COLLECTION/ITEM METADATA RELATIONSHIPS

BY

KAREN MICHELLE WICKETT

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 Allen H. Renear, Chair and Director of Research
Research Associate Professor David Dubin
Associate Professor Jonathan Furner, University of California, Los Angeles
Professor Carole L. Palmer

Abstract

In information organization systems, metadata is often attached to both collections and items. Collection metadata and item metadata are related: one can infer facts about items from descriptions of collections, and facts about collections from descriptions of items. This sort of reasoning, which is important to finding, understanding, and using information, is guided by specific, if usually only implicit, inference rules. This dissertation explores the general nature of these rules and develops a logic-based framework of categories for collection/item metadata rules. The resulting framework has 28 rule categories related by two logical relationships. This framework has practical applications in metadata vocabulary development, metadata-enabled search and retrieval, and metadata quality and completeness. A number of foundational questions are also discussed, including the ontological nature of collections, the logic of the collection membership relationship, the semantic and logical nature of collection/item inference rules, and difficulties in the translation of colloquial metadata records into a logic-based knowledge representation language.

Acknowledgments

Portions of this work reflect earlier research carried out primarily with Allen Renear, Richard Urban, and Dave Dubin, and supported by a 2007 IMLS NLG Research & Demonstration grant as part of the IMLS Digital Collections and Content project (DCC), Principal Investigator, Carole L. Palmer, Center for Informatics Research in Science and Scholarship (CIRSS). That work also benefitted from discussions with other DCC Collection/Item Metadata Group (CIMR) members, including Katrina Fenlon, Jacob Jett, Wu Zheng, and Larry Jackson. Additional funding was provided by the University of Illinois Graduate College.

Portions of this dissertation have been previously published as articles on which I was lead author.

> Adapted as Chapter 2: Wickett, K. M., Renear, A. H., and Furner, J. (2011). "Are collections sets?" In *Proceedings of the 74th ASIS&T Annual Meeting.*
>
> Adapted as portions of Chapter 3: Wickett, K. and Renear, A. (2012). "The logical form of a metadata record." In *Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries.*
>
> Adapted as portions of Chapter 4: Wickett, K. M. (2009). "Logical expressiveness of semantic web languages for bibliographic modeling." In *Proceedings of the iConference*, and Wickett, K. M. (2011). "Expressiveness requirements for reasoning about collection/item metadata relationships." In *Proceedings of the iConference.*
>
> Adapted as portions of Chapter 1 and Chapter 5: Wickett, K. M., Renear, A. H., and Urban, R. J. (2010). "Rule categories for collection/item metadata relationships." In *Proceedings of the 73rd ASIS&T Annual Meeting.*

In addition, portions of the introduction are based on Renear, A. H., Wickett, K. M., Urban, R. J., Dubin, D., and Shreeves, S. (2008). "Collection/Item metadata relationships." In *Proceedings of the International Conference on Dublin Core and Metadata Applications.*

I am deeply grateful to my family, my spouse Adam D. Miller, my many wonderful friends, my committee members, and the GSLIS community for their support and encouragement throughout this process.

Table of Contents

Chapter 1

Introduction

[1] In information organization systems, metadata is often attached to both collections and items. Collection metadata and item metadata are related: one can infer facts about items from descriptions of collections, and facts about collections from descriptions of items. For instance, one might conclude from the fact that a collection is owned by a particular person that a given item in that collection is owned by that person as well. This sort of reasoning can be extremely important to finding, understanding, and using information. Like any reasoning, reasoning about collection and items is guided by inference rules, such as, in this case: whoever owns a collection owns each of its items. These rules are often only implicit, but they provide the basis for the perceived validity of the inference.

This dissertation explores the general nature of these inference rules and develops a logic-based framework of categories for collection/item metadata rules. The resulting framework has 28 rule categories related by two logical relationships. As explained later in this chapter this framework has practical applications in metadata vocabulary development, metadata-enabled search and retrieval, metadata quality and completeness, and in other areas as well. I also explore key foundational questions including the ontological nature of collections, the logic of the collection membership relationship, the semantic and logical nature of collection/item inference rules, and difficulties in the translation of colloquial metadata records into a logic-based knowledge representation language.

Collections are a prominent entity in library, museum, and archival practices. The concept of a *collection* might even be considered fundamental to information organization

[1] Portions of the introduction were adapted from Renear, A. H., Wickett, K. M., Urban, R. J., Dubin, D., and Shreeves, S. (2008). "Collection/Item metadata relationships." In *Proceedings of the International Conference on Dublin Core and Metadata Applications*

in general. Recently researchers in library and information science have recognized the importance of collections in information organization systems and have begun study them from a variety of perspectives, including the role of particular kinds of collections in scholarly activities (Palmer, 2004), their general features as informational artifacts (Lee, 2000, 2005), and how traditional archival notions of collecting are evolving (Yeo, 2012). However there is still no systematic understanding of collections in terms of the logical semantics of collection membership or the ontological status of collections themselves. The research presented here addresses this need. These are first steps towards a more adequate understanding of a foundational concept in information science.

1.1 Collection/Item Metadata Relationships

By supplying structured descriptions of features such as the subject, audience, or purpose of a collection, collection-level metadata provides essential contextual information that can be used in retrieval, access, and evaluation. Thoroughly developed collection-level descriptions may also include facts about the method of selection used in building a collection, descriptions of the nature of items, an account of the geographic or temporal coverage, indications of the completeness of the collection, and many other features (Heaney, 2000; Lagoze et al., 2006; Palmer et al., 2006). The availability of information of this sort enables collections to fulfill their distinctive role in the research process (Brockman et al., 2001; Palmer, 2004).

However, the structure and context provided by collection membership has been underutilized in many information systems. Many retrieval and browsing systems fail to exploit collection-level metadata, which reduces their effectiveness and because researchers are not able to use the contextual information provided by collection descriptions (Foulonneau et al., 2005; Wendler, 2004). For example, systems that aggregate descriptions from multiple sources often focus solely on item-level descriptions and do not incorporate information available in collection-level metadata (Christenson and Tennant, 2005; Foulonneau et al., 2005; Lagoze et al., 2006; Warner et al., 2007). The failure to fully conceptualize collections and collection membership also means that information systems designers are missing op-

portunities for the creation, maintenance, and validation of metadata at both the collection and item levels in digital libraries and repositories.

To support tools that make the most of both item and collection metadata we need a much better understanding of the kinds of logical relationships between collection-level and item-level metadata. The precise formalization of those relationships is required so that they may be used by retrieval and browsing systems. Toward this end, the Collection/Item Metadata Relationships (CIMR) group developed a logic-based framework of relationship rule categories for collection/item metadata (Renear et al., 2008b; Wickett et al., 2010). The CIMR group was formed in 2007 as part of the IMLS Digital Collections and Content (DCC) project[2], with the goal of exploring how a formal description of collection/item metadata relationships could help users of the DCC Collections Registry locate and use digital items.

The primary result of the CIMR research was a method for expressing relationships between collection-level and item-level descriptions as *propagation rules*, and a framework for organizing rules according to their logical features. Propagation rules are inference rules where the properties of one individual entail conclusions about another individual in virtue of a particular relationship those individuals bear to each other (Brachman et al., 1991; Seidenberg and Rector, 2006). While propagation may seem similar to inheritance, it differs in being grounded not on subclass or membership relationships, but on specific non-logical relationships, such as "part of" or in the case of collections, "is gathered into". As a consequence, propagation is more challenging than inheritance to identify and illuminate, but potentially a rich source of insight into the distinctive features of the grounding relationship, in this case collection membership.

[2]Funded by a 2007 IMLS NLG Research Demonstration grant; Principal Investigator, Carole L. Palmer, Center for Informatics Research in Science and Scholarship (CIRSS). Project documentation: http://imlsdcc.grainger.uiuc.edu/about.asp.

1.2 The CIMR Framework

Earlier work presents the basic strategy for a framework of rule categories and discusses some of the technical issues involved (Renear et al., 2008b; Wickett et al., 2009). The potential of this approach to expressing metadata relationships has been noted by metadata researchers (Greenberg, 2009; Lourdi et al., 2009), although the preliminary example categories offered in Renear et al. (2008b) cover only a small portion of the semantics of metadata in common use. Further analysis and an empirical examination of actual metadata assignments led to the larger, more fine-grained framework presented in Wickett et al. (2010), which consists of a total of 18 rule categories. The fully developed framework presented in Chapter 5 of this dissertation consists of 28 rule categories and three levels of specificity. The framework reveals the logical entailments among rule categories and identifies key formal characteristics of metadata constraint relationships that provide the foundation for inferencing.

Metadata relationships are represented as rules, where a rule is a logical conditional, an assertion that if something is the case then something else is the case. For example, the metadata attribute *marcrel:own*, used to identify the owner of a resource, and the relevant concept of ownership might imply a rule that states that if someone owns a collection then they own each item in the collection. Whether or not a rule like this holds will depend on the social and legal environment in which the attribute is being used. The purpose of this framework is not to decide which rules apply in what circumstances, but to provide for an explicit representation of this kind of semantic determination for a particular domain vocabulary or repository system. Additionally, as demonstrated in Wickett et al. (2010), the framework allows an empirical exploration of patterns and relationships in a set of metadata without prior commitments or expectations for relationships.

The ownership rule can be expressed with regimented English as:

For any x and for any y,

if x is a collection and is owned by y,

then for any z,

if z is gathered into x (is a member of the collection),

then z is owned by y.

And the rule can be represented in first order logic as:

$$\forall x \forall y((ownedBy(x,y)\,\&\,Collection(x)) \supset \forall z(isGatheredInto(z,x) \supset ownedBy(z,y))) \tag{1.1}$$

The logical characterization uses the full resources of first order predicate logic, including a functionally complete set of connectives, universal and existential quantifiers, and polyadic predicates. Monadic predicates are used to express class membership ("y is a collection" is expressed as "$Collection(y)$"), and dyadic predicates are used to express descriptive attributes (as in "$ownedBy(y,z)$" to express "y is owned by z"). The predicate $isGatheredInto(x,y)$, used to express the collection membership relationship, is based on the Dublin Core Collections Application Profile (Dublin Core Collection Description Task Group, 2007).

Full first order logic is an expressive logical language and presents difficulties for efficient (or even decidable) reasoning. However, the work of understanding the logical relationships between descriptions is largely conceptual and not bound to particular reasoning problems. Implementation of rule-based systems will have to take efficiency problems into account, but for conceptual analysis the expressiveness of first order logic is a good match for representing descriptions and their possible relationships.

The statements that express metadata relationships are about properties of things, and not necessarily how they are recorded in metadata. For example, ownership may be encoded in a record in a number of ways, with a structured XML element like "`<marcrel:OWN>`", with a "`<dcterms:rightsHolder>`" element, or with a free-text description field in a record.

Chapter 3 provides a detailed exposition of the translation from typical metadata records in XML or other similar languages into first order logic.

Regardless of the form any particular metadata record takes, the property expressed is the dyadic ownership relation between a resource and the owner of that resource. In some cases, the property may not be expressed directly in a record at all. This is frequently the case with collection membership, which can be inferred from the ways resources are organized or described, but may not be stated directly at the collection or item levels.

The framework developed here is based on the fact that rules can be grouped together according to their logical form. This logical form is determined by a number of features discussed in detail in Chapter 5. These features include:

- the "direction" of reasoning – from items to collections, or from collections to items

- quantification of items in rules – either existential or universal quantification

- number of properties used in the antecedent of rules – single or multiple.

Rules that have the same logical form belong to the same *rule category*. Rules belonging to the same rule category have the same kinds of logical relationships and the categories themselves have systematic interrelationships, as described in Chapter 5.

The members of a particular rule category vary only in the predicate constants that appear in the rules. Therefore it is possible to generate candidate rules from the framework by replacing the schematic predicate symbols that appear in the framework (e.g. A or B) with predicate symbols based on the attributes from a particular metadata vocabulary (e.g. $ownedBy$, $itemType$, or $type$). This systematic generation of candidate rules will allow repository managers and the developers of metadata vocabularies to propose and assess potential rules for their use with a vocabulary or repository.

1.3 Research Problems

The framework developed in previous work on collection/item metadata relationships was promising and demonstrated the potential of using first order logic to represent relationships

between collection-level and item-level description. However, several important issues were not addressed.

The previous work did not discuss a number of relevant issues that seem foundational to understanding collection/item metadata relationships. These include: the logical properties of the collection membership relationship, the ontological status of collections, the logical and computational nature of propagation rules, and the semantics of colloquial metadata records that provide the assertions about collections and items. Exploration of these basic topics is necessary in order to understand the assumptions being made for the development of a rule framework.

Perhaps most importantly, the framework of rule categories presented in Wickett et al. (2010) was not comprehensive in coverage. It did not address drawing conclusions about collections on the basis of item descriptions, or rules that operate on the basis of a combination of attributes at either the collection level or the item level. These gaps also meant that the entire set of relationships between rule categories had not been explored. As outlined in detail in Section 1.5, this dissertation addresses these research problems in order to supply a systematic foundation for the representation and exploitation of relationships between collection-level and item-level metadata.

1.4 Applications

A formal framework for collection/item metadata relationships could have a variety of applications, such as improving the specification of metadata vocabularies, enhancing the performance of retrieval systems, supporting the creation and validation of of metadata and both levels of description.

1.4.1 Metadata Vocabulary Development

Metadata vocabularies are often specified by giving definitions and usage notes, or by the use of application profiles, which allow selection and refinement of terms from known vocabularies. In either of these methods for vocabulary development, the semantics of attributes

are typically described using natural language prose. The incorporation of inference rules that express the expected relationships between attributes can support a more explicit specification of this relational aspect of the semantics of the attributes involved. Currently, these relationships are usually informal and implicit, latent in the natural language prose of the specification, or just existing as general background assumptions with little or no documentation.

The framework of rule categories presented in this dissertation can support the specification of intended relationships between attributes, by providing a system of regularized categories of rules. This framework of categories is organized according to the logical form of rules, and will allow metadata developers to indicate which rules match the attribute semantics required for a particular application. Metadata schemas could be annotated with a controlled vocabulary of rule names to support automated processing, adding some formal semantics to an otherwise purely syntactic schema. In addition, when the metadata vocabulary is already in use, the rules will allow repository managers to determine whether current practice confirms or refutes their suppositions about the relationships between attributes. Repository managers and the developers of metadata vocabularies can annotate their schemas and definitions by including specific rules derived from the framework, or with references to the applicable rule forms from the framework.

While the framework of rules described in the previous work from the CIMR group and the extended version developed here are targeted at rules that operate between collection-level and item-level description, the method of using rules to express relationships between attributes could be extended to other kinds of relationships. Rules could express relationships between attributes at a single level of description, such as constraints between creation dates of resources and resource formats (e.g. TIF) that were introduced at some particular date.

1.4.2 Improving Search and Retrieval

Rules that express collection-level and item-level metadata can be used to enhance search for both collections and items. By displaying information that has been propagated between levels of description, search systems can supply valuable contextual information that can aid users in navigation and selection of resources.

Digital library systems and aggregations that bring together descriptions of resources may index very large numbers of item records. These records often come from a number of sources, and vary significantly in terms of the completeness and depth of description, and aggregations may include item-level records that are very sparse. In cases where collection-level description is available, it could be used to enhance search for items. For example, the IMLS Digital Collections and Content aggregation [3] takes a basic approach to this kind of enhancement by displaying information about collections that an item is a member as an element of the display of item records. These collection records supply contextual information about an item that users can use in their selection of resources.

The specification of collection/item metadata relationships can be used to refine this method for supplying contextual information, by targeting specific collection-level properties that are deemed likely to contain the kinds of information most likely to aid users. In addition, users themselves could indicate properties of particular interest for propagation to the item level. For example, a teacher may want to have information about the intended audience of collections displayed as an element in item-level search results, in order to quickly assess the relevance of items for a particular educational use.

Collection membership can also be used to structure search results into more meaningful displays. Large aggregations may include many items with sparse descriptions that include very similar titles (e.g. "Number 51" from a series of photographs). Depending on how retrieval is implemented, this can lead to results lists where the initial results are extremely homogenous, organized only by uninformative titles. In order to help users better navigate through such resources, the display of results can incorporate collection-level information.

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

This can currently be achieved with an option to switch to a view of the collections that contain retrieved items (as can be found in the most recent IMLS DCC interface), or by organizing results into blocks by collection membership. A fully operational system of propagation rules can further support this kind of enhancement by allowing collection-level information that may further distinguish groups of items to be propagated to those items. This would allow retrieval and ranking that factors in aspects of collection-level description such as topical or geographic coverage, which may not have originally been part of the item records.

1.4.3 Improving Metadata Quality and Completeness

The managers of digital library aggregations can use relationships between collection-level and item-level metadata to improve the quality of their metadata by discovering inconsistencies in metadata assignments between the collection and item levels, and by assisting in the creation of records at both levels of description.

Rules that express relationships between descriptions of items and descriptions of collections that those items are members can be used for the validation of metadata. Suppose that a repository manager has selected a propagation rule which states that if a date range is given for the creation dates of items within a collection, then every item in the collection has creation dates that are within that range. This rule could be used to detect item-level records that describe items as having creation dates outside of the range given by the associated collection-level record. These discrepancies may indicate a simple errors at the item or collection levels, or may reflect previous use of an attribute for some non-standard purpose. If a large number of items display unexpected values for attributes, but seem to be accurate and to follow current best practice, this may indicate that the collection has shifted focus and that reassessment of the collection policy is called for. Similar methods could be used to detect inconsistencies at the collection level based on item records.

Relationships between collection-level and item-level description may also be used to facilitate the creation of new records. Propagation rules that support inferences about

collections on the basis of descriptions of items could be used to support the creation of collection records. Given item records and rules, it would be possible to build basic collection records by propagating values from the item records or computing values (such as a date range) where appropriate. These records may not be of sufficient quality without review, but could act as a valuable intermediate step for repository managers with limited time for the development of detailed collection descriptions. A similar method could be used to build initial item-level records in cases where only collection-level records are currently available, or to supplement sparse item-level records.

1.5 Organization of Chapters

Four chapters make up the main body of this dissertation.

