken up later in the course of developing my analysis. Similarly, empirical studies that attempt to identify violations are also sources of evidence for us and are discussed later as appropriate (Shreeves et al., 2005; Hutt & Riley, 2005).

1.3 Organization of this Study

1.3.1 Origins of the Dublin Core *1:1 Principle*

Chapter 2 explores a fundamental question raised by the presence of the *1:1 Principle* in DCMI documentation: Why did the Dublin Core community find it necessary to articulate the *1:1 Principle* requiring that "related but conceptually different entities, for example a painting and a digital image of the painting, are described by separate metadata records" (Woodley et al., 2005)? A review of the literature reveals that libraries, archives, and museums have, in fact, been engaged in a long struggle to define best practices for representing the relationships between originals, reproductions, and multiple versions of resources within knowledge organization systems—something that has "proven elusive through all the cataloging codes of the twentieth century" (Knowlton, 2009). Problems and solutions that emerged with the introduction of analog reproduction technologies—such as microforms, photography, xerography, etc.—to LAM collections extended into the rapidly expanding digital environment of the World Wide Web (WWW). Although these approaches conformed with accepted knowledge organization (KO) practices, within the Dublin Core community they confronted alternative approaches to knowledge representation (KR) that emerged from the computer science and artificial intelligence communities.

1.3.2 Representation Paradigms

Recognizing that the *1:1 Principle* is situated at the intersection of distinct conceptual domains complicates the objective of providing a clear and coherent operational definition as originally proposed by Urban (2009). Instead, in Chapter 3 I identify and describe the two paradigms that frame interpretations of the *1:1 Principle:*

knowledge organization paradigm: encompasses traditional library, archive, and museum cataloging practices based in information retrieval of surrogate representations constructed with formal syntaxes, but informal semantics.

knowledge representation paradigm: emerged from the needs of the artificial intelligence community, which needed representation languages built on formal semantics to enable computational intelligent reasoning.

I describe how each paradigm approaches the construction of descriptive representations differently, each according to its own objectives, fundamental principles, and underlying ontological commitments. The different views through these lenses inhibit the articulation of a single, coherent conceptual definition of the *1:1 Principle*. Articulating the features of each paradigm lays a foundation for understanding both the confusion about the *1:1 Principle* among metadata creators and the apparent prevalence of violations.

1.3.3 Description and Reference for Cultural Heritage Resources

What does it mean for metadata to be about one, and only one, resource? Although this injunction appears throughout Dublin Core documentation, in the form of the *1:1 Principle* and more strongly in the formal *Dublin Core Abstract Model* (Powell et al., 2007), the motivation for this requirement is not articulated within DCMI. Building on our analysis of representation paradigms in Chapter 3, Chapter 4 demonstrates that the use of descriptions that uniquely refer to a resource is a core concept of knowledge representation languages, such as the *Resource Description Framework*. In contrast, the knowledge organization paradigm uses colloquial bibliographic languages. These allow the construction of representations that refer to multiple resources or require pragmatic contextual clues supplied by human readers for successful interpretation. I articulate how these distinctions between formal and informal approaches to description and reference help explain the observed *1:1 Principle* problems in practice.

1.3.4 Revisiting *1:1 Principle* Violations

In Chapters 3 and 4, we demonstrated that there is not a single coherent definition of the *1:1 Principle* because it can be understood differently from the perspective of the knowledge organization and knowledge representation paradigms. Chapter 5 analyzes the heuristics used by metadata quality studies that identify *1:1 Principle* violations through the lens of each paradigm. I find that, while useful for certain kinds of quantitative and qualitative analysis, these heuristics are difficult to formalize into knowledge representation axioms that reliably supply interpretations of colloquial metadata records.

1.3.5 Contributions & Future Research

This research reveals that the *1:1 Principle,* although appearing to be a simple rule with some problematic definitions, actually has a rich and complex background. Although this work began with the objective of supplying a single coherent operational definition of the *Principle,* it instead revealed the existence of two distinct paradigms for representation languages. Because of the different objectives and fundamental assumptions of each paradigm, multiple interpretations of the *Principle* are in operation. While each may be valid from one perspective, considered together they raise important questions about how research on the *Principle* should proceed. Because violations appear to depend as much on the nature of representation languages as on metadata creators' behavior, new solutions are needed to move us beyond merely detecting violations. To solve the problem of ambiguous metadata we need new ways to represent cultural heritage information. Knowledge representation languages are one possible solution; but to take full advantage of their power, cultural heritage professionals must understand the formal models on which they are based. Preparing cultural heritage professionals to take advantage of the Semantic Web and Linked Data languages presents challenges for educators and for interface developers. The first step is to recognize the fundamental paradigm shift that these languages represent.

Chapter 2

Origins of the Dublin Core *1:1 Principle*

> **The 1:1 Principle**: In general, Dublin Core metadata describes one manifestation or version of a resource, rather than assuming that manifestations stand in for one another. For instance, a jpeg image of the *Mona Lisa* has much in common with the original painting, but it is not the same as the painting. As such the digital image should be described as itself, most likely with the creator of the digital image included as a Creator or Contributor, rather than just the painter of the original Mona Lisa. The relationship between the metadata for the original and the reproduction is part of the metadata description, and assists the user in determining whether he or she needs to go to the Louvre for the original, or whether his/her need can be met by a reproduction.
> (Hillmann, 2003)

> "One to one" is a many-headed snake, and it has bitten us often over the years.
> (Weibel, 2010)

2.1 Introduction

The Dublin Core Metadata Initiative standard is a cornerstone for constructing representations about cultural heritage resources. Built on a foundation of simple principles, Dublin Core is intended to be a *lingua franca* that facilitates interoperable exchange of information (Baker, 2000; Hillmann, 2003). An important concept for the Dublin Core language is the *1:1 Principle,* "whereby related but conceptually different entities... are described by separate metadata records" (Woodley et al., 2005). Yet, according to Park & Childress (2009), the metadata creators indicate "a great deal of confusion" about what the *1:1 Principle* is and how it should be applied in the metadata creation processes. Hutt & Riley (2005), Shreeves et al. (2005), Han et al. (2009), and Miller

(2010) suggest that the *Principle* is routinely ignored or willfully violated in practice, especially for cultural heritage collections where representations describe both analog and digital resources. It is puzzling that these violations occur despite recommendations from metadata textbooks (Caplan, 2003; Hillmann & Westbrooks, 2004; Zeng & Qin, 2008), best–practice guidelines (Moen, 1998; Digital Library Federation & National Science Digital Library, 2007), and original DCMI documentation (Hillmann, 2003; Woodley et al., 2005; Powell et al., 2007). Publications such as Han et al. (2009) and Miller (2010) point towards limitations of metadata management systems, and as Hutt & Riley (2005) observe:

> Representing complexities in the OAI DC environment obviously presents a challenge. . . . This leaves data providers with two choices. Create records that adhere to the 1:1 rule and omit pertinent information, or violate the rule. We observed many cases in which data providers chose to violate the rule and combine data about the original intellectual object as well as its digital manifestation.

What remains unclear from these studies and Dublin Core documentation itself, however, is why the *1:1 Principle* is needed at all. Why is it necessary that Dublin Core metadata is about one, and only one, resource? Furthermore, the many accounts of the *1:1 Principle* present a confusing picture of which syntactic and semantic structures defined by Dublin Core should be about one, and only one, resource (i.e., "a metadata?" "a record?" "a description?"). In order to sort through these confusions, this chapter presents an account of the *1:1 Principle*'s origins. My account draws from knowledge organization (KO) literature about rules for cataloging related resources such as reproductions, multiple versions, and surrogates for museum objects (e.g., a painting and a photograph of the painting). I find that within the Dublin Core Metadata Initiative these traditional practices encounter emerging methods for specifying the formal semantics of knowledge representation (KR) languages. The *1:1 Principle* emerges to communicate a fundamental feature of KR languages in terms that address the concerns of the KO and cultural heritage community. Understanding that the *1:1 Principle* is thus situated between two paradigms for constructing representation languages will lead us to explore each in turn in later chapters. The background developments that resulted in the articulation of the *1:1 Principle* will also lead us to revisit our understanding of what it means to violate it.

2.2 Description on the Web: Emergence of the Dublin Core Metadata Initiative

The Dublin Core Metadata Initiative (DCMI) began as the result of informal conversation at the 2nd International World Wide Web Conference in October 1994.[1] Struggling with how to address the explosion of resources on the World Wide Web led the National Center for Supercomputing Applications (NCSA) at the University of Illinois and the Online Computer Library Center (OCLC) to sponsor a Metadata Meeting in March 1995 (Weibel, 1995, 2010). The meeting's location at OCLC's headquarters in Dublin, Ohio provided the name for a new lightweight core vocabulary for the description of networked "document-like objects" (Weibel, 1995). This vocabulary would enable document creators to describe their own works without having to use a complex set of cataloging rules like the *Anglo-American Cataloging Rules (AACR2)* (American Library Association et al., 1988). Instead, DCMI adopted several "principles" to guide its development (Weibel, 1995):

1. Intrinsicality: Dublin Core describes the properties of an "item at hand"

2. Extensibility: Dublin Core may be adapted for specialized uses by adding new intrinsic properties

3. Syntax Independence: Dublin Core does not specify a syntax; rather, an implementer may select a format appropriate for his/her environment.

At the same time that Dublin Core was getting off the ground, cultural heritage agencies such as the Research Libraries Group (RLG), the Getty Information Institute (GII), and the Arts and Humanities Data Service (AHDS) engaged in multiple projects exploring online access to their members' materials—especially to reformatted special collections and archival materials (Erway, 1996; Miller & Greenstein, 1997; Fink, 1999). These organizations advocated for cultural organizations in the discussions about Dublin Core's future and tested DC's statement of principles. First, RLG challenged the notion that Dublin Core was only to be used for online documents, pointing out that many repositories wished to publish metadata about offline physical materials (Erway, 1996). Second, RLG became a strong proponent of extending DC's definition of "document-like

[1] Table 2.1 provides a quick roadmap to developments related to the *1:1 Principle*.

objects" to include images. Initial conversations began at the 3rd Dublin Core "Workshop on Metadata for Networked Images" hosted by the Center for Networked Information (CNI) and OCLC (Cromwell-Kessler & Erway, 1997; Sutton & Miller, 1997).[2] But this meeting left an important issue unresolved—*how should Dublin Core records represent digital images that served as surrogates for offline physical collections?* (Cromwell-Kessler & Erway, 1997; Sutton & Miller, 1997). A subsequent Metadata Summit meeting sponsored by the Association of Library Collections and Technical Services (ALCTS) and RLG developed specific recommendations for future Dublin Core refinements (Research Libraries Group, 1997b). Among the participants' concerns was the distinction between non-networked resources and online surrogate representations:

> The Dublin Core elements were designed to describe Web documents. Consequently, they do not distinguish a relationship between object and source since the definitions of specific Dublin Core elements pertain to the "item" only as it is represented in its electronic form. The Dublin Core elements DATE, PUBLISHER and SOURCE are especially problematic. The frequently suggested solution of using the SOURCE element for all information about the original item is insufficient to resolve the confusion effectively. Lumping all information about the original into SOURCE will not be helpful to researchers when they are concerned about the "thing" rather than with the Web page describing it. (Research Libraries Group, 1997b)

The major outcome of the Task Force on Meta Access workshop was a proposal to extend Dublin Core in a way that would address this concern:

> **II. Basic principle and related definitions:**
>
> In the strategic Web application of the Dublin Core elements, for the most part, it is assumed the documents are Internet-accessible original documents. The basic distinction between the strategic application and the generic application under discussion is that in the generic application it is important to indicate:
>
> 1. if a document is the original or not
> 2. if a document is Internet-accessible or not

[2] Also known as The Center for Networked Information (CNI)/OCLC Image Metadata Workshop.

> The nature of the document is defined here as either:
> “Original” is used to mean the first manifestation, e.g., a journal published only on the web or a painting in a museum.
> “Surrogate” is used to mean a version that stands for an original – not necessarily an exact reproduction. A copy on microfilm is an example of a surrogate. This use of the word includes lesser versions (e.g., thumbnail images) and reformatted versions (e.g., a digital audio version of an analog recording), but not a part of the whole (e.g., a detail from a photograph). (Research Libraries Group, 1997a)

Although Dublin Core did not specify a standard syntax, RLG’s proposal also recommended adding a mechanism to “indicate when neither an original nor a surrogate is Internet-accessible” or “indicate when only a surrogate is being described,” in order to allow indexing systems to include or ignore different *record types* (Research Libraries Group, 1997a). The Visual Arts Data Service (VADS) developed a similar set of recommendations that also included Dublin Core refinements to describe whether a reproduction was analog or digital (Miller & Greenstein, 1997).

The RLG proposal became a central point of discussion at the 1997 DC-4 Workshop in Helsinki, Finland. Rather than adopt the proposed changes in the RLG *Guidelines* (Research Libraries Group, 1997a), workshop participants discussed the relationship between “logical clusters of metadata. . . that reference one, and only one, state of the information resource,” an approach that became the nucleus of the *1:1 Principle* (Weibel & Hakala, 1998; Bearman et al., 1999). The DCMI community’s response to RLG’s proposal was grounded in work on more formal data models for Dublin Core that began at the 1996 DC-2 meeting in Warwick, England (Dempsey & Weibel, 1996; Weibel & Hakala, 1998). Under the *Warwick Framework,* information could be exchanged through the use of a container architecture that allowed for the representation of discrete metadata packages (Lagoze, 1996). In part, the Warwick Framework’s architecture would allow packages to use different standards (e.g., a Dublin Core package and a MARC package) that may semantically overlap. However, it could also be used to address the problem of packaging descriptions about an original and its surrogate reproductions. Discussing this problem of these relationships at the 1997 DC-4 meeting, Erik Jul paraphrased Ranganathan’s laws of library science: *“to each resource, its [own] description”* (Weibel, 2010).

The response to RLG's proposal also represented a tension within the DCMI community between *minimalists,* who argued that Dublin Core should provide a simple set of descriptive terms, and *structuralists,* who wished to allow Dublin Core to be extended to meet the needs of each particular application domain (for example, by adding new elements for the cultural heritage community or by allowing semantic reinterpretation of existing elements) (Weibel et al., 1997). Further definition of the *Principle* was delegated to the One-to-One Working Group and the Relation Working Group, because the `relation` element was seen as a "logical means" of linking distinct descriptions of originals and surrogates (Weibel & Hakala, 1998).

2.2.1 Discussion

From this overview, we find that DCMI needed to articulate the *1:1 Principle* in part to bridge the gap between structuralists and minimalists within the community. For the cultural heritage community, the simplicity of the Dublin Core vocabulary did not allow for the description of related resources. Between the first discussions about Dublin Core in 1995 and the appearance of the *1:1 Principle* in Hillmann (2003), the cultural heritage community organized multiple workshops and working groups to discuss and inform the development of the Dublin Core standard, as is shown in Table 2.1. Their solution to this problem was to add additional elements to the standard to enable users to make more explicit the kind of object being described. For minimalists, these kinds of extensions only added undesired complexity and inhibited Dublin Core's interoperability goal. The minimalist solution to the problem could be found in the emerging container architecture of the Warwick Framework, where discrete, but related, sets of statements could be organized. This overview, therefore, explains how an explicit articulation of the *1:1 Principle* came to be a part of the Dublin Core suite of standards. However, the treatment of minimalists and structuralists as pragmatic camps within DCMI (Weibel et al., 1997) lacks theoretical depth. In order to understand and assess the *1:1* debate, the following sections explore the precursors to LAM approaches for describing related resources (originals, reproductions, surrogates, etc.) and the computational methods that necessitated simple, logical clusters of metadata.

Date	Workshops	1:1 Events
1995	DC-1 (Dublin, OH, USA)	Participants in DC-1 leave unresolved how Dublin Core should handle different versions of the same resource. (Weibel et al., 2000)
1996	DC-2 (Warwick, UK)	Warwick Framework. (Lagoze, 1996) Participants in DC-2 leave unresolved how Dublin Core should handle the representation of surrogates. (Cromwell-Kessler & Erway, 1997; Sutton & Miller, 1997)
	DC-3 CNI/OCLC Image Workshop (Dublin, OH, US)	General relationships among resources discussed. (Sutton & Miller, 1997)
1997	DC-4 (Canberra, AUS)	Eric Jul introduces 1:1 Concept. (Weibel, 2010)
	RLG/ALCTS Metadata Summit (Mountain View, CA)	
	DC-5 (Helsinki, FIN)	RLG presents proposal for how to handle originals and reproductions. (Research Libraries Group, 1997a) *1:1 Principle* and Relation Working Groups established. (Weibel & Hakala, 1998) AHDS Report. (Miller & Greenstein, 1997).
1998	DC-6 (Washington, D.C., USA)	Discussions about the relationship between originals/reproductions continues on the *dc-one2one* community listserv. (Dublin Core Metadata Initiative, 2000a) RDF emerges from earlier metadata proposals. (Miller, 1998)
1999	DC-7 (Frankfurt, DEU)	Bearman et al. (1999) discusses “logical clusters of metadata” used for the INDECS project.
	Santa Fe Convention of the Open Archives Initiative	
2000	DC-8 (Ottawa, CAN)	Relation and One-to-One Working Groups disbanded. (Dublin Core Metadata Initiative, 2000a)
2001	DC-9 (Tokyo, JPN)	OAI-PMH 1.1 requires simple Dublin Core. (Lagoze & Van de Sompel, 2001)
2003	DC-2003 (Seattle, WA, USA)	*1:1 Principle* first appears in *Using Dublin Core.* (Hillmann, 2003).

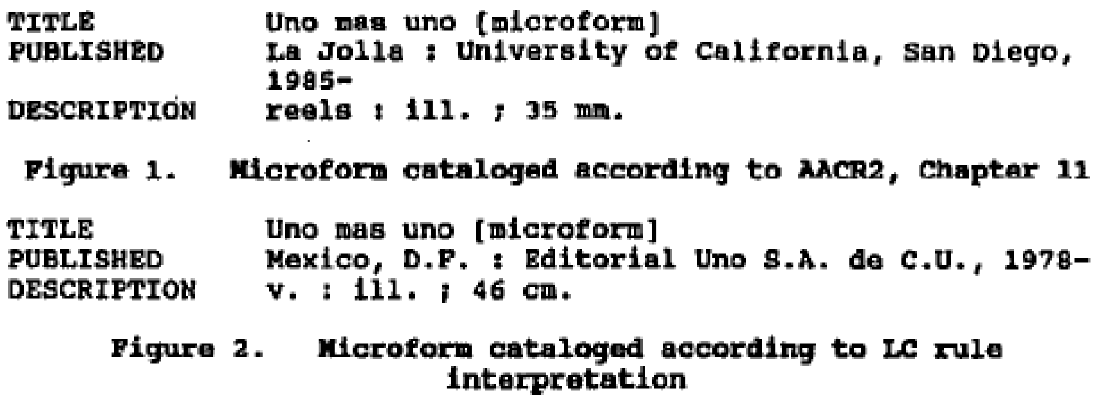
Table 2.1: Key events in the development of the Dublin Core *1:1 Principle.*

2.3 Principle Precursors: Cataloging Cultural Heritage Materials

The cultural heritage community's concerns about "document-like objects" and the relationships between "original" and "surrogate" resources within the Dublin Core standards were not new. Rather, they came after an important period of development that saw existing bibliographic standards adapted to include non-book materials, such as photographs, graphic materials, audio/visual resources, museum artifacts, and archival records. These discussions were informed by concerns in the broader library community about how to represent the plethora of media formats appearing in collections. These formats ranged from microforms to emerging electronic resources.

2.3.1 Cataloging Reproductions

Early bibliographic cataloging codes were primarily concerned with book-like materials held in library collections. However, as readily available photomechanical reproduction techniques came into use (especially the use of microfilm and microfiche in the mid-twentieth century), "confusion reigned when it came to describing the reproduction" (Knowlton, 2009). Simonton (1962) completed a survey for the Association of Research Libraries (ARL) in which he proposed two different approaches to the problem. Each solution recommended that catalogers create a new catalog entry to represent the reproduction (resulting in two entries—one for the original and one for the reproduction), but the approaches differed in what they treated as the object of the description. The "Facsimile Theory" approach privileged the intellectual content of an item by making the "original" resource the focus of the record representing a reproduction. Following the long-standing practice of "dash entries," a description of the reproduction itself would be included as a note. Alternatively, Simonton's "Edition Theory" required a record to represent the physical features of the reproduction, using a note to provide a description of the "original" resource (Simonton, 1962; Graham, 1992). The first edition of the *Anglo-American Cataloging Rules (AACR1)* (American Library Association, 1967) incorporated Simonton's facsimile theory (1962) by requiring a new entry for a reproduction that described the "original" with a note describing the resource's status as a reproduction (Graham, 1992). However, the second edition of the rules, *AACR2* (American



TITLE	Uno mas uno [microform]
PUBLISHED	La Jolla : University of California, San Diego, 1985-
DESCRIPTION	reels : ill. ; 35 mm.

Figure 1. Microform cataloged according to AACR2, Chapter 11

TITLE	Uno mas uno [microform]
PUBLISHED	Mexico, D.F. : Editorial Uno S.A. de C.U., 1978-
DESCRIPTION	v. : ill. ; 46 cm.

Figure 2. Microform cataloged according to LC rule interpretation

Figure 2.1: Examples of reproduction records from Graham (1992)

Library Association et al., 1988), reversed this decision, requiring that entries for reproductions adhere to the "edition theory" (see Figure 2.1 an example of each approach). The justification for this change was *AACR2's* cardinal principle:

> The starting point for description is the physical form of the item at hand, not the original or any previous form in which the work has been published. (American Library Association, 1967; Graham, 1992)

The cataloging community did not welcome this reversal and "assailed [it] as 'an obsession with principle to the exclusion of common sense'" (Graham, 1992). Catalogers felt that the edition theory prevented entries about reproductions from co-locating with entries about originals in the catalog, inhibiting the ability of users to find them. Reproductions also possessed features distinct from those of the original, like place and date of publication. This could confuse patrons and make it difficult for them to recognize that the reproduction was the resource they sought. The change also had an economic impact. Under the facsimile theory *(AACR1),* existing records for original resources could be cloned easily in electronic bibliographic catalogs. Simply adding a note about the reproduction is all that would be required. Under the edition theory *(AACR2),* however, a cataloger would have to "start over" and create a new entry for the reproduction, adding significant cost to cataloging procedures (Graham, 1992).

Challenges to the rule change came from several quarters. After studying the problem from various perspectives, the Library of Congress, the National Library of Medicine, and the Agricultural

Library (all heavy users of microfilms) rejected the changes. The Library of Congress released a Rule Interpretation maintaining the facsimile theory approach for LOC cataloging (Graham, 1992; Library of Congress, 2010). The push to adopt these rules in academic libraries came primarily from the members of the preservation community who were engaged in projects to transfer brittle books and newspapers onto microfilm. The change in rules meant that significant portions of funding were being used for re-cataloging efforts rather than converting additional materials. In response to concerns from the National Endowment for the Humanities (NEH), RLG modified its cataloging system, the Research Libraries Information Network (RLIN), to allow the use of older rules for reproductions. In September 1990, the ARL released *Guidelines for Bibliographic Records for Preservation Microfilm Masters*. The adoption of these guidelines by LOC, RLG, RLIN, and OCLC made the facsimile theory the de facto national standard for describing reproductions (Association of Research Libraries, 1990; Graham, 1992).

The Multiple Versions Problem

While these discussions were under way, a parallel conversation was taking place about how to handle the *multiple versions problem* as a proliferation of new media formats entered library collections. The same musical work might be released on a vinyl record or a tape cassette; a film could appear on 35mm stock, or as a Beta or VHS videotape; a journal could be published in both print and electronic forms. “The gradual result of all this was proliferation of bibliographic records for multiple versions of serials, especially in shared cataloging databases, such as OCLC’s. The content of these records was virtually identical with that of records for the original printed serials, with access points that were largely redundant across all records” (Jones, 1997).

As with rules for describing microforms, pressure came from the preservation community. The United States Newspaper Program (USNP) was an NEH project that aimed to create a comprehensive national union list of all U.S. newspaper publications and their reproductions.

> Multiple versions were endemic in the union list. The USNP found that the needs of its users were best served by consolidating information for all versions onto a single catalog record representing the original (even when the original no longer survived). (Jones, 1997)

Instead of creating multiple records for different physical versions, the USNP used a single descriptive record and attached information about variants using the standard for “holdings” information.