Chapter 2 presents an analysis of the concept of a collection as it is commonly used in information organization systems and digital libraries. This analysis focuses in particular on the relationship between collections and their members, as this relationship is the basis for reasoning about collections and items between levels of description. The collection membership relationship *isGatheredInto* is then analyzed in detail in terms of its relation properties, possible axioms to describe the behavior of relation in the context of information organization, and the similarity to other relations that give structure to collective entities. The question of whether collections are mathematical sets is considered in detail, and a proposal to define a collection as a set of objects in a particular curatorial role is introduced.

Chapter 3 systematically explores the semantics of metadata records. In digital library systems information about collections and items both are typically expressed with metadata records. If we are to connect our formal framework for collection/item inferencing with these actual assertions about collections and items then we must have a system for translating ordinary metadata records into a formal knowledge representation language, such as first order logic. In this chapter three approaches to analyzing the semantics of ordinary metadata records are explored and the appropriateness of these different approaches to the formal representation of information about collection and items, and to the formal

representation of collection/items inferencing rules, is discussed.

Chapter 4 describes the general form of rules for collection/item metadata relationships in terms of the logic of *propagation rules*. The nature of propagation rules as conditionals and their expression in logic is discussed. The general logical form of propagation rules is given, along with the four general categories of propagation rules for reasoning about collection-level and item-level description. The technique of pre-emptive modal exclusion is introduced a method to handle the "paradoxes" of material and strict implication. The expressive requirements of knowledge representation languages that might be used to reason about metadata on the basis of propagation rules are examined. It is shown that propagation rules for collection/item metadata relationships are equivalent to Horn Clauses, which is a promising result in terms of building inference systems for reasoning about metadata. However, the kind of general inference rule expressed in a propagation rule is outside the scope of the Web Ontology Language (OWL).

Chapter 5 presents a comprehensive framework of rules for collection/item metadata relationships based on the logical form of the rules. The top-level categories are based on the "direction" of reasoning; whether rules support reasoning about items on the basis of collection description, or reasoning about collections on the basis of item description. Further categories are based features such as quantification at the collection and item levels, and the relationships between the attributes involved in a rule. General rule schemas are given for each category of rules, and the relationships between the categories are examined.

Together these chapters not only support improvements in our ability to reason about information objects and create rigorous documentation for metadata systems, but they are also first steps towards a more adequate understanding of several foundational concepts in information science.

Chapter 2

The Logical Nature of Collections

Collection/item inferencing is about collections and the items that they contain. Therefore, as preparation for the analysis of relationships between collection-level and item-level description, we begin with an examination of the concept of *collection* and the *isGatheredInto* relationship. The primary focus in this chapter is on one specific, but critical, question: are collections sets? More precisely the question is whether collections are sets (in the common mathematical sense of "set") of their members.

This question of whether collections are mathematical sets is addressed in what follows in two ways. The first is by consideration of axioms that describe the *isGatheredInto* relationship, which is the membership relationship between items and collections. Conjecturing and evaluating axioms about the behavior of this distinctive relationship provides a generalized picture of the relationship, and reveals its relation properties. This detailed account of the *isGatheredInto* relationship supports the comparison of the relationship with other relationships that describe membership in collective entities (such as part-hood and set membership) with known relation properties.

Next the relationship between collections and sets is addressed with a detailed analysis of the possible interpretations of the claim that collections are sets and the consequences that arise from those interpretations. The claim can be understood as asserting (i) that the collection and set are equivalent concepts, or (ii) that a collection is a kind of set. Counterexamples to both of these interpretations are presented and discussed, with particular attention to the fact that identifying collections as a kind of set will imply that collections have their members essentially.

Finally, an approach to the ontological status of collections that defines a collection

as a set in a particular informational or curatorial role is presented. This is a theory which is refinement of the claim that collections are a kind of set, and states that being a collection is a property that sets have only in certain contingent circumstances. Although the results are not conclusive with respect to either the ontological nature of sets, nor the axioms governing *isGatheredInto*, they do provide a better understanding of the constraints that different approaches entail with respect to collection description and reasoning about collections and items.

2.1 Introduction

[1]Many empirical studies in library and information science have addressed the role of particular kinds of collections in scholarly activities (Palmer, 2004) as well as the general features of collections as informational artifacts (Lee, 2000, 2005). However, there has been very little attention given to determining the ontological status of collections or the semantics of collection membership. It is not unusual to hear collections described variously as "sets" (Gonçalves et al., 2004), "groups" (Galton, 2010), "aggregations" (Le Boeuf et al., 2012), or "selections" (Lagoze and Fielding, 1998). Yet it is unclear what exactly is intended by these terms when they are used casually, or even whether anything very specific and definite is intended at all.

Although it is reasonable to think that a notion this fundamental to information science deserves a systematic analysis for that reason alone, there are practical applications as well. The category framework that was developed in the work that preceded this dissertation (Renear et al., 2008b; Wickett et al., 2010) was largely determined by fairly obvious relationships between properties at different levels, it became clear that there would be issues requiring a more precise understanding of the concept of a collection than was currently available. The way that these uncertainties are resolved could make a difference to the emerging account of collection, shifting the pattern of inferences in one direction or an-

[1]Material in this chapter was adapted from Wickett, K. M., Renear, A. H., and Furner, J. (2011). "Are collections sets?" In *Proceedings of the 74th ASIS&T Annual Meeting.*

other. The particular question attended to in this chapter is whether modeling collections as mathematical set of their members accurately reflects the nature of collections as entities in information organization systems.

Collections are certainly often described as sets. This is not only a frequent casual characterization (e.g. Lagoze and Fielding (1998)), but also found explicitly in formal models of digital library systems (Gonçalves et al., 2008; Meghini et al., 2010). This is not surprising. The colloquial notion of set seems to have ubiquitous application, the corresponding mathematical concept is well-defined and widely deployed, and upper-level formal ontologies frequently have sets as a fundamental kind of thing (e.g. Niles and Pease (2001)). But are collections sets? And how do we tell?

2.1.1 The Concept of a Collection

The introduction of digital resources into library catalogs provided a valuable opportunity to review how collection development and management functions were addressed in libraries (Buckland, 1995; Atkinson, 1998). This was also an opportunity to "reconceptualize collection" in terms of how collections serve particular purposes for stakeholders, instead of relying on traditional notions of collections rooted in physical proximity (Lee, 2000; Casserly, 2002). This shift in focus had a corresponding impact on evaluating user needs and how they can be met (Covi and Cragin, 2004; Kaczmarek, 2006).

For the purpose of developing a general conceptual account of collections, these reconceptualizations to accommodate digital collections provide deliberate and general characterizations of collections. Digitization projects continue to drive this work, both in the context of individual projects (e.g. M'kadem and Nieuwenhuysen (2010)) and in general (Lynch, 2002; Moss and Currall, 2004). However, these efforts have not, so far, resulted in a consistent accepted definition of collections. This gap has been noted several times since digital collections brought the issue into focus (Hill et al., 1999; Lee, 2000; Currall et al., 2004).

In spite of the lack of clarity on what precisely collections are, arguments in favor of

the usefulness of collection description point to benefits of collection description not for just management of collections, but for supporting scholarship and providing contextual information (Brack et al., 2000; Sweet and Thomas, 2000; Foulonneau et al., 2005; Palmer et al., 2006). Collection descriptions are designed to provide the contextual information (e.g. information about locations, times and related events, provenance, collection method) that aid scholars in making sense of the items within a collection.

Heaney (2000) provides an entity-relationship model of collections that has informed several schemas for collection description (Powell et al., 2000; Shreeves and Cole, 2003; Dublin Core Collection Description Task Group, 2007). In specifying what properties a collection can have and what relationships it can enter into such descriptive schemas and, where they exist, conceptual models (like Heaney's) provide insights into what collections are, as well as a useful starting point for terminology and informal definitions. These models also suggest examples of the sort of problems the lack of a robust definition of collection might cause when automatic inferencing and semantic technologies are common: an examination of collection-level description reveals some attributes that are consistent with some possible concepts of collection and some with other concepts of collection.

However, the restricted expressiveness of ER and UML modeling languages, and the specific systems design focus of most of these conceptual models, limits their usefulness for the project here. The analysis below will instead use first order logic, which provides additional expressiveness, precision, and clarity, and better supports identification of entailments.

2.2 The *isGatheredInto* Relation

The predicate $isGatheredInto(x, y)$ is used in what follows for the relation that stands between an item (x) and a collection (y) of which the item is a member. This name for the relationship comes from Heaney's model of collections (Heaney, 2000), and is used in the Dublin Core Collections Application Profile (Dublin Core Collection Description Task Group, 2007). Until there is compelling reason found to believe otherwise, it is assumed that this is a distinct relationship that is not necessarily equivalent with any other known

relationships from logic or ontology.

A one-place predicate $Collection(x)$ is used to represent the property of being a collection. The predicates $isGatheredInto(x, y)$ and $Collection(x)$ are clearly closely related, and discussion of the possibility of reducing one of these concepts to the other is presented later. The following informal definitions appear in the Dublin Core Collections Application Profile (Dublin Core Collection Description Task Group, 2007):

> Collection: An aggregation of Items.
>
> Item: A physical or digital resource.

And the remark:

> an Item is-Gathered-Into one or more Collections.

2.2.1 Axioms Relating *Collection* and *isGatheredInto*

We now consider some possible axioms for $isGatheredInto(x, y)$. The axiom below seems to follow immediately from the informal definitions of *isGatheredInto* just mentioned – if something y has something x gathered into it, then that thing y is a collection.

$$\mathbf{A1} : \forall y(\exists x\, isGatheredInto(x, y) \supset Collection(y)) \tag{2.1}$$

On the other side of the *isGatheredInto* relationship are items, which can be characterized with the predicate $Item(x)$, and another axiom that states that if something is gathered into a collection, then it is an item.

$$\mathbf{A2} : \forall x(\exists y\, isGatheredInto(x, y) \supset Item(x)) \tag{2.2}$$

The converse of the first axiom states that if a resource is a collection, then it has something gathered into it.

$$\mathbf{A3} : \forall y(Collection(y) \supset \exists x\, isGatheredInto(x, y)) \tag{2.3}$$

Accepting this axiom will, in effect, deny the existence of "empty collections", collections with no members. Whether or not this is a reasonable axiom for $isGatheredInto(x, y)$ is taken up later.

We can ask whether it is possible to gather a collection into a collection as a member by considering a domain restriction axiom.

$$\mathbf{A4} : \forall x \forall y(isGatheredInto(x, y) \supset \neg Collection(x)) \tag{2.4}$$

While there are often salient sub-collections within a collection, it is not clear that a collection that is part of a larger collection has been gathered into a collection in the same sense that items are. Given a case where collections are brought together to form a larger collection, we can make a distinction between understanding the curator to be gathering those collections into a larger collection and understanding the curator to be gathering the individual items from those collections into the larger collection on the basis of their membership in some previous collection.

2.2.2 Relation Properties of *isGatheredInto*

Intuitions about the various relation properties of the *isGatheredInto* relation can be used to assess further potential axioms with respect to collection membership. Since relation properties describe how a relation acts over a domain, it is necessary to consider the domain for the formal characterizations of the relationship between collections and items. Previous work on collection/item metadata relationships (Wickett et al., 2010) developed rules that use a basic universal domain (the variables can take anything as values), and we take the same approach here. Cases that require further restriction of the domain can be handled by adding conditions to axioms as needed.

Reflexivity

A reflexive relation is one that every member of the domain stands in with itself. For example, greater-than-or-equal-to ($\geq$) is a reflexive relation over the integers, since every

integer is equal to itself.

A reflexivity axiom for $isGatheredInto$ would be:

$$\mathbf{R1}(Ref) : \forall x(isGatheredInto(x, x)) \tag{2.5}$$

R1 states that everything is gathered into itself. Since that implies, by A1, that everything is a collection , R1 is clearly not a plausible axiom.

An irreflexive relation is one that no member of the domain stands in with itself. The relation greater than is irreflexive over the integers since no integer is greater than itself.

Irreflexivity of $isGatheredInto$ would result in this axiom:

$$\mathbf{R2}(Irr) : \forall x \neg(isGatheredInto(x, x)) \tag{2.6}$$

R2 states that no collections are members of themselves. This axiom is much more plausible. It is not only difficult to imagine a collection that is a member of itself in a normal course of the use and development of collections of information resources, but it is hard to see what it would even mean for a collection to be gathered into itself.

Symmetry

A symmetric relation is one where, given any x and y in the domain, if x bears the relation to y, then y bears the relation to x. For example, the relation $marriedTo$ is symmetric, since if x is married to y, then y is married to x.

Symmetry of $isGatheredInto$ would give us the axiom:

$$\mathbf{R3}(Sym) : \forall x \forall y(isGatheredInto(x, y) \supset isGatheredInto(y, x)) \tag{2.7}$$

which would mean that every member of a collection has that collection as a member. This is certainly not plausible as a general axiom for collections.

An asymmetric relation is one where, given any x and y in the domain, if x bears the

relation to y, then y does not bear the relation to x. Asymmetry of $isGatheredInto$ would result in the following axiom:

$$\mathbf{R4}(Asym) : \forall x \forall y(isGatheredInto(x, y) \supset \neg isGatheredInto(y, x)) \tag{2.8}$$

which states that if something is a member of a collection, then that collection is not member of it in turn. Since it is difficult to imagine under what circumstances we would ever consider a collection to be gathered into one of its own members this seems a reasonable axiom for $isGatheredInto$.

An antisymmetric relation is one where, for any x and y, if x bears the relation to y and y bears the relation to x, then x and y must be the same thing. Greater-than-or-equal-to ($\geq$) is an example of an antisymmetric relation.

$$\mathbf{R5}(AntiSym) : \forall x \forall y((isGatheredInto(x, y) \,\&\, isGatheredInto(y, x)) \supset x = y) \tag{2.9}$$

If $isGatheredInto$ is asymmetric, then antisymmetry will follow as well, due to the "trivial" satisfaction of the conditional – the standard semantics for the material conditional ($\supset$) has the conditional true whenever the antecedent false. Since the entire conditional will be true only because the asymmetry of $isGatheredInto$ means that the antecedent of the conditional is always false, antisymmetry would not seem to give us a particularly useful axiom for our theory, though it is true nonetheless.

Transitivity

A transitive relation is one where, for any x, y and z, if x bears the relation to y and y bears the relation to z, then x bears the relation to z.

$$\mathbf{R6}(Trans) : \forall x \forall y \forall z((isGatheredInto(x, y) \,\&\, isGatheredInto(y, z)) \supset isGatheredInto(x, z)) \tag{2.10}$$

If collections can be gathered into collections, and if collection membership is transitive, then whenever collection A is gathered into collection B, every member of A will also be a member of B (along with the collection A as an individual).

Transitivity again raises the question posed by A4: whether collections can themselves be gathered into collections. If it is not possible for a collection to be an individual member of a collection, then transitivity follows, although only by trivial satisfaction of the conditional. At this point it is most revealing to reject A4 and allow collections to be gathered into collections to see whether transitivity sensibly holds.

Allowing collections to be members of collections with transitivity would distinguish collections and $isGatheredInto$ from sets and set relationships. Sets can be members of other sets, but set membership is not transitive: if set S is a member of set T, this does not mean that the members of S will themselves be members of T (although some or even all may happen to be).

While the $subsetOf$ relationship between sets is transitive, it also has a distinct structure from the kind of transitivity for collections described above. If the set S is a subset of T, then every member of S is a member of T, but here S itself is not a member of T (although, again, it may happen to be).

If collections can be gathered into collections, as opposed to the items of some distinguished collections being gathered into a collection, then this creates a hierarchical structure within the collection. Allowing $isGatheredInto$ to be transitive then collapses this structure. In order to preserve the intentions of curators who choose to gather whole collections instead of individual items from collections, we can consider transitivity not to hold for $isGatheredInto$.

These axioms are intended to give a generalized account of collection membership, so the existence cases where transitivity does not seem appropriate is sufficient to reject the axiom. However, the rejection of a transitivity axiom would not conflict with transitivity being implemented in specific cases. Non-transitivity will mean that $isGatheredInto$ aligns with the set membership ($memberOf$) relation.

2.3 Comparison of *isGatheredInto* with Other Relations

There are some well-known relations that describe the structure of collective entities, most notably mereological (part/whole) relations and the set theoretic relations just mentioned. Examining how our intuitive understanding of *isGatheredInto* compares to the features of these relations can give us insight into collections. Table 2.1 shows a comparison between the relation properties of *partOf*, *properPartOf*, *memberOf*, *subsetOf*, and *isGatheredInto*.

Table 2.1: Comparison of Relation Properties

+/-	*partOf*	*properPartOf*	*memberOf*	*subsetOf*	*isGatheredInto*
Reflexivity	+	-	-	+	-
Irreflexivity	-	+	+	-	+
Symmetry	-	-	-	-	-
Asymmetry	-	+	+	-	+
AntiSymmetry	+	+	+	+	+
Transitivity	+	+	-	+	-

The relation properties for *partOf* and *properPartOf* are based on the Classical Extensional Mereology as presented in Varzi (1996). The *partOf* relation is reflexive (everything is a part of itself), antisymmetric (if x is part of y and y is part of x, then x is y), and transitive (if x is part of y and y is part of z, then x is part of z). The *properPartOf* relation is irreflexive (nothing is a proper part of itself), asymmetric (if x is a proper part of y, then y is not a proper part of x), antisymmetric (as a trivial consequence of asymmetry), and transitive (if x is a proper part of y and y is a proper part of z, then x is a proper part of z). Neither of these merological relations match with *isGatheredInto*, which is irreflexive and asymmetric (like *properPartOf*) but is not transitive.

The relation properties for *memberOf* and *subsetOf* are based on a standard Zermelo-Fraenkel axiomatization of set theory (ZFC) (Fraenkel and Bar-Hillel, 1958). The set membership relation *memberOf* is irreflexive (no set can be a member of itself), asymmetric (if x is a member of set y then y cannot be a member of x), antisymmetric (again, a trivial consequence of asymmetry) and not transitive. The *subsetOf* relation is reflexive (every set is a subset of itself), anti-symmetric (if x is a subset of y and y is a subset of x then x

is y) and transitive (if x is a subset of y and y is a subset of z, then x is a subset of z).