While this approach was successful within the USNP, finding broad consensus for the USNP solution to the multiple versions problem was more elusive. An initial survey conducted by the Library of Congress Cooperative Serials Program (CONSER) recommended continuing the practice of creating a new record for each format, but providing links in those records to others for related forms(Jones, 1997). The Multiple Versions Forum was convened in December 1989 to sort out the available options. The Forum rejected both consolidated records and the use of separate, but linked, records. This was based on the limitations of existing bibliographic databases, especially in the case of shared networks of bibliographic records. Instead, the *Guidelines for Bibliographic Description of Reproductions* recommended a “two-tiered hierarchical technique” that mediated between these approaches (Association of Research Libraries, 1990). Because no record syntaxes were developed to support this technique, a fractured set of approaches was left in place (Jones, 1997).

Cataloging Electronic Resources

Unfortunately, the precedents set by the treatment of analog reproductions and multiple versions informed the discussions about how to handle emerging digital formats. Unlike earlier practices, “once digitized the surrogate of the original is subject to a very different set of cataloging rules. It is cataloged according to the ‘class’ of materials to which the item belongs, ‘Electronic Resources’ ” (Copeland, 2002). The Library of Congress extended its earlier rule interpretation for reproductions (Library of Congress, 2010) to include digital reproductions:

> The Prints and Photographs Division regards its digital reproductions, even the high-resolution images being prepared under the current conversion contracts, as surrogates.... The records describe the intellectual expression and the original form of the material and provide a link to the corresponding digital reproductions. Information about the digital files is not recorded in the MARC bibliographic records since these are considered surrogates for reference purposes rather than separate works. (Arms, 1999)

2.3.2 Representing Visual Documentation

The developments in cataloging for reproductions and multiple versions discussed so far mainly dealt with traditional bibliographic materials such as newspapers, journals, and audio/visual resources in various formats. Independently of this work, librarians, archivists, and curators struggled with how best to provide access to non-book materials such as photographs, architectural drawings, museum artifacts, and photographic reproductions that depicted such objects. Building on existing rules for book materials, professionals responsible for these collections began developing specialized rules for different types of material used in different contexts. One strand of development adapted the MARC bibliographic data format for use with visual materials (MARC-VM) or Archival and Manuscript Control (MARC-AMC). For many communities, newer approaches to constructing databases and document–markup techniques provided the opportunity to establish community-specific data standards and content rules. One example of this is the *Categories for the Description of Works of Art (CDWA).* In these standards we can see directly the examples that informed the recommendations found in *Guidelines for Extending the Use of Dublin Core Elements* (Research Libraries Group, 1997a).

Cataloging Visual Materials

Cataloging rules for visual materials initially focused on describing original items found in special collections and archives (Parker & Library of Congress, 1981; Shatford, 1984; White-Hensen & Library of Congress, 1984). But the growing visual resources movement began developing approaches for describing secondary sources found in slide libraries used by art historians, architects, and historians. An immediate concern for these efforts was the question of what, exactly, a MARC record represented and what properties should be represented using the format. Traditionally some of these concerns had been handled through the creation of physically distinct card catalog systems that provided context to distinguish between:

- Actual Daguerreotypes
- 35mm slide reproductions of Daguerreotypes
- Books about Daguerreotypes (Dooley & Zinham, 1990)

The introduction of new genres and formats of materials into cataloging practices, along with the unifying effect of the MARC format and integrated library systems, resulted in loss of this context. The structure of MARC itself, based in book cataloging and subject classification, disguised the distinctions as well. To help restore some of the contextual information afforded by physically distinct catalogs, MARC-VM and MARC-AMC introduced a control field (`006`) to indicate "the type of record." A brief code list corresponded with different material types:

b Archival and manuscripts control (AMC format only)

g Projected medium

k Two-dimensional non-projectable graphic

o Kit

r Three-dimensional artifact or naturally occurring object

(Evans & Will, 1988)

Types of records were loosely associated with different cooperative cataloging systems, such as OCLC, RLIN, or other specialized projects such as the Getty Art History Information Program (AHIP). Within the contexts of these cataloging services, the type of record could also indicate an association with a set of content rules or best practices for description. In some cases, for example the RLIN database, records of different types could be treated differently by indexing and retrieval systems.

The new formats and rules required rethinking the properties available in a MARC record. Initially MARC only provided a single tag (`650`) to express "topical, iconographic, genre type or physical descriptions" which did not allow a record to distinguish *a set of drawings* from *a book about drawings* (Barnett & Petersen, 1990). MARC-VM and MARC-AMC introduced two new fields to indicate genre type (`655` Index Term-Genre/Form) and physical description of the cataloged item (`755` Added Entry-Physical Characteristics). However, "[t]he distinctions between field `655` and `755`...often remain ambiguous.... [W]hen form of material is understood as 'format' or 'type of object' the intellectual-physical boundary between fields `655` and `755` disappears" (Dooley & Zinham, 1990). For example, materials such as scrapbooks, diaries, posters, and broadsides might be defined as much by their physical characteristics as by their genre.

This distinction also became important when describing resources that reproduced other resources. For example, a portrait (a genre of image) reproduced on a microfilm would remain a portrait. However, as a tangible object, what was a Daguerreotype (`755` physical characteristics) now has the characteristics of a microfilm (Dooley & Zinham, 1990). Sorting out the distinctions between genre and physical formats was the responsibility of controlled vocabularies such as the *Thesaurus for Graphic Materials (TGM)* or the *Art and Architecture Thesaurus (AAT).* However, different thesauri approached the question of reproductions differently:

> The *AAT* considers reproductions of works of art to be surrogates for original works and will recommend that they be indexed in a similar fashion. For example, `PAINTING (655)` would be used to describe both Leonardo's *Mona Lisa* and a slide reproduction; `SLIDE (655)` would also be used in the latter case. This holds serious implications for effective retrieval, as Shatford has warned. In an integrated database containing both of these media, searchers interested only in examples of actual paintings might have to learn to exclude slides, microfilm, and other reproduction media in their search queries to retrieve only records for original paintings.... One solution might be the addition of a "reproduction" facet to indexing strings for object surrogates so that they would be differentiated from "originals" in a browse display. (Dooley & Zinham, 1990)[3]

In Dooley & Zinham's quote (1990) we see almost the same examples and concerns represented more than ten years later in Hillmann's account (2003) of the *1:1 Principle.* Discussions about appropriate representations for reproductions and surrogates are frequently concerned with the information needs of users. Just as advocates of Simonton's "facsimile theory" noted the usefulness of collocating its "reproductions" alongside originals represented in the catalog, recommendations for reproductions and surrogates noted that users were not looking for features of reproductions, but were looking for "originals." This was particularly true for slide libraries, where entire collections consisted of nothing but "reproductions" (Most, 1998). "The question is whether we are cataloging slides, or the item depicted. We catalog the building shown in the slide, not the photographer of the building. In 99 percent of the cases, a photographer is not important to the slide librarians or the

[3] The *AAT* did include explicit classifications for facsimiles and reproductions; however, it "did not provide a way to link them with the characteristics of the originals" (Dooley & Zinham, 1990).

patrons" (Snow, 1990). For Shatford (1984), these use cases meant that rules for cataloging pictorial works should be based on the *objectives* of bibliographic catalogs (e.g., Cutter and Lubetsky), not necessarily on the rules that were derived from them for book cataloging. Shatford (1984) recommended developing a new set of rules that also more closely examined the entities that were the focus of subject access and descriptive cataloging. Shatford (1984) explored proposals for both narrow and broad definitions of "represented works" that would help meet the objectives of the catalog.[4]

Visual Documentation of Museum Artifacts

Museums and other non-library cultural organizations that had not yet adopted a standardized syntax like MARC had the advantage of being able to start with a clean slate. A key part of the Getty Information Institute's agenda was to develop a better understanding of the information needs of humanities scholars who relied on visual resources and artworks for their research.[5] This work reinforced what collection professionals who worked with these scholars already knew—that consumers of art information had a keen awareness of the differences between the original works and various analog and digital reproductions that depicted them (Blackwell et al., 1988; Sundt, 2002). Both scholars and general museum users sought items based on the features of original works depicted by reproductions, not of the reproductions themselves (Spinazze, 2002). Emerging descriptive format standards, such as the *Categories for the Description of Works of Art (CDWA),* explicitly provide distinct sets of properties assigned to either "objects/works" or to "images/visual documentation" (Harpring & Baca, 2009).

Discussion

In Section 2.2, we learned that the *1:1 Principle* emerged out of the needs of cultural heritage institutions to clearly represent information about both physical and digital resources. Here, we learn that these concerns did not present themselves because of the new digital environments. Instead, they are a continuation of a decade of development as libraries, archives, and museums struggled

[4] I.e., a broad definition would include all copies, reproductions, and derivatives of a visual work whereas a narrow definition would make all such entities new works in their own right.

[5] Then known as the Art History Information Project (AHIP).

to represent the same kinds of relationships using existing bibliographic languages. In most cases, the specification of rules tended to acknowledge that regardless of the actual state of affairs, users sought resources based on the properties of the original resource. Rules for bibliographic catalogs encouraged representations that allowed descriptions of originals and reproductions to collocate when they shared the same catalog. The context in which a record appeared also provided important clues to interpreting its meaning. A catalog devoted entirely to reproductions (e.g., in a slide library) allowed users of records to infer the relationship between originals and reproductions without making that relationship explicit in the records. In emerging electronic services, developers felt it was important to preserve these contextual clues by adding elements that explicitly declared what kind of record followed. This not only communicated information to the user about the resource being described. It also allowed computer systems to handle records according to their type.

Because initial drafts of Dublin Core lacked features that had already been introduced in MARC formats, the cultural heritage community wished to ensure that Dublin Core would reflect those practices already in place. However, these recommendations were also heavily based in an environment that relied on fixed document-like syntaxes for records. The approaches to creating metadata representations at DCMI were moving towards less fixed approaches that allowed for more dynamic kinds of representations. These approaches were also not new, but were derived from developments in computer science and artificial intelligence.

2.4 Principle Precursors: Representation Semantics and Syntaxes

Conversations about “logical clusters of metadata” in Helsinki were informed by developments at the three previous DC workshops and by discussions in the larger Web metadata community. At the DC-1 meeting, questions about syntaxes for Dublin Core were intentionally tabled. Participants feared that a discussion of implementation formats would be premature until the vocabulary was more established. At the second Dublin Core workshop, held in Warwick, England, delegates began working out issues related to Dublin Core syntaxes that would enable interoperable exchange of descriptions (Dempsey & Weibel, 1996). The recommendations focused on how to embed Dublin Core elements within HTML documents or as a standalone description using an Standard Generalized Markup Language (SGML) Document Type Definition (DTD). Developers believed that the standalone SGML approach would afford "explicit semantics of each Dublin Core element"; however, "discrete packages of metadata cannot be identified and the semantics of repeated elements are not specified" (Burnard et al., 1996). In order to address the latter problem, the participants at the workshop emerged with the Warwick Framework, which defined "containers" for "packages" of metadata (Burnard et al., 1996; Dempsey & Weibel, 1996; Lagoze, 1996). A package might include Dublin Core metadata, or metadata in other formats.

> Metadata is data, no more and no less. . . . A better approach is to consider the information architecture as a collection of inter-related resources. While these resources may have a type, such as PostScript, HTML, or a Java program, this type is orthogonal to whether the resource is acting as data vs. metadata in some context. That contextual information is specified by the relationships between the resources. We can model these inter-related resources using directed graphs, where nodes represent the resources and the labeled arrows between nodes represent the relationships. Since a resource may be related to many other resources, nodes may have many arcs originating from or terminating at them. Looking at the direction of an arrow, it is easy to see whether a resource is playing the role of data or metadata in the context of that particular relationship. We can easily accommodate such a model by generalizing the Warwick Framework so that it may contain any resources, not just those considered "metadata". Thus, we can use the Warwick Framework Catalog to specify the relationships between various resources, both inside and outside the container. (Lagoze, 1996)

The Warwick Framework discussions become an influential part of conversations about metadata for web resources being developed by the new World Wide Web Consortium (W3C). In particular, W3C was also addressing technical concerns raised by laws concerned with filtering adult content on the Internet. One such initiative was the *Platform for Internet Content Selection (PICS)* that would embed ratings and content labels into HTML pages. PICS would enable software to prevent minors from seeing adult material, or would make assertions about the authority of an author. In order to accomplish this, the PICS working group developed several metadata object models to express the labels assigned to webpages. The PICS Model and Syntax used identifiers (such as the URL of a web page) to make statements about a resource (Lassila, 1997). Although designed for content filtering, it was recognized that the general approach recommended by PICS could be used elsewhere. The development of PICS coincided with the submission of several other proposals for metadata formats. Microsoft's *XML Web Collections (XMLWC)* used a "collection" of statements to describe a resource. Such collections could be embedded as part of a resource or use URIs to refer to the resource being described (Hopmann et al., 1997). A similar proposal from Apple was called the *Meta Content Framework (MCF).* Created by R.V. Guha, MCF was based on earlier work on "structure description languages" for the Cyc knowledge base. MCF was intended to translate the concepts of that research into an application for the Web. As did Lagoze's suggestion for the Dublin Core Warwick Framework (Lagoze, 1996), MCF based its language model on directed labeled graphs (DLG) consisting of a set of labels, nodes, and arcs—a triple. MCF also rejected the distinction between "data" and "metadata," noting that an object might play different roles in different situations. Using DLGs meant that MCF could express multiple relationships between many different kinds of objects—both "data" and "non-data" resources. MCF also provided a basic set of nodes that would allow designers to specify their own vocabularies and properties. MCF's specifications used the new XML standard to represent the structure of the DLGs that made up a description, rather than the tree-like hierarchy of traditional documents. In order to organize statements into DLGs, MCF expected "units" to use canonical URIs as unique identifiers, making them part of a network of nodes and arcs represented by properties, relationships to other objects, and expressions of values (Guha & Bray, 1997).

Rather than developing each of these recommendations separately, the W3C rolled them together into a new initiative known as the *Resource Description Framework (RDF).* RDF was not based on any single technology. Rather, it borrowed from XMLWC, MCF, and the Dublin Core Warwick Framework (Miller, 1998). RDF's emergence was also tightly tied to the new *eXtensible Markup Language (XML)* specification that aimed to provide a bridge between the complexities of the *Standard Generalized Markup Language (SGML)* and *HyperText Markup Language (HTML).* "By exploiting the features of XML, RDF imposes structure that provides for the unambiguous expression of semantics, and, as such, enables consistent encoding, exchange, and machine-processing of standardized metadata" (Miller, 1998). The development of the RDF specifications encouraged DCMI to begin work on providing a more formal data model for Dublin Core (Weibel & Hakala, 1998; Weibel, 2010).[6]

2.5 A Principle in Practice: Developing Implementation Guidelines for the *1:1 Principle*

Following the Helsinki meeting, the two working groups (One-to-One and Relations) set to work. The discussions were frequently contentious debates between members in different camps. For cultural heritage participants, the issues with the *1:1 Principle* were primarily questions about the kinds of resources that could be described using Dublin Core. Drawing on their experiences with previous standardization efforts, this camp felt it necessary to provide guidance. However, there was a strong resistance to Dublin Core getting into the cataloging rules business. The members of this opposing camp preferred to let Dublin Core remain a simple standard for resource discovery. Acknowledging the concerns of cultural heritage professionals, they argued that the kind of discrimination they sought could be handled by more robust local standards (Miller & Greenstein, 1997). Furthermore, discussions on the one-to-one listserv:

> . . . made absolutely clear that there is no consensus on what 1:1 really means in practice. In the end, people will describe what *they* want to describe, for their purposes and the purposes of their user community. That means they may describe a TIFF of an Ansel Adams photograph as having been created by Ansel Adams. Who's to say they're wrong? (Wendler, 1999)

By the end of 1999, this online discussion dwindled. The group was formally disbanded at the 2000 DC Workshop without having reached a clear consensus or definition of the *Principle*.

At the time of these discussions, Dublin Core still had not published any definitive guidelines for encoding syntaxes. Assumptions about what the *1:1 Principle* applied to were largely based on assumptions about what a record was. Among the suggestions for how to handle the *1:1 Principle*

[6] This would ultimately result in the DCMI Abstract Model (DCAM).

was the concept that a "record" might consist of different logical packages of information, each of which described a different related resource (for example, see Figure 2.2 from Weinheimer (1999)).

```
<set 1>
        <Title=Mona Lisa>
        <Format=oil painting>
        <Creator=Leonardo da Vinci>
</set 1>
<set 2>
        <Title=Photo of Mona Lisa>
        <Format=35 mm>
        <Creator=Stan>
</set 2>
<set 3>
        <Title=Digitized image of Mona Lisa>
        <Format=jpeg>
        <Creator=Joe>
</set 3>
```

Figure 2.2: Example of "logical clusters" of metadata from Weinheimer (1999)

Discussions in the Relation Working group focused more on developing "logical clusters" of metadata that could be linked together. The discussions echoed concerns aired in regard to earlier MARC-based solutions to representing the related "facets" of originals and reproductions. In particular, there were concerns that separate "records" for an original or a reproduction could be separated, resulting in a loss of information. The suggestion of separate records also raised concerns about how such records would be displayed to the user, with a sense that independent representations of "originals" and "reproductions" would make the task harder. Proponents of "keeping Dublin Core simple" suggested that the use of atomic statements about resources could enable better discovery of resources without the additional complexity being suggested by models that identified resource types. Instead, statements about resources could be dynamically organized into logical packages for particular uses such as retrieval or display for a user (Lagoze, 1997, 2001a). One method that offered promise was the emerging RDF semantic model.

As noted above, the emergence of RDF as a parallel set of developments to Dublin Core spurred DCMI to begin thinking about a data model for representing Dublin Core metadata. Initially the development focused on expressing Dublin Core as a variant of RDF. However, within the implementer community, there was a great deal of initial resistance to RDF in favor of simpler "plain" XML representations. This was due in part to a lack of practice and of software tools that could understand RDF. Because the XML serialization of RDF represented a graph structure, it was also less human-readable than straightforward encoding of records consisting of element/value pairs. Push back against RDF came from the Open Archives Initiative (OAI) community, which was developing a protocol for exchanging "packages" of metadata along the lines of the Warwick Framework. The lack of broad uptake of RDF was of particular concern for the OAI developers. "It may be that the vast majority of data providers don't need (or even understand) RDF and are mainly interested in exposing metadata as simple attribute-value pairs or simple trees for which XML is perfectly appropriate" (Lagoze, 2001b). In order to conform both to the simple Dublin Core model and to provide a low barrier to use (i.e., by using well-supported technologies), the OAI community adopted a simple Dublin Core XML schema as a required part of the protocol (Lagoze et al., 2008). Initially this schema was developed as one of many approaches to specifying Dublin Core. This multiplicity was possible because there were no formal recommendations from DCMI.

In a variety of ways, DCMI addressed calls for greater guidance for implementing Dublin Core. In 2001, DCMI released official guidance for encoding Dublin Core in XML. This included rudimentary definitions of an abstract model that specified a one-to-one relationship between a record and a resource (Powell & Johnston, 2002).[7] *Using Dublin Core* provided an informative introduction to both metadata and Dublin Core specifically. Although early drafts of this document do not refer to the *1:1 Principle,* it did appear in the 2003 version (Hillmann, 2003). Hillmann's account offers

[7] This definition would mature into the separate specification for the Dublin Core Abstract Model (Powell et al., 2007). The DCAM provides a clearer specification of how the *1:1 Principle* could have be implemented by Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH).

> A *description* is made up of one or more statements (about one, and only one, resource) and zero or one resource URI (a URI reference that identifies the resource being described).... This is commonly referred to as the *1:1 Principle*.
>
> However, real-world metadata applications tend to be based on loosely grouped sets of descriptions (where the described resources are typically related in some way), known here as *description sets.* For example, a description set might comprise descriptions of both a painting and the artist. Furthermore, it is often the case that a description set will also contain a description about the description set itself (sometimes referred to as "admin metadata" or "meta-metadata").
>
> Description sets are instantiated, for the purposes of exchange between software applications, in the form of metadata records, according to one of the DCMI encoding guidelines (XHTML meta tags, XML, RDF/XML, etc.) (Powell et al., 2007).

general guidance about the *1:1 Principle* through the example of the *Mona Lisa* and a surrogate image of the famous painting in a way that closely follows Dooley & Zinham's example (1990).

In regard to recommendations, however, DCMI was not in a position to enforce any particular approach to implementing Dublin Core. Each community that adopted it found their own solutions to the problems at hand, often using practical workarounds to solve problems. For example, the Consortium for the Computer Interchange of Museum Information (CIMI) project constructed a Dublin Core testbed and found:

> The element set is not effective for discovery of complex, museum information. Attempts to extend the DCMES using Semantic Refinements resulted in violation of both the 1:1 and dumb-down principles. There is a tension between the type of information that museums are used to sharing and the fundamental principles of the Dublin Core. (Spinazze, 2000)

2.6 Conclusion

The developers of Dublin Core intended it to be a simple vocabulary that could be broadly applied to emerging Internet resources. The introduction of cultural heritage material also introduced more complex kinds of relationships between online and offline resources or "originals" and "reproductions." Faced with this problem, the cultural heritage community proposed solutions based on many years of practice using *document surrogates* in information retrieval systems. Human-understandable cataloging rules made such documents meaningful for both metadata creators and users. However, users of traditional cataloging systems also struggled with defining best practices for describing reproductions and multiple versions. Conflicting interpretations meant that document surrogates could appear in two forms based on the object of description (i.e., facsimile/edition theory approaches). Within the Dublin Core Metadata Initiative, these developments in descriptive cataloging encountered new approaches to representing descriptions as "metadata." While emerging technologies such as XML enabled the creation of document-like data models, the development of Dublin Core was also influenced by more formal modeling techniques that required a one-to-one relationship between entities and their descriptions. Because this requirement conflicted with the cultural heritage community's recommendations for handling reproductions, it was necessary to articulate it in DCMI documentation as the *1:1 Principle*.

Unfortunately, what lies behind the *1:1 Principle* are fundamentally different approaches to description. The conversations and debates that led up to the *1:1 Principle* tended to obscure the theoretical differences between these approaches in the course of working towards pragmatic consensus. In order to further understand the *1:1 Principle*, Chapter 3, Representation Paradigms, will examine these differences through the lens of two *representation paradigms: knowledge organization* and *knowledge representation.* In Chapter 4, Description and Reference for Cultural Heritage, I examine the methods used by each paradigm to establish a relationship between described entities and their representations. In Chapter 5, Revisiting *1:1 Principle* Violations, these observations lead me to question how metadata quality research has characterized *1:1 Principle* errors and their detection.

Chapter 3

Representation Paradigms

> The representation of knowledge in symbolic form is a matter that has preoccupied the world of documentation since its origin.... The structure of records and files in databases; data structures in computer programming; the syntactic and semantic structure of natural language; knowledge representation in artificial intelligence; models of human memory: in all these fields it is necessary to decide how knowledge may be represented so that the representation may be manipulated. (Vickery, 1986)

> The ability of a system to solve the problems for which it was designed is strongly influenced by the choice of representation. (Smith, 1980)

3.1 Introduction

Baker (2000) describes Dublin Core as "a small language for making a particular class of statements about resources." In his view, Dublin Core is a *pidgin language* intended to serve as a *lingua franca* for the exchange of information about resources in a "linguistically diverse" Internet commons. As with any language, common rules of grammar allow speakers and listeners to construct and understand what is expressed using Dublin Core. The *1:1 Principle* seems to be exactly the kind of grammatical rule needed for clear communication, yet the available evidence suggests that metadata creators frequently have ignored it. What explains this apparent disconnect between Dublin Core recommendations and practice?

Park & Childress (2009) and Urban (2009) suggest that confusions related to the *1:1 Principle* may arise because the concepts that the *Principle* embodies have not been clearly stated in available Dublin Core documentation. This lack of clarity has resulted in confusion among metadata creators tasked with conforming to the Principle. Urban's suggestion (2009), then, was to provide an improved conceptual definition of the *Principle* that might help alleviate this confusion and also serve as a basis for assessing metadata quality. However, in Chapter 2, we found that the *Principle* emerged from discussions among members of the Dublin Core community with diverse perspectives about the fundamentals of Dublin Core's grammar. If we are to articulate a clear, coherent definition of the *1:1 Principle*, in which of these perspectives should it be grounded? Finally, how can paradigms illuminate our understanding of *1:1 Principle* violations?

This chapter serves as a foundation for answering these questions by introducing two representation language paradigms:

- The Knowledge Representation Paradigm uses *knowledge representations* that express propositional statements using formal semantics for automated reasoning. *(artificial intelligence, ontology, formal semantics, Semantic Web)*

- The Knowledge Organization Paradigm uses *document surrogates/records* that represent features of documents optimized for relevant retrieval and organization in closed information retrieval systems. *(bibliographic classification, information retrieval, colloquial syntax)*

While each of these paradigms is based on the practices of particular communities, it is unclear from the available evidence whether they can be associated with specific actors involved with DCMI specifications. Weibel et al. (1997) noted the emergence of *structuralists* and *minimalists* during debates at DC-4, with the following caution:

> It is an oversimplification to suggest that these two groups are distinct and mutually exclusive; in fact, they are two poles of a continuum, and conferees were distributed throughout this continuum.