As Table 2.1 shows, the relation properties of $isGatheredInto$ align with the relation properties of the set theoretic relation $memberOf$. Although this result supports the claim that collections are sets, it does not automatically provide a decisive answer to our question. As we argue later in this paper, it is still difficult to reconcile all of the properties we commonly assign to collections with the properties of sets.

2.4 $Collection(x)$ Defined by $isGatheredInto(x, y)$

The predicates $Collection(x)$ and $isGatheredInto(x, y)$ are based on the closely related notions of being a collection and being an item in a collection. Therefore it may be possible to reduce our discussion of collections to a discussion of $isGatheredInto$ by defining $collection(x)$ in terms of $isGatheredInto(x, y)$.

A plausible definition of collection is that something is a collection if and only if there is something gathered into it.

$$\textbf{D1} : \forall x(Collection(x) =_{df} \exists y\, isGatheredInto(y, x)) \tag{2.11}$$

The notation "$=_{df}$" is read "if and only if" as definitions of this sort are typically understood as implying logical equivalence.

Because D1 is just the joint assertion of A1 and A3 it has the consequence, (from A3) that there are no empty collections. Can we be confident this is so? It is easy to imagine cases where curators deem a collection to have been created (in terms of allocating resources, or giving a collection a name or preliminary description) without having, at that time, gathered any items into the collection.

In order to handle these cases we can modify the definition to say that something is a collection if and only if it is *possible* that there is something gathered into it.

$$\textbf{D2} : \forall x(Collection(x) =_{df} \Diamond\exists y\, isGatheredInto(y, x)) \tag{2.12}$$

D2 is read "for all x, x is a collection if and only if it is (logically) possible that there exists a y such that y is gathered into x". Or alternatively, "... if and only if x could have something gathered into it." This definition uses the modal operator for possibility ("$\Diamond$"), and so requires the use of a more expressive language than first order logic. Or one might rather forgo empty collections and say that while curators may have created something when they designate resources for the development of a future collection, they have not (yet) created a collection.

Of course even if either of these definitions is accepted, our understanding of collections is not advanced much. The concepts of *collection* and *isGatheredInto* are so closely related that any definition of one in terms of the other is unlikely to provide many satisfying analytical insights.

In addition these definitions do not, as they stand, give us identity conditions for collections. That is, while such a definition will tell us whether or not something is a collection, they do not provide a way to distinguish different collections, or to identify a specific collection as continuing to exist despite changes over time, and that is a critical part of providing an account of the ontological nature of collections. In the next section we consider whether or not collections are sets, and here a positive answer will in fact provide identity conditions.

2.5 Collections as Sets

Table 3.1 shows that *isGatheredInto* appears to have the same relation properties as the set membership relation: both are irreflexive, asymmetric, and non-transitive. And not surprisingly sets are indeed commonly used to represent collections in models of digital library systems (Lagoze and Fielding, 1998; Gonalves et al., 2004; Meghini and Spyratos, 2010). However, it is not always clear if these models are treating collections and collection as equivalent notions, or if collections are being represented as a kind of set. If the former then all collections are sets and all sets are collections. But if the latter then while all collections are sets, not all sets are collections.

2.5.1 Collections as Equivalent to Sets

In a formal logic-based model for digital libraries (Gonalves et al. 2004), we have the definition:

> A collection $C = \{do_1, do_2, \ldots, do_k\}$ is a set of digital objects.

Although this definition does limit collections to a kind of set (sets of digital objects), it seems to imply that any arbitrary set of digital objects is a collection, regardless of whether that set has received any sort of recognition or attention from a curatorial agent. That is, regardless of whether or not any collecting has occurred.

Lagoze and Fielding (1998) define a collection as "a set of criteria for selecting resources from the broader information space." Taken literally, this phrasing implies that a collection is a set of criteria, which seems peculiar. But this account can be read more charitably as intending that any set of resources meeting a set of criteria is a collection ("set membership ... is ... criteria-based").

These definitions leave us with the same basic problem. In Gonalves et al. (2004), all sets of digital objects, however arbitrary, are considered collections, and since every set of resources meets some set of criteria or other, the same is true with the account in Lagoze and Fielding (1998). This argument is presented in more detail below.

2.5.2 Collections as a Kind of Set

The addition of further restrictions (e.g. "a set identified by curators") will provide a more intuitive view, with some but not all sets being collections, although the identification and precise formulation of such a criterion will not be easy.

However, even if we identify a criterion that seems to cut the cases correctly, giving us only, and all, the sets that are collections, a separate issue will pose a challenge regardless of what criterion is chosen. It is natural to say that collections can change their membership items are often removed from collections, and collections are often expanded to include new topics or creators. Sets, on the other hand, cannot lose or gain members. So how could

collections be sets if collections can have properties (becoming larger or smaller) that no set has?

2.6 Arguments Against Collections as Sets

This section presents in detail the two arguments against collections being sets that were mentioned briefly above. The exposition will be methodical and detailed not just because the belief that a collection is a set is widespread, but also because there seem to be few good alternatives to collections being sets, and doing without sets as collections will almost certainly pose substantial modeling and ontology challenges. In addition, although the arguments here may seem too simple to warrant such methodical treatment, Sharvey (1968) and van Cleve (1985) have shown that in this area simple and apparently valid arguments have received wide endorsement before their fallacies were identified.

Counterexamples are presented against both of the claims mentioned above: (i) that the collection and set are equivalent concepts, and (ii) that a collection is a kind of set. Each of these two claims are replaced with weakened versions that are entailed by the corresponding originals. Arguments are the presented against the weaker version directly. Since the weaker versions are entailed by the originals any successful arguments against the weaker versions will also be effective against the original claims.

We begin with:

E1: Collection and set are equivalent concepts

Understanding what claim is being made by E1 is complicated by the several possible senses of "equivalent", ranging from strict identity of both meaning and extension ("collection" and "set" have the same meaning and apply to the same things; i.e. they are synonyms), to some sort of definitional or analytical equivalence, to logical equivalence, to material equivalence. Although material equivalence (all collections are sets and all sets are collections, but just as a matter of fact, at the moment of assertion) may seem too weak to be the intended sense of E1, it is logically implied by any relevant variety of equivalence that might be the intended

sense of E1. Consequently if this weak sense of equivalence, material equivalence, fails – because either some collections are not sets, or some sets are not collections – then there is no relevant sense of (logical) equivalence in which collections and sets are equivalent.

The weakened form of E1, then, is:

$$\textbf{E2} : \forall x(Collection(x) \supset Set(x)) \,\&\, \forall x(Set(x) \supset Collection(x)) \tag{2.13}$$

E2 states that if something is a collection then it is a set and if something is a set then it is a collection

2.6.1 Against Collection and Set Being Equivalent Concepts

Our criticism of E1 takes the form of counterexamples to E2. If those are successful then E1 is false, regardless of the intended equivalence. The argument against E2 will take the form of counterexamples against its second conjunct:

$$\textbf{E3} : \forall x(Set(x) \supset Collection(x)) \tag{2.14}$$

E3 states that if something is a set then it is a collection. A counterexample against E3 will be a case where something is a set but not a collection.

E3 does indeed seem to have serious counter-intuitive consequences when combined with any standard axiomatization of mathematical set theory. The most important of these would have collections existing without intentional activity on the part of curatorial agents. In any set theory with "unrestricted comprehension" there is, for any two things in the world, a set that has just those two things as members. So there is, for instance, a set that contains the planet Mars and the tallest redwood tree in California. There is such a set, but is there such a collection? If we think not, if we believe that some sort of curatorial agency is required for the existence of collections over and above the existence of sets, then this is a reason to reject E3 and conclude that not all sets are collections.

It might be argued that once Mars and the redwood tree have been indicated by someone

(as it has been indicated just now) then the collection does indeed exist and so this is not a case where a set is not collection. However the set in question also existed ten years ago (by, again, the unrestricted axiom of comprehension), well before this text, or anyone, indicated that set. If ten years ago that set existed but was not a collection then not all sets are collections.

Naive set theory, with unrestricted comprehension, is of course prey to Russell's paradox[2]. The set theories that avoid Russell's paradox typically define new sets in terms of existing sets and so for those axiomatizations the general form of the argument just given is not available (Fraenkel and Bar-Hillel, 1958). Nevertheless it is routine in mathematics and ontology to assume that for any two objects there is a set that contains them both and that assumption, at least when restricted to objects other than sets, is consistent with ZFC and does not generate Russell's paradox – and does provide a counterexample to E3.

In any case it is a consequence of ZFC, which does not allow unrestricted comprehension, that if one set exists then an infinite number of sets exist (by repeated applications of the axiom of pairing). But it would be wildly at odds with an ordinary notion of collection to say that if one collection exists then there are an infinite number of collections – few if any of which having received any curatorial attention whatsoever.

One might argue that implications of this latter sort are consequences of axiomatizations that do not match our pre-theoretical notion of a set, and that an improved axiomatization could be developed, one that reflected what real sets (sets for ordinary people, not the sets of mathematicians) really are. But it is hard to see how any axiomatization could simultaneously capture most of the common assumptions about mathematical sets (for instance, that for any two things there is a set that has just those two things as members) and yet not have counterintuitive consequences for the assertion that all sets are collections.

2.6.2 Against Collections as a Kind of Set

The second sense of an affirmative answer to "Are collections sets?" is:

[2]Is the set of all sets that are not elements of themselves an element of itself? If it is then it isn't and if it isn't then it is.

K1: A collection is a kind of set.

Whereas the claim that sets and collections are equivalent may have seemed implausible on its face (implying as it does that all sets are collections), the claim that a collection is a kind of set (which allows that some sets are not collections) is quite common, and appears in LIS literature and formal models of digital libraries, as we have shown.

Although it might be difficult to say exactly what the "is a kind of" relationship amounts to, by any reasonable construal it implies at least that all collections are sets, and specifically it implies the material conditional form of that assertion. And so our weakened version of K1 states that if something is a collection then it is a set.

$$\textbf{K2} : \forall x(Collection(x) \supset Set(x)) \tag{2.15}$$

Arguments that K2 is false and that therefore a collection is not a kind of set are now considered. As with E1 the criticism of K1 will take the form of counterexamples against K2. If those counterexamples are successful then they are counterexamples against K1 as well, regardless of the precise meaning of "is a kind of".

Since E2 also implies K2, any arguments against K2 will also be arguments against E2, and therefore arguments against E1. That is, the evidence against "a kind of" is also evidence against equivalence.

For every collection there exists a set that has as members all and only those things gathered into that collection.

$$\textbf{S1} : \forall x \forall y \forall t(Collectiont(x,t) \supset \exists z(memberOf_t(y,z,t) \equiv isGatheredInto_t(y,x,t))) \tag{2.16}$$

S1 asserts that for every collection x at a time t there is a set z that has as its members all, and only, the items in the collection. The predicate "$memberOf(x,y)$" is used here only in the strict sense of set membership, and time indices are introduced.

If a collection is a set, what set might it be? Obviously one candidate is the set just identified, the set of all things that are items in that collection. Certainly that set exists, and certainly the set and the collection have a lot in common (their items/members for one thing).

That the set of any collection's items exists is not in dispute. The question is not whether there is some set S that has the same members as a collection C, but whether C is that set, that is, whether S and C are identical.The claim that a collection and the set of its items are identical can be made by adding "$x = z$" to the biconditional in S1.

$$\mathbf{S2} : \forall x \forall y \forall t(Collectiont(x,t) \supset \exists z(memberOf_t(y,z,t) \equiv isGatheredInto_t(y,x,t) \,\&\, x = z)) \tag{2.17}$$

This claim is inconsistent with common beliefs about sets and collections. Specifically it is inconsistent with affirming Member Essentialism (van Cleve, 1985) for sets while denying Item Essentialism for collections. Member Essentialism can be expressed this way:

$$\mathbf{ME} : \forall x \forall y \forall t_1[(Set_t(y,t_1) \,\&\, memberOf_t(x,y,t_1)) \supset \Box \forall t_2(exists_t(y,t_2) \supset (existst(x,t_2) \,\&\, memberOft(x,y,t_2))] \tag{2.18}$$

Member Essentialism says that if some set y has x as member at some time t_1, then y has x as a member, necessarily that is, y has x as a member at every time x exists and in any possible alternative circumstance (or "possible world").

Member essentialism is widely accepted by mathematicians, philosophers, and ontologists. The basic idea is that a set, e.g. $\{a, b, c\}$, could not have had in the past, or have in the future, or have in some alternative circumstance (any other possible world) any other members than those it actually has, which are, in our example: a, b, and c. Consider the temporal case first. How can set $\{a, b, c\}$, for instance, come to have, at some point in the future, only two members, a and c, and so become the set $\{a, c\}$? Exactly what would the persisting entity be that at one time had a, b, and c as its (only) members and then at a

later time had just a and c as members?

Since only sets have members only a set will do for the thing that once had b as a member and then later does not. But exactly what set would it be?

Obviously it can't, on pain of immediate contradiction, be the set $\{a, b, c\}$, as that set does have b as a member and so doesnt meet the criterion of not now having b as a member. So we are left with $\{a, c\}$ as a candidate for the set that once had b as a member and now does not. It is indeed true that $\{a, c\}$ does not now have b as a member, so that half of the requirement is satisfied. But is it really the case that the set $\{a, c\}$ might at some point have had b as a member? It is true that we say things like "if the value 42 is removed from the set S the result of applying the formula will be different", but this seems just a manner of speaking. We may say that we removed b from the set $\{a, b, c\}$, and even that the numerical size of some set changed (in this case decreasing by one). But it seems more accurate to make the same point by saying instead that we turned our attention from one set, a set with three members (a, b, and c), to another set, one with just two members (a and c).

The claim that sets cannot lose or gain members is a claim specifically about sets and their members. It is not deduced from the (implausible) claim that nothing can lose or gain a property, nor does it imply that thesis. There is no difficulty in assuming that a person can be happy at one time and not happy at a later time, or that a leaf can be green at one time and not green at later time. In both cases we easily identify an underlying enduring object. Nor do we assume that Member Essentialism for sets implies or is implied by mereological essentialism, the thesis that composite physical objects have their parts essentially.

In ordinary discourse we also say things like "the set of people living at 303 Main St is larger than it used to be." To make the problem acute let's name that set F and define it using by a rule, using "set builder" notation:

B1: $F = \{x : livesAt303Main(x)\}$.

On Monday just John and Jill live at 303 Main; so let

$M = \{\text{John, Jill}\}$.

On Tuesday John, Jill, and Mike live at 303 Main, so let

$T = \{\text{John, Jill, Mike}\}$.

Is the set F larger on Tuesday than it was on Monday? No. There are two ways to understand what set "F" refers to at a given time, but in neither case does the set F "get larger".

1) If B1 is understood as assigning to "F" the set of objects that satisfy $livesAt303Main(x)$ at the time B1 is asserted, and B1 is asserted on, Monday, then on Monday "F" refers to M, i.e. {John, Jill} and on Tuesday "F" continues to refer to M, John, Jill. By this convention "F" refers to the same two person set on two different days, and not to some single set that got larger between Monday and Tuesday.

2) Alternatively B1 may be understood to define "F" as referring to whatever set satisfies $livesAt303Main(x)$, at the time "F"is used. In that case "F" refers on Monday to set M {John, Jill} and on Tuesday to set T John, Jill, Mike. But since these are two different sets there is not, again, any set that once had two members and then later had three.

Since Member Essentialism for sets is believed by many to follow directly from the ZFC axiom of extensionality it may seem that this point is emphasized too much. However Sharvey (1968) and van Cleve (1985) have shown that ME does not follow from that ZFC axiom (alone) and in fact its basis remains obscure. The arguments just given are intended to reiterate and confirm the intuitive plausibility of Member Essentialism, not decisively deduce it from other principles. Fortunately, why ME is true need not be of concern here – it is almost universally affirmed and does indeed appear to be part of the common understanding of sets.

Now let's consider a corresponding principle for collections, Item Essentialism.

$$\mathbf{IE}: \forall x \forall y \forall t_1 [(Collection_t(y, t_1) \,\&\, isGatheredInto_t(x, y, t_1)) \supset \Box \forall t_2 (exists_t(y, t_2) \supset (exists_t(x, t_2) \,\&\, isGatheredInto_t(x, y, t_2))] \quad (2.19)$$

Item Essentialism says that if some collection y has something x as an item at some time t_1, then that collection has x as an item whenever x exists and in every possible alternative circumstance.

While ME is universally accepted as a necessary truth about sets, IE seems to conflict with our settled conviction that (i) items may be added to collections and removed from collections and (ii) collections could have had items other than the items they do have – imagine a failed attempt to acquire an item: if the attempt had succeeded the collection would have had an item it does not have.

In summary, if collections are sets, with the items gathered into them as the members of those sets, then Item Essentialism would be part of our concept of what a collection is. But if Item Essentialism is not part of our concept of what a collection is, then collections are not sets.

There are just two ways out: deny ME for sets or affirm IE for collections. But denying ME for sets would seem to be out of the question, so that leaves affirming IE for collections, that is, holding that collections cannot add or gain members. That seems a high price to pay, conceptually, for classifying collections as a kind of set.

2.7 Collection as Set-in-a-Role

Understanding collections to be a kind of set was initially promising and natural. But if collections aren't a kind of set, then what are they? It is not likely that the answer to this question will be easy either to develop or to defend. Moreover, while there are substantial counterintuitive consequences if collections are understood as a kind of set, perhaps the counterintuitive consequences of the alternative treatments are worse. It is worth making the strongest case possible for collections being sets of their members. This will be a case that accepts Item Essentialism as an unavoidable, though counterintuitive, consequence.

This approach to the ontological status of collections holds that a collection is a set in a particular informational or curatorial role. Being a collection is a property that sets have only in certain contingent circumstances. So on this account, sets that have not

received any kind of curatorial attention (e.g. the set that contains the planet Mars and the tallest redwood tree in California) exist, but do not qualify as collections. Although a set exists whenever its members exist, a set is not a collection unless it is treated as such in the appropriate social circumstances. Therefore sets that aren't collections can become collections, but nothing that is not a set can be collection.