I believe the same caution holds here. The goal of this chapter, then, is to introduce the paradigms as a conceptual and analytic lens through which we can understand how each paradigm goes about the task of constructing representations. As such, this chapter is grounded in the technical specifications and research literature associated with each paradigm, not the attitudes or beliefs of metadata creators themselves (as represented by Park & Childress (2009)).

3.1.1 The Concept of a Paradigm

The concept of scientific paradigms was introduced by Thomas Kuhn's *The Structure of Scientific Revolutions* (1970) as a way to understand how scientific theory advances. In Kuhn's approach, a scientific paradigm shapes how scientists think about the research problems they address in their work and prompts new questions when those modes of thinking encounter anomalies. Tjörnebohm (1974), as translated by Hjørland (2003), presents the following attributes of paradigms:

- ideals and beliefs about science, such as epistemic goals, methods, and criteria in the production and evaluation of scientific results inside the discipline;
- world view hypotheses, including basic social ontological assumptions about the part of the world studied inside the discipline, and;
- ideals concerning the extra-scientific significance of knowledge produced inside the discipline, such as significance for society and culture, for practical use, and for enlightenment.

Although Kuhn's original conceptualization of paradigms was directed at scientific paradigms, the concept has been usefully applied elsewhere, including the social sciences. Here, the kinds of features that are the basis of Kuhn's criteria are useful for teasing out confusions around the *1:1 Principle* and sharpening the distinctions that appear in its evolution, as found in Chapter 2.

Representation Paradigms

Kuhn's conceptualization of paradigms has been especially useful for understanding the internal divisions of research within the library and information science community (Brown, 1987; Miksa, 1991; Bates, 1999; Hjørland, 2000, 2003, 2005) and the artificial intelligence (AI) community (i.e., "scruffy" vs. "neat" approaches) (Halpin, 2004).

For the most part, however, these uses of paradigms have focused inwardly on different approaches to a common research agenda. In this way, they may be more closely tied to Kuhn's conceptualization of paradigms as part of scientific progress. Through these discussions we can see the evolution and branching of research within each of these disciplinary areas. However, as Smith (1976) suggests, there is also value in importing concepts across disciplinary paradigms as a way to ask new questions within one's own domain. While Smith (1976) did this for early artificial intelligence research, conversations around the Dublin Core *1:1 Principle* suggest that a similar comparison is needed for contemporary representation languages.

In order to outline how knowledge representation and knowledge organization operate as distinct paradigms, this chapter builds on Tjörnebohm's characterization (1974) of Kuhn's paradigms and definitional criteria of representation languages (drawing from definitions of knowledge representation in Davis et al. (1993) and Sowa (2000), and description of bibliographic languages in Svenonius (2000)). For both knowledge representation and knowledge organization, I present the following paradigm-defining criteria:

- The objectives and principles of representation languages;
- The methods used by each paradigm to achieve stated objectives;
- The ontological commitments inherent in those objectives, principles, and methods;
- The significance of each paradigm for the Dublin Core *1:1 Principle*.

3.2 Knowledge Representation

3.2.1 Overview

The *knowledge representation paradigm* emerged from the research and design of *artificial intelligence (AI) systems* in the 1970s and 1980s. Although the field is most closely associated with computer science, it traces its ancestors to the earliest methods of systematically reasoning about the world around us (Davis et al., 1993; Sowa, 2000; Halpin, 2004). "Over the long time after Aristotle, logicians such as Frege, Russell, Leibniz, Gödel, Tarski and others have fully developed formal logic systems such as propositional logic and predicate logic, which formalize the thinking and reasoning process of humans" (Chen, 2010). Broadly, artificial intelligence research focuses on a series of problems related to *pattern recognition, machine learning, problem solving*, and especially relevant to this discussion, *representation* (Smith, 1976, 1980). Although not all of AI's agenda has been achieved, the research has benefited applications in medicine, bioscience, and machine translation. This research also has enabled large-scale search and retrieval on the World Wide Web.

Google, Yahoo! and other tools use related techniques for knowledge extraction, indexing, neural networks, and genetic algorithms for organizing the Web (Chen, 2010).

Within AI, the field of *knowledge representation* emerged to develop formalizations of knowledge needed for intelligent reasoning. Early representation languages for artificial intelligence systems matured into Semantic Web languages, such as the Resource Description Framework (RDF). RDF is directly descended from large-scale knowledge base projects, like Cyc, that attempted to represent "common-sense" knowledge needed for reasoning (Halpin, 2004). Such representations rely not just on syntaxes They also use formalized interpretations of theories of description and reference. Because of a core concern with description, the RDF community has influenced (and been influenced by) the Dublin Core community, as can be seen in recommendations such as the Warwick Framework, the DCMI Abstract Model, and specific syntax guidelines for Dublin Core.

3.2.2 Objective of Relevant Reasoning

According to Davis et al. (1993), knowledge representations operate within a "theory of intelligent reasoning" that specifies what counts as reasoning, what inferences are licensed by the theory, and what other inferences it recommends. For many knowledge representation languages this leads to a formal semantics that corresponds to First-order Logic (FOL) (Sowa, 2000). Statements that rely on formal semantics excel at expressing declarative sentences that can be assigned a truth value (true/false) and can be modeled by formal *interpretations*.

For example, RDF adopts *model-theory* to specify its representation language semantics. A precise mathematical theory allows the language to represent an interpretation of a world by:

> ...describing the minimal conditions that a world must satisfy in order to assign an appropriate meaning for every expression in the language.... The chief utility of a formal semantic theory is not to provide any deep analysis of the nature of the things being described by the language or to suggest any particular processing model, but rather to provide a technical way to determine when inference processes are valid, i.e., when they preserve truth. This provides the maximal freedom for implementations while preserving a globally coherent notion of meaning." (Hayes, 2004)

Another way to understand such a model is that it is the designer who provides his or her interpretation of what *true statements* can be made about some world. When a statement is expressed using these kinds of semantics, it says something true about something that exists in the model of the world. How well the interpretation of the world corresponds to the "real world" is part of the design challenge. However, a reasoning system using the model remains ignorant of this correspondence. It can only reason about what it knows.

> The meaning of a knowledge base derives from features and relationships that are common to all possible models. If, for example, the interpretation of a class must always be the empty set, then that class is said to be inconsistent, while if there are no possible interpretations, the knowledge base itself is said to be inconsistent. If the relationship specified by a given axiom must hold in all interpretations of a knowledge base, then that axiom is said to be entailed by the knowledge base, and if one knowledge base entails every axiom in another knowledge base, then the first knowledge base is said to entail the second knowledge base. (Horrocks et al., 2003)

In other words, a knowledge representation environment supplies representations with a *formal semantics* that enables reasoning systems to understand the meaning of propositional statements to a degree that is sufficient to derive inferences from them. KR does not supply broader understanding of "semantics" as representations that are meaningful for human consumption and reasoning.

Principles of Knowledge Representation

As the focus of this study is on the *1:1 Principle,* it is also helpful here to consider the principles each paradigm uses. This begs the question: "what are principles and how do they help us understand representation languages?" In answering this question, we find that there are, in fact, different senses of the term "principle" that are also helpful in thinking about each paradigm. Consider:

principle: A fundamental truth or proposition on which others depend; a primary assumption forming the basis of a chain of reasoning. (Oxford English Dictionary, 2011)

principle: A general law or rule adopted or professed as a guide to action; a settled ground or basis of conduct or practice; a fundamental motive or reason for action, esp. one consciously recognized and followed. (Oxford English Dictionary, 2011)

In pursuit of the objectives above, knowledge representations are built on the fundamental principles (i.e., fundamental propositions) of logical reasoning as represented by predicate logic. In developing models for reasoning, then, a fundamental concept like "representations are about one, and only one, resource" can be seen as an essential proposition of a KR model. Any reasoning that takes place within a model that has adopted this as a principle will depend on its truth to evaluate the mathematical truth of other statements. It is not a principle that can be set aside to accommodate a particular contextual situation, although different interpretations may be constructed where a different fundamental principle applies.

This is not to suggest that the second sense of principle has no role in KR. Similar principles of guidance inform the translation of natural language sentences into a formal representation.[1] These translations often favor "naturalness" and simplicity over the complexities of other possible translations. Likewise, principles of guidance can help developers of an ontology represent the messiness of the real world in a functional model by suggesting appropriate simplifications. However, in most cases these guiding principles stand outside of the formal system and precede any formal reasoning tasks. In an RDF-like environment, once statements and interpretations are established, only principles of reasoning act to determine their validity within a model.

3.2.3 Knowledge Representation Methods

In order for knowledge representation to achieve these objectives, it seeks to provide methods for encoding knowledge in ways that support computational reasoning.

At the root of any KR language is the specification of the logical rules that form the basis of the reasoning system. For many KR languages, these rules are based on some fragment of first-order logic that enables reasoning to be both completable and efficient. This set of rules will specify the primitive entities that will serve as the building blocks for an interpretation that represents some domain of discourse. At a basic level, the rules specify the fundamental components of propositional statements (i.e., the RDF concepts of subject, predicate, and object, or more complex statements) and the ability to designate other classes. The Resource Description Framework (RDF) is an example of this kind of KR language. Its basic mathematical rules of logic are specified by a

[1] Thanks to Karen M. Wickett for this example.

model-theoretic approach to semantics (Hayes, 2004; Halpin, 2009).

An important feature of KR languages is the need to specify a primitive entity that functions in a way that is equivalent to how proper names function in natural language. These entities perform an important function in KR languages as logical constants that allow reasoning systems to recognize when propositions are referring to the same entity. Early KR systems relied on simple terms for this purpose. While this worked well within locally controlled systems where names could be carefully managed, researchers found that this approach caused confusions when different knowledge bases were merged together. To solve this problem for the Internet, RDF uses Uniform Resource Identifiers (URIs) that can be used to refer unambiguously to the subjects of propositions and the concepts that serve as predicates (Hayes, 2004; Halpin, 2009).

Having provided the principle rules for reasoning and primitive representation concepts, it is necessary to specify the entities needed by an application domain. At the root level, most KR languages have a primitive concept of “things” or “resources” that is sufficiently general to encompass both concrete and abstract entities as objects of reasoning. For some domains, however, it is necessary to formalize our conceptualization of the world by supplying an ontology. An ontology not only allows us to indicate what kinds of things are included, but also what properties should be associated with each ontological kind. That is, an ontology may also include a specification of the vocabulary a particular reasoning system will use to express propositions about formal classes.

So far what we have been discussing are abstract constructs that are not associated with any concrete approach to encoding representations within a computational system. Because KR languages are firmly grounded in these abstract concepts, it is possible to use multiple approaches to constructing, storing, and manipulating representations that conform to a particular interpretation. A KR language, therefore, may be used with a variety of data models provided that they accurately represent the basic structures of the language and that the syntax has been mapped to the interpretation.

Because the ultimate objective is intelligent reasoning, the last part of any KR system is the ability to specify reasoning tasks that will be applied to representation instances. Again, languages like RDF provide a basic set of rules (i.e., logical entailment) that can bootstrap more complex reasoning tasks. The capabilities of reasoning, however, are always constrained by the expressivity of a particular language.

Thus a fully semantic language system requires not only syntactic representations of propositional content, but also a specification of an interpretation that provides statements with formal meanings. Such interpretations do have some correspondence to "a state-of-affairs in a world" (Hayes, 2004), but the interpretation may simplify or distort the world in the service of particular reasoning tasks. The form that knowledge representations take is not accidental. Rather, it is closely tied to a holistic reasoning system. Changing the features of a knowledge representation syntax requires making corresponding changes to the semantic model on which it is based.

3.2.4 Knowledge Representation Ontological Commitments

An important part of Kuhn's paradigms is "world view hypotheses, including basic social ontological assumptions about the part of the world studied inside the discipline" (Tjörnebohm, 1974; Hjørland, 2000). As noted above, contemporary knowledge representation languages, such as RDF, have made strong commitments to the fundamental principles of predicate logic and interpretations based in model-theory. Such commitments entail other requirements of knowledge representation languages. For example, predicate logic mainly deals with declarative propositions that can be assigned a truth value (true/false). It does not handle other kinds of knowledge (modal statements, questions, commands, belief, etc.) well.

In order to achieve reasoning objectives, certain computational constraints also affect what kinds of expressions are allowable in a KR language. For example, when specifying the fundamental logical rules for an interpretation, most KR languages only allow conjunction of propositions (this AND that). Disjunctive statements (this OR that) have proven to increase problems with the completion and decidability of reasoning. Likewise, most KR environments do not allow statements of negation (NOT). The limitations on disjunction and negation mean that the formalizations of KR languages frequently treat any statements within a knowledge base as functionally true statements.[2]

As noted above, the semantics of RDF seek to be "metaphysically and ontologically neutral" (Hayes, 2004) with regard to the entities we reason about, beyond the requirement that some entity

[2] This is not to say that statements, due to their lack of correspondence to some state of affairs, cannot be considered false. However, within an interpretation, any statements considered to be a model of the interpretation are treated as true.

has identity and can be supplied with a name. However, based on the desired reasoning tasks, the use of KR language does require accepting the constraints of the language in addition to making further ontological commitments to represent a domain:

> If, as we have argued, all representations are imperfect approximations to reality, each approximation attending to some things and ignoring others, then in selecting any representation we are in the very same act unavoidably making a set of decisions about how and what to see in the world. That is, selecting a representation means making a set of ontological commitments. The commitments are in effect a strong pair of glasses that determine what we can see, bringing some part of the world into sharp focus, at the expense of blurring other parts.
>
> These commitments and their focusing/blurring effect are not an incidental side effect of a representation choice; they are of the essence: a KR is a set of ontological commitments. It is unavoidably so because of the inevitable imperfections of representations. It is usefully so because judicious selection of commitments provides the opportunity to focus attention on aspects of the world we believe to be relevant. (Davis et al., 1993)

3.2.5 Significance for the *1:1 Principle*

Because of the formal semantics adopted by knowledge representation languages, anything that is considered to be a statement is, by definition, about one and only one thing. In addition, the use of a formal theory of reference, one that uses *names* as constants in a statement, means that any set of statements that uses the same name also describes one and only one thing. Within an interpretation, it is not possible to create a set of statements that isn't about one and only one thing. Representations that are a model of an interpretation, therefore, cannot violate a rule such as the *1:1 Principle*. This, however, does not mean that such representations would remain valid if provided different interpretations that used the same representations. In Chapter 4, we will take a closer look at how formal theories of description and reference work, and in Chapter 5 we will consider how they may lead to problems of reference different from those that have been identified by existing studies of the *1:1 Principle*.

3.3 Knowledge Organization

3.3.1 Overview

Knowledge organization is an interdisciplinary area of work within the field of library and information science. The knowledge organization paradigm is broadly concerned with the description, arrangement, and classification of "information-bearing messages in recorded form" (Svenonius, 2000).[3] For the purposes of this comparison, we are interested in the development of traditional bibliographic catalogs, rule systems, and *document surrogates*, such as MARC or Dublin Core XML *records*, and associated information retrieval systems. My treatment of the knowledge organization paradigm is informed by Elaine Svenonius's work *The Intellectual Foundation of Information Organization* (2000), as well as Biger Hjørland's numerous articles on the epistemological and philosophical foundations of knowledge organization (Hjørland, 1998, 2003, 2007).

The knowledge organization paradigm has played a significant role in the development and deployment of Dublin Core. As noted in Chapter 1, concerns about the use of Dublin Core representations for cultural heritage collections initially emerged from the library community. These concerns were informed by the fundamental objectives, principles, and rules for defining *bibliographic representation languages,* manifested as the Paris Principles, the *Anglo-American Cataloging Rules (AACR)*, MAchine Readable Cataloging (MARC), and Dublin Core XML schemas used by the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH).

3.3.2 Objective of Relevant Retrieval

The knowledge organization paradigm specifies four primary objectives for bibliographic languages (IFLA Study Group on the Functional Requirements for Bibliographic Records, 2009):

1. using the data to ***find*** materials that correspond to the user's stated search criteria (e.g., in the context of a search for all documents on a given subject, or a search for a recording issued under a particular title);

[3] I use *knowledge organization* here, understanding that there is some terminological confusion with *information organization.* Information organization may be understood as having a broader set of concerns related to "information" rather than "knowledge"; however, these terms are frequently used interchangeably in the LIS literature. There are also divisions within the community regarding what the proper concerns for information organization vs. knowledge organization should be. I use knowledge organization here without fixed connection to a single community of practice.

2. using the data retrieved to ***identify*** an entity (e.g., to confirm that the document described in a record corresponds to the document sought by the user, or to distinguish between two texts or recordings that have the same title);

3. using the data to ***select*** an entity that is appropriate to the user's needs (e.g., to select a text in a language the user understands, or to choose a version of a computer program that is compatible with the hardware and operating system available to the user);

4. using the data in order to ***acquire or obtain*** access to the entity described (e.g., to place a purchase order for a publication, to submit a request for the loan of a copy of a book in a library's collection, or to access online an electronic document stored on a remote computer).

The objectives of the IFLA Study Group on the Functional Requirements for Bibliographic Records (2009) are very closely tied to knowledge-organization research on user information-seeking tasks. The dominance of the MARC format in library information retrieval systems has focused attention on *document surrogates* that bear some relationship to the documents that they represent, but this relationship has not been grounded in a formal semantic theory. Formal approaches in knowledge organization have, instead, focused on developing *operational definitions* of user tasks that are "formulated in such as way that their achievement (or non-achievement) can be ascertained" (Svenonius, 2000). For example, a fundamental understanding of the *find* objective is represented by defining *relevance* as the measurement of *recall* (how many representations are retrieved) and *precision* (the representations retrieved meet the user's need).

Knowledge Organization Principles

Complementing the objectives of knowledge organization systems is a broad spectrum of bibliographic principles. According to Lubetsky, principles establish the "cardinal points" by which to navigate in the production of bibliographic entries in a catalog (cited in Svenonius 2000). These principles are organized into a loose hierarchy of importance, in which certain principles may be considered to override other principles. At the pinnacle of this hierarchy is the *Principle of User Convenience*, which frequently overrides such others as the *Principle of Accuracy.*

In light of our earlier observation of two senses of "principles," we may consider bibliographic principles to be *principles of guidance*:

> **principle:** A general law or rule adopted or professed as a guide to action; a settled ground or basis of conduct or practice; a fundamental motive or reason for action, esp. one consciously recognized and followed. (Oxford English Dictionary, 2011)

However, because such principles do not logically follow from one another, "it is recognized that these principles may contradict each other in specific situations and a defensible, practical solution should be taken" (Tillett & Christan, 2009). Svenonius (2000) attempts to provide a neat analysis of contemporary bibliographic principles to demonstrate that they form an important link between objectives and individual cataloging rules. However, Spanhoff (2002) finds that rules do not necessarily proceed logically from principles and objectives. "The rules were drawn up based, not on theory, but upon practices already fixed by usage.... [S]ystematizing and coordinating the rules...has resulted in bringing to light a few principles" (Pettee, 1985). The decision to promote one principle over another is part of the subjective judgement of an individual cataloger (or shared local conventions), who is able to assess local information needs and contexts.

The implications of this approach to principles can be seen in the debates around how to treat reproductions in current cataloging rules. In general, *AACR2* rules specify that the object of description should be the item itself—that is, that one should describe a reproduction as a reproduction. But because of concerns about user needs (and economics of cataloging production), the practice in many cultural heritage environments is to describe the original, with a note specifying the resource's status as a reproduction. Again, local decisions about user needs may dictate which approach is taken. Unfortunately this may lead to later ontological confusion when attempting to interpret an individual representation. "As yet the input standards themselves are not standardized, so that what is recognized as the object of description by one agency may not be so recognized by another" (Svenonius, 2000).

3.3.3 Knowledge Organization Methods

Knowledge organization has built up a large body of practice regarding the organization of information. This body of practice supplies a variety of methods. At the forefront of these developments has been the specification of user tasks, which catalog systems are intended to address as described above. The specification of methods to achieve those objectives, in Svenonius's view (2000), is further refined through the application of bibliographic principles.

These objectives and principles inform the creation of a set of rules that guide the production of bibliographic entries. Such rules are intended to be interpreted by human agents who examine a document and extract from it the relevant features that the document's surrogate needs to represent. For bibliographic materials, the core set of rules is defined by the *Anglo-American Cataloging Rules* (American Library Association et al., 1988). However, a variety of related rule sets also exist. Which rules are applied to create a specific description may depend on the genre and format of materials being described, the nature of the collecting institution (i.e., public library or academic library), or the technical environment in play (variation among rules by cataloging service). Specific domains have also adopted the rule-based approach of *AACR2* to produce guidelines that are parallel with, or extend, *AACR2*. For example, the museum and visual resources communities developed *Cataloging Cultural Objects* (Baca et al., 2006) and archives use *Describing Archives: A Content Standard* (Society of American Archivists & Hensen, 2004).

From the perspective of knowledge organization professionals, these rule systems provide a kind of *semantics* that make bibliographic entities meaningful. However, according to Hjørland (2007), information science "has not yet addressed semantics in a systematic fashion," leaving the field "very fragmented and without a proper theoretical basis." Where work has been done, it has largely focused on the meaning of classification systems and thesauri (Khoo & Na, 2006; Hjørland, 2007). Because relevance is defined as a relationship between a document surrogate and a query, interpreting the meaning of searchers' queries has also received attention under the rubric of semantics (Hjørland, 1998; Blair, 2003). However, the interest in users' conceptualizations of information needs has led KO researchers towards Wittgenstein's later theories of semantics as "language games." In this theory of semantics, words do not have a definite meaning, but may change their meaning when used in different sentences or different contexts. The "language games" theory has proven useful for understanding how searchers use terms in a query and how retrieval systems might mediate their terms with those used in a representation (Hjørland, 1998; Blair, 2003).

Knowledge organization research has also focused on the semantics of classification systems and taxonomies of terms. For example, Svenonius's treatment of bibliographic languages (2000) only deals with the relational, referential, and categorical semantics found in organization vocabularies. While the focus on query language has adopted a language-game orientation, the bases for thesauri have focused more heavily on Wittgenstein's earlier "picture theory" of meaning, whereby terms have a fixed definite meaning. Rules for the construction of taxonomies therefore seek to disambiguate homonyms and polysemes through qualifiers that make them unique terms.

In contrast to KR languages, KO representation syntaxes, such as MARC or various flavors of XML encoding, have not invested heavily in understanding the formal semantics that underlie their representations. For the objective of retrieval of relevant document surrogates, KO's focus has been more heavily weighted towards the concerns listed above. Representation syntaxes for document surrogates therefore are largely based on available data models that also enable efficient storage and presentation (especially presentation that reifies practices found in cataloging codes—e.g., requirements for ISBD punctuation, etc.). Research on KO representation syntaxes demonstrates that they often lack a strong data model and that they intertwine representations of various entities into a single document surrogate (Attig, 1989; Heaney, 1995; Fattahi, 1997; Thomale, 2010). Although we may think of production rules such as *AACR2* as supplying "meaning" to these representations (e.g., Tillett 2003), this uses a different sense of "meaning" than that supplied by the formal semantics in KR languages.[4] Colloquial understandings of "semantic" as "meaningful" also inform studies of metadata quality which focus on definitional issues and confusions of metadata elements, not overall grammars (Park, 2002; Hutt & Riley, 2005; Park, 2005; Park & Childress, 2009). As a whole these studies illustrate Hjørland's sense (2007) that our understanding of semantics is deeply fragmented.

It is also important to recognize that many of these syntaxes are designed for the transport or display of document surrogates and may not be directly used in the course of performing particular information retrieval tasks. As noted by the recent report from the Library of Congress, *A Bibliographic Framework for the Digital Age,* MARC "was seldom used by systems as an internal format" (Library of Congress, 2011). Instead, processes for extracting relevant information from document surrogates are used to create indexes, etc., to which information retrieval queries are

[4] This attitude also appeared as an undercurrent in Dublin Core listserv discussions. Dublin Core and RDF were seen as lacking semantics because they did not have a similar set of rules.

actually applied. If the application of information retrieval/relevance algorithms is the kind of intelligent reasoning used within KO, it is reasoning that takes place one step removed from the document surrogate itself. Because such indexes rely on the use of extracted terms, this requires the kind of relational semantics discussed by Svenonius (2000, 2004).