On this account the property of being a collection is what Guarino and Welty (2000) refer to as a *non-rigid* property. Guarino and Welty define rigidity using modal logic and model theory (the notion is based on the idea of *de re* necessity used in ME and IE), but the basic idea is simple: a property is rigid if and only if nothing that has that property could have both (i) existed and (ii) failed to have that property. For example, being a person is rigid because the things that are persons could not have been anything but persons, but being a student is not rigid because the things (persons) that are students might not have been students. Becoming a student or ceasing to be a student (without ceasing to be entirely) is only possible if being a student is a non-rigid property. According to Guarino and Welty rigid properties indicate *types*, fundamental kinds of things, while non-rigid properties indicate *roles* that things of some particular type may enter into.

On the set-in-a-role view a collection is not a type of thing, but a role that things of some type or other have in particular circumstances. What sort of thing is it that can enter into the role of a being a collection? *Sets*. Being a collection is a role that sets have in the right contingent social circumstances. That being a collection is a role is suggested by the fact that being a collection is not rigid: the thing (the set) that is a collection might not have been a collection – it might not have received any curatorial attention. On the other hand the property of being a set is rigid: nothing that is a set could have been anything other than a set, and things that are sets in any possible world are sets in every possible world. Being a set then is a type and therefore a candidate for bearing a role.

On this account of collections, collections are sets, and so they have the identity conditions of sets: they cannot lose or gain items. This is the counterintuitive consequence that is the price of affirming Item Essentialism in order to have collections be sets. But

there is some consolation in noting that there are alternative ways to characterize the fact that collections can lose and gain members. When we say that a collection has undergone a change in membership, we mean that a different set has been selected for the purpose at hand by some particular person or persons. The expectations and practices for the development, management, and description of collections, in combination with social conventions and practices, create a system for recording and communicating the set that is currently distinguished as the relevant collection.

Even though this approach does imply that items cannot be literally added to or removed from persisting collections, it does not require treating sentences such as "We added an item to the collection" as false. Such sentences may be considered idioms that while not literally true, are true when given the appropriate elucidation. This strategy for defining collection and changes in collection size is similar to one proposed for defining "document" and understanding changes in documents (Renear and Wickett, 2009). A remaining challenge will be to provide an account of cases where two sets are considered the "same collection" if they cannot literally be the same collection.

Any theory of collections as a kind of set will contradict the widely held belief that collections can lose or gain members, as well as the related belief that collections might have had members other than the ones they do have. Revised notions of membership change and collection identity may not be such a high price to pay to allow the well-studied and common notions of sets and set membership to function as a critical elements in a theory of collections.

2.8 Conclusion

This chapter examines the *isGatheredinto* relationship, the grounding relationship for collection/item propagation rules, and identifies a number of plausible axioms for that relationship. These axioms illuminate the nature of the relationship and can be combined with collection/item propagation rules to allow additional inferencing.

Arguments for and against collections being mathematical sets are also examined. Al-

though the analysis is not conclusive it does provide a better understanding of the constraints that different conceptualizations entail with respect to collection description. For instance, if collections are sets then there are certain properties that strictly speaking collections cannot have, such as mass or location (on the assumption that sets are abstract), and apparent attributions of such properties will be understood to be idioms that can be translated into literal equivalents. However, as this examples suggests, alternative conceptions of collection seem to be primarily consequential for *instance data* (actual metadata assertions), or for specific rules, and not for rule categories.

Membership essentialism, if applicable to collections, might at first glance to be pose problems for rule categories. But here again the consequences would seem to be for instance data or specific rules rather than for rule categories. The rule formulas that define categories should be understood as being in the present tense, or, alternatively, time-indexed to a single time. An actual rule may introduce a time-based attribute that conflicts with membership essentialism, but that rule may simply be accepted or rejected depending on what notion of collection is assumed.

The analysis of what collections really are was only advanced to a preliminary stage in this chapter, and there are several opportunities for continued work on the logical nature of collections. Fully developing an account that states certain sets are collections given the appropriate contingent circumstances will require saying more about those circumstances. What, precisely, is the curatorial role that a set plays that allows us to say it is a collection? Research into this question may reveal how the curatorial practices of particular domains contribute to how collections are understood as objects in those domains, which may end up showing that the concept of collections varies significantly between domains. Since the framework developed in later chapters provides categories of rules from which practitioners in a domain can generate specific propagation rules that reflect their needs and expectations, a later discovery that there are different types of collections will not interfere with the use of the framework to support the expression of collection/item metadata relationships.

Chapter 3

The Logical Form of a Metadata Record

This chapter systematically explores the semantics of metadata records. In digital library systems information about collections and items both are typically expressed with metadata records. If we are to connect our formal framework for collection/item inferencing with these actual assertions about collections and items then we must have a system for translating ordinary metadata records into a formal knowledge representation language, such as first order logic. In this chapter three approaches to analyzing the semantics of ordinary metadata records are explored. The appropriateness of these different approaches to the formal representation of information about collection and items, and to the formal representation of collection/items inferencing rules, is discussed.

3.1 Introduction

[1]Performing inferences on the basis of relationships between collection-level and item-level description requires that the descriptions be articulated explicitly and in comparable formats. This does not mean that the descriptions need to have been recorded in the same knowledge representation language originally, or even in any kind of formal representation language. But what is necessary to support integration and inferencing between the collection and item levels is the ability to translate from a given item-level or collection-level metadata description into a representation scheme with a clear formal semantics.

In many cases the information we would want to use for the basis of collection/item metadata propagation is stored in "colloquial" metadata records. Such records are a ubiq-

[1]Portions of this chapter appeared in Wickett, K. M. and Renear, A. H. (2012) "The Logical Form of a Metadata Record" in *Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries.*

uitous and critical component of many information systems. These are metadata records that are presented in a serialization language that has delimiter conventions and a vocabulary of attributes and values, but which does not have an associated formal semantics that determines how the information carried by the delimited tokens within a record should be expressed in a logic-based knowledge representation language (such as RDF or some dialect of first order logic). The intended semantics of such metadata languages is typically given only informally and partially in natural language prose documentation. Of course the syntax of such languages is often rigorously defined by a formal meta-grammars such as BNF or XML/XSD, but there is rarely any comparably formal specification of what the terms mean and precisely how they are to be used to make informative statements.

Colloquial metadata records appear to be simple and uniform enough in structure that they would seem to present little difficulty for formalization. These records are typically recognized as being a sequence of delimited attribute/value pairs, usually with one pair providing an "identifier" for the resource. This is a structure so straightforward and apparently obvious in intended interpretation that determining semantics would seem to be a trivial matter. But this turns out not to be the case.

In this chapter I argue that, despite appearances, it is not clear what the correct understanding of the logical semantics of a metadata record actually is. The number of expressive first order constructs necessary to secure the anticipated entailments is greater and more varied than one might think, and involves some famously troublesome notions. More specifically, the identifier attributes commonly found in records, although they have the appearance of any other attribute, pose a particular challenge to formalization. There is no clear convention for treating this critical attribute and the alternatives are substantially different in nature.

3.2 Related Work

The general problem of explicitly and formally representing the information carried by serializations that were developed without formal semantics has been identified and studied

with the XML document markup community (Renear et al., 2002). Markup in XML documents is obviously in some sense meaningful, asserting for instance that a bit of text is a citation, that a section is part of a particular chapter, that an author has such and such affiliation, and so on. However it is only in virtue of shared conventions that this meaning can be identified and exploited. An XML schema (a DTD or XSD) defines vocabulary and syntax for a markup language, but there is no comparable schema mechanism for formally representing the semantics of the language. Some of the semantics may of course be described in prose documentation, but these descriptions are typically informal, incomplete, and not computer-processable. Since the late 1990s there have been efforts underway to develop techniques for attaching semantics to XML markup languages (Dubin et al., 2003; Marcoux et al., 2009).

The development of precise accounts of the semantics of descriptive metadata in the library information organization community has previously been focused on the identification of the entities that are the focus of records, the properties those entities can be said to have, and the relationships between them. The CIDOC Conceptual Reference Model (Doerr, 2003) and the Functional Requirements for Bibliographic Records (IFLA, 2009) give conceptual models that are intended to reflect the information found in museum and catalog records respectively, and provide some aspects of the semantics of metadata records by supplying a picture of the things that are described in records.

While modeling the bibliographic or cultural heritage domains provides a more precise account of descriptive metadata, it does not supply any analysis of how the structures (e.g. MARC fields or Dublin Core elements) function to ascribe properties to those entities. Svenonius (2000) takes a language-oriented approach to descriptive metadata that treats these structures as elements in bibliographic languages, but does not directly address how, for example, the particular structures in a record are assigned objects and properties. Wickett (2010) has taken this approach one step further by treating the use of a metadata record as a form of communicative discourse and modeling how the structures in XML metadata records make systematic contributions to how a record is interpreted within an information

system.

The problem of reference in descriptive metadata records has been recognized in the digital library community, where ambiguity between an original resource (e.g. a photograph) and a digital scan of that resource can introduce problems for correctly interpreting the statements in a record. This issue motivated the introduction of the "1:1 Principle," which is intended to constrain a Dublin Core metadata record to only having only a single object of description (Hillmann, 2003). Recent work has investigated the problems that arise in the application of this principle (Miller, 2010; Urban, 2012), which have become more pressing as digital libraries have started evaluating and implementing semantic technologies (like RDF and OWL) to encode descriptions and reason over those descriptions. When approaches that rely on computational reasoning and systematic reference to individual things are put into place, there will be no opportunity for human beings to intervene and apply the background knowledge and context necessary to refine, disambiguate, and correct the semantics of these records.

Development and promotion of the Resource Description Framework (RDF) and the Web Ontology Language (OWL) has drawn attention to the possibilities of translating metadata created in different formats into RDF triples (Sperberg-McQueen and Miller, 2004; Bowen, 2010). The expression of metadata in knowledge representation languages like RDF can support search and browse systems that use reasoning over descriptions together with encoded knowledge about the world to produce semantically enhanced results for users (Wickett et al., 2010; Haslhofer and Isaac, 2011). However, models of digital libraries tend to focus on an integrative view of digital libraries as systems within a broader context (e.g. Gonçalves et al. (2008)) and give little guidance as to the treatment of colloquial metadata records within a knowledge representation system. Furthermore, the use of URIs as proper names in distributed information systems like semantic web introduces issues for their consistent logical interpretation (Hayes and Halpin, 2008; Halpin, 2011). Recent work by Meghini et al. (2010) does supply an approach to deriving information from metadata within a knowledge representation framework, but does not provide an analysis of how the

information content that goes into a knowledge representation system can be reached from an existing colloquial metadata record.

3.3 Preliminaries

3.3.1 Colloquial Metadata Records

A metadata record is a sequence of symbolic tokens that is understood as expressing information about a resource. In treating a metadata record as a symbolic expression that expresses information about a resource I defer taking a position on the nature of representation in general. A full account will draw some further distinctions, but these are not needed for the points made here. For some preliminary work in this area see Wickett (2010), Sacchi et al. (2011), and Dubin et al. (2011).

Defined in this way *metadata records* includes both natural language sentences and formulas in knowledge representation systems. The interest here however is with metadata records that lie between those two poles. This class includes the majority of metadata found in the digital library applications, where metadata takes the form of expressions in languages such as OAI-PMH, VRACore, MARC21, METS, MODS, etc.

Metadata languages such as these are unlike natural languages in that they typically have a well-defined *syntax*, perhaps expressed formally in meta-grammars such as XML Schema, BNF, or ASN.1. And while in this regard metadata languages are like formal knowledge representation languages, they differ from knowledge representation languages in rarely having a well-defined *semantics*. There may be some prose documentation, but there is usually no rigorous account of class or property relationships, domain and range structures, compositional semantics, axioms, or other familiar approaches to characterizing meaning or truth maintenance. I refer to these metadata records as "colloquial metadata records", to indicate that despite the semi-formality provided by a specified syntax, the use and understanding of such records is much the same in nature as the use and understanding of natural language sentences.[2] Moreover, it is often not fully explicit even what predications

[2]Here I follow (Renear et al., 2002) who distinguish "colloquial XML", such as TEI, that has an in-

are intended, what entities are being referred to, or how that reference is being accomplished, as is shown below.

3.3.2 Method

The information expressed by a metadata record is assumed to be propositional in nature, where a proposition is the language-independent content expressed by a symbol structure. Propositions in this sense are the things that may be considered true or false strictly speaking, whereas metadata records are only true or false in a derivative sense, depending on whether the proposition they express is true or false. Alternatively propositions might be considered the sorts of things that are the proper objects of epistemic attitudes such as belief or doubt.

To determine the logical form of the proposition expressed by a metadata record I present a characteristic metadata record as an example. The important logical entailments of the proposition expressed by that record are then listed and represented in a fragment of first order logic supplied with the relevant vocabulary of predicates and individual constants. Next translations of the metadata record itself into first order logic expressions are conjectured. If translation has all of the same entailments formally (according to the rules of first order logic) that the colloquial metadata record has intuitively, and no entailments that the colloquial record does not have, then it is reasonable to conclude that the translation reveals the *logical form* of the proposition expressed by that record.

3.3.3 Example Record

Below is an example metadata record, given in a general form. It could, for example, be expressed in Dublin Core and encoded in XML, but for simplicity and readability it is presented here in formatted natural language and with no intended relationship to any

formal semantics, from XML that serializes a formally defined knowledge representation language such as XML/RDF.

specific metadata vocabulary or representation encoding.

$$\begin{aligned} \mathbf{R}: \quad & \text{Identifier : LNG42121.4} \\ & \text{Title : Mother Rose} \\ & \text{Creator : Dorothea Lange} \\ & \text{Date : 1941} \end{aligned} \tag{3.1}$$

This metadata record was extracted from a record found in the University of California Calisphere portal[3].

3.3.4 Entailments

From the record and the standard interpretation of identifier elements, it is natural to conclude that with respect to some particular context, there is exactly one resource with the identifier "LNG42121.4" (E1), and that the resource described by the record is that resource (E2).

E1: exactly one entity has the identifier "LNG42121.4".

E2: LNG42121.4 has the identifier "LNG42121.4".

Natural readings of these metadata elements also imply that the title of the resource that is described by the record is "Mother Rose" (E3) and that it was created by Dorothea Lange (E4) in 1941 (E5).

E3: LNG42121.4 was created by Dorothea Lange.

E4: LNG42121.4 has the title "Mother Rose".

E5: LNG42121.4 was created in 1941.

The entailments are expressed in first order logic in order to prepare for determining whether they are logically implied by formalizations of the metadata record. In all of the formalizations that follow, an attribute-value pair that is associated with a resource is represented

[3]http:// www.calisphere.universityofcalifornia.edu/

by a two-place predicate. The predicate represents the attribute, the first argument place indicates the entity that the record describes, and the second argument place gives the value for the attribute. The predications within an assertion can be combined conjunctively with the truth-functional operator "&".

To express the uniqueness constraint on the identifier, the formula states that there is some individual x such that x is identified by the string "LNG42121.4", and if any thing y is identified by that string, then y is, in fact, identical to x.

$$\mathbf{P1}: \quad \exists x[\, identifier(x, \text{“LNG42121.4”}) \;\&\; \forall y(\, identifier(y, \text{“LNG42121.4”}) \;\supset\; x = y)] \tag{3.2}$$

The domain for this and all of the formalizations that follow is a general universal domain. That is, the entities that go into the argument places of any predicates can be anything in the world. More specifically, entities may be (among other things) physical objects, persons, cultural objects, or abstract objects like strings. To disambiguate between a string (often used as a name for a thing) and a thing itself in our formalizations, we adopt a convention wherein strings are shown enclosed in quotation marks. It should be noted however that although this is intuitive and improves the readability of the formula, in standard first order logic individual constants are opaque tokens, with no logically relevant internal structure. This convention can be seen in action with the second entailment. It is natural to use the value of the identifier element as a proper name referring to the resource.

$$\mathbf{P2}: \quad identifier(LNG42121.4, \text{“LNG42121.4”}) \tag{3.3}$$

The last three entailments are given by the attribute value pairs for title, creator and date. The title is treated as a string, the creator is treated as a person, and the date is treated

as a year.

$$\textbf{P3}: \quad creator(LNG42121.4, Dorothea\,Lange) \tag{3.4}$$

$$\textbf{P4}: \quad title(LNG42121.4, \text{“Mother Rose”}) \tag{3.5}$$

$$\textbf{P5}: \quad date(LNG42121.4, 1941) \tag{3.6}$$

The next section presents three translation schemes for expressing the record R in first order logic, and examines whether the resulting expressions formally imply these entailments.

3.4 Formal Analyses

3.4.1 Identifier Uniqueness

The uniqueness of an identifier in an information system is a fact about the system as a whole, not a fact about a single resource. Therefore this fact is expressed as the Unique Identifier Axiom (UIA), which acts as a constraint on identifiers.

The axiom in regimented English:

For any x and for any z,

if x has the identifier z,

then for any y, if y has the identifier z,

then x and y are identical.

The Unique Identifier Axiom (UIA) in first order logic:

$$\textbf{UIA:} \quad \forall x \forall z[\,identifier(x,z) \supset \forall y(identifier(y,z) \supset x = y)] \tag{3.7}$$

The uniqueness of identifier values suggests using them in ways that go beyond the

recording of an attribute-value pair, and is part of the motivation for using of the identifier value as a logical constant in the versions of the entailments given above – in the standard model theory for first order logic interpretations of formulas map each individual constant to a single object in the interpretation. This approach to naming is explored in the analysis that follows.

3.4.2 S Analysis

In this account a descriptive metadata record is an open sentence that is a conjunction of binary predicates with the same unbound variable in the entity place of each predication. This approach is similar to the one taken in Meghini et al. (2010) and Meghini and Spyratos (2010), which explicitly treats a description as having the logical form of an open sentence.

$$\begin{aligned} \mathbf{R_S}: \quad & [identifier(x, \text{“LNG42121.4”}) \ \& \\ & creator(x, Dorothea\,Lange) \ \& \\ & date(x, 1941) \ \& \\ & title(x, \text{“Mother Rose”})] \end{aligned} \tag{3.8}$$

In first order logic, open sentences do not themselves have truth values. Rather, an open sentence is said to be *satisfiable* if it is possible to assign constants to the individual variables in a way that results in a true sentence. Such an assignment is said to satisfy the open sentence. It is plausible that colloquial metadata records have this form, and that the association of a record with a resource in the context of an information system amounts to a statement that the associated resource satisfies the open sentence.

In order to consider an assignment of constants to the individual variables in some formula, we must first consider an interpretation. For the kinds of formulas used here (function-free formulas with no propositional constants), an interpretation in first order logic consists of a domain (a set whose members the formulas will be about), an assignment of constants to members of the domain, and an assignment of each predicate symbol (e.g.

"*creator*" or "*date*") to a relation that is defined for objects in the domain.

We can construct an interpretation I_1 where the domain contains (at least) all of the objects referred to in the record and names are assigned as they appear in P1-P5. Then in I_1, S1 will be satisfied when x is assigned to the object of description, which in this case is the resource named by "LNG42121.4". However, none of the entailments can be derived from the open sentence R_S itself.

3.4.3 Q Analysis

Instead of treating a description as an open sentence, one might conclude that the record is asserting the fact that a resource exists and it has all of the properties indicated in the record. This corresponds with the notion that a metadata description is a description of some particular object (Svenonius, 2000) and with the idea that a metadata record describes a situation in which a resource with the indicated properties exists (Wickett, 2010). If this is so, then the record has the logical form of an existential claim, and the sentence is closed by binding the variable x under the scope of a single existential quantifier.

On this account a metadata records corresponds to what the DCMI Abstract Model (Powell et al., 2007) calls a *description* (a set of attribute-value pairs that are all about a single resource). The notion that a record describes a single object also aligns this account with the *describes* relation in Gonçalves et al. (2008). However, the use of an existential quantifier to bind the statements does not enforce uniqueness of that individual, since existential quantification means that the statement may be true of more than one individual. This is in contrast to Gonçalves et al. (2008), and Powell et al. (2007), where there are explicit restrictions that a description describes *only* a single object.

As in R_S above, an individual variable is used in the entity argument place of each predication. In this analysis, the entire conjunction is within the scope of an existential

quantifier that binds the variable x to at least one member of the domain.

$$\begin{aligned}\mathbf{R_Q}: \quad & \exists x\,[\,identifier(x, \text{“LNG42121.4”}) \;\&\\ & \quad creator(x, Dorothea\,Lange)\;\&\\ & \quad date(x, 1941)\;\&\\ & \quad title(x, \text{“Mother Rose”})]\end{aligned} \tag{3.9}$$

Although R_Q does not formally entail any of our intuitive entailments on its own, together with the uniqueness axiom UIA, R_Q does formally entail P1.

Recall P1, which is an existentially quantified conjunction:

$$\begin{aligned}&\exists x[\,identifier(x, \text{“LNG42121.4”})\;\&\\ &\quad \forall y(\,identifier(y, \text{“LNG42121.4”})\;\supset\; x = y)]\end{aligned} \tag{3.10}$$

The first clause of the conjunction follows directly from R_Q by conjunction elimination. The second clause of P1 follows because the identifier clause in R_Q makes the antecedent of the conditional true when we substitute $LNG42121.4$ for z in the unique identifier axiom (UIA). Therefore the consequent of the axiom under the same substitution – which is the same as the second clause of P1 – follows by modus ponens.

However, despite the inclusion of the identifier statement, none of the other intuitive entailments are formally entailed by R_Q. If the identifier statement is treated in this logical analysis the same way as all other attribute-value pairs (as a two-place predicate), then it does not support using the value of the identifier attribute as a logical constant.

Although it follows from R_Q that there is some single thing that has each of the properties ascribed to the resource by our record, this claim operates at the level of a general existential claim. It does not follow as a matter of logic that it is the individual resource LNG42121.4 which has these properties; only that there is such a resource. A human being, of course, can conclude that LNG42121.4 is this resource, but there are no rules of inference for first order logic that will let us conclude, for example,

"$creator(LNG42121.4, Dorothea\,Lange)$" from R_Q.

In other words, R_Q does not give us the remaining entailments because uniqueness of the identifier does not, by itself, deliver "LNG42121.4" as an individual constant that can be substituted for x within the scope of the existential quantifier. R_Q is a general statement that something with these properties exists, but it does not tell us which individual thing has these properties.

One way to address this issue is to add an identity statement that directly asserts that the thing that we have been describing generally as existing is, in fact, LNG42121.4.

Adding this additional clause gives R_{QI}:

$$\begin{aligned} \mathbf{R_{QI}}: \quad & \exists x\,[\,identifier(x, \text{“LNG42121.4”})\ \& \\ & creator(x, Dorothea\,Lange)\ \& \\ & date(x, 1941)\ \& \\ & title(x, \text{“Mother Rose”})\ \& \\ & x = LNG42121.4\,] \end{aligned} \tag{3.11}$$

This formalization does deliver all of the intuitive entailments, because the final equality clause allows us to substitute LNG42121.4 for x within the scope of the existential quantifier. Then each of P2, P3, P4, and P5 will follow by rules of logic.

The identity statement here will restrict the existentially quantified statement to being true of exactly one thing. So it would seem that R_{QI} fully aligns with accounts of descriptions that require them to be true of just one thing, as in Gonçalves et al. (2008), and Powell et al. (2007).

However, the identity statement, while well-formed in any version of first order predicate logic that includes an identity operator, is functioning to control reference in a way that is usually handled in the meta-theory for a logical language. If the statements in the record are all about one thing, and we know the name of that thing, we might have a more accurate account of the proposition expressed by the record if we simply use the name as a constant

in our predications. This is what the next analysis does.

3.4.4 C Analysis

This account treats the proposition expressed by a metadata record as a conjunction of binary predications, all referring to the object indicated by the identifier element. This approach corresponds to the semantics of an RDF graph, where each subject-predicate-object triple is understood as a binary predication of the form $P(s, o)$ where s and o are each individual constants. Treating the entire expression as a conjunction is in line with Klyne and Carroll (2004): "the assertion of an RDF graph amounts to asserting all the triples in it, so the meaning of an RDF graph is the conjunction (logical AND) of the statements corresponding to all the triples it contains".

In this formalization, there are no individual variables or quantifiers. Instead the name that appeared in the identity statement in R_{QI} is used as a constant to refer to the object of description in each predication. Therefore R_C will effectively be a substitution instance of R_Q, where the individual variable x is consistently replaced with the name "LNG42121.4" within the scope of the existential quantifier.

$$\begin{aligned} \mathbf{R_C}: \quad & [identifier(LNG42121.4, \text{“LNG42121.4”}) \ \& \\ & creator(LNG42121.4, Dorothea\,Lange) \ \& \\ & date(LNG42121.4, 1941) \ \& \\ & title(LNG42121.4, \text{“Mother Rose”})] \end{aligned} \tag{3.12}$$

This formalization yields P2 by conjunction elimination, and together with the axiom UIA, the uniqueness statement given by P1 follows as shown in the previous section. This formalization also implies P3, P4, and P5.

3.4.5 Modifying Entailments

The main differences between the three accounts described above is in how they handle reference to the object of description of a metadata record. The S analysis uses an (unbound) individual variable, the Q series uses an existentially bound variable, and the C analysis uses an individual constant. The natural language version of the entailments shown in E1-E5 use the identifier value from the record as a common proper name. The formalization of the entailments shown in P1-P5 go from the common proper name to the individual constant provided by the identifier value. However, this is not the only option for explaining the logic of names as they appear in sentences like E1-E5.

While R_{QI} and R_C deliver our entailments, both formalizations interpret the identifier attribute/value pair in R as simultaneously providing (i) an individual constant and (ii) the name of that constant. This suggests that the meaning of R involves both object language and metalanguage assertions, something which is typically problematic in formal languages.

An alternative, which would avoid individual constants in the first argument position, is to reconsider the logical form of the entailments from the record. Instead of being treated as statements using an individual name, they can be treated as existential statements. In this approach the identifier element from the record is not interpreted as supplying an individual constant that can be used to refer to the object of description. Instead, the identifier element is seen as providing the definite description "the object with the identifier LNG42121.4". Following Russell (1905), this definite description has a logical form that does not correspond to a singular term, but is instead a compound existential claim:

$$\begin{aligned} &\exists x[\, identifier(x, \text{“LNG42121.4”}) \;\& \\ &\quad \forall y(\, identifier(y, \text{“LNG42121.4”}) \;\supset\; x = y)] \end{aligned} \tag{3.13}$$

The second line of the above formula will follow from the first line and the Unique Identifier Axiom (UIA) by modus ponens. This means that it can be omited from a re-factored version of (e.g.) E3, which is taken to be asserting that the object with the identifier

"LNG42121.4" was created by Dorothea Lange:

$$\mathbf{D3}: \quad \exists x[\, identifier(x, \text{“LNG42121.4”}) \;\&\; creator(x, Dorothea\,Lange)] \tag{3.14}$$

This refactoring provides a new set of formalized entailments (assuming UIA):

$$\mathbf{D1}: \quad \exists x[\, identifier(x, \text{“LNG42121.4”}) \;\&\; \tag{3.15}$$

$$\forall y(\, identifier(y, \text{“LNG42121.4”}) \;\supset\; x = y)] \tag{3.16}$$

$$\mathbf{D2}: \quad \exists x[\, identifier(x, \text{“LNG42121.4”})] \tag{3.17}$$

$$\mathbf{D3}: \quad \exists x[\, identifier(x, \text{“LNG42121.4”}) \;\&\; \tag{3.18}$$

$$creator(x, Dorothea\,Lange)] \tag{3.19}$$

$$\mathbf{D4}: \quad \exists x[\, identifier(x, \text{“LNG42121.4”}) \;\&\; \tag{3.20}$$

$$title(x, \text{“Mother Rose”})] \tag{3.21}$$

$$\mathbf{D5}: \quad \exists x[\, identifier(x, \text{“LNG42121.4”}) \;\&\; \tag{3.22}$$

$$date(x, 1941)] \tag{3.23}$$

This form for the entailments closely matches the form of a response to a database query, as it gives an identifier element that is unique and therefore may be used as a primary key. This version of the entailments may be the best match for what we expect to be entailed by a colloquial metadata record.

Recall R_Q:

$$\mathbf{R_Q}: \quad \exists x\,[\, identifier(x, \text{“LNG42121.4”}) \;\&\; creator(x, Dorothea\,Lange) \;\&\; date(x, 1941) \;\&\; title(x, \text{“Mother Rose”})] \tag{3.24}$$

Each of the entailments D2-D5 follow directly from R_Q by conjunction elimination, and D1 follows by standard rules of first order logic as an instance of the Unique Identifier Axiom (UIA). If D1-D5 are the most accurate formalization of the entailments we expect to be able to reach from the record R, then it would seem that R_Q best captures the logical form of the proposition expressed by a metadata record.

3.4.6 Relationships Between the Formalizations

Figure 3.1 shows the logical relationships between the formalizations of the record and the entailments. We assume that the Unique Identifier Axiom (UIA) holds throughout.

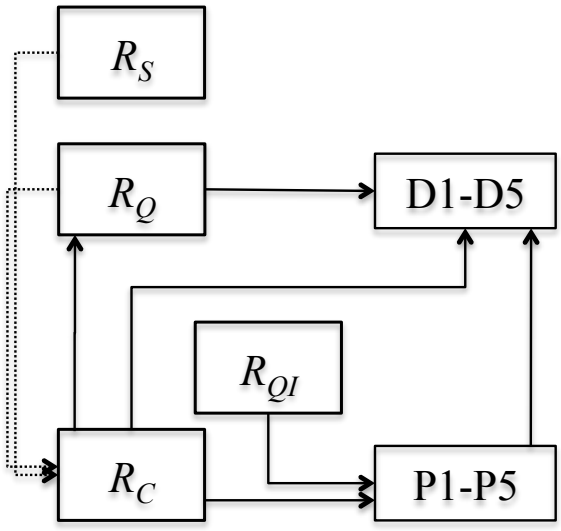

Figure 3.1: Logical relationships between the formalizations, assuming UIA. Dashed arrows show substitution of $LNG42121.4$ for x. Solid arrows show implication.

R_C can be generated as a substitution instance of both R_S and R_Q by consistently replacing the variable x with the name "LNG42121.4" throughout (and in the case of R_Q, dropping the existential quantifier). From the specific statement in R_C, it is possible to deduce the existential claim in R_Q through existential generalization and UIA. Existential generalization is a rule of most first order logic systems: intuitively, if a specific thing has a property, then it is true that there exists some thing with that property. The modified entailments $D1 - D5$ can also be reached from R_C and $P1 - P5$ with existential general-

ization.

R_{QI} implies the entailments P1-P5. The final equality clause in R_{QI} allows us to substitute $LNG42121.4$ for x within the scope of the existential quantifier. Then each of P2, P3, P4, and P5 will follow by rules of logic. The entailments D2-D5 follow directly from R_Q by conjunction elimination, and D1 follows by standard rules of first order logic as an instance of the Unique Identifier Axiom (UIA).

3.5 Discussion

The three formal accounts given above are closely related. All three use the mechanisms of first order logic and binary predication to represent several attribute-value pairs for the same object. R_S uses the logical form of an open sentence, where the object of description is represented with an individual variable, and any object for which the attribute-value statements are all true will satisfy the description. In R_Q and R_{QI}, the same basic form is followed for the attribute-values pairs, and the entire conjunctive statement is bound by a single existential quantifier. This gives the Q Series the logical form of an existential claim.

The S analysis uses the logical form of an open sentence. In addition to being a plausible account of metadata records that aligns with the account given by Meghini et al. (2010), the open sentence form corresponds in many ways to a query over an information system. The set of individuals that satisfy the open sentence will correspond to the members of an answer set that is returned in response to a query. This account of descriptions therefore aligns with the account of bibliographic entities and bibliographic languages developed by Svenonius (2000). For example, an expression in what Svenonius calls a *work language* describes all of the members of a work, where a work is understood to be a set of documents that embody essentially the same intellectual or artistic content. Then it follows that any document that satisfies the open sentence that we generate from a description in a work language will be a member of the work set.

The notion that a description is an open sentence that can be satisfied in a given interpretation suggests a constraint on the logical formulas that correspond to descriptions;

they must be *satisfiable*. A formula is satisfiable if and only if it is possible to specify an interpretation in which the formula is satisfied. Similarly, a set of formulas is simultaneously satisfiable if and only it is possible to specify an interpretation in which every member of the set can be satisfied.

Treating a query as an open sentence that is satisfied by the members of a result set for that query will also mean that valid queries over a repository or catalog must be logically satisfiable. With such a constraint in place it would still be possible for nothing in the repository to match the query and have zero results returned. But it would mean that it must be possible for there to be results that match the query, so queries could not be self-contradictory.

Constraints on the satisfiability of descriptions and queries are particular relevant to scenarios where logic-based knowledge representation systems are used to support information organization systems for retrieval or browsing. In such a system, determining entailments correctly will depend on the sentences in the knowledge base being simultaneously satisfiable. If descriptions are self-contradictory or stand in contradiction to other statements in the knowledge base, it may make correct automated reasoning over descriptions impossible.

The logical form of an existential claim for an object that has certain specific characteristics, seen in R_Q and R_{QI}, is familiar from accounts of *definite descriptions* in the philosophy of language (Russell, 1905). This form also aligns with the account of metadata descriptions in the DCMI Abstract Model (Powell et al., 2007), which forms the basis for many definitions in current introductions and texts on metadata (see, for example, Zeng and Qin (2008), Liu (2007)). R_Q and R_{QI} are accounts that take a metadata record to be an assertion that at least one thing with the specified properties exists. R_{QI} takes the claim one step further with the identity statement, which means that the entire conjunctive statement is true of exactly one individual: the object LNG42121.4.

The R_C has the logical form of a conjunction of binary predications about the object of description. Therefore it is, in a sense, more specific than either R_Q (the existential claim without the binding identity statement that appears in R_{QI}) or R_S (the open sentence). In

fact, the existential claim in R_Q can be deduced from R_C in most systems for first order logic by existential generalization since if we know of a specific object with the given properties, we can conclude the general fact that such an object exists.

As mentioned above, R_C is a substitution instance of R_Q. This close connection between a general description and a conjunction of binary predications may mirror a potential process that is of current interest for many digital library researchers and managers: the process of transforming XML metadata records into a set of RDF statements. We can understand the generation of RDF triples from a description contained in, say, a Dublin Core XML record as the production of substitution instances of the existential formula, where the appropriate identifier for the object described by the record is substituted in for the existentially quantified variable.

This kind of process was used by the Collection/Item Metadata Relationships (CIMR) group to build an RDF testbed to examine patterns between collection-level and item-level description in records aggregated by the IMLS Digital Collections and Content project (Wickett et al., 2010). Collections from the IMLS DCC Collection Registry were selected on the basis of the availability of item descriptions and the appearance of certain properties in those item descriptions in order to support an examination of patterns of values for type, format, temporal, and geographic elements. Item-level and collection-level OAI-PMH XML records were harvested and processed with XML stylesheets to produce RDF that described the "complete collection graph". This RDF graph contained RDF descriptions of collections, items, and collection membership.

In building the testbed, unique identifiers were created for each collection and item described by an XML record, and RDF statements were created using those identifiers. So the CIMR testbed construction followed R_C, but without using an identifier found in the record.

3.6 Conclusion

A systematic study of the semantics of metadata records is a natural part of an investigation into metadata relationships. While their simple surface structure may lead us to think that the semantics of a metadata record is unproblematic and easily discerned, our analysis of an example record suggests otherwise.

The semantics of the common metadata record are somewhat elusive. The main source of this problem is the identifier attribute. Although the identifier attributes commonly found in records have the syntactic appearance of any other attribute this syntactic similarity conceals their potential for assuming a distinctive semantic role, one which involves the simultaneous use and mention of individual constants and appears to challenge the traditional boundary between metalanguage and object language that is respected by first order languages. This suggests that the translation of metadata records into first order languages is not a process that is first order throughout.

These problems may appear to have been avoided by a reanalysis of the intuitive entailments. There, the identifier values that functioned like natural language proper names in the prose representations of entailments are not expressed directly as individual logical constants, but rather re-factored as definite descriptions. But in fact whether this is the correct analysis isn't clear.