3.3.4 The Ontological Commitments of Knowledge Organization

The strongest commitment that knowledge organization makes is towards the relationship between a user's information need and some document that exists, one that may be represented for the purposes of searching by a document surrogate. While codes of rules guide the production of such surrogates, there is not an expectation that the *meaning* of records will be interpreted by machine. To date, these commitments have entailed a dedication to document-like representation surrogates with fixed syntaxes (e.g., the MARC format). By making relevance the prime objective, knowledge organization has also committed itself to the assumption of a closed world. For example, relevance judgements assume that if a relevant document exists, it will have a representation in the system. If a valid query does not retrieve any documents, it is inferred that no relevant document exists. This is the opposite of the open world assumption which enables one only to confirm the existence of a representation. In the absence of a representation, it is not possible to assume that something doesn't exist. A closed world is a necessary commitment of most retrieval systems.

Rule systems for descriptive cataloging are frequently organized around ontological distinctions among the types of materials being described. *AACR2*'s *Part 1: Description* divides the bibliographic universe into eleven distinct categories, each with its own rules that may change the meaning and/or syntax of a MARC instance. The distinctions are based on the *format* of materials (e.g., printed materials, electronic resources, microfilms), type of artistic content (e.g., moving images, graphic materials, sound recordings), or mode of issuance (continuing resources, serials, etc.) (American Library Association et al., 1988). For example, if a cataloger decides that the item-at-hand is a *reproduction*, it requires the addition of a "reproduction note" (designated by the `533` tag). While *AACR2* provides definitions that describe the kinds of materials that fit these categories, there is neither a formal representation of these ontological classes nor a mechanism to directly relate a particular description to these classes of materials (e.g., to know that the item-at-hand is a reproduction). A cataloger knowledgeable about the rules may infer from the appearance of a 533 "reproduction note" that the item described is a reproduction. However, in other cases, inferring the class of materials from patterns in the representation may be unreliable.[5] Decisions regarding which set of cataloging rules to apply to a description of a particular resource are not universally consistent. A cataloger may choose to apply different rules depending on the context of the collection or purpose of the description, following high-level principles such as the *Principle of User Convenience*. The same kind of object may be treated differently by different libraries, within different systems, or when represented in different syntaxes.

An ongoing development within knowledge organization has been the refinement of other ontological entities inherent in bibliographic languages, currently represented by the *Functional Requirements for Bibliographic Records (FRBR)* (IFLA Study Group on the Functional Requirements for Bibliographic Records, 2009). FRBR distinguishes between four primary entities: works, expressions, manifestations, and items. While the recognition of these entities has helped efforts to rewrite cataloging rules, FRBR's development has mostly been independent of formal ontologies found in KR. Much of the work on FRBR has focused on the vertical relationships between primary entities. Horizontal relationships, such as those that involve equivalent, derivative, or descriptive relationships, are less well understood and modeled (Tillett, 2001; Knowlton, 2009). While FRBR may represent an important step towards recognizing entities that are implicit in cataloging codes, it remains an essentially ontological distinction that is divorced from a formal semantics for document surrogates.

A further complication in knowledge organization, broadly construed, is that different domains also have distinct ontological commitments regarding different kinds of entities. This is especially true in regard to how libraries, archives, and museums define which entities may be considered to be "reproductions" or what the properties of reproductions are. While many approaches to the *1:1 Principle* identified by Han et al. (2009) seem rooted in the facsimile theory (whereby a representation is about the original, with notes about the reproduction), museum and visual resource rules treat images as distinctly different resources. For example:

> In its least complex form, the surrogate might be compared to the bibliographic equiv-

[5] For example, understanding the implications of certain codes that appear as part of a 007 field.

> alent to the reproduction. . . . Many visual documents, however, are not copies. With three-dimensional views of art, vantage point alone turns the visual document into something other than just a copy in another form. (Most, 1998)

This attitude towards the relationships between works of art and surrogate images is an important ontological distinction made by XML syntaxes based on the *Categories for the Description of Works of Art (CDWA)* (Harpring & Baca, 2009) and *VRACore* (Visual Resources Association, 2007). These ontological differences in how to treat conceptually different kinds of entities are mapped onto common syntaxes, such as Dublin Core. This likely contributes to semantic confusion.

3.3.5 Significance for the *1:1 Principle*

Traditionally, knowledge organization principles have been viewed as a means of ensuring the primary objective of a KO system—to provide the user with materials relevant to their information needs. Because of this overriding principle, approaches to descriptive cataloging have been willing to accept representation models that more accurately reflect a user's expectations than an actual state of affairs. This, at least, has been a primary motivation for developing a solution to the *multiple versions problem* and specifications for describing reproductions. While we may understand cataloging rule systems as meaningful guides for producing records, the relationship between the rules and the representations is indirect—so it is not possible to know which set of rules to use when interpreting a record. While this may not pose a challenge for those steeped in the production rules, their meaning may be opaque to other users of a representation. Because the meaning of a representation is distinct from its role in an information retrieval process, however, it may succeed in communicating information that is relevant to a user's needs.

3.4 Conclusion

In Chapter 2 my account of the origins of the *1:1 Principle* revealed that there were multiple views of why such a principle was needed and how it should be implemented within the Dublin Core Metadata Initiative. This chapter introduced two representation paradigms that present alternative accounts of how representation languages work. Both are important to understanding the distinctions raised in DCMI conversations about the *Principle.*

From the perspective of the knowledge representation paradigm, formal semantics that use names to refer to resources provide the mechanism needed to make metadata statements about one, and only one, resource. Because these names are logical constants in a mathematical graph, statements organize themselves into coherent, logical clusters of metadata. Whether a set of statements consists of coherent propositions or not will depend on the formal interpretation used to evaluate those statements.

From the perspective of knowledge organization, the *1:1 Principle* is one among many competing and inconsistent principles metadata creators must balance. In the case of cultural heritage repositories, records that satisfy the needs of users seeking cultural heritage resources have been afforded a higher priority than records that are logically coherent. The *1:1 Principle* is also in direct conflict with rule interpretations for "reproductions" that specifically require the creation of records that are never about one, and only one, thing. Metadata creators are forced to make this choice, in part, because of constraints imposed by the knowledge organization paradigm's conceptualization of "document surrogates" with a fixed syntax, but informal semantics.

The identification of the paradigmatic divide has significant implications for the identification of *1:1 Principle* violations. Existing studies have relied on a colloquial understanding of the *Principle* applied using qualitative methods. How colloquial heuristics of metadata quality should be formalized remains an open question. However, before we can understand violations, it is necessary to paint a clearer picture of how representations are "about" anything, let alone one and only one thing. In the following chapter, we will explore how the knowledge representation domain has formalized theories of description and reference that are lacking in the knowledge organization domain.

Chapter 4

Description and Reference for Cultural Heritage Resources

> Description is revelation. It is not
> The thing described, nor false facsimile.
>
> It is an artificial thing that exists,
> In its own seeming, plainly visible,
>
> *Description without Place* (Stevens, 1954)

4.1 Introduction

Libraries, archives, and museums holding cultural heritage resources commonly characterize their role as not just storing artifacts, but providing *meaning* and context for the materials in their collections. One of the ways they accomplish this is by making *descriptions* of those artifacts available. However, the preceding analysis of the *1:1 Principle* now raises an important question—do we really understand the formal nature of how descriptions of cultural heritage objects succeed in describing? Understanding and developing formal theories of description and reference has been a major theme in research in linguistics, the philosophy of language, and, more recently, computer science. Despite the knowledge organization paradigm's core interest in description, the literature on the relationship between bibliographic description and formal theories of description is extraordinarily thin (Smith, 1976; Blair, 2003). The prevalence of apparent failure to comply with the *1:1 Principle,* and the difficulties of understanding the nature of those failures, reveals the dangers of ignoring research on formal theories of description. This is especially true as the libraries, archives, and museums (LAM) community considers replacing existing descriptive syntaxes with representation languages grounded in formal semantics.

This chapter will explore the development of theories of description and reference in philosophical circles and how these theories became necessary for computational reasoning using semantic knowledge representations. So far, bibliographic languages have succeeded while using less formal approaches. In some cases this has been *because* they have used less formal approaches that accommodate necessary ambiguities inherent in the cultural heritage domain. However, new expectations and opportunities are inexorably moving description practices in the direction of logically explicit, computer-processable formalizations. The *1:1 Principle* problems observed by metadata-quality researchers demonstrate the implications for translating these colloquial languages into more formal representations.

4.2 Theories of Description and Reference

Over the course of the twentieth century, linguists and philosophers of language have investigated approaches to formalizing our natural understanding of sentences—that is, to formalizing the *semantics* of natural language. At the heart of this endeavor are theories of *description* and *reference* (whether through the use of *descriptions* or *proper names*). While we routinely use descriptive sentences in everyday life to successfully communicate meaning, these investigations have found that it is surprisingly difficult to provide a formal account of how communication actually takes place. The notions of reference, description, and meaning provide many enduring puzzles. Because this work took place within communities that took a formal approach to these problems, emphasizing symbolic logic as both a model and a method, it influenced the artificial intelligence community's efforts to represent our knowledge of the world in languages that would support computer-based reasoning systems. These theories of description and reference, along with portions of mathematical logic, became the important parts of the background for the development of *formal semantics* for knowledge representation languages.

4.2.1 Using Logical Definite Descriptions to Refer

In response to several puzzles about how *definite descriptions* (such as "the author of *Waverley* is Scotch"[1]) contribute to the meaning of the sentences in which they occur, Bertrand Russell (1905)

[1] *Pace* to readers who are Scots. "Scotch" is the classic example used by Russell in his writings.

offered the following analysis:[2]

$$\text{The author of } \textit{Waverley} \text{ is Scotch.} \tag{4.1}$$

is to be understood as asserting:

$$\text{At least one person authored } \textit{Waverley} \tag{4.2a}$$

$$\text{At most one person authored } \textit{Waverley} \tag{4.2b}$$

$$\text{Whoever authored } \textit{Waverley} \text{ is Scotch} \tag{4.2c}$$

or more formally expressed in first-order logic:

$$\exists x(F(x) \wedge \forall y(F(y) \rightarrow x = y) \wedge G(x)) \tag{4.2d}$$

On this account, the referent of the definite description "The author of *Waverley*" is whoever is the sole author of *Waverley,* if there is a sole author of *Waverley.* If there is no sole author of *Waverley*, then the definite description has no referent. The sentence "The author of *Waverley* is Scotch" is true if there is a sole author of *Waverley* and that person is a Scot, and false otherwise. So if there is no author of *Waverley,* or there is more than one, then the description "The author of *Waverley*" has no referent and the sentence "The author of *Waverley* is Scotch" is false. Note that neither having a truth value nor having meaning requires that the description has a referent.

4.2.2 Using Names to Reference

Mona Lisa, Leonardo da Vinci, and *Herman Melville* are familiar proper names that we seem to use effortlessly and meaningfully to refer to things in the world. But how does this reference actually take place? Russell held that corresponding to every meaningful proper name was a definite description which fixed the referent of the name (a view also held by Göttlob Frege). The definite description would then in turn be analyzed as described above. Approaches of this sort, as exemplified by Russell and Frege, have been referred to as *descriptivist* theories of proper names.

[2] *On Denoting* is considered by many philosophers to be one of the most influential philosophy articles of the 20th century

However, Saul Kripke (1980) has convincingly challenged the *descriptivist* view of names as surrogates for descriptions. Among other things, Kripke points out that we seldom seem to be able to provide a reliable definite description of the sort Russell's account demands for the proper names that in fact we successfully use. After all, many people with hardly any knowledge of Renaissance art history, or any definite sense of who Leonardo was, manage to refer to him successfully using his name ("da Vinci was a famous general, I believe." "No, he was not, he was an artist."). Here it seems that *both* participants are referring successfully to Leonardo da Vinci, although the first may well have no knowledge of the sort that could be expressed in a typical definite description that would actually uniquely pick out Leonardo da Vinci and no one else.

Kripke offers an alternative *causal-historical* approach, whereby names refer to their referents through a causal chain of verbal communication and behavior that typically extends back to an original immediate encounter with the named entity, often in what Kripke calls a *baptismal event*. In this way, I can, after four hundred years, use the name "Leonardo da Vinci" to refer directly to a particular individual without relying on a definite description to pick him out uniquely. Whether a name refers to an entity depends on the causal chain through which I received the name, not whether it satisfies some particular description I am using to fix its referent.

There are variations on these two theories of how descriptions and names refer. However, they serve as the central influential approaches to reference used in contemporary knowledge representation languages on the Web.

4.3 Formal Semantic Representation Languages

Efforts to build formal knowledge representation languages emerged from the larger efforts to develop methods of intelligent reasoning for artificial intelligence systems. According to the proponents of knowledge representation within AI, in order to engage in intelligent reasoning, artificial agents needed some representation of a world. This attitude distinguishes KR from other *procedural* approaches to artificial intelligence (Davis et al., 1993; Sowa, 2000).

The agenda of the artificial intelligence community ran into a number of hard problems. First, in order to reason successfully about the real world, a very large amount of knowledge needs to be represented. This was the motivation behind projects such as Cyc, which represents one of the largest knowledge bases of basic facts (Lenat & Guha, 1990). Second, reasoning over this information was limited by hard problems in the scalability of reasoning systems. And finally, the limited expressiveness of first-order logic is challenged by the representation demands of many common, modal, adverbial, and intensional notions. However, as noted in Chapter 2, research on these problems informed approaches to describing resources on the Web, through the Warwick Framework, Platform for Internet Content Selection, and Meta Content Framework. These streams came together under the rubric of the Resource Description Framework. RDF and the work of the knowledge representation paradigm were also the foundations for Tim Berners-Lee et al.'s thinking about the Semantic Web (2001).

The Resource Description Framework represented a departure from the document-like structures that were being discussed in the early development of Dublin Core. Although XML may be understood as a kind of directed graph (a tree), it affords developers many different ways to represent the same kind of assertions. That is, I may define several different XML schemas that represent a structured document containing information about your postal address. By examining the XML graph, parsers are able to answer questions about the document itself: e.g., which elements are present in the document, which are contained within other elements, etc. However, "there are in general a large number of ways in which the XML maps onto the logical tree," creating a great deal of complexity (Berners-Lee, 1998, 2009). RDF may also be represented as a directed graph; however, this reverses the logical relationship with XML. RDF's model represents a very simple logical structure that can be represented in many different ways in XML. Regardless of the particular XML syntax (or non-XML syntax), it results in the same RDF graph (Berners-Lee, 1998, 2009).

```
<rdf:Description rdf:about="http://louvre.fr/works/779#">
        <dcterms:title>Mona Lisa</dcterms:title>
</rdf:Description>
```

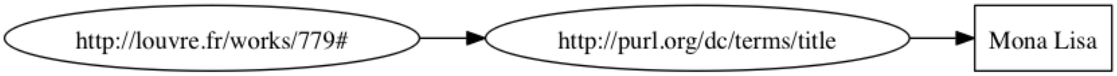


Figure 4.1: Example of an RDF Statement and the graph it represents

An important part of logical languages, however, is the use of logical constants or names that refer to the entity an assertion is about. Early knowledge representation systems each included their own method of tracking referents within the system, often using natural language terms (Lenat & Guha, 1990). However, when knowledge representation researchers attempted to use these constants outside of local systems (e.g., by merging two knowledge bases) natural language terms that represented different concepts could collide. The Semantic Web and RDF solve this problem by using Uniform Resource Identifiers (URIs)[3] as names for resources, defined as "anything that has identity. Familiar examples include an electronic document, an image, a service (e.g., 'today's weather report for Los Angeles'[4]), and a collection of other resources" (Halpin, 2009). When used as logical constants by the RDF model-theory (Hayes, 2004), URIs function as names that denote and appear as nodes in the RDF graph model. In a simple XML representation of RDF assertions (a triple consisting of a URI subject, URI predicate, and some literal or non-literal value, as in Figure 4.1 and Example A.4), it is not necessary to organize assertions into a hierarchical record structure. Instead, the abstract graph that the XML syntax represents can organize nodes around URIs. When the DCMI community contributed to the emergence of RDF, it was these kinds of "logical clusters of metadata" its members had in mind (Lagoze, 1996, 1997; Weibel & Hakala, 1998; Bearman et al., 1999).

As conceived by Hayes (2004), URIs function much like Russell's account of names, with the important requirement that a formal *interpretation* is required that specifies what a name denotes. However, Hayes's approach is not accepted by Berners-Lee (2002) and some other members of the Semantic Web community. Instead, Berners-Lee prefers an account which is closer to Kripke's *causal-historical* approach to names (Berners-Lee, 2002; Hayes & Halpin, 2008; Halpin, 2011). Although this is a debate about how URIs come to denote resources, it does not change their fundamental role as constants within logical statements in RDF. In the following section, we explore how traditional approaches to bibliographic description take a very different approach to developing descriptions.

[3] http://tools.ietf.org/html/rfc3986

[4] http://tools.ietf.org/html/rfc2396

4.4 Bibliographic Description

In the sections above we have demonstrated how knowledge representation languages, like RDF, have a formalized theory of description and reference. While we also talk about bibliographic records as "describing" resources, it is unclear whether we can apply these theories directly to existing descriptions. In this section we explore the nature of bibliographic representations in relation to more formal approaches to description. As we saw in Chapter 2, both the MARC format and XML syntaxes have been used in similar ways by the LAM community. We find that additional work will be needed to supply implicit semantics to these formal syntaxes, especially in cases where a record refers to multiple versions or reproductions of resources.

Because the *1:1 Principle* seems to be concerned primarily with distinguishing between different "manifestations or versions" of resources, my discussion here focuses on the practice of *descriptive cataloging* as opposed to subject cataloging or indexing. Definitions of the objectives and roles of descriptive cataloging within the process of bibliographic control contain informal understandings of how description works and its relationship to bibliographic syntaxes.

> Identification and description are interrelated processes in descriptive cataloging. Identification consists of the choice of conventional elements, guided by a set of rules based on agreed-upon international standards. When the cataloger has properly identified the conventional elements, they are described in a catalog record in such a fashion that the description is unique and can be applied to no other entity in the collection. In other words, each resource should be distinguished from everything with which it could be confused. (Taylor & Joudrey, 2010)

The international standards, such as the *International Standard for Bibliographic Description (ISBD)* mentioned by Taylor & Joudrey (2010), are primarily oriented towards the production of uniform, consistent bibliographic records that may be shared across cataloging utilities (Rodriguez, 2010). To ensure consistency they prescribe standards for punctuation, abbreviations, what sources of evidence could be used for description, and the overall arrangement of bibliographic elements in a record. Because the focus is on the exchange of information, ISBD and other descriptive cataloging standards are mostly concerned with *manifestation*-level properties of resources. For general library collections, "[t]he detailed physical description has always been criticized as not necessary and less economic, but it is important information for library activities such as storage, circulation, and conservation" (Rodriguez, 2010).

Descriptive cataloging is tied most directly to the *identify* and *select* functions of a catalog. First, it allows users searching for known items to discover whether the resource is available within a particular bibliographic utility. If multiple entries match this initial request, a description should be sufficiently precise to allow the user to select a resource that best meets their needs from among the available manifestations (Taylor & Joudrey, 2010).

4.4.1 Records and Bibliographic Surrogates

Catalogs represented an important innovation in the early development of the library. Rather than relying on the organization of the physical collection, the catalog could allow librarians and users to navigate the collection according to different intellectual facets. Ultimately, however, the catalog needed to provide a relationship between the catalog representation and the object sought by the user (i.e., the *obtain* function). However, because the catalog was seen as a substitute for the actual collection, bibliographic records are frequently referred to as "surrogates" (Svenonius, 1989, 2000, 2004; Greenberg, 2010; Rodriguez, 2010).

> In other words, bibliographic control is the process of describing information resources and providing name, title, and subject access to the descriptions, resulting in records that serve as surrogates for the actual items of recorded information. These surrogate records (sometimes called *entries*, *bibliographic records*, or simply *metadata)* are then placed into information retrieval tools, where the records act as pointers to the actual information resources. The comprehensive descriptions found in the records provide users with enough information to determine the potential value of the resources without actually having to view the items directly. Surrogate records are stored in a variety of retrieval tools including bibliographies, catalogs, indexes, finding aids, museum registers, bibliographic databases, and search engines. (Taylor & Joudrey, 2010)

Among the various definitions of bibliographic records and their function within cataloging systems, I have been unable to identify any associations with the formal theories of description and reference that are so closely associated with knowledge representation approaches. As noted in Chapter 3, the information retrieval literature has been more interested in Wittgenstein's later theories of language games (Blair, 1992; Svenonius, 2000; Blair, 2003; Svenonius, 2004). While language games may be useful in determining what resources (or representations of resources) are relevant to a user's query, they have not been sufficiently applied to explain the relationship between a representation and the resource to which it refers. While discussions of the semantics of thesauri and controlled vocabularies do take up concepts of referential semantics, these are limited to discussions about the relationships between terms and the classes of resources they represent (Svenonius, 2000; Khoo & Na, 2006; Hjørland, 2007).

The lack of a clear theory of description and reference has shaped the development of cataloging codes, functional requirements, and syntaxes used in the knowledge organization domain. The direct re-application of these rules and expectations for record syntaxes to metadata appears to be complicit in the confusions surrounding the *1:1 Principle*. In the following sections, I will show that current syntaxes lack a relationship to formal data or semantic models and will explore the difficulties of associating these syntaxes with existing theories of description and reference.

4.4.2 Data Model or Document Model

The original intent of the MARC format was to support the transport of records between cataloging systems using linear storage technologies such as punch cards or magnetic tapes. The format was not originally designed to represent a model of the bibliographic universe, but rather to enable the production of physical card catalogs or card-like online catalog displays. It provides "a mechanism for organizing and *labeling the data* in a record [emph. mine]" (Attig, 1989). "ISO 2709 [MARC] only specifies tags as three character numeric, but does not designate the semantics of any tag number" (McCallum, 2010). MARC's principle orientation is towards preserving the conventions in existing standards such as the *ISBD* and the *Anglo-American Cataloging Rules* (Spanhoff, 2002). As a data storage model developed in the mid-1960s, MARC precedes all of the major advances that introduced formalizations that explicitly identified and related formal semantics to data representations.[5]

[5] Such as Codd's Relational Database Theory (1970) or Chen's Entity-Relationship models (1976).

Rather than considering MARC as encoding formal semantic structures, in the vein of knowledge representation languages, it is more appropriate to treat it as a markup language for bibliographic *documents* (Jul, 2009). The MARC format lacks a formal data model and is not associated with any unambiguous formal semantic interpretation (Heaney, 1995; Thomale, 2010). The difficulties of attempts to directly translate MARC's informal syntax into more formal representations demonstrate its weaknesses (Coyle, 2011).

Furthermore, the move from MARC to established, syntactically well-defined markup languages, such as XML, does not necessarily ensure the sort of explicit semantics associated with knowledge representation languages.

> . . . even though SGML/XML is thought of as providing access to a document's meaningful structure, current SGML/XML methods cannot represent the fundamental semantic relationships amongst document components and features in a systematic machine-processable way. . . . The result is that SGML/XML markup language users must guess at the semantic relationships the markup language designer had in mind, but no way to formally express. (Renear et al., 2002)

For both MARC records and Dublin Core XML representations, developing an interpretation of the intended semantics often requires an understanding of context within the document. "In MARC, punctuation that appears in one subfield can subtly change the meaning of data in a different subfield; changing subfield order can also subtly change the data's meaning" (Thomale, 2010). When documents are translated into different syntactic schemas or interpreted in different contexts, the meaning found in original documents and contexts may be lost. The problem can be particularly acute when polysemes and synonyms are introduced across schemas (Renear et al., 2000; Park, 2002; Dubin et al., 2003; Park, 2005).

4.4.3 Understanding Descriptive Documents

As document markup, bibliographic representations differ little from the markup of other kinds of documents—for example, an HTML page, a Text Encoding Initiative (TEI) representation of a piece of correspondence, etc. In this sense, a bibliographic record can be seen as a *surrogate document* that "stands in" for an actual resource in a knowledge organization system (whether that representation is an analog card catalog or an electronic file). When we ask what a record is "about," it may be in the same sense that *Moby Dick* is "about" *whaling* (what Ryle (1933) calls "about-linguistic") (Wilson, 1968). The question of whether a bibliographic document must always *refer to* one, and only one, resource seems unclear from the knowledge organization literature (i.e., as a Russellian *definite description* or what Ryle (1933) calls "about-referential").

The problems surrounding the description of reproductions and multiple versions highlight the problematic nature of assuming that *all records* function in this manner. The current recommendations remain divided between *facsimile theory* and *edition theory* approaches that designate what the *primary* resource is that is being described. However, both approaches appear to be referring in different ways. For example, consider what each approach suggests in terms of a definite description, along the lines of Russell's "The $\mathcal{F}$ is $\mathcal{G}$"

facsimile theory: The resource entitled *Moby Dick* and written by Herman Melville in 1851 ***is reproduced by*** a PDF file

edition theory: A PDF file ***reproduces*** the resource entitled *Moby Dick* and written by Herman Melville in 1851.

Current fixed syntaxes require catalogers to choose one or the other approach. This choice is influenced by a number of factors, such as user expectations and the type of collection being described (e.g., collections that consist wholly of reproductions prefer the facsimile approach). While the syntaxes of bibliographic records are fixed, their implicit semantics allow them to be used in either way. In aggregation environments, sophisticated rules would be necessary to understand which of the two approaches was used in each aggregated record. This may be possible for syntaxes like MARC that are sufficiently granular; but in simple Dublin Core, the subtle distinctions in the meaning of elements are lost.

Additional theoretical work is needed to help understand whether formal theories of description are in operation at a different level of representation within bibliographic systems (e.g., the indexes, etc. derived from bibliographic records). It is unclear from this investigation whether it will be possible to supply bibliographic records with the kinds of semantics implied by these theories.

4.5 Conclusion

In this chapter I have identified two fundamental theories of description and reference that are the basis of knowledge representation semantics. While these theories are not untroubled by hard philosophical problems, knowledgeable members of the KR community can defend their positions based on well-defined philosophical debates. These theories are not just theories of descriptions which are informative, but rather descriptions that uniquely refer to one, and only one, thing. As the theoretical basis for knowledge representation semantics, this translates into formal descriptions that are logically sound references to resources. In the case of RDF, reference is established through the use of URIs as names. Any properties that are attached to one such name are therefore a description of the resource referred to by the URI.

In contrast, knowledge organization practices lack a formal theory of description, substituting an informal understanding of "surrogates" that stand in for resources. Yet, definitions of records suggest that there is an expectation for records to "point at" the resources for which they are surrogates. However, formal theories of description do not appear to be part of emerging conceptual models in this area. These attitudes towards surrogates have persisted across several technological shifts, from cards, to MARC, and onward to conceptions of metadata. Although the *1:1 Principle* attempts to confront these attitudes, the relationship between the *Principle* and fundamental theories of description and reference is obscured in Dublin Core documentation oriented towards knowledge organization practices.

The identification of the formal semantics and formal theories of description and reference in the knowledge representation paradigm appears to provide a grounding for detecting *1:1 Principle* violations. However, in this chapter we have also found that knowledge organization practices do not share these same assumptions and formal models. In Chapter 5, I will explore the implications these theories have for existing methods of detecting *1:1 Principle* violations that appear in cultural heritage Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) repositories.

Chapter 5

Revisiting *1:1 Principle* Violations

> It is perhaps worth while saying that semantics as it is conceived in this paper (and in former papers of the author) is a sober and modest discipline which has no pretensions of being a universal patent-medicine for all the ills and diseases of mankind, whether imaginary or real. You will not find in semantics any remedy for decayed teeth or illusions of grandeur or class conflicts. Nor is semantics a device for establishing that everyone except the speaker and his friends is speaking nonsense. (Tarski, 1944)

5.1 Introduction

An objective that both knowledge organization languages and contemporary Semantic Web knowledge representation languages share is the ability for systems to usefully exchange and use representations. Within the knowledge organization domain, emphasis has been placed on creating interoperable metadata that conforms to community standards of quality. Studies of metadata quality have led KO researchers to discover that cultural heritage metadata frequently fails to conform to their interpretations of the Dublin Core *1:1 Principle*.

However, because definitions of interoperability and quality are broadly defined, operationalizing our understandings often depends on the framework of a paradigm. This chapter examines how contemporary metadata quality research, based in the knowledge organization paradigm, has developed techniques to identify *1:1 Principle* violations. However, when viewed through the lens of knowledge representation, I find that the differences between paradigms make formalizing KO accounts of *1:1 Principle* violations into KR languages difficult, if not impossible.

5.2 Metadata Interoperability and Quality

An important objective for knowledge organization standards has been to enable the distribution of representations among individual systems and cataloging environments. Standardization for the purposes of record exchange has been an important area of work in the library community, from the launch of the Library of Congress's card distribution program through contemporary metadata exchange protocols, including the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) (Svenonius, 2000). Interoperability not only requires a standardization of formats, but also a commitment to community-defined standards of quality.

5.2.1 Interoperability

Haslhofer & Klas (2010) cite metadata interoperability as "a prerequisite for uniform access to media objects in multiple autonomous and heterogenous information systems." Interoperability is a broadly defined concept that ranges from low-level capabilities of computer systems to understand files through higher-level abilities to preserve the meaning of the information that is exchanged. The DCMI also provides a hierarchy of interoperability levels for users of Dublin Core, as shown in Table 5.1 (Nilsson et al., 2009).

4	**Description Set Profile interoperability** • Shared formal vocabularies and constraints in records
3	**Description Set syntactic interoperability** • Shared formal vocabularies in exchangeable records
2	**Formal semantic interoperability** • Shared vocabularies based on formal semantics
1	**Shared term definitions** • Shared vocabularies defined in natural language

Table 5.1: DCMI Interoperability Levels based on Nilsson et al. (2009)

Because many of the studies about the *1:1 Principle* use metadata exchanged through OAI-PMH, it is also useful to understand how the protocol specified a "low-barrier approach to enable metadata harvesting" (Van de Sompel & Lagoze, 2000). OAI-PMH provides two kinds of interoperability:

- syntax: an XML schema that provides a container architecture for descriptive metadata.
- transactional: a communication protocol that allows metadata providers and metadata harvesters to engage in exchange transactions.

Because the developers of OAI-PMH intended it to be useful across many different kinds of repositories and genres of materials, the container architecture allows a repository to include domain-specific metadata in a container. That is, OAI-PMH allows repositories to define both the vocabularies and the semantics for locally relevant metadata. To ensure a baseline of syntactic interoperability across all repositories, OAI-PMH does require a simple set of Dublin Core properties that provides a shared vocabulary of terms (Lagoze et al., 2008)—i.e., Level 1 in the hierarchy in Nilsson et al. (2009). The focus of OAI-PMH, therefore, is on enabling the successful transfer of a metadata package from one repository to another. The preservation of the meaning of the metadata exchanged in this way depends on outside specifications. These include Dublin Core and other community-specific standards that provide guidelines, rules, and vocabulary definitions. As we have seen in earlier chapters, how that meaning is supplied varies between paradigms. In knowledge organization, a representation's meaning is bound to human-readable rule systems that are not represented formally. In order to evaluate a representation, it is therefore necessary to define operational definitions of quality.

5.2.2 Quality

How to define metadata quality has been the subject of ongoing effort within the digital library community. Park (2009) provides a comprehensive account of recent trends. Relevant to our discussion about the *1:1 Principle* is a series of quality-centered investigations that examined OAI-PMH metadata from cultural heritage aggregations (Bruce & Hillmann, 2002; Dushay & Hillmann, 2003; Hillmann et al., 2004; Stvilia et al., 2004; Hutt & Riley, 2005; Shreeves et al., 2005, 2006; Hillmann, 2008; Han et al., 2009).

Intrinsic Quality	**Relational Quality**
Accuracy/Validity	Accuracy
Cohesiveness	Completeness
Complexity	Complexity
Semantic Consistency	Latency/Speed
Structural Consistency	Naturalness
Currency	Informativeness
Informativeness	Relevance (Aboutness)
Naturalness	Precision
Precision	Security
	Verifiability
Reputational Quality	Volatility
Authority	

Table 5.2: Information Quality Categories & Dimensions from Shreeves et al. (2005)

Stvilia et al. (2004, 2007) and Shreeves et al. (2005) provide a framework for metadata quality that is based on features shown in Table 5.2.2. These attempt to measure the "fitness for use" of metadata. These features can be organized in three categories: intrinsic information quality (IQ), relational/contextual quality (RQ), and Reputational Quality. Intrinsic information quality can be objectively measured by assessing "attributes of information items themselves in relation to a reference standard, such as spelling mistakes and conformance to a date encoding standard" (Shreeves et al., 2005). Attributes of relational/contextual quality:

> measure how well an information object reflects some external condition (e.g., actual accuracy of addresses in an address database). Since metadata records are surrogates for information objects, many relational dimensions apply in measuring metadata quality. (Shreeves et al., 2005)

For Stvilia et al. (2004, 2007) and Shreeves et al. (2005), these "uses" of metadata are the traditional *find, identify, select,* and *obtain* objectives found in the knowledge organization paradigm. The framework used by Stvilia et al. (2004) measured the impact of poor-quality metadata on both the *find* activity (i.e., inaccurate metadata would prevent users from finding objects they needed) and the *obtain* activity (i.e., incorrect metadata could prevent a user from obtaining an item, for example if an invalid URL were provided).

Part of the ability of metadata to meet these needs is its *semantic consistency* or "the extent to which the collections use the same values (vocabulary control) and elements for conveying the same concepts and meanings throughout" (Shreeves et al., 2005). Such semantic consistency is also important for the interoperability of metadata in aggregated environments, where automated processes may be applied to normalize metadata to prepare it for use with metadata from other repositories (Shreeves et al., 2005). This sense of *semantics* is based on the natural language definitions provided by standards documentation, not on a set of formal semantic specifications that would be found in a KR language (recalling, again, the hierarchy of interoperability in Nilsson et al. (2009)).

5.2.3 *1:1 Principle* Violations

Within the account of quality provided by Shreeves et al. (2005) are both intrinsic and relational aspects related to the *1:1 Principle*:

intrinsic cohesiveness: the extent to which the content of an object is focused on one topic.

relational accuracy: the degree to which an information object correctly. represents another information object or a process in the context of a particular activity.

However, each of these measures of quality is grounded within the contexts of some particular activity (i.e., the *find* activity) or through the specification of some "outside reference source," such as a set of cataloging rules or a descriptive vocabulary, like Dublin Core (Shreeves et al., 2005).

In their analysis of individual metadata records, Shreeves et al. (2005) found that a particular source of ambiguity that reduced metadata quality consisted of metadata records that appeared to include information about both an "original" resource and a "digital" resource.

> ... no collection maintained a consistent one-to-one mapping between the metadata and a single resource.... These practices represent a very real and understandable tension between the need for standardized, accurate description of digital objects and description that meets the needs of end users. (Shreeves et al., 2005)

In practice, Shreeves et al. (2005) found that between 57% and 100% of records in their sample included properties for both physical and digital manifestations of a resource. Not only did these ambiguities impair *find* tasks. They also limited the ability of aggregators to provide services relevant to users' needs. One such example mixed physical/analog dimensions with electronic file attributes, as shown in Figure 5.1.

```
<description>100 x 70 cm</description>
<description>image/tiff</description>
<format>image/jpeg</format>
<format>Any machine capable of running
        a graphical Web browser, 640x480
        minimum monitor resolution
</format>
```

Figure 5.1: Example of *1:1 Principle* inconsistencies identified by Shreeves et al. (2005)

These findings were confirmed by Hutt & Riley (2005):

> The *1:1 Principle*, which requires DC records to describe exclusively one version of a resource, is particularly problematic for cultural heritage institutions where the majority of digital objects are not born digital, but are instead created from the digitization of existing analog materials. Therefore, it is common for multiple versions of an intellectual object to exist within a single institution, often including the original analog materials, the master digital file, and at least one derivative digital file. Representing this complexity in the OAI DC environment obviously presents a challenge. . . . This leaves data providers with two choices, create records that adhere to the *1:1* rule and omit pertinent information, or violate the rule. We observed many cases in which data providers chose to violate the rule and combine data about the original intellectual object as well as its digital manifestation.

In their study of a single digital repository system (CONTENTdm), Han et al. (2009) also found that many descriptions included statements about both digital and physical manifestations. In the local CONTENTdm environment, the meaning of local properties may indicate what kind of entity the property intends to describe (for example, a property such as `digital.format`). Following on concerns raised during Dublin Core's early development (see Chapter 2), Han et al. (2009) note that "no element in Dublin Core can distinguish between different manifestations... usually conflicting fields are mapped to the same Dublin Core element."

Because local projects frequently use a mix of local standards (such as MARC, Encoded Archival Description (EAD), Dublin Core, Metadata Object Description Standard (MODS), etc.) many opportunities for these kinds of mapping problems occur when their content is translated into simple OAI-PMH Dublin Core. For large-scale aggregation projects this can increase the cost of providing needed services (Foulonneau & Cole, 2005). Furthermore, "nonstandard use of fields seemed to be more prevalent with Dublin Core" (Palmer et al., 2007).

Miller (2010) notes that *1:1 Principle* problems also result from "database and user interface systems [that] do not have the capacity to adequately link separate records and to display them together in a clear and meaningful way for end users." Systems such as CONTENTdm (studied by Han et al. 2009) also base their primary information models around digital assets, such as single files, making it difficult to independently represent non-digital resources from which those files were derived. Such systems enable metadata creators to create specialized, locally specified metadata elements on a collection-by-collection or project-by-project basis. The ease which these systems allow the addition of new properties encourages ad-hoc modeling of metadata optimized for display in one local context, rather than more formal and rigorous methods of information modeling.

Despite the formal specifications provided by the *DCMI Abstract Model* (Powell et al., 2004), metadata creators find the *1:1 Principle* "particularly confusing" (Park & Childress, 2009):

> The problem with DC aside from its ambiguity is the failure of the *1:1 Principle*. Inevitably implementors use a single record to describe both original (what users are really interested in) and digital. DC is still really only for digital manifestations.... What's needed in DC is a simple subset of elements clearly designated for the digital manifestation with the primary element reserved for the original object.

In the previous chapters we found that the Dublin Core *1:1 Principle* is grounded in requirements that are found in knowledge representation languages, but that these requirements are not essential for knowledge organization document surrogates, for which practitioners have developed a conflicted set of approaches for describing originals and reproductions. In light of the analysis so far, the findings of metadata quality studies seem consistent with a culture of practice based in KO approaches. When seen as a *principle of guidance* rather than a *principle of reasoning*, creating metadata documents that meet users' needs is paramount. Because many of the rule systems based on *AACR2* are bound to principles of guidance, they often encourage metadata creators to make pragmatic choices in representations. When they are also bound to the expectations of a particular syntax (e.g., MARC) or information retrieval system (e.g., CONTENTdm), applying them to alternative syntaxes/systems (e.g., Dublin Core or OAI-PMH) may result in the inconsistencies identified by metadata quality studies. In the sections that follow, we will look more closely at the presuppositions inherent in these studies, with an eye towards what would be required to translate them into a formal knowledge representation language.

5.3 Principle Presuppositions

The example provided by Hillmann (2003) is a caution against "metadata" that describes both the properties of the *Mona Lisa* (a painting created in the 16th century) and a digital image (a JPEG). An example of a record that has been interpreted as "violating" the *1:1 Principle* is provided in Example A.1, a fragment of which is included below.

```
<title>Mona Lisa</title>
<date>2008</date>
<date>1501-1519</date>
<source>TIFF</source>
<type>image</type>
<format>oil on poplar board</format>
<format>H. 77 cm; W. 53 cm</format>
<format>image/jpeg</format>
<format>16781 bytes</format>
```

Figure 5.2: Example of a Dublin Core record that "violates" the *1:1 Principle*

In order to help understand which records count as violations of the *1:1 Principle* according to current studies, we will first unpack the criteria these studies have used. This will entail examining their assumptions about the syntax of records, the meaning of individual Dublin Core properties, the meaning of property/value statements, and the larger contexts in which these judgments are being made (Hutt & Riley, 2005; Shreeves et al., 2005; Park & Childress, 2009).

5.3.1 Interpreting the Formal Semantics of Syntaxes

In Chapter 2, we saw that many early discussions centered around a need to create "logical clusters of metadata" that might be organized into "packages" (Lagoze, 1996; Weibel & Hakala, 1998; Bearman et al., 1999). However, definitions of what, exactly, the *1:1 Principle* applies to are ambiguous within different DCMI documents. Hillmann's account of the *1:1 Principle* (2003) ambiguously suggests that "metadata" is the object of the *Principle* while Woodley et al. (2005) suggest this the object is "records." The more formal *DCMI Abstract Model* (Powell et al., 2007), which was developed after the articulation of the *1:1 Principle,* states:

> The abstract model presented above indicates that each DC metadata description describes one, and only one, resource. This is commonly referred to as the one-to-one principle.
>
> However, real-world metadata applications tend to be based on loosely grouped sets of descriptions (where the described resources are typically related in some way), known here as description sets. For example, a description set might comprise descriptions of both a painting and the artist. (Powell et al., 2007)

Despite the underspecification of what the *1:1 Principle* applied to, various studies have adopted "records" as the unit of analysis for metadata quality (Stvilia et al., 2004; Hutt & Riley, 2005; Shreeves et al., 2005; Stvilia et al., 2007; Han et al., 2009). However, as is demonstrated by Example A.1, this does not refer to an OAI-PMH `record`, but rather to all of the element/value pairs that are children of the OAI-PMH `metadata` element as is shown in Figure 5.3.

```
<oai:record>
        <metadata>
                <!-- Only XML Elements that appear at this
                level of the XML document are considered
                to be "records" -->
        </metadata>
</oai:record>
```

Figure 5.3: Selecting the `metadata` node as the object of *1:1 Principle* violations

The OAI-PMH specification itself does not directly license the assumption that metadata contained in a `metadata` node is about one, and only one, thing. In fact, OAI-PMH asserts that it is not concerned with the nature of resources being described (Lagoze et al., 2008). The OAI XML schema[1] merely provides us with a specification for the structural organization of an OAI document. It explicitly leaves the specification of what `metadata` is "about" to the semantics of the enclosed metadata. This attitude is also reflected in the original Warwick Framework, which preferred to defer the meaning of metadata packages to individual implementations:

> Each logically distinct metadata set may represent the interests of and domain of expertise of a specific community; for example, catalogers should create and maintain descriptive cataloging sets and parties with legal and business expertise should oversee terms and conditions metadata sets. The syntax and notation of each should be determined by the responsible party and fit the semantic requirements of the type of metadata. (Lagoze, 1996)

For example, an OAI-PMH item might be associated with a VRACore, MODS, or CDWA representation that allows for a single record to represent several related objects (see Example A.2). Because OAI-PMH does not specify the relationship between an OAI item and any particular resource, it can support a representation about both the *Mona Lisa* itself and an image of the *Mona Lisa*. The OAI specification leaves undecided how such rich representations should be "dumbed-down" into the required simple Dublin Core. A complex representation like Example A.2 could easily result in a representation that looks like Example A.1.

> In the end, the semantics of the metadata associated with an object need to be understood by the "consumers" of the metadata—the clients and agents that access objects and the users that configure these clients and agents. We run the danger, with the full expressiveness of the Warwick Framework, of creating such complexity that the metadata is effectively useless. Finding the appropriate balance is a central design problem. (Lagoze, 1996)

A main interoperability requirement of the OAI-PMH protocol was that each OAI Item would be associated with a simple Dublin Core representation. While the overall structure of the OAI-

[1] http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd

PMH record syntax did not prevent the inclusion of more discrete "packages" of Dublin Core elements within the `metadata` container, at the time that OAI was developed DCMI had not provided a recommended XML syntax. Initially the OAI community provided a flat, simple Dublin Core syntax until it was later superseded by the official DCMI recommendation (Lagoze & Van de Sompel, 2001). However, the official DCMI recommendation for this simple syntax distanced itself from specifying a formal grammatical semantics, stating "there is no formal linkage between a simple DC record and the resource being described. Such a linkage may be made by encoding the URI of the resource as the value of the DC Identifier element, however this is not mandatory" (Powell & Johnston, 2003).[2] While this schema associates XML elements with individual Dublin Core property definitions, it does not provide a formal semantic model for "records," noting:

> . . . it is anticipated that *records* will be encoded within one or more container XML element(s) of some kind. This document makes no recommendations for the name of any container element, nor for the namespace that the element should be taken from. (Powell & Johnston, 2003)

The inclusion of the newly emerging abstract data model in Powell & Johnston (2003) suggests that *records* are "made up of one or more *properties*, each *property* is an attribute of the resource being described" (Powell & Johnston, 2003). Because this model defers the relationship between a resource and a record to a container syntax, the initial abstract model stops short of claiming that *records* are about "one, and only one" resource. Since OAI-PMH defers the relationship between a representation and an object to the enclosed metadata standard, the descriptive relationship is left underspecified in its own use of simple Dublin Core.

From this discussion, we learn that the semantics of an OAI-PMH XML document are underspecified. In order to analyze OAI-PMH documents for metadata quality, researchers have treated the elements contained in `metadata` nodes as a single unit with the expectation that as a whole, all elements are about the same resource (in conformance with their interpretation of the *1:1 Principle).*

[2] This statement is unclear about whether the mere presence of a URI established a formal relationship. For more on the implications of this ambiguity, see Renear et al. (2011) and Wickett et al. (2012).

Semantic Relationships among Elements

A second concern raised by metadata quality researchers is the relationship between repeated elements found in an OAI-PMH `metadata` container. The assumption provided by researchers is that there exists a conjunctive relationship between element/value pairs, at least when the same Dublin Core element is repeated within the `metadata` element. For example, the meaning of the following set of XML elements could be interpreted as asserting "the resource has the format oil on board AND image/jpeg":

```
<dc:format>oil on board</dc:format>
<dc:format>image/jpeg</dc:format>
```

However, because the semantics of the Dublin Core XML syntax are underspecified, this is not the only interpretation that could be supplied for the same element/value pairs. Early Dublin Core schemas noted that "the semantics of repeated elements are not specified" (Burnard et al., 1996). While a container syntax like OAI-PMH might demarcate one set of metadata elements from another, different applications might wish to transform it into different kinds of representations that might supply different kinds of relationships between the elements (e.g., to a mapping for a particular data model or into first-order predicate logic) (Renear et al., 2000; Sperberg-McQueen & Miller, 2004). Rather than assuming that the relationship between repeated elements is conjunctive, we might supply an interpretation that treated sibling elements as disjunctive (for example, if we were creating a list of terms that we wished to be unique).

Figure 5.4 illustrates that all we may know about an XML document is that a hierarchical relationship exists between elements. The semantics of these relationships are unknown. The manner in which we treat other XML properties expressed in this OAI document illustrates the point. Although we may wish to understand an OAI Item that contains metadata "about a single resource," OAI does not specify what may count as a resource (i.e., it is also possible to provide a description of some complex resource that has multiple parts). We also do not interpret the entire OAI `record` to be "about an OAI Item"; rather, it has some other relationship to some Item (and thereby a resource). The OAI document is merely a constituent of an OAI repository. The inclusion of elements such as `oai:responsedate` and `oai:datestamp` are reflexively about the OAI document itself (see Example A.1). While human readers may be able to intuitively shift

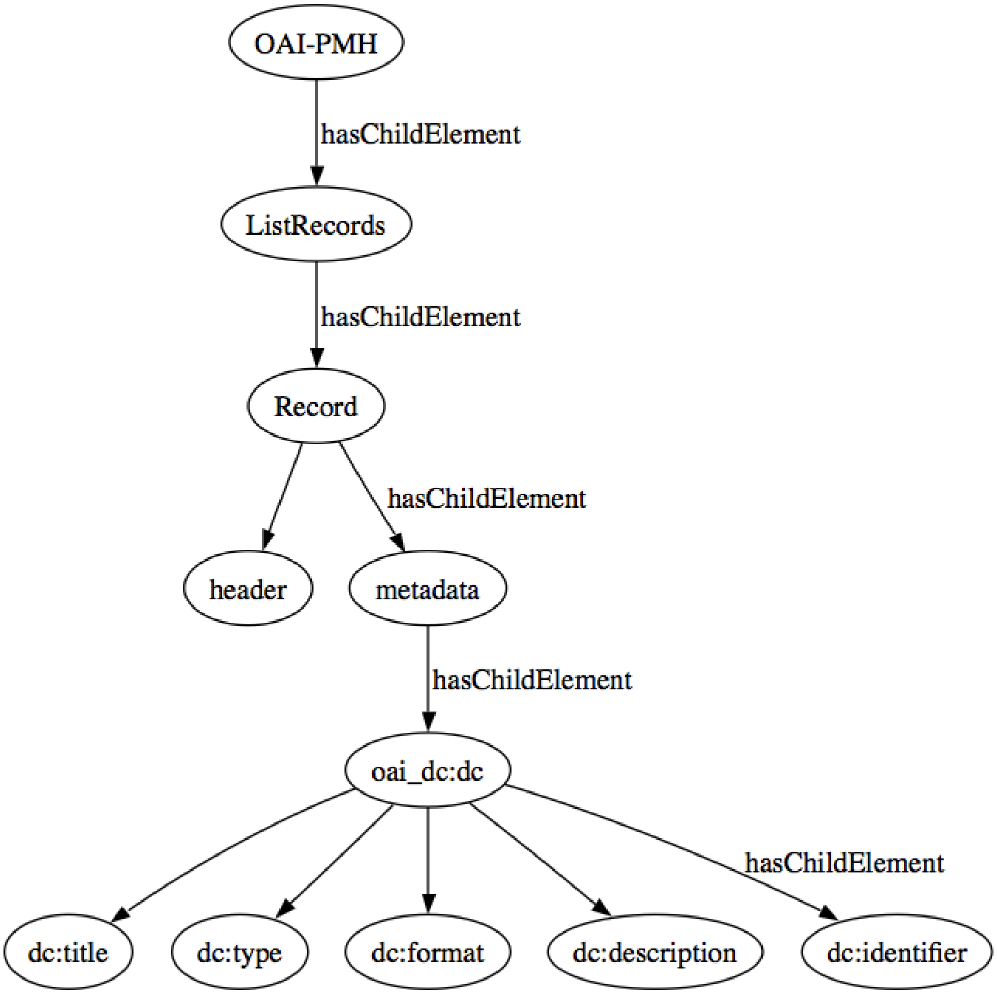


Figure 5.4: XML Tree of an OAI-PMH representation

their attention from the OAI Item context, to OAI `record` context, to `metadata` context, a formal *semantic* relationship between the document and a described resource is not supported by the OAI-PMH syntax, nor by the OAI Dublin Core XML schema.