Is it really the case that in logic-based representations of the propositions expressed by metadata records, the proper-name based references to resources are to be re-expressed as definite descriptions rather than as individual constants? If so this is a result apparently at odds with the direction being taken by the Linked Open Data and Semantic Web Communities, where URIs clearly function as individual constants referring to resources. Or are these URIs to be understand as being "really" unanalyzed definite descriptions? Moreover it seems unlikely that we can get by without any individual constants at all: R_Q and the D entailments don't avoid individual constants altogether, they only avoid them in the first argument place.

The problems here are deep and not specific to metadata records – they are in fact at the

heart of logic-based knowledge representation. When, in logic-based representation, should logical constants be used and when not? This is not a problem that arises in first order logic *per se*, where individual constants are arbitrary tokens that allow us to explore the mathematics or logical consequence. However it does arise when knowledge representation languages that are based on first order logic are used with logical constants that are intended to actually refer to things in the world.

* * *

I have argued in this chapter that there are at least three plausible possibilities for the logical form of the proposition expressed by a metadata record, and not only are all three substantially different in the first order constructs utilized, but no two can be recognized as equivalent for the purposes of information organization. Despite the puzzles discussed here, applying a method similar the C analysis to collection-level and item-level metadata records has been shown in Wickett et al. (2010) to be an effective approach to prepare for inferencing on the basis of collection/item metadata relationships. This is the approach that will be assumed in what follows.

Chapter 4

Propagation Rules

[1]In order for the relationships between collection-level and item-level description to be useful for information organization, these relationships need to be expressed in a way that supports inferencing about collection and item descriptions. The form this expression takes in this study is a *propagation rule*. Propagation rules are inference rules where the properties of one individual entail conclusions about another individual in virtue of a particular relationship those individuals bear to each other.

This chapter examines the general nature of propagation rules for collection/item metadata relationships. The approach to expressing metadata relationships is examined, along with certain consequences of this method of formalization. Propagation rules are conditionals expressed in first order logic, so the nature of conditionals and their expression in logic is discussed, along with a look into how some well-known challenges for conditionals might be handled for collection/item propagation rules. In addition, techniques for expressing propagation rules in semantic web languages are explored, since these knowledge representation systems are likely candidates for the management of collection/item metadata relationships with propagation rules.

[1]Portions of this chapter were adapted from:
Wickett, K. M. (2009). "Logical expressiveness of semantic web languages for bibliographic modeling." In *Proceedings of the iConference*,
Wickett, K. M. (2011) "Expressiveness requirements for reasoning about collection/item metadata relationships." In *Proceedings of the iConference*, and
Renear, A. H., Wickett, K. M., Urban, R. J., Dubin, D., and Shreeves, S. (2008) "Collection/Item metadata relationships." In *Proceedings of the International Conference on Dublin Core and Metadata Applications.*

4.1 Collection/Item Metadata Relationships and Propagation Rules

Relationships between collection-level and item-level description can be expressed as rules. For example, the relationship between the collection-level attribute *itemType* and the item-level attribute *type* can be expressed as a rule stating that if a collection has a particular value for the *itemType* attribute, then there will be some item in the collection that has that value for the *type* attribute. In first order logic, this kind of rule is expressed as a universally quantified conditional.

These rules are *propagation rules* – inference rules where the properties of one individual entail conclusions about another individual in virtue of a particular relationship those individuals bear to each other[2]. In what follows, I refer to this relationship as the *supporting relationship* for a propagation rule. Propagation rules that operate between collections and items have the supporting relationship *isGatheredInto*, which is the collection membership relationship analyzed in Chapter 2.

Rules that express collection/item metadata relationships support drawing a conclusion about items on the basis of collection description, or drawing a conclusion about a collection on the basis of item description. These are metadata relationships in the sense that they are grounded in descriptive metadata, and are expected to be used in the automation, enhancement, and validation of metadata in information systems. Furthermore, rules like these can be used as axioms to explicitly represent the semantics of metadata attributes. The analysis of metadata records in terms of their logical form that was presented in Chapter 3 provides the basis for a method of moving from colloquial metadata records like those commonly found in digital library systems to the kind of logical representation that we use in the discussion of rules that follows.

[2]This term is used here in the sense introduced by Brachman et al. (1991)

4.2 The General Form of a Propagation Rule

The general case for a propagation rule is one where if an individual y has some property F and some thing x bears the relation R to y, then we can conclude that x has some property G. We express this general rule form in first order logic as a universally quantified conditional with a conjunctive antecedent.

$$\forall x \forall y((F(y) \ \& \ R(x,y)) \supset G(x)) \tag{4.1}$$

A specialization of the general case is when F and G are the same property.

$$\forall x \forall y((F(y) \ \& \ R(x,y)) \supset F(x)) \tag{4.2}$$

This specialized form is sometimes called *inheritance*, since x can be seen as "inheriting" the property F from y via the relationship R. We call R the *supporting relationship* for a propagation rule. The *is-a* relationship is commonly considered to be a supporting relationship for inheritance.

A collection/item propagation rule is one where the supporting relationship is the $isGatheredInto$ relationship that stands between an item and a collection that the item is a member of. For example, we might have a propagation rule that states that if the University of Illinois owns a collection, then every item in the collection is owned by the university. This rule falls under the specialized form of propagation mentioned above if we take R to stand for $isGatheredInto$, and F to stand for the property "is owned by the University of Illinois". A rule stating that if a collection has items created within the date range "1850-1899" then every item in the collection has a creation date within the range falls under the general form of propagation if we understand R as $isGatheredInto$, F as "$dateItemsCreated = 1850 - 1899$" and G as "$1850 \leq date \leq 1899$".

The account of propagation above expresses the properties in question as unary (one-place) predicates. When this approach is applied to metadata relationships, it requires the use of complex definitions for predicates, as seen above with the predicate "is Owned by the

University of Illinois". A more natural approach to characterizing descriptive metadata in first-order logic was shown in Chapter 3, and uses binary predicates for attribute value pairs. Giving explicit representation for the value for an attribute (as opposed to building it in to the definition of a unary predicate) allows us to use binary predicates with more general definitions like "x is owned by v" or "x has creation date v". Representing attribute values as terms in a binary predicate supports a more natural and useful account of propagation rules for collection/item metadata relationships that specifies an attribute (as a predicate symbol) without restricting the rule to applying to a single value for the attribute.

Previous work on collection/item metadata relationships has used binary predicates to represent descriptive metadata, and a framework of categories for this kind of propagation rule was proposed in Wickett et al. (2010) and is developed further in Chapter 5. There are two main divisions in the framework; one between *collection-to-item* propagation and *item-to-collection* propagation, and one between *universal* propagation and *existential* propagation. These divisions result in the four top-level categories of propagation rules defined below.

4.2.1 Collection-to-item Propagation Rule Schemas

Attributes A and B propagate **universally** from collections to items $=_{df}$

> If a collection y has the value z for the attribute A, then every item in the collection has some value w for the attribute B such that w is related to z by the constraint C.

$$\forall y \forall z((A(y,z)\,\&\,Collection(y)) \supset$$
$$\forall x(isGatheredInto(x,y) \supset \exists w(B(x,w)\,\&\,C(w,z))) \tag{4.3}$$

Attributes A and B propagate **existentially** from collections to items $=_{df}$

> If a collection y has the value z for the attribute A, then there is some item in the collection that has some value w for the attribute B such that w is related

to z by the constraint C.

$$\forall y \forall z((A(y,z) \,\&\, Collection(y)) \supset$$
$$\exists x(isGatheredInto(x,y) \,\&\, \exists w(B(x,w) \,\&\, C(w,z))) \tag{4.4}$$

4.2.2 Item-to-collection Propagation Rule Schemas

Attributes A and B propagate **universally** from items to collections $=_{df}$

If every item x in a collection y has a value w for the attribute A that is related to z by the constraint C, then the collection has the value z for the attribute B.

$$\forall y \forall z[\forall x(isGatheredInto(x,y) \supset$$
$$\exists w(A(x,w) \,\&\, C(w,z))) \supset (Collection(y) \,\&\, B(y,z))] \tag{4.5}$$

Attributes A and B propagate **existentially** from items to collections $=_{df}$

If there is some item x in a collection y that has a value w for the attribute A that is related to z by the constraint C, then the collection has the value z for the attribute B.

$$\forall y \forall z[\exists x(isGatheredInto(x,y) \,\&\, \exists w(A(x,w) \,\&\, C(w,z))) \supset$$
$$(Collection(y) \,\&\, B(y,z))] \tag{4.6}$$

4.2.3 Logical Form of Collection/Item Rule Schemas

Each of the four general rule schemas that form the basis of the rule framework is a universally quantified conditional. The rule schema for universal collection-to-item propagation (formula 4.3) has a conjunctive antecedent and a universally quantified conditional as the consequent. The rule schema for existential collection-to-item propagation (formula 4.4) has a conjunctive antecedent and a existentially quantified conjunction as the consequent.

The rule schema for universal item-to-collection propagation (formula 4.5) has a universal statement as the antecedent and an existentially quantified conditional as the consequent. Finally, the rule schema for existential item-to-collection propagation (formula 4.6) has an existentially quantified conjunction as the antecedent and a conjunction as the consequent.

There are several ways that collection/item propagation rules can be used in information systems. Propagation rules could be used to support reasoning about collections on the basis of item descriptions (or vice versa). The rules could also be used to classify attributes according to their propagation behavior as part of metadata vocabulary development.

4.3 Semantics

4.3.1 The Material Conditional

It is common to assert propositions that have a particular *conditional* structure. We say that if something is the case then something else is the case. These conditional propositions are typically expressed in natural language sentences that have distinctive form, such as the "if ... then ... " construction in English. These sentences which express conditional propositions are themselves called "conditionals" by grammarians.

Conditionals are not only frequently found in science, mathematics, and everyday life, but in fact seem to play a very significant, and possibly irreducible, role in describing and explaining the world. Here are some examples:

1. If this is water then it will vaporize at 100° Centigrade.
2. If this is a 1:1 right triangle then there is no pair of integers that will represent the ratio of a side to the hypotenuse.
3. If this is the sandwich I made then it is falafel.
4. If John is a professor at the University then he is an employee at the University

Although there is a common structure to these assertions, it is also evident that they vary widely in nature. For example, some follow from natural laws (e.g. 1 and 2 above), while

other simply follow from conventions of language (e.g. 4 above), or are reports about a particular state of affairs (e.g. 3 above).

In propositional logic conditional assertions of the form "If P then Q" can be expressed by the formula "$P \supset Q$" with the binary operator "$\supset$" assigned truth-functional semantics according to Table 4.1. The conditional is false only in the case where the antecedent P is true and the consequent Q is false; in all other cases it is true. In other words, the material conditional, "$P \supset Q$" is logically equivalent to "$\neg P \vee Q$", or alternatively, "$\neg(P\&\neg Q)$", both of which are obviously satisfied whenever, and only whenever, $\neg P$ is true or Q is false.

Table 4.1: Truth table for "$\supset$"

P	Q	$P \supset Q$
T	T	T
T	F	F
F	T	T
F	F	T

In predicate logic, the semantics of the conditional has a similar more general definition. A universally quantified conditional will be false in an interpretation if there is a substitution instance for the conditional that makes the antecedent true and the consequent false and true if every substitution instance that makes the antecedent true makes the consequent true as well.

A conditional that is assigned truth functional semantics in this way is often referred to as a case of *material implication* following Russell (1903). This terminology refers to evaluation of the conditional solely in terms of actual truth values of the components, without regard to whether there is any natural connection between the antecedent and the consequent of the conditional [3]. This indifference to relevance between antecedent and consequent can be a source of confusion when moving between natural language and a formal system like first order logic, since a particular connection between these components is frequently a critical part of what is meant by a natural language conditional.

The problem, in short, is that given the semantics of conditionals just presented, condi-

[3] See Rescher (2007) for further discussion on the various types of conditionals.

tionals are counted as true if there are no substitution instances that make the antecedent true, regardless of the truth value of consequent, or no substitution instances that make the consequent false, regardless of the truth value of the antecedent.

If a conditional like "if there is rain in the forecast, then Sally will carry an umbrella" is handled as a material conditional in a logical language, it will be evaluated as false only in the case where there is rain in the forecast and Sally does not carry an umbrella. In all other cases it will be considered true. Clearly, this evaluation is not troubling in the case where there is rain in the forecast and Sally does not take an umbrella. It is natural to say that the conditional is false in those circumstances. But what about the cases where there is no rain in the forecast? Sally choosing to take an umbrella when rain is not forecast does not seem intuitively to contradict the "if/then" statement, nor does Sally choosing not to take an umbrella when rain is not forecast. Neither case provides evidence against the truth of the conditional, and for a truth-functional definition of a conditional in a language with two truth values, it is sufficient to evaluate these cases as true.

More unsettling examples can be constructed when scientific or mathematical conditionals are translated into conditionals in first order logic. Consider the (false) assertion:

A) If there is water in beaker A then it will vaporize at 80°C 1A.

The corresponding conditional in first order logic is true if beaker A does not contain water, and it is also true if the substance it does contain *does* boil at 80°C 1A. If the material conditional is intended to capture the full meaning of this conditional about physics, the truth-functional evaluation seems counterintuitive. The conditional appears to make an assertion about water that is, in fact, false (since water at 1A will not vaporize at that temperature).

Still stranger (although not necessarily more compelling) examples can be given since regardless of what assertions are substituted for the antecedent and consequent, if either the antecedent is false or the consequent is true the conditional assertion is true. There need be not be a relevant connection whatsoever between the antecedent and consequent. "If beaker A contains water then the Red Sox will win the next World Series" is true if the

beaker doesn't contain water, or if the Red Sox win the next World Series.

As another example, consider:

> *B*) If beaker A contains water then beaker A does not contain water.

This peculiar conditional will be true, on the material interpretation, when either the antecedent is false or the consequent true, which is to say, whenever the beaker does not contain water. And yet it seems odd to most of us to affirm as true a conditional that asserts that if something is true then it is false.

Mathematical conditionals fare no better when represented as material conditionals. Consider this assertion.

> *C*) If figure A is a right triangle with equal sides then the ratio of one of its sides to its hypotenuse is 2:3.

This may appear to be a false mathematical assertion. However the corresponding material conditional is true if figure A is not a right triangle and also true if the figure has a hypotenuse of 2:3 – regardless of what sort of figure it is. Just as above this also seems counterintuitive if the material conditional is intended to capture the full meaning of the mathematical conditional.

These results are sometimes called the "paradoxes of material implication". They are not true paradoxes but rather unexpected peculiar results. In the standard discussion typically found in elementary logic textbooks, it is noted that real world conditionals of science, mathematics, and everyday life are semantically complex and sophisticated and so cannot be fully captured by the logician's material conditional. However the translation into material conditionals, the explanation goes, can still be justified because the material conditional does capture *part* of the meaning of *all* real world conditionals – and so it is still legitimate, and useful, to represent real world conditionals as material conditionals when we wish to reason using formal logic. After all, we need to represent as much as we can in the formal languages available to us, even if we cannot represent everything in those languages alone.

So while the truth of a material conditional may not imply the truth of the corresponding real world conditional, the truth of a real world conditional does entail the truth of the corresponding material conditional. This means that translating real world conditionals into material conditionals to support logical deduction is a safe strategy: if the real world conditional is true then the corresponding material conditional is as well. What remains forbidden, however, is translating a material conditional (perhaps deduced as the conclusion of an argument) into a real world conditional; that, as the paradoxes of material implication make clear, may indeed take you from a truth to a falsehood. But with that restriction (and another involving nested real world conditionals) the truth functional account of the conditional preserves the validity of conditional reasoning, since it will prevent going from a true premise to a false conclusion.

This standard response to the problems of the material conditionals has not satisfied some logicians and has motivated the development of alternative approaches to conditionals and conditional reasoning. This work is indeed relevant to the broader agenda of understanding propagation as rule-based reasoning about collections, but for the most part beyond the scope of the current project. One area however is particular important and directly related to our later discussions.

4.3.2 The Strict Conditional

Concerned with the representation of mathematical and, particularly, logical conditionals (e.g. "If P and Q, then P") C.I. Lewis (1918) defined a restricted form of implication, "strict implication". On this account P (strictly) implies Q if and only if it is *impossible* for P to be true and Q false. This is a higher bar than material implication, which requires only that, as a matter of fact, it is not the case that P is true and Q false.

Consider again a conditional of the form "If P then Q". If we know that Q is true then we know the conditional, understood as a material conditional, is true. But now consider a logician asserting "If (P and Q) then Q" and intending to make this logical point: if a conjunction is true each of its conjuncts are true. This claim cannot be proved simply by

noting that Q is as a matter of fact true (if Q stands for some actual assertion). Nor is an appropriate rejoinder: "well sometimes it is and sometimes it isn't, if Q is true it is, or if one of P or Q is false it is, but otherwise it isn't". The point is that "If (P and Q) then Q" cannot be false under any circumstances, because it is a tautology: no assignment of truth values to P and Q consistent with the truth tables for "&" and "$\supset$" can make "$(P\&Q) \supset Q$" false.

Lewis introduced a special symbol ($P \dashv\!\!\!\supset Q$) to indicate strict implication. He also introduced the use of the letter L as an appended operator to mean "necessarily" (later writers have used the symbol $\Box$ for this purpose). Since a strict conditional is a material conditional that cannot be false (i.e. is necessarily true), Lewis gave this equivalence:

$$P \dashv\!\!\!\supset Q = \Box(P \supset Q). \tag{4.7}$$

He then explored several axiom systems for reasoning with these "modal" notions.

For instance, one obvious plausible axiom for modal logic of this sort is:

$$\mathbf{T}: \quad \Box P \supset P \tag{4.8}$$

That is, if P is necessary then P is true.

Another is:

$$\mathbf{K}: [\Box P \,\&\, \Box(P \supset Q)] \supset \Box Q \tag{4.9}$$

That is, if P is necessary and $(P \supset Q)$ is necessary, then Q is necessary.

And an important definition is

$$\Diamond P = \neg\Box\neg P \tag{4.10}$$

$\Diamond P$ is read "possibly P".

Axiom systems for modal logic make it possible reason systematically about necessity in

general as well as strict conditionals in particular. At several points in what follows modal axioms will be used in reasoning about the logic of propagation rules.