Implicit Meaning of Values

An analysis of cultural heritage Dublin Core records provided by Hutt & Riley (2005) adds additional requirements to being able to identify *1:1 Principle* violations that depend on the interpretation of expressed *values*. In their observations, they classified values that appeared in statements according to "pseudo qualifiers that refine the meaning of an element" and "semantic types." For example, a `dc:date` statement could include a value such as "Digitized: 2005-01-19" to offer an implicit narrowing of the general meaning of "a date in the lifecycle of a resource."[3] Even when such pseudo-qualifiers were not present, Hutt & Riley (2005) were able to classify values into three broad categories, shown in Figure 5.5.

	Unique Values		All Values	
Digitization Date	5286	40.68%	8026	24.33%
Creation Date	7585	58.37%	24624	74.64%
Copyright Date	124	0.95%	339	1.03%
Total	124995	100%	32989	100%

Figure 5.5: "Semantic Types" for date values found by Hutt & Riley (2005)

The assignment of these values to these classes was conducted as part of a qualitative analysis that used unspecified "high level assumptions" about the nature of date values (Hutt & Riley, 2005). While this provides a useful high-level classification of date values, it is unclear how the assumptions might be represented in a more formal analysis of date values.

Hutt & Riley's approach (2005) to assigning "semantic types" to values is not limited to dates and creators. Their approach could equally be applied to a significant property of Dublin Core, the `identifier` property. Currently the library community relies on manifestation-level identifiers such as ISBNs and ISSNs. Emerging identifier systems, such as Digital Object Identifiers (DOIs) may be assigned to different conceptual entities. Many institutions rely on locally maintained

[3] Although this statement is not necessarily associated with more specific refinements provided by DCMI Terms, e.g., `dcterms:created`.

numbering schemes, such as museum accession numbers (year.accessionOrdinal.itemOrdinal). As readers we may also understand the "meaning" of such values based on orthographic features and whether they are associated with a particular kind of entity.[4] However, the LAM community has been reluctant to adopt universal identification schemes, which have largely originated from outside the community (Vitiello, 2004). In the case of URLs used to access online resources, their assignment frequently relies on underlying system architectures of content management systems (e.g., CONTENTdm). Han et al. (2009) found that local schemas ofter additional granularity to define local identifier values that are often mapped to Dublin Core `identifier`. Figure 5.6 provides examples of the kinds of identifiers that are found within the same type of `metadata` node in an OAI Record.

```
<metadata>
        <dc:identifier>Case Pullman 05/02/03 Folder 9 Sheet 509</dc:identifier>
        <dc:identifier>NL002641.tif</dc:identifier>
</metadata>
<metadata>
        <dc:identifier>ALA-9901014-5-1</dc:identifier>
        <dc:identifier>ala-9901014-5-064</dc:identifier>
        <dc:identifier>http://images.library.uiuc.edu:8081/u?/ALA,58</dc:identifier>
</metadata>
<metadata>
        <dc:identifier>lccn:09033358</dc:identifier>
        <dc:identifier>http://name.ex.edu/AEY0004.0001.001</dc:identifier>
</metadata>
<metadata>
  <dc:identifier>X68.2890</dc:identifier>
  <dc:identifier>http://content.ex.org/ark:/13030/hb6779p502/</dc:identifier>
</metadata>
```

Figure 5.6: Example record fragments that illustrate the repetition of `dc:identifiers`

A number of studies have taken a closer look at identifiers and identification systems with the aim of identifying "persistent" identifiers for resources (Vitiello, 2004; Coyle, 2006; Campbell, 2007a). Activities such as the Persistent Identifier Linking Infrastructure (PILIN) Ontology or other discussions of identifier systems could provide the basis for constructing an approach to classifying identifier instances.[5]

While we may be able to classify different types of identifier strings, it is unclear whether the LAM community understands them as functioning as *referential names* in the way that URIs are used in the model-theoretic semantics of RDF representations. Stvilia et al. (2004) found that

[4] At least in general, the division between ISBNs that are applied to books and ISSNs that are applied to serial publications, etc.

[5] http://www.pilin.net.au/Project_Documents/PILIN_Ontology/PILIN_Ontology.pdf

identifier was present in 100% of cultural heritage repository records, often with more than one identifier present. These identifier values exhibited a high degree of uniqueness that allowed representations to be distinctive, especially when compared to elements such as subject or type that are intended to collocate representations (Stvilia & Gasser, 2008). However, having a *unique* value does not necessarily entail that an identifier *refers uniquely* to one, and only one, resource. Stvilia & Gasser's assessment of metadata quality (2008) assumed that a URL provided access to a resource and dramatically reduced the value of a metadata record if it failed to provide access (i.e., if a URL was inaccurate, or a system was unable to deliver a described resource). Within the present metadata quality literature, the question of whether a URI successfully refers to the described resource is left unmeasured, especially for the use of URIs that do not provide access to offline resources, but may successfully refer to them. This suggests that another kind of *1:1 Principle* violation may occur if a URI is used to refer to more than one resource.

Semantic Interpretation through Record Context

Implicit in Hutt & Riley's analysis of values (2005) is an understanding of the contexts in which metadata was produced and expected to be used. In this case metadata was selected from a known cultural heritage repository where it was understood *a priori* that representations are related to digitized cultural heritage materials with physical and digital manifestations. Likewise, the analysis in Shreeves et al. (2005) draws on metadata from the IMLS Digital Collections and Content project (IMLS DCC),[6] an aggregation specifically targeted at cultural heritage organizations. Provided this context, it is possible for a qualitative researcher to infer that two dates, such as "1901" and "2008," may represent different kinds of events in the lifecycle of a resource (e.g., creation and digitization).

Likewise, being aware that metadata represents cultural heritage resources heightens our awareness of incoherent format statements that seem to describe the properties of both physical and digital resources. In a heuristic evaluation of metadata records, qualitative researchers bring a great deal of background knowledge to their assessments as knowledgeable cultural heritage professionals. They may understand that terms like image/jpeg and glass plate negative are properties that are unlikely to be shared by the same resource. They also may understand that JPEGs are

[6] http://imlsdcc.grainger.illinois.edu

the kinds of the things that may "reproduce" something like a glass plate negative, but rarely will glass plate negatives "reproduce" a JPEG. They also understand that JPEGs are the kind of thing that might be associated with a date like "2008" and are not things that could have been created in "1901."

Not all aggregation environments understand these contextual cues, which offer helpful guidance in supplying an implicit semantics to XML metadata representations. To a machine-aggregator, many of these important contextual elements are not accessible in a machine readable form—and frequently they are not publicly documented for human consumption. For example, Han et al. (2009) found that many CONTENTdm repositories provided locally defined properties or interpretations of existing Dublin Core properties. While some of these repositories may have followed published best-practice guidelines, information about which guidelines were used is not provided by the OAI-PMH repository and may not be represented elsewhere on a repository's website. Examination of XML representations merely provides links to more documentation of the syntax (via the OAI-PMH XML schema, DTDs, etc.).

For human interpreters sensitized to these contexts and armed with *a priori* notions of records that should be about one, and only one, resource (with a clear ontological sense of resources being different manifestations), it is possible to supplement colloquial XML records with a meaning that leads to the conclusion that records are violating the *1:1 Principle.* In the next section we will explore the difficulties in formalizing these many colloquial heuristics in order to automate judgments about metadata records.

5.4 Formalizing *1:1 Principle* Violations

In order to develop an automated method for identifying *1:1 Principle* violations, many of the assumptions outlined here would need to be represented formally. Attempting to do so reveals just how difficult it is to translate professional intuitions into formal, machine-understandable languages.

First, colloquial XML metadata records need to be translated into a syntax that supplies more formal semantics, such as the Resource Description Framework (RDF) or the Dublin Core Description Set XML (Johnston & Powell, 2008). Haslhofer & Schandl (2008) have provided a service which directly translates the OAI-PMH syntax into a simple Dublin Core RDF syntax (see

Example A.6.[7] However, as Haslhofer (2008) notes:

> Unfortunately, most [metadata schema mappings] do not consider the whole heterogeneity spectrum but focus mainly on schema level mappings and disregard the instance level.... Although these kind of mappings suffice for human interpretation, the question remains how machines should interpret them in order to provide uniform access to the sources. They need exact information about relationships between concepts and precise processing instructions for dealing with the instances or data originating from heterogeneous sources..... The problem is that even within a single metadata integration scenario, data sources as well as the mappings created between their metadata schemes and instances may embody different assumptions on how information should be interpreted.

An important part of making successful semantic translations is the inclusion and representation of the contexts that make representations understandable.

As many of the records that appear to "violate" the Dublin Core *1:1 Principle*seem also to conform to widely adopted practices for describing "reproductions" or "visual documentation," part of the context that would be necessary to formalize this would be to determine exactly how to identify what it means to be "a reproduction" or "visual documentation." What is needed is an ontology of the kinds of things that are reproductions or visual documentation. As noted before, current cataloging practice is based on fundamental ontological commitments about the nature of resources. These commitments then guide the application of particular rule sets for description.

One approach to solving these mapping problems is to use ontologies and/or meta-ontologies that provide formal representations of contexts. In the case of cultural heritage materials, what is needed to replicate the kinds of judgments made by human evaluators is an ontology that defines the classes of resource formats. At present, many metadata instances draw terms from one or more thesauri, such as the *Thesaurus for Graphic Materials*[8] or the *Art and Architecture Thesaurus*.[9] These thesauri provide a structured organization of terminology and mechanisms to ensure that terms uniquely represent concepts (i.e., that they specify term values that avoid

[7] A similar utility is provided by the SIMILE oaiRDFizer at: http://simile.mit.edu/wiki/RDFizers

[8] http://www.loc.gov/rr/print/tgm1/

[9] http://www.getty.edu/research/tools/vocabularies/aat/

problems of polysemy). While the *AAT* and *TGM* provide natural-language definitions for the meanings of terms and specify what types of resources a term describes, they do not provide any formal way to represent the properties of the resources they describe (Wielinga et al., 2001). For example, a term might denote a stylistic convention that was prominent during a particular time period (e.g., "Late Georgian"). Unfortunately, a representation of a date range ("1760–1811") that could be used to evaluate and classify a resource based on an instance of metadata is not available.

Furthermore, thesauri like *AAT* and *TGM* provide organizational structures for terms, but they are not associated with the kinds of formal logic present in languages such as the Web Ontology Language (OWL). While we might be able to understand that two terms are distinct or have different taxonomic parents, we are unable to declare that they represent disjoint kinds of resources. In an ontology, we may also wish to include additional information that could help us make judgements about resources—for example, that nothing created before 1992 can be an image/jpeg. Even with the power of OWL, some of the intuitive assumptions we use in current *1:1 Principle* heuristics may be outside the inherent capabilities of FOL-based knowledge representation systems and model theory (for example, see Wickett, 2011).

5.5 Conclusions

Studies that identify *1:1 Principle* violations are motivated by concerns about metadata quality and interoperability. While these studies adopt concepts from the KO paradigm concerning what syntactical structures count as a record to be evaluated, they apply the KR-based assumption that these structures must therefore be about one and only one resource. Because the representations being evaluated are based in colloquial XML syntaxes that are not associated with formal semantic models, the *1:1 Principle* may be unwarranted if other local standards are used as reference sources for quality evaluation. The heuristics and context that make these studies valid within the KO paradigm may not easily translate into formal KR methods.

Although Urban (2009) suggested that it would be possible to develop more formal operational definitions of the *1:1 Principle,* this analysis suggests that applying KO heuristics more formally may be unwarranted. Because the violations that have been identified from the KO perspective are partially the result of the colloquial semantics of available standards, what is needed are not more formal ways of detecting violations, but rather suggestions and guidance for constructing and using more formal knowledge representation languages that avoid *1:1 Principle* problems by the very nature of their semantic models. While this was an unexpected result of this research, it lays important groundwork for future research.

Chapter 6

Contributions & Future Research

> It is easier to produce ten volumes of philosophical writing than to put one principle into practice. —Leo Tolstoy

6.1 Introduction

The Dublin Core *1:1 Principle* appears to offer a simple dictum: metadata is about one, and only, one resource. Despite its apparent simplicity, prior metadata quality research suggests that widespread violations of the *Principle* occur, in part, due to the behavior of metadata creators responsible for conforming to the *Principle.* This research demonstrates that the violations are not mere errors, but rather that they conform to long-standing rules for the description of reproductions within the knowledge organization domain. These rules are enmeshed in a network of conflicting principles of guidance that require metadata creators and catalogers to balance the needs of users against the accuracy of their descriptions. Because user needs are more highly valued within the knowledge organization community, many of the infelicities encountered by metadata quality research are easily dismissed as being in the service of users. While these kinds of problems (i.e., multiple date and format values) may impede performance in large-scale aggregations, they are often seen as necessary to achieve local *find, identify, select,* and *obtain* objectives. In these contexts, the shift in reference between an original and a reproduction is easily navigated by human consumers of document surrogates.

Unfortunately, the same cannot be said for machine-understanding of these representations. To a reasoning system, these document surrogates appear incoherent and contradictory. In order to prevent such ambiguities, knowledge representation approaches have adopted formal theories of description and reference as an essential part of their semantics. These semantic models do not merely provide a definition of a descriptive vocabulary. They also provide mechanisms that allow statements to refer to resources. In the formal knowledge representation semantics, the *1:1 Principle* is not a suggestion that can be ignored in favor of perceived user needs. It is an essential feature required for intelligent reasoning tasks.

Because the Open Archives Initiative-Protocol for Metadata Harvesting (OAI-PMH) XML syntax is more closely aligned with knowledge organization attitudes towards document surrogates, and less so with referential knowledge representations, it is ill-equipped to prevent the kinds of confusion found by metadata quality studies. Instead, the OAI-PMH syntax enables the kind of pragmatic solutions that traditionally have been applied within the knowledge organization domain. Pragmatic rules for describing reproductions using the MARC syntax create the same kinds of representation conundrums when implemented using OAI-PMH XML (Miller, 2010). Because these kinds of flexible solutions are supported by the internal data models used by local metadata management systems, such as CONTENTdm (Han et al., 2009), the problem with *1:1 Principle* violations does not lie with metadata creators. Rather, if we must identify a violation, it is in the OAI-PMH DC syntax. As a descriptive language, OAI-PMH DC was not designed to conform to the expectations of knowledge representation semantics. Indeed, the developers of OAI-PMH made a conscious choice to adopt a human-readable, document-like syntax over the emerging RDF specification (Lagoze, 2001b). The challenge may not be in detecting *violations,* but rather in finding any OAI-PMH representations that actually conform to the *1:1 Principle* as an axiom of knowledge representation.

The development, application, and misunderstanding of the *1:1 Principle* serve as a cautionary tale for the cultural heritage community as we move forward with the adoption of Semantic Web languages. Although this study has not provided solutions for correcting the kinds of violations identified by metadata quality studies, it identifies many of the complexities that we, as a community, will have to confront as we shift paradigms.

6.2 Contributions

Although this study does not provide any immediate solutions to resolving the ambiguities found in individual metadata instances, it makes the following contributions to our general understanding of metadata.

1. Chapter 2 provides a detailed account of the development of the Dublin Core *1:1 Principle.* This was absent from earlier metadata quality research. My account demonstrates that the Dublin Core Metadata Initiative (DCMI) community found it necessary to articulate the *1:1 Principle* when the cultural heritage community proposed fundamental changes to Dublin Core based on existing library, archive, and museum (LAM) description practices. The DCMI community's response to these proposals suggests a fundamentally different understanding of how representation languages could be constructed, one that drew from more recent developments in computer science and artificial intelligence. The difficulties in articulating these differences encountered by members of the DCMI community reveal a paradigm divide between knowledge organization and knowledge representation approaches to representation languages.

2. In order for us to more clearly understand the confusions, Chapter 3 outlines the features of the knowledge organization (KO) and knowledge representation (KR) paradigms. An examination of these features demonstrates that KO and KR adopt distinct methods to achieve fundamentally different objectives. Among those methods are the theoretical considerations that define relationships between a resource and a representation. While the *1:1 Principle* is closely aligned with fundamental KR theories of description and reference, it may not correspond to colloquial understandings in KO. This presents a challenge for defining a single, coherent conceptual definition of the *Principle.*

3. Chapter 4 identifies the formal theories of description and reference that require knowledge representation descriptions to be "about one, and only one, resource." We find that colloquial bibliographic languages lack a corresponding theory for the relationship between resources and document surrogates. Because the objectives of each paradigm differ, this is not necessarily a problem within the framework of each paradigm. However, the differences between these approaches help to explain the *1:1 Principle* problems observed by metadata quality research. This is especially true when representations created within one paradigm are evaluated using the criteria from the other.

4. In Chapters 3 and 4 we find that the *1:1 Principle* is associated with theories of description and reference used by the KO and KR paradigms. Because each paradigm handles this problem differently, Chapter 5 reexamines the assumptions made by previous studies that have identified *1:1 Principle* violations. We find that while metadata quality studies have articulated valid interpretations of *Principle* violations from the knowledge organization perspective, these are not based on fundamental knowledge representation principles. Translating the implicit knowledge used by metadata quality researchers into formal knowledge representation interpretations will require additional research.

Although the unexpected complications that arose from identifying and exploring representation paradigms prevented the development of the concrete operational definitions and the testbed implementations suggested by Urban (2009), this study has revealed a fertile area for new research. The importance of exploring the divide between knowledge representation and knowledge organization will become even more critical as the cultural heritage community moves to adopt Semantic Web technologies. In what follows, I explore some of the implications of this study for future work.

6.3 Beyond *1:1 Principle* Violations

The Dublin Core community created the *1:1 Principle* in order to provide guidance for using what at the time was merely a set of properties of very general resources. The existence of the *Principle* in the documentation drew the attention of researchers, who found that it partially explained the metadata quality problems they observed. Unfortunately, available syntaxes for expressing Dublin Core, such as the OAI Dublin Core or CONTENTdm's internal data model, make it difficult to follow the *Principle* in practice. Because of these limitations in current approaches, it is unlikely that encouraging metadata creators to follow the *1:1 Principle* will prevent violations from continuing to occur.

6.3.1 User-oriented Interpretations

One of the reasons given by metadata creators for ignoring the *1:1 Principle* is that *users'* information needs are expressed in terms of the properties of original, analog materials (Shreeves et al., 2005; Hutt & Riley, 2005; Park & Childress, 2009; Miller, 2010). These arguments follow from earlier debates between facsimile- and edition-theory approaches to describing reproductions in the catalog (Graham, 1992; Svenonius, 2000; Knowlton, 2009). In this scenario, the *1:1 Principle* is overridden by a more important one—the *Principle of User Convenience.* Within a knowledge organization system, a user's need is essentially met by a representation that accurately models his or her understanding, not a state of affairs in the actual world. Additional research is needed to verify that the intuitions represented by these arguments are supported by the relevant information-seeking behavior literature. For example, Lee et al. (2006) explore the problems inherent in known-item searching when the user possesses inaccurate information.

Ingwersen & Jarvelin (2005) charge that information retrieval (IR) research has largely focused on laboratory-based approaches that tune systems for maximum relevance using computer science methodologies. This characterization of IR research is not unlike much of the knowledge representation research that provided foundational theories for the Semantic Web. Cognitive-based approaches were also explored by AI researchers. However, understanding how that work is incorporated into contemporary Web-based KR languages like RDF was outside the scope of this study.

This raises two questions. First, if the facsimile/edition theories represent models that meet users' needs, how do we measure these theories' impact on KO/KR objectives? Second, if these are useful perspectives, should our formal interpretations directly reflect the users' views of the world? If so, it may be necessary to supply multiple interpretations and sets of derived representation instances. For example, we might supply an interpretation that is valid for internal management of assets or for inter-repository exchange of representations. Each of these might also be accompanied with interpretations for the presentation and use of representations for particular user communities.

Such translations would move our understanding of interoperability from being a merely syntactic problem to one that is tied to clear objectives for our representations. To date, interoperability efforts have placed an emphasis on how representations are constructed, not how we might express rules and ontological commitments as formal models that can be shared along with representation instances. While sharing our interpretations may not decrease variation in metadata, it could ease the work and reduce the costs associated with integrating representations in large-scale aggregations where developers must now guess at the intentions of data providers. Shared interpretations may provide a more successful approach to achieving metadata interoperability.

6.3.2 Dublin Core Application Profiles for Cultural Heritage

Subsequent to the articulation of the *1:1 Principle,* the DCMI community has provided recommendations for both record syntaxes (the DCMI Abstract Model (Powell et al., 2007)) and their use within a particular domain (the DC Application Profiles (Nilsson et al., 2008; Coyle & Baker, 2009)). To date, few of the systems used by cultural heritage professionals to express metadata have adopted these recommendations in their internal representations (Han et al., 2009; Miller, 2010). A potential solution to this problem would be for the cultural heritage community to provide a baseline *Application Profile (DCAP)* that would specify how Dublin Core syntaxes should be created for cultural heritage materials. While other application communities (e.g., the Scholarly Works Application Profile[1]) provide such models, none has been specified for cultural heritage. Such a profile could provide a template for accurately representing the relationships between "originals" and "reproductions" by requiring separate, but related, DCAM description sets for each. Both the CDWALite and VRACore syntax specifications facilitate these kinds of representations that could easily be replicated as a DCAP.

Problems are compounded by the failure of OAI-PMH Dublin Core to move beyond a relatively simple syntax while adoption of the protocol within cultural heritage remains steady. Although the Open Archives community has provided an alternative that supports richer representations (OAI-Object Reuse and Exchange, OAI-ORE[2]), it has not been widely implemented for the exchange of cultural heritage metadata. An important exception to this observation is the use of OAI-ORE as the basis of the *Europeana Data Model (EDM)* (Doerr et al., 2010). Given the issues surrounding OAI-PMH and *1:1 Principle* confusions, an analysis of EDM data may reveal whether it has effectively solved the problems in practice by using a more robust data model.

[1] http://dublincore.org/scholarwiki

[2] http://www.openarchives.org/ore/

6.3.3 Disambiguation of Existing OAI-PMH Data

This study's original objective was to be able to define a conceptual definition of the *1:1 Principle,* a definition that would lead to operational definitions for detecting violations. However, merely being able to detect violations does little to improve services to users. Beyond the ability to detect violations would be the ability to translate offending representations into new representations that conform to the *1:1 Principle.* Along with the difficulties faced by detection, translation also faces several hurdles.

First, the language of the current *1:1 Principle* and of the studies that identify violations is focused on two entities: "originals" and "reproductions" (or analog/digital kinds). The developments that have led to the *Functional Requirements for Bibliographic Records,* however, demonstrate that there may be multiple layers of entities to which a bibliographic description refers. The development of an Application Profile, based on FRBR, could provide guidance on the construction of DCAM descriptions and description sets. However, additional research is needed in order to understand how to translate existing statements into properly referring descriptions. If we do wish to use more formal knowledge representation semantics, with their formal theories of reference, constructing representations based on FRBR would also entail providing mechanisms to provide URIs for each of the entities referenced by a representation.

As noted in Chapter 5, disambiguating existing OAI-PMH metadata will also require a framework through which to interpret attribute/value pairs where a value changes the interpretation of the predicate (i.e., through the presence of pseudo-qualifiers in the value). In part this may be accomplished by providing a mapping between strings in present metadata and the concepts/things the string value is intended to represent. Organizing these relationships may be accomplished through the application of more formal ontologies (for example, in the case of file formats/genres). Once mapped to an ontology, additional axioms may help identify where inconsistencies exist.

However, because of open-world assumptions, such approaches may only succeed at identifying inconsistencies that are intrinsic to available metadata. For example, such axioms would not identify representations that follow a strict facsimile/edition-theory approach by including a note in the `description` or `source` properties. For this reason, additional research is necessary to systematically evaluate formal rules for disambiguation that proved to be outside the scope of this preliminary study.

However, by combining methods of disambiguating ontologically incoherent metadata with improved representation semantics and syntax, we may be able to move beyond merely identifying *1:1 Principle* violations and towards solutions that help metadata achieve our objectives for it. What these techniques will not do, though, is prevent the creation of further incoherent metadata. Achieving that objective will require new ways of thinking about how we construct bibliographic descriptions, train and educate metadata creators, and develop systems and workflows that support knowledge representation objectives. This also will require consensus, or at least widespread agreement, that knowledge representation's objectives should supplement and augment existing knowledge organization approaches.

6.4 The Future of Bibliographic Description

In March 2011, the Library of Congress (2011) announced that it was beginning work to develop an RDF-based replacement for the MARC format. While that document is a preliminary presentation of the work ahead, its orientation towards RDF as a *syntactic* replacement for MARC with a corresponding discussion of overall bibliographic objectives is worth noting. Attempts to provide a mapping from existing MARC data elements to RDF demonstrate the difficulties that lie ahead (Coyle, 2011), but the first steps in this direction demonstrate the importance of this work if we are to understand new bibliographic models that use knowledge representation languages. At present, there seem to be very few conversations about the fundamental differences in paradigm objectives and methods, and in what each entails with regard to syntax. As highlighted by Park & Childress (2009), metadata creators already suffer from confusion about how to apply fundamental KR principles like the *1:1 Principle.* However, this is just the tip of the iceberg.

Parry & Pratty (2008) ask if we, the cultural heritage community, "understand and *need* the semantic web." For them, a core part of the mission of cultural institutions is to make their collections meaningful—in part by providing descriptions of our holdings. However, their approach to semantics illustrates what a slippery concept it is and how it applies to representation formats. In their estimation, the kind of semantics that cultural organizations supply is a broader sense of meaningfulness, not the narrow, rigorous semantics that undergirds the formal models of knowledge representation languages. Moving forward, we need to understand better how formal semantic models can be used to represent other colloquial intuitions about the meaning of cultural heritage metadata.

The questions that arise from contemplating RDF-based approaches to bibliographic data indicate that knowledge organization practitioners will need additional training and support to understand the strengths, weaknesses, and opportunities that they hold. Too often, we immediately see RDF only from our own perspectives within knowledge organization, and we tend to treat different concepts as if they are equivalent (i.e., proceeding as though RDF graphs are equivalent to records). Training cultural heritage information professionals to understand knowledge representation concepts first requires a recognition of how these concepts differ from knowledge organization concepts.

6.4.1 Do We Need and Understand the Semantic Web?

Identifying the distinctions between the knowledge representation and knowledge organization paradigms also raises an important question, one posed by Parry & Pratty (2008): do we need and understand the Semantic Web? The account provided by Parry & Pratty (2008) and others in the knowledge organization community suggests that we do not. As noted above, from Parry & Pratty's perspective, a core part of cultural heritage institutions' mission is to provide *meaning* for our collections. As they note, systematizing rules for cataloging, defining controlled vocabularies, and standardizing data models and syntaxes to exchange information are examples of how the community currently provides meaning to our representations. Like Parry & Pratty, Campbell (2007b) too easily equates the *meaningfulness* of existing knowledge organization methods with the formal semantics found in knowledge representation languages such as RDF. While knowledge representation approaches can be studied and critiqued from a Foucauldian perspective, to do so ignores the fundamental paradigm differences that allow knowledge representation semantics to function as formal computational systems for intelligent reasoning.

The power that such formal systems promise, however, comes at a cost that the knowledge organization paradigm—and the World Wide Web—may not be willing to bear. In order to understand how knowledge organization professionals can best take advantage of more formal approaches, we need greater awareness of what these entail. Furthermore, it may be useful for us to understand the struggles among paradigms that occur within the artificial intelligence and knowledge representation communities. For example, the Semantic Web formalisms discussed in my account of the knowledge representation paradigm have struggled to gain wide adoption on the World Wide Web. Marshall & Shipman (2003) note that even within the Semantic Web community there are disagreements about overall objectives and methods needed to achieve them. In a related article, Shipman & Marshall (1999) outline the difficulties that increasing formalism presents to users of knowledge-base systems. Not only do such formalisms place a heavier burden on users to understand the relationships between syntax and semantics, but they often decrease the efficiency of description tasks by requiring a heavier cognitive load to translate objectives into a formal language (Shipman & Marshall, 1999). While software agents and interfaces may help reduce this load, this will also require translating between user tasks and the needs of a formal representation. Without a thorough understanding of the differences between knowledge organization and knowledge representation paradigms, these kinds of translations will not be possible.

McDonough (2008) suggests that perhaps the digital library community "cares less about interoperability than we thought," as is born out by the metadata quality literature. In that literature, metadata developers "strongly favor local control over encoding practice [that] insur[es] interoperability between institutions" (McDonough, 2008). Here, McDonough is speaking only of syntactic interoperability, not the more complex kinds of semantic interoperability entailed by formal knowledge representation languages.

That the formalisms of the Semantic Web have not succeeded socially is borne out by the growing momentum behind the Linked Open Data (LOD) movement (Berners-Lee, 2006). What has made LOD successful is that it has set aside many of the more formal objectives (such as intelligent reasoning) of the Semantic Web while retaining the data structures, such as RDF, that resulted from this line of research. In many cases, more familiar methods from the knowledge organization paradigm, such as classical information retrieval techniques, have replaced formal reasoning.

A set of problems, known as the *Semantic Web Identity Crisis* or *http range-14 problem* (Hayes & Halpin, 2008; Halpin, 2011), closely parallels the problems of the *1:1 Principle.* At the heart of the problem is the question of whether a URI can refer to both an information object that describes an entity (i.e., a surrogate representation) and the entity being described. Hayes & Halpin (2008) provide the example of a URI that may refer to the Eiffel Tower itself (the structure in Paris designed by Gustave Eiffel) and a photograph of the Eiffel Tower (or equally, a set of RDF statements the Eiffel Tower). According to Hayes & Halpin (2008), what a URI refers to may be specified by the formal interpretation associated with it. In one interpretation, the URI may refer to the surrogate representation (the photo); in another, it may refer to the entity the surrogate stands for (the Eiffel Tower itself). Hayes & Halpin (2008) conform, more or less, to Russell's Theory of Description (as interpreted by Tarski and model-theoretic semantics). In contrast, Berners-Lee (2002) argues that URIs refer to one, and only one, resource, as determined by the agent responsible for "minting" the URI (in part through the authority bestowed by the owner of a domain name). To date, World Wide Web consortium recommendations have supported Berners-Lee's approach (Sauermann & Cyganiak, 2008).

While this issue remains a hot topic of debate within the Semantic Web/Linked Data communities, it also illustrates the subtleties of interpretation that result from adopting these kinds of formalisms. Because this case so closely parallels *1:1 Principle* issues, it deserves closer attention as a way of exploring how the knowledge organization and knowledge representation domains may find ways to work successfully together to achieve shared objectives. As suggested by McDonough (2008), "we need to cease viewing this purely as a technical problem, and acknowledge that it is the result of the interplay of a variety of technical and social factors." If the cultural heritage community is to properly assess where and how to use knowledge representation formalisms, we will need greater understanding of representation paradigms—as sociotechnical phenomena.

References

American Library Association (1967). *Anglo-American cataloging rules, North American text.* Chicago: American Library Association.

American Library Association, Australian Committee on Cataloguing, Canadian Committee on Cataloguing, British Library, & Library of Congress (1988). *Anglo-American cataloging rules.* Chicago: American Library Association, 2nd revised edition.

Arms, C. R. (1999). Getting the picture: Observations from the Library of Congress on providing online access to pictorial images. *Library Trends*, *48*(2), 379–409.

Association of Research Libraries (1990). *Guidelines for bibliographic records for preservation microfilm masters.* Washington, D.C.: Association of Research Libraries.

Attig, J. C. (1989). Descriptive cataloging rules and machine-readable record structures: Some directions for parallel development. In Svenonius, E. (Ed.) *Conceptual foundations of descriptive cataloging.* San Diego: Academic Press.

Baca, M., Harpring, P., Lanzi, E., & Whiteside, A. (2006). *Cataloging cultural objects: A guide to describing cultural works and their images.* Chicago: American Library Association.

Baker, T. (2000). A grammar of Dublin Core. *D-Lib Magazine*, *6*(10). Retrieved from: http://www.dlib.org/dlib/october00/baker/10baker.html.

Barnett, P. & Petersen, T. (1990). Extending MARC to accommodate faceted thesauri: The AAT model. In Molholt, P. & Petersen, T. (Eds.) *Beyond the book: Extending MARC for subject access.* Boston: G.K. Hall, 7–23.

Bates, M. J. (1999). The invisible substrate of information science. *Journal of the American Society for Information Science*, *50*, 1043–1050.

Bearman, D. (1997). Relation element working draft. Dublin Core Metadata Initiative. Retrieved from: http://dublincore.org/documents/1997/12/19/relation-element/.

Bearman, D., Rust, G., Miller, E., Trant, J., & Weibel, S. (1999). A Common Model to support interoperable metadata: Progress report on reconciling metadata requirements from the Dublin Core and INDECS/DOI Communities. *D-Lib Magazine*, *5*(1). Retrieved from: http://www.dlib.org/dlib/january99/bearman/01bearman.html.

Berners-Lee, T. (1998). Why RDF model is different from the XML model. Retrieved from: http://www.w3.org/DesignIssues/RDF-XML.html.

Berners-Lee, T. (2002). What do URIs identify? W3C. Retrieved from: http://www.w3.org/DesignIssues/HTTP-URI.html.

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

Berners-Lee, T. (2009). Axioms of Web Architecture: Using XML for data. W3C. Retrieved from: http://www.w3.org/DesignIssues/XML-Semantics.html.

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic Web. *Scientific American*, *284*(5), 34–44.

Bishoff, L. & Garrison, W. A. (2000). Metadata, cataloging, digitization and retrieval: Who's doing what to whom: The Colorado digitization experience. In *Proceedings of the Bicentennial Conference on Bibliographic Control for the New Millennium, Washington, D.C.* Library of Congress, Cataloging and Distribution Service. Retrieved from: http://lcweb.loc.gov/catdir/bibcontrol/bishoff_paper.html.

Blackwell, E., Beeman, W. O., & McMichael Reese, C. (1988). *Object, image, inquiry: The art historian at work*. Santa Monica: J. Paul Getty Trust.

Blair, D. C. (1992). Information retrieval and the philosophy of language. *The Computer Journal*, *35*(3), 200.

Blair, D. C. (2003). Information retrieval and the philosophy of language. *Annual Review of Information Science and Technology*, *37*(1), 3–50.

Brown, A. D. (1987). *Towards a theoretical information science: Information science and the concept of a paradigm.* Ph.D. thesis, University of Sheffield, Dept. of Information Studies, Sheffield, UK.

Bruce, T. R. & Hillmann, D. I. (2002). The continuum of metadata quality: Defining, expressing, exploring. In Hillmann, D. I. & Westbrooks, E. L. (Eds.) *Metadata in practice*. Chicago: American Library Association, 238–256.

Buckland, M. (1997). What Is a "Document"? *Journal of the American Society for Information Science*, *48*(9), 804–809.

Buckland, M. K. (1991). Information as thing. *Journal of the American Society for Information Science*, *42*(5), 351–360.

Burnard, L., Miller, E., Quin, L., & Sperberg-McQueen, C. (1996). A syntax for Dublin Core metadata. Retrieved from: http://dublincore.org/workshops/dc2/report-19960401.shtml.

Campbell, D. (2007a). Identifying the identifiers. In *Proceedings of the 2007 International Conference on Dublin Core and Metadata Applications*. Dublin Core Metadata Initiative, 74–84.

Campbell, D. G. (2007b). The birth of the new Web: A Foucauldian reading of the Semantic Web. *Cataloging & Classification Quarterly*, *43*(3-4), 9–20.

Caplan, P. (2003). *Metadata fundamentals for all librarians.* Chicago: American Library Association.

Chen, J. (2010). Artificial intelligence. In *Encyclopedia of library and information sciences.* Taylor & Francis, third edition, 289. Retrieved from: http://www.informaworld.com/10.1081/E-ELIS3-120043680.

Collaborative Digitization Program (2006). Dublin Core metadata best practices: Version 2.1.1. Lyrasis. Retrieved from: http://www.lyrasis.org/Products-and-Services/Digital-and-Preservation-Services/Digital-Toolbox/˜/media/Files/Lyrasis/Products%20and%20Services/Digital%20Services/dublin%20core%20bp.ashx.

Copeland, A. (2002). Works and digital resources in the catalog: Electronic versions of Book of Urizen, the Kelmscott Chaucer and Robinson Crusoe. *Cataloging & Classification Quarterly*, *33*(3), 161–180.

Coyle, K. (2006). Identifiers: Unique, persistent, global. *Journal of Academic Librarianship*, *32*(4), 428–431.

Coyle, K. (2011). MARC21 as data: A start. *The Code4Lib Journal*, (14). Retrieved from: http://journal.code4lib.org/articles/5468.

Coyle, K. & Baker, T. (2009). Guidelines for Dublin Core Application Profiles. Retrieved from: http://dublincore.org/documents/profile-guidelines/.

Cromwell-Kessler, W. & Erway, R. (1997). Metadata summit. Retrieved from: http://web.archive.org/web/19971110211243/www.rlg.org/meta9707.html.

Cronin, C. (2008). Metadata provision and standards development at the Collaborative Digitization Program (CDP): A history. *First Monday*, *13*(5). Retrieved from: http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/2085/1957.

Davis, R., Shrobe, H., & Szolovits, P. (1993). What is a knowledge representation? *AI Magazine*, *14*(1), 17.

Dempsey, L. & Weibel, S. (1996). The Warwick Metadata Workshop: A framework for the deployment of resource description. *D-Lib Magazine.* Retrieved from: http://www.dlib.org/dlib/july96/07weibel.html.

Digital Library Federation & National Science Digital Library (2007). Best practices for shareable metadata. Retrieved from: http://webservices.itcs.umich.edu/mediawiki/oaibp/index.php/ShareableMetadataPublic.

Doerr, M., Gradmann, S., Hennicke, S., Isaac, A., Meghini, C., & van de Sompel, H. (2010). The Europeana Data Model (EDM). In *World Library and Information Congress: 76th IFLA General Conference and Assembly.* 10–15.

Dooley, J. M. & Zinham, H. (1990). The object as "Subject": Providing access to genres, forms of materials, and physical characteristics. In Molholt, P. & Petersen, T. (Eds.) *Beyond the book: Extending MARC for subject access.* Boston: G. K. Hall, 43–80.

Dubin, D., Renear, A. H., Sperberg-McQueen, C., & Huitfeldt, C. (2003). A logic programming environment for document semantics and inference. *Literary and Linguistic Computing*, *18*(2), 225–233.

Dublin Core Metadata Initiative (2000a). DCMI 1:1 Principle/Physical Object Description Working Group. Dublin Core Metadata Initiative. Retrieved from: http://dublincore.org/groups/one2one/.

Dublin Core Metadata Initiative (2000b). DCMI Datamodel Working Group. Dublin Core Metadata Initiative. Retrieved from: http://dublincore.org/groups/datamodel/.

Dublin Core Metadata Initiative (2000c). DCMI Relation Working Group. Dublin Core Metadata Initiative. Retrieved from: http://dublincore.org/groups/relation/.

Dublin Core Metadata Initiative (2012a). DC-General Home Page. JISC. Retrieved from: https://www.jiscmail.ac.uk/cgi-bin/webadmin?A0=dc-general.

Dublin Core Metadata Initiative (2012b). DCMI Architecture Forum. Dublin Core Metadata Initiative. Retrieved from: http://dublincore.org/groups/architecture/.

Dushay, N. & Hillmann, D. I. (2003). Analyzing metadata for effective use and re-use. In *Proceedings of the 2003 International Conference on Dublin Core and Metadata Applications*. Dublin Core Metadata Initiative.

Erway, R. (1996). Digital initiatives of the Research Libraries Group. *D-Lib Magazine*. Retrieved from: http://www.dlib.org/dlib/december96/rlg/12erway.html.

Evans, L. J. & Will, M. O. (1988). *MARC for archival visual materials*. Chicago: Chicago Historical Society.

Fattahi, R. (1997). AACR2 and catalogue production technology: Relevance of cataloging principles to the online environment. In Weihs, J. (Ed.) *International Conference on the Principles and Future of AACR (October 23–25, 1997, Toronto, Canada)*. Chicago: American Library Association, 17–43.

Fink, E. (1999). The Getty Information Institute: A retrospective. *D-Lib Magazine*, *5*(3). Retrieved from: http://www.dlib.org/dlib/march99/fink/03fink.html.

Foulonneau, M. & Cole, T. W. (2005). Strategies for reprocessing aggregated metadata. In Rauber, A., Christodoulakis, S., & Tjoa, A. (Eds.) *Research and advanced technology for digital libraries*, volume 3652 of *Lecture Notes in Computer Science*. Springer Berlin/Heidelberg, 290–301. Retrieved from: http://www.springerlink.com/content/ehfgfc0fwjup9nva/abstract/.

Furner, J. (2006). Conceptual analysis: A method for understanding information as evidence, and evidence as information. *Archival Science*, *4*(3–4), 233–265.

Graham, C. (1992). Microform reproductions and multiple versions. *The Serials Librarian*, *22*(1), 213–234.

Greenberg, J. (2010). Metadata and digital information. In *Encyclopedia of library and information sciences*. Taylor & Francis, third edition, 3610–3623. Retrieved from: http://www.tandfonline.com.proxy2.library.illinois.edu/doi/pdf/10.1081/E-ELIS3-120044415.

Guha, R. & Bray, T. (1997). Meta Content Framework using XML. Retrieved from: http://www.w3.org/TR/NOTE-MCF-XML-970606.

Halpin, H. (2004). The Semantic Web: The origins of artificial intelligence redux. Retrieved from: http://www.ibiblio.org/hhalpin/homepage/publications/airedux.pdf.

Halpin, H. (2009). *Sense and reference on the Web.* Ph.D. thesis, University of Edinburgh, Edinburgh. Retrieved from: http://www.era.lib.ed.ac.uk/bitstream/1842/3796/1/Halpin2009.pdf.

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

Han, M., Cho, C., Cole, T., & Jackson, A. (2009). Metadata for special collections in CONTENTdm: How to improve interoperability of unique fields through OAI-PMH. *Journal of Library Metadata*, *9*(3), 213–238.

Harpring, P. & Baca, M. (Eds.) (2009). *Categories for the description of works of art.* Santa Monica: J. Paul Getty Trust. Retrieved from: http://www.getty.edu/research/publications/electronic_publications/cdwa/index.html.

Haslhofer, B. (2008). *A Web-based mapping technique for establishing metadata interoperability.* Dissertation, Universitat Wien, Vienna, Austria. Retrieved from: http://eprints.cs.univie.ac.at/307/1/phd_haslhofer_final.pdf.

Haslhofer, B. & Klas, W. (2010). A survey of techniques for achieving metadata interoperability. *ACM Computer Surveys*, *42*(2), 7:1–7:37.

Haslhofer, B. & Schandl, B. (2008). The OAI2LOD Server: Exposing OAI-PMH metadata as linked data. In *Proceedings of the WWW2008 Workshop on Linked Data on the Web, Beijing, China, April 22, 2008.*

Hayes, P. (2004). RDF Semantics. W3C. Retrieved from: http://www.w3.org/TR/2004/REC-rdf-mt-20040210/.

Hayes, P. J. & Halpin, H. (2008). In defense of ambiguity. *International Journal on Semantic Web and Information Systems*, *4*(2), 1–18.

Heaney, M. (1995). Object-oriented cataloging. *Information Technology and Libraries*, *14*(3), 135–153.

Hillmann, D. (2003). Using Dublin Core. Dublin Core Metadata Initiative. Retrieved from: http://dublincore.org/documents/2003/08/26/usageguide/.

Hillmann, D. (2008). Metadata quality: From evaluation to augmentation. *Cataloging & Classification Quarterly*, *46*(1), 65–80.

Hillmann, D. I., Dushay, N., & Phipps, J. (2004). Improving metadata quality: Augmentation and recombination. In *Proceedings of the 2004 International Conference on Dublin Core and Metadata Applications.* Dublin Core Metadata Initiative.

Hillmann, D. I. & Westbrooks, E. L. (2004). *Metadata in practice.* Chicago: American Library Association.

Hjørland, B. (1998). Information retrieval, text composition, and semantics. *Knowledge Organization*, *25*, 16–31.

Hjørland, B. (2000). Library and information science: Practice, theory, and philosophical basis. *Information Processing & Management*, *36*(3), 501–531.

Hjørland, B. (2003). Fundamentals of knowledge organization. *Knowledge Organization*, *30*(2), 87–111.

Hjørland, B. (2005). Empiricism, rationalism and positivism in library and information science. *Journal of Documentation*, *61*(1), 130–155.

Hjørland, B. (2007). Semantics and knowledge organization. *Annual Review of Information Science and Technology*, *41*(1), 367–405.

Hopmann, A., Berkun, S., & Hatoun, G. (1997). Web Collections using XML. W3C. Retrieved from: http://www.w3.org/TR/NOTE-XMLsubmit.

Horrocks, I., Patel-Schneider, P. F., & van Harmelen, F. (2003). From SHIQ and RDF to OWL: the making of a Web Ontology Language. *Web Semantics: Science, Services and Agents on the World Wide Web*, *1*(1), 7–26.

Hutt, A. & Riley, J. (2005). Semantics and syntax of Dublin Core usage in Open Archives Initiative data providers of cultural heritage materials. In *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries.* JCDL '05, New York: ACM, 262–270.

IFLA Study Group on the Functional Requirements for Bibliographic Records (2009). Functional Requirements for Bibliographic Records: Final Report. Technical report, International Federation of Library Associations and Institutions.

Ingwersen, P. & Jarvelin, K. (2005). *The turn.* New York: Springer.

Johnston, P. & Powell, A. (2008). Expressing Dublin Core Description Sets using XML (DC-DS-XML).

Jones, E. A. (1997). Multiple versions revisited. *The Serials Librarian*, *32*(1-2), 177–198.

Jul, E. (2009). MARC and mark-up. *Cataloging & Classification Quarterly*, *36*(3-4), 141–153.

Khoo, C. S. & Na, J. C. (2006). Semantic relations in information science. *Annual Review of Information Science and Technology*, *40*, 157.

Knowlton, S. A. (2009). How the current draft of RDA addresses the cataloging of reproductions, facsimiles, and microforms. *Library Resources and Technical Services*, *53*(3), 159–165.

Kripke, S. (1980). *Naming and necessity.* Boston: Harvard University Press.

Kuhn, T. S. (1970). *The structure of scientific revolutions.* Chicago: University of Chicago Press.

Lagoze, C. (1996). The Warwick Framework: A container architecture for diverse sets of metadata. *D-Lib Magazine*, *2*. Retrieved from:
http://www.dlib.org/dlib/july96/lagoze/07lagoze.html.

Lagoze, C. (1997). From static to dynamic surrogates: Resource discovery in the digital age. *D-Lib Magazine*, *3*. Retrieved from: http://www.dlib.org/dlib/june97/06lagoze.html.

Lagoze, C. (2001a). Keeping Dublin Core simple: Cross-domain discovery or resource description? *D-Lib Magazine*, *7*(1). Retrieved from: http://dlib.anu.edu.au/dlib/january01/lagoze/01lagoze.html.

Lagoze, C. (2001b). RE: RDF, OAI, and application within libraries. DCMI General Listserv, May 17, 2001.

Lagoze, C. & Van de Sompel, H. (2001). The Open Archives Initiative: Building a low-barrier interoperability framework. Roanoke, VA: ACM.

Lagoze, C., Van de Sompel, H., Nelson, M., & Warner, S. (2008). Open Archives Initiative Protocol for Metadata Harvesting. Retrieved from: http://www.openarchives.org/OAI/openarchivesprotocol.html.

Lassila, O. (1997). PICS-NG metadata model and label syntax. Retrieved from: http://www.w3.org/TR/NOTE-pics-ng-metadata.

Lee, J. H., Renear, A., & Smith, L. C. (2006). Known-Item search: Variations on a concept. In *Proceedings of the American Society for Information Science and Technology*, volume 43. American Society for Information Science and Technology, 1–17.

Lenat, D. B. & Guha, R. (1990). *Building large knowledge-based systems: Representation and inference in the Cyc project.* Reading, MA: Addison-Wesley.

Library of Congress (2010). 1.11A facsimiles, photocopies, and other reproductions. Library of Congress. Retrieved From: http://www.loc.gov/cds/PDFdownloads/lcri/LCRI_2010-03.pdf.

Library of Congress (2011). A bibliographic framework for the digital age. Retrieved from: http://www.loc.gov/marc/transition/pdf/bibframework-10312011.pdf.