The question remains: what exactly is the *semantics* for $\Box$, that is, when is $\Box P$ true and when false? Or more informally, what does "$\Box P$" mean? One simple formal approach to the semantics of modal propositional logic that is consistent with Lewis's original intention is based on the notion of a tautology. A compound proposition is a tautology if and only if no assignment of truth values to its constituent propositions that is consistent with the definitions of the connectives of truth-functional propositional logic will make it false; it is a contradiction if and only if no such assignment of truth values will make it true. It follows that the negation of a tautology is a contradiction and the negation of contradiction is a tautology. A proposition is necessarily true if and only if it is a tautology, necessarily false if and only if it is a contradiction, and contingent if and only if it is neither a tautology nor a contradiction.

The semantics for modal predicate logic is a little more complicated, but can proceed basically along the same lines using the notion of an interpretation: a proposition is necessarily true if and only if it is true in every interpretation, necessarily false if an only if it is true in no interpretation, and so on.

The notion of logical necessity, and therefore modal logic generally, can be extended to claims that cannot be easily reduced to truths based on tautologies and interpretations of first order logic. Most of us would agree that it is impossible for b to occur (temporally) later than a and c to occur later than b without c occurring later than a, or that something can be both a number and a person, or that one can own a collection without owning the items in the collection, even though there is no direct basis for these claims in the semantics of first order logic alone. Apparent truths of this sort may be added as formal axioms that apply to all possible states of affairs.

It was noted early on that strict implication had its own "paradoxes", exactly parallel to those of material implication. If the antecedent of a conditional is itself not only a falsehood but *necessarily* false (impossible) then that conditional is true regardless of the truth value

of the consequent, and similarly if the consequent is necessarily true then the conditional is true regardless of the truth value of the antecedent.

So the use of modal notions to define strict implication does eliminate some of the counterintuitive cases and provide a more discriminating conditional, but it does not eliminate all counterintuitive cases. We will see in a following section that the remaining cases do create problems for defining propagation characteristics of attributes, but that these problems have remedies.

4.3.3 Intended Semantics for Propagation Rules

In light of the preceding discussion of conditionals we should now ask what sort of conditional a propagation rule is, and whether representing collection/item metadata relationships as a material conditional creates any problems. We begin by considering what sort of claim a propagation rule is intended to make. That is, what sort of claim is it that we are proceeding to formalize in first order logic conditionals.

Material factual assertions

Collection/item propagation rules might be intended only as simple factual statements about some particular collection at some particular time and without any suggestion of a connection between the antecedent and consequent. In this case they would appear to be intended as nothing more than material conditionals, and so representing them as material conditions would not be a problem. However the example rules we have been considering do not appear to of this sort. In each case there appears to be some intended connection that goes beyond the mere assertion that either the antecedent is false or the consequent is true. We can broadly divide the intended connection into two sorts: empirical and logical, with logical having two subdivisions in turn.

Empirical Assertions

Some propagation rules may be understood as conditionals that make empirical claims based on some natural or social fact. For example: "If a collection supervised by a graduate of The Institute, then it will have no item in it of dubious provenance". There appear to be many different varieties of empirical conditionals. For instance some empirical conditionals, such as those in the natural sciences ("If water is raised to 100°C it will boil"), appear to be based on fundamental physical laws. Others appear to be expressions of some tendency or likelihood "all other things being equal", such as the first example above. Some of our example propagation rules may be understood as empirical assertions, but most are probably general or definitional logical assertions.

Logical Assertions – Definitional

Many of our examples of collection/item propagation rules are most naturally understood as making a stronger claim than mere factual assertion, or even an empirical assertion. They appear to be making claims that are grounded in some sense on the concepts involved, and that are true, if they are true, regardless of any particular contingent circumstances. Consider:

> A) If a collection has a value x for *cld:itemtype* then at least one item in the collection has a value of x for *dc:type*.

A) follows as an immediate consequence of the DCMI definition of *itemtype* – it is in fact *part* of the definition of *itemtype*. As as result A) is true, one might say, by *fiat*, by an explicit stipulation of the definition of a technical term (*itemtype*) in a formal language. Formally we can reflect this by taking A) to be an *axiom* that is assumed to be true in all interpretations of formulas where *itemtype* is a predicate intended to have the meaning specified by DCMI. Some of our exemplary propagation rules seem to be of this sort and that is in fact the source of one application of the rule system: it provides a framework for interpreting and exploiting the definitions of metadata vocabularies in formal standards.

Logical Assertions – General

The above example suggested that a propagation rule could be taken as axiomatic, and appropriate for guiding reasoning, when it was part of, or followed logically from, a stipulated definition of a technical term such as a formally defined attribute. But we can also add as axioms assertions that we wish to assume as (necessarily) true even when they are not explicitly part of, or a deductive consequence of, the formal definition of a technical metadata term. For instance, the attribute "owns" may be defined in some metadata standard by using phrases and synonyms that serve only to direct our attention to the appropriate familiar concept, but that do not explicitly say that whoever owns a collection owns each of its items. If we wish to assume that the relevant concept of ownership does indeed entail propagation then we can add propagation as an axiom.

From the perspective of inferencing, the framework provided in Chapter 5 will accommodate propagation rules of any of the four sorts described above. It will also support the use of propagation rules that are not actually being asserted, but that are articulated and formalized in order to be tested. However the "paradoxes" for conditionals that were described in the preceding section will complicate our effort to give a formal definition of propagation and propagation rule that matches our intuitions exactly. The next section shows how the problems with conditionals generate counterintuitive results that must be carefully managed in order to avoid substantively flawed definitions. The technique used for this management, modal exclusion, appears to be successful in eliminating problems for logical assertions (of both kinds), but may not be successful in eliminating counterintuitive results for empirical assertions.

4.3.4 Using Modal Exclusion to Avoid Trivial Satisfaction

Since the formalizations of propagation in first order logic use truth-functional material conditionals ("$P \supset Q$") to express conditional assertions they will have the counterintuitive consequences described in Section 4.3.1 as the "paradoxes of material implication." These consequences become particularly awkward in definitions that classify attributes according

to the kinds of propagation they support. The discussion below focuses on the case of universal attribute/value propagation from collections to items, but the modal exclusion technique used to handle the paradoxes can be applied to any of the categories discussed in Chapter 5.

The following example is from Renear et al. (2008a).

> Consider the attribute, *acme:collIdentifier*, whose value is intended to be a collection identifier assigned by a particular identifier assignment agency, the ACME collection identifier agency. This attribute obviously should not value propagate from collections to items: one cannot conclude from the fact that a collection has a value for *acme:collIdentifier* that the items in the collection have that value (or even any value) for *acme:collIdentifier*. However before the assignment of any of these collection identifiers by the ACME agency there will be no collections with a value for *acme:collIdentifier*. Therefore, the conditional rule for attribute/value propagation will be satisfied ("trivially") and *acme:collIdentifier* will be classified as attribute/value propagating, which it is not.
>
> To avoid this erroneous result, we can use a modal version of the conditional which, in the case of universal collection-item propagation, states that we have universal attribute/value propagation for an attribute A if and only if it is *impossible* for: a collection to have v for A and its items not have v for A.

Revising the definition of attribute/value propagation according to this strategy gives us the following:

An attribute A universally value propagates from collections to items $=_{df}$

$$\Box\forall y\forall z((A(y,z)\,\&\,Collection(y)) \supset \forall x(isGatheredInto(x,y) \supset A(x,z)) \tag{4.11}$$

Where the "$\Box$" is read "necessarily".

However, as mentioned in Section 4.3.2, modalized conditionals are themselves suscep-

tible to "the paradoxes of *strict* implication": if the antecedent of a modal conditional is *necessarily* false, then the conditional is true regardless of the consequent; and if the consequent is *necessarily* true, then the conditional is true, regardless of the antecedent.

This problem can be addressed by modal restrictions that exclude the further counterexamples.

An attribute A universally value propagates from collections to items $=_{df}$

1. a) It is possible for a collection to have some value z for the attribute A; &

   b) It is possible for a collection member to not have some value z for the attribute A; &

   c) It is possible that some value for A is had by one thing and lacked by another; &

2. Necessarily, if some item is a member of a collection which has some value for A, then that item has that value for A.

Or, in first order modal logic:

An attribute A attribute-value propagates from collections to items $=_{df}$

1. a) $\Diamond\exists y\exists z[Collection(y)\,\&\,A(y,z)]\,\&$

   b) $\Diamond\exists x\exists z[Member(x)\,\&\,\neg A(x,z)]\,\&$

   c) $\Diamond\exists x\exists y\exists z[A(x,z)\&\neg A(y,z)]\,\&$

2. $\Box\forall x\forall y\forall z[(IsGatheredInto(x,y)\,\&\,A(y,z))\supset A(x,z)]$.

Where "$\Diamond$" is read it is possible that and is equivalent to "$\neg\Box\neg$", Similar modal definitions can be developed for other categories of rules in the framework.

As we observed in Renear et al. (2008a) the sort of problem created here by trivial satisfaction has been noted in the information retrieval literature. Van Rijsbergen (1986) and Lalmas (1998) describe the problem in the context of logic-based approaches to information retrieval concepts and argue that it is indeed a serious problem. Sebastiani (1998) in turn argues that it is not a problem because the conditionals in question do not nest at the level where genuine problems – rather than simply strange results – are generated. Our analysis

above of the *acme:collIdentifier* case seems to support van Rijsbergen and Lalmas. Apparently when conditionals are used in definitions they are indeed nesting at a problematic level. Consequently the technique of preemptive modal inclusion seems to be required for counterexample-free definitions of propagation categories.

4.4 Expressiveness

The development of the formal system of inference rules exposes the features of inference rules that express collection/item metadata relationships. Some work has been done to characterize the expressiveness that is necessary to fully express those rules (Wickett, 2009, 2011). These requirements help situate the framework in terms of general reasoning tasks (Brachman and Levesque, 2004), and understanding them is an essential part of building systems to use these rules directly in a knowledge representation language (Wickett, 2009, 2011).

4.4.1 Propagation Rules as Horn Clauses

Many logic programming systems that might be used to manage a knowledge base of collection-level and item-level metadata operate on the basis of a reasoning technique known as resolution. Resolution is a procedure to determine entailment of formulas for a knowledge base by searching for individuals that simultaneously satisfy some set of formulas (Robinson, 1965). A resolution procedure searches for a way for a knowledge base to make all of the statements in some set of propositions and first order formulas true. For example, suppose we have a knowledge base that contains ownership information about collections and information about collection membership. If we also include an ownership rule that states that whoever owns a collection owns every item in the collection, then we would expect our reasoning system to allow us to conclude that the owner of some particular item is the owner of the collection which contains the item.

This kind of reasoning technique will only be possible if the rules for managing collection/item metadata relationships can be expressed in a form that reasoning procedures like

resolution can work with. Typically, resolution systems require the expression of formulas in conjunctive normal form (CNF). Every formula in first order can be expressed in CNF, but in order to avoid problems with the computational complexity of reasoning with full first order logic, resolution systems are restricted to Horn clauses. Horn clauses are formulas in conjunctive normal that contain at most one positive literal. Descriptive statements about resources like items and collections that are expressed with binary predicates correspond to Horn clauses, since they will appear as a single positive literal. Conditionals with a single positive predicate atom as the consequent also correspond to Horn clauses. But the rules that have been developed for expressing collection/item metadata relationship have a complex structure, so it is not immediately obvious whether they will correspond to Horn clauses.

Standard techniques were used to transform the rules given in Wickett et al. (2010) into clausal form, where each rule is expressed as a universally quantified conjunction of disjunctions. The clauses are shown with each formula separated by curly braces, and each disjunction delimited by square brackets.

UP-AP-VP:

$$\{[\neg A(x_1, x_2), \neg Collection(x_1), \neg isGatheredInto(x_3, x_1), A(x_3, x_2)]\} \tag{4.12}$$

UP-AP-VC:

$$\begin{aligned}&\{[\neg A(x_4, x_5), \neg Collection(x_4), \neg isGatheredInto(x_6, x_4), A(x_6, f_1(x_4, x_5, x6))],\\ &\ [\neg A(x_4, x_5), \neg Collection(x_4), \neg isGatheredInto(x_6, x_4), C(f_1(x_4, x_5, x6), x_5)]\}\end{aligned} \tag{4.13}$$

UP-AD-VP:

$$\{[\neg A(x_8, x_9), \neg Collection(x_8), \neg isGatheredInto(x_{10}, x_8), B(x_{10}, x_9)]\} \tag{4.14}$$

UP-AD-VC:

$$\{[\neg A(x_{11}, x_{12}), \neg Collection(x_{11}), \neg isGatheredInto(x_{13}, x_{11}), B(x_{13}, f_2(x_{11}, x_{12}, x_{13}))],$$
$$[\neg A(x_{11}, x_{12}), \neg Collection(x_{11}), \neg isGatheredInto(x_{13}, x_{11}), C(f_2(x_{11}, x_{12}, x_{13}), x_{12})]\} \tag{4.15}$$

EP-AP-VP:

$$\{[\neg A(x_{15}, x_{16}), \neg Collection(x_{15}), isGatheredInto(f_3(x_{15}, x_{16}), x_{15})],$$
$$[\neg A(x_{15}, x_{16}), \neg Collection(x_{15}), A(f_3(x_{15}, x_{16}), x_{16})]\} \tag{4.16}$$

EP-AP-VC:

$$\{[\neg A(x_{18}, x_{19}), \neg Collection(x_{18}), isGatheredInto(f_4(x_{18}, x_{19}), x_{18})],$$
$$[\neg A(x_{18}, x_{19}), \neg Collection(x_{18}), A(f_4(x_{18}, x_{19}), f_5(x_{18}, x_{19}))],$$
$$[\neg A(x_{18}, x_{19}), \neg Collection(x_{18}), C(f_5(x_{18}, x_{19}), x_{19})]\} \tag{4.17}$$

EP-AD-VP:

$$\{[\neg A(x_{22}, x_{23}), \neg Collection(x_{22}), isGatheredInto(f_6(x_{22}, x_{23}), x_{22})],$$
$$[\neg A(x_{22}, x_{23}), \neg Collection(x_{22}), B(f_6(x_{22}, x_{23}), x_{23})]\} \tag{4.18}$$

EP-AD-VC:

$$\{[\neg A(x_{25}, x_{26}), \neg Collection(x_{25}), isGatheredInto(f_7(x_{25}, x_{26}), x_{25})],$$
$$[\neg A(x_{25}, x_{26}), \neg Collection(x_{25}), B(f_7(x_{25}, x_{26}), f_8(x_{25}, x_{26}))],$$
$$[\neg A(x_{25}, x_{26}), \neg Collection(x_{25}), C(f_8(x_{25}, x_{26}), x_{26})]\} \tag{4.19}$$

It is clear from this conversion that the rules from Wickett et al. (2010) correspond to definite Horn clauses, wherein there is exactly one positive literal in each clause. This means

that the rules will support reasoning in systems that operate on resolution, like many logic programming systems.

4.4.2 Propagation Rules in Semantic Web Languages

Rules to express collection/item metadata relationships have been developed in this study using the expressive capabilities of first order logic, including a functionally complete set of connectives and universal and existential quantifiers. Since the semantic web ontology language OWL (Horrocks et al., 2003) also incorporates class constructors that correspond to the full set of connectives from first order logic, this might give the impression that the language (or at least OWL Full, which is the most expressive level) can represent anything that can be said with first order logic. However, this is not always the case.

Propagation rules that link collection and item descriptions operate on the basis of relationships between values for some collection-level property and some item-level property. The values themselves are not known for the general case of a rule, so the connection is specified using individual variables. For example, the following formula is the rule schema for attribute value propagation from collections to items.

$$\forall y \forall z((A(y,z)\,\&\,Collection(y)) \supset \\ \forall x(isGatheredInto(x,y) \supset A(x,z)) \tag{4.20}$$

In this rule schema, the individual variable z is used to represent the attribute value that is expected to propagation from the collection level to the item level.

While OWL is known to be expressively equivalent to a description logic, and as such it is possible to express facts about individuals, classes, and relationships between classes, it does not allow the use of an unknown individual (i.e. a variable) to describe a class (Horrocks et al., 2005). Thus the kind of general inference rule expressed in a propagation rule is outside the scope of OWL. One approach to expressing propagation rules for semantic web environments is to use a semantic web rule language Wickett (2009).

Below is a propagation rule for ownership, which states that if a collection has some

value for the *owner* attribute, then every item in the collection has the same value for *owner*.

$$\forall x \forall y (owner(x, y) \supset \forall z (isGatheredInto(z, x) \supset owner(z, y)) \tag{4.21}$$

Following logical equivalences, we can transform the rule into an equivalent form that is more amenable to expression in a rule language:

$$\forall x \forall y \forall z ((owner(x, y) \,\&\, isGatheredInto(z, x)) \supset owner(z, y)) \tag{4.22}$$

The following XML code shows how this rule can be expressed in the Semantic Web Rule Language (Horrocks et al., 2004).

```
<ruleml:imp>
  <ruleml:_body>
    <swrlx:individualPropertyAtom
                swrlx:property="&imlsdcc;isGatheredInto">
      <ruleml:var>x</ruleml:var>
      <ruleml:var>y</ruleml:var>
    </swrlx:individualPropertyAtom>
    <swrlx:individualPropertyAtom swrlx:property="&dccap;owner">
      <ruleml:var>y</ruleml:var>
      <ruleml:var>z</ruleml:var>
    </swrlx:individualPropertyAtom>
  </ruleml:_body>
  <ruleml:_head>
    <swrlx:individualPropertyAtom swrlx:property="&dccap;owner">
      <ruleml:var>x</ruleml:var>
      <ruleml:var>z</ruleml:var>
    </swrlx:individualPropertyAtom>
```

```
  </ruleml:_head>
</ruleml:imp>
```

Rule languages operate on a basis drawn from logic programming. The consequent of a conditional rule is called the "head" of the rule and the antecedent is called the "body". Each of the "individualPropertyAtom" elements in the XML statement corresponds to a two-place predicate atom in the formula shown above. This expression of the ownership rule conforms to a rule language for semantic web applications, but support for these rule languages is very limited.