Marshall, C. C. & Shipman, F. M. (2003). Which semantic web? In *Proceedings of the Fourteenth ACM Conference on Hypertext and Hypermedia.* HYPERTEXT '03, New York: ACM, 57–66.

McCallum, S. H. (2010). Machine Readable Cataloging (MARC): 1975–2007. In *Encyclopedia of library and information sciences.* Taylor & Francis, third edition, 3530–3539. Retrieved from: http://www.tandfonline.com.proxy2.library.illinois.edu/doi/pdf/10.1081/E-ELIS3-120044392.

McDonough, J. (2008). Structural metadata and the social limitation of interoperability: A sociotechnical view of XML and digital library standards development. In *Proceedings of Balisage: The Markup Conference 2008.* Retrieved from: http://www.balisage.net/Proceedings/vol1/print/McDonough01/BalisageVol1-McDonough01.html.

Miksa, F. (1991). Library and Information Science: Two paradigms? In Vakkari, P. & Cronin, B. (Eds.) *Conceptions of library and information science: Historical, empirical and theoretical perspectives.* Los Angeles: Taylor and Graham.

Miller, E. (1998). An introduction to the Resource Description Framework. *D-Lib Magazine*, *4*(5). Retrieved from: http://www.dlib.org/dlib/may98/miller/05miller.html.

Miller, P. & Greenstein, D. (Eds.) (1997). *Discovering online resources across the humanities: A practical implementation of Dublin Core.* London: UKOLN.

Miller, S. J. (2010). The One-to-One Principle: Challenges in current practice. *International Conference on Dublin Core and Metadata Applications.* Retrieved from: http://dcpapers.dublincore.org/ojs/pubs/article/view/1043/992.

Moen, W. (1998). CIMI Profile release 1.0H: A Z39.50 profile for cultural heritage information.

Most, G. P. (1998). National Gallery of Art Slide Library, Washington, D.C. In McRae, L. & White, L. S. (Eds.) *ArtMARC Sourcebook: Cataloging art, architecture and their visual images.* Chicago: American Library Association.

National Information Standards Organization (2007). A framework of guidance for building good digital collections. National Information Standards Organization. Retrieved from: http://www.niso.org/publications/rp/framework3.pdf.

Nilsson, M., Baker, T., & Johnston, P. (2008). Singapore Framework for Dublin Core Application Profiles. Dublin Core Metadata Initiative. Retrieved from: http://dublincore.org/documents/singapore-framework/.

Nilsson, M., Baker, T., & Johnston, P. (2009). Interoperability levels for Dublin Core metadata. Dublin Core Metadata Initiative. Retrieved from: http://dublincore.org/documents/interoperability-levels/.

NINCH (2002). The NINCH guide to good practice in the digital representation and management of cultural heritage materials. National Initiative for a Networked Cultural Heritage. Retrieved from: http://www.ninch.org/guide.pdf.

Oxford English Dictionary (2011). principle, n.

Palmer, C. L., Zavalina, O. L., & Mustafoff, M. (2007). Trends in metadata practices: a longitudinal study of collection federation. In *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries.* JCDL '07, New York: ACM, 386–395.

Park, J. (2002). Hindrances in semantic mapping involving thesauri and metadata. *Journal of Internet Cataloging*, *5*(3), 59–79.

Park, J. (2005). Semantic interoperability across digital image collections: A pilot study on metadata mapping. *Lecture Notes in Computer Science*, *3237*, 621–630.

Park, J. (2009). Metadata quality in digital repositories: A survey of the current state of the art. *Cataloging & Classification Quarterly*, *47*(3), 213–228.

Park, J. & Childress, E. (2009). Dublin Core metadata semantics: An analysis of the perspectives of information professionals. *Journal of Information Science*, *35*(6), 1–13.

Parker, E. B. & Library of Congress (1981). *Rules for cataloging graphic materials.* Washington, D.C.: Prints and Photographs Division, Library of Congress.

Parry, R. & Pratty, J. (2008). Semantic dissonance: Do we need (and do we understand) the Semantic Web? In Trant, J. & Bearman, D. (Eds.) *Museums and the Web 2008: Proceedings.* Montréal: Archives & Museum Informatics.

Pettee, J. (1985). The development of authorship entry and the formulation of authorship rules as found in the Anglo-American Code. In Carpenter, M. & Svenonius, E. (Eds.) *Foundations of cataloging: A sourcebook*. Littleton, CO: Libraries Unlimited, 75–89.

Powell, A. & Johnston, P. (2002). Guidelines for implementing Dublin Core in XML. UKOLN. Retrieved from: http://www.ukoln.ac.uk/metadata/dcmi/dc-xml-guidelines/2002-01-31/#DCARCH.

Powell, A. & Johnston, P. (2003). Guidelines for implementing Dublin Core in XML. Dublin Core Metadata Initiative. Retrieved from: http://dublincore.org/documents/dc-xml-guidelines/.

Powell, A., Nilsson, M., Naeve, A., Johnston, P., & Baker, T. (2004). DCMI Abstract Model. Retrieved from: http://www.ukoln.ac.uk/metadata/dcmi/abstract-model/2004-11-24/#sect-2.

Powell, A., Nilsson, M., Naeve, A., Johnston, P., & Baker, T. (2007). DCMI Abstract Model. Dublin Core Metadata Initiative. Retrieved from: http://dublincore.org/documents/abstract-model/.

Renear, A., Dubin, D., & Sperberg-McQueen, C. (2000). Towards a semantics for XML markup. In *Proceedings of the Fifth ACM Workshop on Role-based Access Control*. 47–63.

Renear, A., Dubin, D., & Sperberg-McQueen, C. M. (2002). Towards a semantics for XML markup. In *Proceedings of the 2002 ACM Symposium on Document Engineering*. DocEng '02, New York: ACM, 119–126.

Renear, A. H., Wickett, K. M., & Urban, R. J. (2011). Meditations on the logical form of a metadata record. In *Balisage: The Markup Conference*. Montréal.

Research Libraries Group (1997a). Guidelines for extending the use of Dublin Core elements. Retrieved from: http://www.oclc.org/research/activities/past/rlg/dcmetadata/guidelines.htm.

Research Libraries Group (1997b). Metadata Summit summary. Retrieved from: http://www.oclc.org/research/activities/past/rlg/dcmetadata/summit.htm.

Rodriguez, E. E. (2010). Descriptive cataloging principles. In *Encyclopedia of library and information sciences*. Taylor & Francis, third edition, 1481–1492. Retrieved from http://www.informaworld.com/10.1081/E-ELIS3-120043680.

Russell, B. (1905). On denoting. *Mind*, *14*(56), 479–493.

Ryle, G. (1933). 'About'. *Analysis*, *1*(1), 10–12.

Sauermann, L. & Cyganiak, R. (2008). Cool URIs for the Semantic Web. W3C. Retrieved from: http://www.w3.org/TR/cooluris/.

Shatford, S. (1984). Describing a picture: A thousand words are seldom cost effective. *Cataloging & Classification Quarterly*, *4*(4), 13–30.

Shipman, F. M. & Marshall, C. C. (1999). Formality considered harmful: Experiences, emerging themes, and directions on the use of formal representations in interactive systems. *Computer Supported Cooperative Work*, *8*(4), 333–352.

Shreeves, S. L., Knutson, E. M., Stvilia, B., Palmer, C. L., Twidale, M. B., & Cole, T. W. (2005). Is “quality” metadata “shareable” metadata? The implications of local metadata practices for federated collections. In *Proceedings of the Twelfth National Conference of the Association of College and Research Libraries, April 7–10, 2005.* Association of College and Research Libraries, 223.

Shreeves, S. L., Riley, J., & Milewicz, L. (2006). Moving towards shareable metadata. *First Monday*, *11*(8). Retrieved from: http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/1386/1304.

Simonton, W. (1962). The bibliographic control of microforms. *Library Resources & Technical Services*, *6*(1), 29–40.

Smiraglia, R. P. (2001). *The nature of “a work”: implications for the organization of knowledge.* Lanham, MD: Scarecrow Press.

Smith, L. C. (1976). Artificial intelligence in information retrieval systems. *Information Processing & Management*, *12*(3), 189–222.

Smith, L. C. (1980). Artificial intelligence applications in information systems. *Annual Review of Information Science and Technology*, *15*, 67–106.

Snow, M. (1990). Visual depictions and the use of MARC: A view from the trenches of slide librarianship. In Petersen, T. & Molholt, P. (Eds.) *Beyond the book: Extending MARC for subject access.* Boston: G. K. Hall, 225–236.

Society of American Archivists & Hensen, S. L. (2004). *Describing archives: a content standard.* Chicago: Society of American Archivists.

Sowa, J. (2000). *Knowledge representation: Logical, philosophical, and computational foundations.* Pacific Grove: Brooks/Cole.

Spanhoff, E. d. R. (2002). Principle issues: Catalog paradigms, old and new. *Cataloging & Classification Quarterly*, *35*(1), 37–59.

Sperberg-McQueen, C. & Miller, E. (2004). On mapping from colloquial XML to RDF using XSLT. In *Proceedings of Extreme Markup Languages 2004.* IDEAlliance and Mulberry Technologies.

Spinazze, A. (2000). Collaboration, consensus and community: CIMI, museums and the Dublin Core. *Cultivate*, (1). Retrieved from: http://www.cultivate-int.org/issue1/cimi/.

Spinazze, A. (2002). Museums and metadata: A shifting paradigm. In Hillmann, D. I. & Westbrooks, E. L. (Eds.) *Metadata in practice.* Chicago: American Library Association, 37–50.

Stevens, W. (1954). Description without place. In *The collected poems of Wallace Stevens.* New York: Knopf, 1st collected edition.

Stvilia, B. & Gasser, L. (2008). Value-based metadata quality assessment. *Library and Information Science Research*, *30*(1), 67–74.

Stvilia, B., Gasser, L., Twidale, M., Shreeves, S. L., & Cole, T. W. (2004). Metadata quality for federated collections. In *Proceedings of ICIQ04–9th International Conference on Information Quality.* Cambridge, MA: International Conference on Information Quality, 111–125.

Stvilia, B., Gasser, L., Twidale, M. B., & Smith, L. C. (2007). A framework for information quality assessment. *Journal of the American Society for Information Science and Technology*, *58*(12), 1720–1733.

Sundt, C. L. (2002). The image user and the search for images. In Baca, M. (Ed.) *Introduction to art image access: Issues, tools, standards, strategies.* Los Angeles: Getty Research Institute.

Sutton, S. & Miller, E. (1997). Image description on the internet: A summary of the CNI/OCLC Image Metadata Workshop, September 24–25, 1996, Dublin, Ohio. *D-Lib Magazine*, *3*. Retrieved from: http://www.dlib.org/dlib/january97/oclc/01weibel.html.

Svenonius, E. (Ed.) (1989). *Conceptual foundations of descriptive cataloging.* San Diego: Academic Press.

Svenonius, E. (2000). *The intellectual foundation of information organization.* Cambridge: MIT Press.

Svenonius, E. (2004). The epistemological foundations of knowledge representations. *Library Trends*, *52*(3), 571–587.

Tarski, A. (1944). The semantic conception of truth and the foundations of semantics. *Philosophy and Phenomenological Research*, *4*.

Taylor, A. & Joudrey, D. N. (2010). Cataloging. In *Encyclopedia of library and information sciences.* Taylor & Francis, third edition, 3530–3539. Retrieved from:
http://www.tandfonline.com.proxy2.library.illinois.edu/doi/pdf/10.1081/E-ELIS3-120044500.

Thomale, J. (2010). Interpreting MARC: Where's the bibliographic data? *Code4Lib Journal*, *11*. Retrieved from: http://journal.code4lib.org/articles/3832.

Tillett, B. (2001). Bibliographic relationships. In Bean, C. & Green, R. (Eds.) *Relationships in the organization of knowledge.* Boston: Kluwer Academic Publishers.

Tillett, B. (2003). AACR2 and metadata: Library opportunities in the global Semantic Web. *Cataloging & Classification Quarterly*, *36*(3), 101–119.

Tillett, B. & Christan, A. L. (Eds.) (2009). *IFLA cataloging principles: Statement of International Cataloging Principles (ICP) and its glossary.* Number 37 in IFLA Series on Bibliographic Control, Munich: K.G. Saur.

Tjörnebohm, H. (1974). *Paradigm i vetenskapernas vrld och i vetenskapsteorin.* Göteborg, Sweden: University of Göteborg.

Urban, R. J. (2009). *Principle violations: Revisiting the Dublin Core 1:1 Principle.* Dissertation proposal, University of Illinois at Urbana-Champaign.

Van de Sompel, H. & Lagoze, C. (2000). The Santa Fe Convention of the Open Archives Initiative. *D-Lib Magazine*, *6*(2). Retrieved from http://www.dlib.org/dlib/february00/vandesompel-oai/02vandesompel-oai.html.

Vickery, B. (1986). Knowledge representation: A brief overview. *Journal of Documentation*, *42*(3), 145–159.

Visual Resources Association (2007). VRA Core 4.0 introduction. Library of Congress. Retrieved from: http://www.loc.gov/standards/vracore/VRA_Core4_Intro.pdf.

Vitiello, G. (2004). Identifiers and identification systems. *D-Lib Magazine*, *10*(1). Retrieved from: http://www.dlib.org/dlib/january04/vitiello/01vitiello.html.

Weibel, S. (1995). Metadata: The foundations of resource description. *D-Lib Magazine*, *1*(1). Retrieved from: http://www.dlib.org/dlib/July95/07weibel.html.

Weibel, S. & Hakala, J. (1998). DC-5: The Helsinki Metadata Workshop: A report on the workshop and subsequent developments. *D-Lib Magazine*. Retrieved from: http://www.dlib.org/dlib/february98/02weibel.html.

Weibel, S., Ianella, R., & Cathro, W. (1997). The 4th Dublin Core Metadata Workshop report: DC-4. *D-Lib Magazine*. Retrieved from: http://www.dlib.org/dlib/june97/metadata/06weibel.html.

Weibel, S. L. (2010). Dublin Core Metadata Initiative (DCMI): A personal history. In *Encyclopedia of library and information sciences*. Third edition, 1655–1663.

Weibel, S. L., Godby, J., & Miller, E. (2000). OCLC/NCSA Metadata Workshop Report. Internet Archive. Retrieved from: http://web.archive.org/web/20000816021253/ http://www.oclc.org:5046/oclc/research/conferences/metadata/dublin_core_report.html.

Weinheimer, J. (1999). Re: principles and examples. Dc-one2one Listserv, April 29, 1999. Retrieved from: http://dublincore.org/groups/one2one/.

Wendler, R. (1999). Re: 1:1 debate: What's the goal? Dc-one2one Listserv, April 14, 1999. Retrieved from: http://dublincore.org/groups/one2one/.

White-Hensen, W. & Library of Congress (1984). *Archival moving image materials: A cataloging manual*. Washington, D.C.: Motion Picture Broadcasting and Recorded Sound Division, Library of Congress.

Wickett, K. M. (2011). Expressiveness requirements for reasoning about collection/item metadata relationships. In *Proceedings of the 2011 iConference*. iConference '11, New York: ACM, 796–797.

Wickett, K. M., Urban, R. J., & Renear, A. H. (2012). Towards a logical form for descriptive metadata. In *Proceedings of the 2012 iConference*. iConference '12, New York: ACM, 574–575.

Wielinga, B. J., Schreiber, A. T., Wielemaker, J., & Sandberg, J. A. C. (2001). From thesaurus to ontology. In *Proceedings of the 1st International Conference on Knowledge Capture*. K-CAP '01, New York: ACM, 194–201.

Wilson, P. (1968). *Two kinds of power: An essay on bibliographic control*. Number 5 in Librarianship, Berkeley: University of California Press.

Woodley, M. S., Clement, G., & Winn, P. (2005). DCMI Glossary. Dublin Core Metadata Initiative. Retrieved from: http://dublincore.org/documents/usageguide/glossary.shtml.

Zeng, M. L. & Qin, J. (2008). *Metadata*. New York: Neal-Schuman.

Appendix: Metadata Examples

A.1 Example OAI-PMH Record for the *Mona Lisa*

```
<ListRecords>
        <responseDate>2011-11-14T05:01:44Z</responseDate>
        <record>
                <header>
                        <datestamp>2009-03-02</datestamp>
                        <identifier>oai:www.louvre.fr/id/779</identifier>
                </header>
                <metadata>
                        <title>Mona Lisa</title>
                        <title> Portrait of Lisa Gherardini,
                                wife of Francesco del Giocondo </title>
                        <creator>Leonardo da Vinci</creator>
                        <publisher>Musee du Louvre</publisher>
                        <identifier>Inv. 779</identifier>
                        <identifier>http://is.gd/fFbqI</identifier>
                        <date>2008</date>
                        <source>TIFF</source>
                        <type>image</type>
                        <format>oil on poplar board</format>
                        <format>H. 77 cm; W. 53 cm</format>
                        <format>image/jpeg</format>
                        <format>16781 bytes</format>
                        <rights>Copyright 2008 Musee du Louvre/
                                A. Dequier - M. Bard</rights>
                </metadata>
        </record>
<ListRecords>
```

A.2 Example OAI-PMH for the *Mona Lisa* using VRACore

```
<ListRecords>
        <record>
                <header>
                        <identifier>oai:www.louvre.fr/id/779</identifier>
                </header>
                <metadata>
                        <vra>
                        <work id="w779">
                                <title type="popular">Mona Lisa</title>
                                <title> Portrait of Lisa Gherardini,
                                        wife of Francesco del Giocondo </title>
                                <agentSet>
                                        <name>Leonardo da Vinci</name>
                                </agentSet>
                                <identifier>Inv. 779</identifier>
                                <format>oil on poplar board</format>
                                <format>H. 77 cm; W. 53 cm</format>
                        </work>
                        <image id="779_jpeg" href="http://is.gd/fFbqI">
                                <publisher>Musee du Louvre</publisher>
                                <date>2008</date>
                                <techniqueSet>
                                        <display>image/jpeg</display>
                                </techniqueSet>
                                <measurementsSet>
                                        <display>16781 bytes</display>
                                </measurementsSet>
                                <relationSet>
                                        <relation type="imageOf" refid="w779" />
                                </relationSet>
                                <rightsSet>
                                        <display>Copyright 2008 Musee du Louvre/
                                                A. Dequier - M. Bard</display>
                                </rightsSet>
                        </image>
                        </vra>
                </metadata>
        </record>
<ListRecords>
```

A.3 Example Dublin Core Description Set for the *Mona Lisa*

```
<dcds:descriptionSet>
        <dcds:description>
                <dcds:statement dcds:propertyURI="http://purl.org/dc/terms/alternate">
                        <dcds:literalValueString>
                                Portrait of Lisa Gherardini, wife of Francesco del Giocondo
                        </dcds:literalValueString>
                </dcds:statement>
                <dcds:statement dcds:propertyURI="http://purl.org/dc/terms/title">
                        <dcds:literalValueString>
                                Mona Lisa
                        </dcds:literalValueString>
                </dcds:statement>
                <dcds:statement dcds:propertyURI="http://purl.org/dc/terms/creator">
                        <dcds:literalValueString>
                                Leonardo da Vinci
                        </dcds:literalValueString>
                </dcds:statement>
                <dcds:statement dcds:propertyURI="http://purl.org/dc/terms/identifier">
                        <dcds:literalValueString>
                                Inv. 779
                        </dcds:literalValueString>
                </dcds:statement>
                <dcds:statement dcds:propertyURI="http://purl.org/dc/terms/created">
                        <dcds:literalValueString>
                                1501-1519
                        </dcds:literalValueString>
                </dcds:statement>
                <dcds:statement dcds:propertyURI="http://purl.org/dc/terms/format">
                        <dcds:literalValueString>
                                oil on poplar board
                        </dcds:literalValueString>
                </dcds:statement>
                <dcds:statement dcds:propertyURI="http://purl.org/dc/terms/extent">
                        <dcds:literalValueString>
                                H. 77 cm; W. 53 cm
                        </dcds:literalValueString>
                </dcds:statement>
        </dcds:description>
        <dcds:description dcds:resourceURI="http://is.gd/fFbqI">
                <dcds:statement dcds:propertyURI="http://purl.org/dc/terms/format">
                        <dcds:literalValueString>
                                image/jpeg
                        </dcds:literalValueString>
                </dcds:statement>
                <dcds:statement dcds:propertyURI="http://purl.org/dc/terms/extent">
                        <dcds:literalValueString dcds:sesURI="http://www.w3.org/2001/XMLSchema#date">
                                2008
                        </dcds:literalValueString>
                </dcds:statement>
                <dcds:statement dcds:propertyURI="http://purl.org/dc/terms/created">
                        <dcds:literalValueString>
                                16781 bytes
                        </dcds:literalValueString>
                </dcds:statement>
                <dcds:statement dcds:propertyURI="http://purl.org/dc/terms/rights">
                        <dcds:literalValueString>
                                Copyright 2008 Musee du Louvre/A. Dequier - M. Bard
                        </dcds:literalValueString>
                </dcds:statement>
        </dcds:description>
</dcds:descriptionSet>
```

A.4 Example RDF Representation for the *Mona Lisa*

```
<?xml version="1.0"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:dc="http://purl.org/dc/elements/1.1/">
<rdf:Description rdf:about="http://louvre.fr/works/779" />
        <dc:title>Mona Lisa</dc:title>
        <dc:title>Portrait of Lisa Gherardini,
                wife of Francesco del Giocondo</dc:title>
        <dc:creator>Leonardo da Vinci</dc:creator>
        <dc:type>physical object</dc:type>
        <dc:identifier>Inv. 779</dc:identifier>
        <dc:format>oil on poplar board</dc:format>
        <dc:format>H. 77 cm; W. 53 cm</dc:format>
</rdf:Description>
<rdf:Description rdf:about="http://is.gd/fFbqI">
        <dc:publisher>Musee du Louvre</dc:publisher>
        <dc:date>2008</dc:date>
        <dc:source>TIFF</dc:source>
        <dc:relation rdf:resource="http://louvre.fr/works/779" />
        <dc:type>image</dc:type>
        <dc:format>image/jpeg</dc:format>
        <dc:format>16781 bytes</dc:format>
        <dc:rights>Copyright 2008 Musee du Louvre
                /A. Dequier - M. Bard</dc:rights>
</rdf:Description>
</rdf:RDF>
```

A.5 Example of an Ambiguous RDF Representation for the *Mona Lisa*

```
<?xml version="1.0"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:dc="http://purl.org/dc/elements/1.1/">
<rdf:Description>
        <dc:title>Mona Lisa</dc:title>
        <dc:title>Portrait of Lisa Gherardini,
                wife of Francesco del Giocondo</dc:title>
        <dc:creator>Leonardo da Vinci</dc:creator>
        <dc:type>physical object</dc:type>
        <dc:identifier>Inv. 779</dc:identifier>
        <dc:identifier>http://is.gd/fFbqI</dc:identifier>
        <dc:format>oil on poplar board</dc:format>
        <dc:format>H. 77 cm; W. 53 cm</dc:format>
        <dc:publisher>Musee du Louvre</dc:publisher>
        <dc:date>2008</dc:date>
        <dc:source>TIFF</dc:source>
        <dc:type>image</dc:type>
        <dc:format>image/jpeg</dc:format>
        <dc:format>16781 bytes</dc:format>
        <dc:rights>Copyright 2008 Musee du Louvre
                /A. Dequier - M. Bard</dc:rights>
</rdf:Description>
</rdf:RDF>
```

A.6 Example of OAI-PMH XML Translated into RDF by Haslhofer & Schandl (2008)

```
<rdf:Description rdf:about="http://www.mediaspaces.info:2020/resource/item/
                oai:lcoa1.loc.gov:loc.gdc/gcfr.0018_0163">
        <rdf:type rdf:resource="http://www.mediaspaces.info/vocab/oai-pmh.rdf#Item"/>
        <oai2lod:setSpec rdf:resource="http://www.mediaspaces.info:2020/resource/set/ascfrbib"/>
        <oai2lod:origin rdf:resource="http://memory.loc.gov/cgi-bin/
        oai2_0?verb=GetRecord&identifier=oai:lcoa1.loc.gov:loc.gdc/
        gcfr.0018_0163&metadataPrefix=oai_dc"/>
        <owl:sameAs rdf:resource="http://example.com/resource/item/oai:example.com/itemX"/>
        <dc:title>
                Don Christopher Columbus to his friend, Don Louis de Santangel,
                on his arrival from his first voyage. At the Azores, Feb. 15, 1493.
        </dc:title>
        <dc:creator>
                Columbus, Christopher.
        </dc:creator>
        <dc:subject>
                America--Discovery and exploration--Spanish--Early works to 1800.
        </dc:subject>
        <dc:identifier rdf:resource="http://hdl.loc.gov/loc.gdc/gcfr.0018_0163"/>
        <dc:coverage>
                America
        </dc:coverage>
</rdf:Description>
```