Transitive propagation in OWL

The latest version of the Web Ontology Language (OWL) specification (Grau et al., 2008) includes mechanisms to support *transitive propagation*, which provides for the encoding of inference rules that have a similar structure to collection/item propagation rules, on the basis of a transitive relation that supports the inference. The primary use cases for transitive propagation in OWL are in biomedical ontologies, and are centered around parthood and location relationships. For example, if it is known that an injury is located on a patient's foot, and it is known that the foot is part of the leg, then the reasoner should be able to conclude that the the injury is located on the leg. In this case, the *locatedOn* property is seen as propagating along the $isPartOf$ relationship between the foot and the leg.

If we think about this case in terms of the axioms required to support the reasoning, we would have an axiom that states that if A is part of B, then anything located on A is located on B. We can express this axiom as a logical conditional as follows.

$$\forall x \forall y \forall z((isPartOf(y,z) \,\&\, locatedOn(x,y)) \supset locatedOn(x,z)) \tag{4.23}$$

This conditional statement is similar in structure to our ownership rule:

$$\forall x \forall y \forall z((isGatheredInto(x,y) \,\&\, owner(y,z)) \supset owner(x,z)) \tag{4.24}$$

The similarity between these rules suggests that transitive propagation in OWL could be used to express collection/item metadata relationships. However, it seems that this approach would only be effective for attribute/value propagation rules, and not for the full framework of rules as presented in Chapter 5. For example, the transitive propagation mechanism will not support expressing value constraint relationships between collection-level and item-level descriptions. In addition, transitive propagation is handled in OWL by expressing a chain of sub-property/super-property relationships between the properties being modeled (by stating, for example, that *locatedOn* is a sub-property of $isPartOf$) (Golbreich and Wallace, 2009). While this may correctly approximate the semantics of the properties needed for biomedical inferencing, it does not accurately reflect the relationships used for describing collections and items in descriptive metadata. For example, it would be puzzling (and could cause problems elsewhere in a knowledge base) to treat ownership as a sub-property of collection membership, even if we hope to have ownership of items inferred from ownership of a collection as a whole.

4.5 Conclusion

This chapter discussed the logical nature of propagation rules for collection/item metadata relationships. Propagation rules are conditionals expressed in first order logic. The use of conditionals introduces certain paradoxes that may challenge the use of the rules to classify attributes. However, it appears that these difficulties can be managed by pre-emptive modal exclusion. Propagation rules were shown to be equivalent to Horn clauses, which is a promising result in terms of using knowledge representation languages for managing relationships between collection descriptions and item descriptions. Propagation rules cannot be expressed in directly OWL, but can be expressed in semantic web rule languages, such as SWRL. A limited set of collection/item propagation rules may be expressed using the mechanisms for transitive propagation provided in OWL, but transitive propagation assumes a particular semantic connection between properties that may not be valid for collection membership and descriptive attributes.

Chapter 5

Rule Categories

[1]Logical relationships between metadata that describes collections and metadata that describes items can be expressed as rules, where a rule is a logical conditional stating that if something is true, then some other thing is the case. Rules can be grouped into categories according to their logical form – features such as quantification at the collection- and item-levels, and the relationships between the attributes involved in a rule.

The logical form of these rules and their expression in first order logic and in knowledge representation systems has been in examined in detail in Chapter 4. The framework presented in this chapter represents rules as universally quantified conditionals in a standard version of first order logic that includes all connectives. These rules are based on the meanings of the various attributes that appear in metadata records for collections and items, and intended to operate along with a set of metadata. A systematic account of how colloquial metadata records are translated into a representation of the propositional content expressed by those records was given in Chapter 3.

The identification of the correct rule to express a relationship between collection-level and item-level metadata is a challenging prospect. A framework of rule categories can assist in the identification of rules by allowing the generation of rules on the basis of available attributes, and making the various potential rules explicit. As was argued in Chapter 2, the concept of collections as it is commonly used in libraries and museums has not received much attention in terms of the logical nature of collections or the collection membership relationship $isGatheredInto$. The framework presented here also contributes to our under-

[1]Portions of this chapter were adapted from Wickett, K. M., Renear, A. H., and Urban, R. J. (2010). "Rule categories for collection/item metadata relationships." In *Proceedings of the 73rd ASIS&T Annual Meeting.*

standing of collections as information organization artifacts by providing an analysis of the descriptive aspect of the relationship between collections and items.

Wickett et al. (2010) introduces a framework of rule categories for rules that support reasoning from collection description to item description, but did not provide a comprehensive framework for the relationships between collection-level and item-level metadata. The framework presented here also includes categories for rules that support reasoning from item description to collection description and rules that incorporate multiple attributes at the collection or item levels.

5.1 Reasoning About Items on the Basis of Collection Description

Given a description of a collection as a whole, it is possible to reason about the members of that collection. For example, in a particular legal and social context, we may have a rule that states that if a collection y is owned by an institution z, then every item in the collection is owned by the institution z. We call this kind of reasoning *collection-to-item reasoning*, since we can consider a property of a collection (e.g. ownership) as being propagated to the items in the collection.

5.1.1 Quantification Categories

When reasoning about items on the basis of collection description, the quantification categories in the framework refer to how quantification in the formulas operates at the item-level. Item-level quantification can be either existential (implying that something is true of at least one item in the collection), or universal (implying that something is true of every item in the collection). This gives two general notions of collection-to-item propagation: universal collection-to-item propagation and existential collection-to-item propagation.

Universal collection-to-item propagation

Attributes A and B propagate from collections to items **universally** $=_{df}$

> If a collection y has the value z for the attribute A, then every item in the collection has some value w for the attribute B such that w is related to z by the constraint C.

The general rule schema expressed in first order logic:

$$\forall y \forall z((A(y,z) \,\&\, Collection(y)) \supset$$
$$\forall x(isGatheredInto(x,y) \supset \exists w(B(x,w) \,\&\, C(w,z))) \tag{5.1}$$

For universal collection-to-item propagation, the consequent of the main conditional is itself a conditional that follows a standard pattern for expressing a categorical statement. In general, we can express a categorical statement like "all frogs are green" as a universally quantified conditional statement, "if x is a frog, then x is green." We follow the same pattern to express "every member of the collection y has the property F" as "if x is gathered into the collection y, then x has the property F." We express the property that is shared by the items in the collection as a existential statement that expresses "there is a value w for the attribute A and that value stands in a constraint relationship with the value z." This expansion of the "certain property" allows us to explicitly model attributes, values and the relationships the values might stand in.

As an example of universal collection-to-item propagation, consider a collection that is described with the collection-level property *dateItemsCreated* from the Dublin Core Collections Application Profile[2]. The definition for this property is "A range of dates over which the individual items within the collection were created." On the basis of this collection description, it is reasonable to conclude that every item in the collection was created during the time period indicated at the collection level. Since the item-level property can be applied to *every* item in the collection, this will be a case of universal propagation that can

[2]http://dublincore.org/groups/collections/collection-application-profile/

be characterized with the following rule.

$$\forall y \forall z((dateItemsCreated(y,z)\,\&\,Collection(y)) \supset$$
$$\forall x(isGatheredInto(x,y) \supset \exists(created(x,w)\,\&\,temporallyWithin(w,z))) \quad (5.2)$$

This rule is an instance of the general category defined above, where A is the collection-level attribute *dateItemsCreated*, B is the item-level attribute *created*, and C is the *temporallyWithin* constraint (read as "w is temporally within z").

Existential collection-to-item propagation

Attributes A and B propagate from collections to items **existentially** $=_{df}$

> If a collection y has the value z for the attribute A, then there is some item in the collection that has some value w for the attribute B such that w is related to z by the constraint C.

The general rule schema expressed in first order logic:

$$\forall y \forall z((A(y,z)\,\&\,Collection(y)) \supset$$
$$\exists x(isGatheredInto(x,y)\,\&\,\exists w(B(x,w)\,\&\,C(w,z))) \quad (5.3)$$

The Dublin Core Collections Application Profile defines the property *itemType* as "the nature or genre of one or more items within the collection." This definition, particularly the "one or more items" clause, suggests existential collection-to-item propagation, since it seems to support drawing a conclusion that there is at least one thing with a particular property and this aligns with the logic of existential quantification.

We can characterize existential reasoning about *itemType* with the following rule.

$$\forall y \forall z((itemType(y,z)\,\&\,Collection(y)) \supset$$
$$\exists x(isGatheredInto(x,y)\,\&\,\exists w(type(x,w)\,\&\,equalTo(w,z))) \quad (5.4)$$

This is an instance of the general existential propagation category where A is the collection-level attribute *itemType*, B is the item-level attribute *type* and the constraint C expresses equality between the attribute values.

5.1.2 Specialization Conditions

The possible relations between attributes and conditions on constraints, like the ones seen in the examples above, give rise to further categories that can be used to characterize the propagation of metadata from collections to items.

Conditions on attributes

The general categories of propagation defined above do not place any restrictions on the relationship between the attributes A and B, and the general account allows A and B to be the same attribute. We call cases where A and B are the same attribute *attribute propagation* (AP), and cases where A and B are different attributes *attribute differentiation* (AD).

The ownership example mentioned above is a case of attribute propagation, since it will involve the ownership attribute at both the collection item levels. We can characterize an ownership collection-to-item propagation rule in first order logic as follows.

$$\begin{aligned} &\forall y \forall z((ownedBy(y,z) \,\&\, Collection(y)) \supset \\ &\quad \forall x(isGatheredInto(x,y) \supset \exists(ownedBy(x,w) \,\&\, equalTo(w,z))) \end{aligned} \tag{5.5}$$

This is also a universal rule, so it is an instance of the universal general rule schema where both A and B are the ownership attribute, and the constraint C expresses equality between the attribute values. According to the framework being developed here, this a case of universal attribute propagation.

On the other hand, the rules shown above for *itemType* (formula 5.4) and *dateItemsCreated* (formula 5.2) are cases of attribute differentiation since they feature different attributes at the collection and item levels. The *itemType* rule is a case of existential

attribute differentiation, and the *dateItemsCreated* rule is a case of universal attribute differentiation.

Conditions on attribute values

Rule categories also arise from the nature of the constraint that describes the relationship between values at the collection and item levels. Cases where the constraint relationship C is identity (meaning that values are the same at the collection and item levels) are called *value propagation* (VP), and cases where the constraint relationship implies that values at the collection and item levels may be different (but related by a constraint) are called *value constraint* (VC).

The *itemType* (formula 5.4) and the *ownedBy* (formula 5.5) examples discussed above are both cases of value propagation. The *itemType* rule is a case of existential attribute differentiation with value propagation, while the *ownedBy* rule is a case of universal attribute propagation with value propagation.

Since the constraint between values in these cases is identity, it is possible to give the rules in a simplified form that is logically equivalent to the rules given above.

The *itemType* rule can be expressed as:

$$\forall y \forall z((itemType(y,z)\,\&\,Collection(y)) \supset \exists x(isGatheredInto(x,y)\,\&\,type(x,z))) \quad (5.6)$$

Similarly, the *ownedBy* rule can be expressed as:

$$\forall y \forall z((ownedBy(y,z)\,\&\,Collection(y)) \supset \forall x(isGatheredInto(x,y) \supset ownedBy(x,z))) \quad (5.7)$$

The *dateItemsCreated* (formula 5.2) rule discussed above is an example of a value constraint rule. The constraint between the values at the collection and item levels in that case is temporal containment (*temporallyWithin*), and implies that each value that reflects when an item was created will fall within the range of dates indicated at the collection level.

5.2 Reasoning About Collections on the Basis of Item Description

Given descriptions of the items in a collection it is possible to reason about the collection as an entity. We call this kind of reasoning *item-to-collection reasoning*, since is propagation from descriptions of items to descriptions of collections. For example, we might have a rule that states that if every item in a collection has the same topical subject, then the collection has the topical subject.

5.2.1 Quantification Categories

Similar to the categories described above, the categories here also refer to quantification at the item level. However, in this case, we are concluding something about a collection on the basis of items in the collection. So the item-level quantification appears in the antecedent, rather than the consequent of the conditional.

Once again, we have two general rule schemas; one for existential propagation from items to collections, and one for universal propagation from items to collections. An existential rule for propagation from items to collections will state that if there is *at least one* item in the collection with a certain property, then the collection has some property. A universal item-to-collection propagation rule will state that if *every* item in a collection has a certain property, then the collection has some property.

Universal item-to-collection propagation

Attributes A and B propagate from items to collections **universally** $=_{df}$

> If every item x that is gathered into the collection y has some value w for the attribute A that is related to some z by the constraint C, then the collection has the value z for the attribute B.

We can express this rule schema in first order logic as follows.

$$\forall y \forall z[\forall x((Collection(y) \,\&\, isGatheredInto(x,y)) \supset \exists w(A(x,w) \,\&\, C(w,z))) \supset (B(y,z))] \tag{5.8}$$

For universal propagation from items to collections, the antecedent of the main conditional is itself a conditional that follows the pattern for expressing a categorical statement that was explained in Section 5.1.

We can consider a universal rule for reasoning about the *dateItemsCreated* attribute for a collection on the basis of creation dates for items. Recall the definition of *dateItemsCreated* from the Dublin Core Collections Application Profile: "a range of dates over which the individual items within the collection were created". If each of the items in the collection has a creation date that falls within a certain range, then we can conclude that the collection has that range as the value for *dateItemsCreated*.

We can express this reasoning with the following universal item-to-collection propagation rule.

$$\forall y \forall z[\forall x((Collection(y) \,\&\, isGatheredInto(x,y)) \supset \exists w(created(x,w) \& temporallyWithin(w,z))) \supset (dateItemsCreated(y,z))] \tag{5.9}$$

This rule is an instance of the general category defined above, where A is the item-level attribute *created*, B is the collection-level attribute *dateItemsCreated*, and C is the constraint *temporallyWithin*.

Existential item-to-collection propagation

Attributes A and B propagate from items to collections **existentially** $=_{df}$

> If there is some item x that is gathered into the collection y that has some value w for the attribute A and w is related to some z by the constraint C, then the collection has the value z for the attribute B.

The general rule schema in first order logic:

$$\forall y \forall z[(Collection(y) \,\&\, \exists x(isGatheredInto(x,y) \,\&\, \exists w(A(x,w) \,\&\, C(w,z)))) \supset (B(y,z))] \tag{5.10}$$

For existential item-to-collection propagation, the antecedent of the main conditional is a conjunction that states that y is a collection and at least one of the items that is gathered into it has a certain property. As with universal item-to-collection propagation, we use a existentially quantified conjunction to express the property as an attribute value pair where the value w stands in the constraint relationship with the value z.

An existential propagation rule for reasoning about collections on the basis of item descriptions will allow us to draw a conclusion about a collection when at least one item in the collection has some property. As an example we can again consider the property *itemType* from the Dublin Core Collections Application Profile, which has the definition "the nature or genre of one or more items within the collection". So if we have a collection and at least one item in the collection has a particular genre or nature (i.e. *type*), then we can conclude that the collection has that type value for the *itemType* attribute. We can express this reasoning with the following existential rule.

$$\forall y \forall z[(Collection(y) \,\&\, \exists x(isGatheredInto(x,y) \,\&\, \exists w(type(x,w) \,\&\, equalTo(w,z)))) \supset itemType(y,z)] \tag{5.11}$$

This rule is an instance of the general category defined above for existential item-to-collection propagation, where A is the item-level attribute *type*, B is the collection-level *itemType* attribute, and C expresses identity between the attribute values.

5.2.2 Specialization Conditions

As with collection-to-item propagation, the possible relations between attributes and conditions on constraints, like the ones seen in the examples above, give rise to further categories

that can be used to characterize the propagation of metadata from items to collections.

Attribute Conditions

As with the general schemas for propagation from collections to items discussed above, these schemas place no restrictions on the attributes A and B. Once again, we call cases where A and B are the same attribute *attribute propagation* (AP), and cases where they are different attributes *attribute differentiation* (AD). The *dateItemsCreated* (formula 5.9) and *itemType* (formula 5.11) rules for item-to-collection propagation have different attributes at the collection and item levels and are therefore cases of attribute differentiation.

Suppose we have a collection where everything has a topical subject in common. Then we might want to conclude that the collection has the same topical subject.

$$\forall y \forall z[\forall x((Collection(y) \,\&\, isGatheredInto(x, y)) \supset \exists w(subject(x, w) \,\&\, equalTo(w, z))) \supset subject(y, z)] \tag{5.12}$$

This rule is an example of universal attribute propagation from items to collections, since we have the same attribute (*subject*) in the antecedent and the consequent of the rule.

Value Conditions

Rule categories also arise from the nature of the constraint that describes the relationship between values at the item and collection levels. Cases where the constraint relationship C is identity (meaning that values are the same at the collection and item levels) are called *value propagation* (VP), and cases where the constraint relationship implies that values may be different at the collection and item levels are called *value constraint* (VC).

The *dateItemsCreated* example (formula 5.9) for propagation from items to collections discussed about is a case of value constraint, where the constraint relationship is temporal containment (*temporallyWithin*). The collection-level attribute value for *dateItemsCreated* will be a date range that contains each of the item-level values for the *created* attribute.

The *itemType* (formula 5.11) and *subject* (formula 5.12) examples are both cases of

value propagation, since the constraint relationship in those rules is identity. Due to the identity between the values at the item and collection levels, it is possible to give the rules a simplified form.

The *itemType* rule can be expressed as:

$$\forall y \forall z[(Collection(y) \,\&\, \exists x(isGatheredInto(x,y) \,\&\, (type(x,z)))) \supset (itemType(y,z))] \tag{5.13}$$

The *subject* rule can be expressed as:

$$\forall y \forall z[\forall x((Collection(y) \,\&\, isGatheredInto(x,y)) \supset subject(x,z)) \supset (subject(y,z))] \tag{5.14}$$

5.3 Reasoning About Combinations of Attributes

The rules discussed up to this point have concerned only single attributes at both the collection and item levels. A complete framework needs to account for rules that allow us to come to useful conclusions about a collection based on a combination of item-level attributes, or about items on the basis of multiple collection-level attributes. We will only handle conjunctive additions of properties to the antecedent of propagation rules. This kind of rule will continue to conform to the restriction on addressing propagation rules that are equivalent to Horn Clauses.

Geopolitical entities like cities, countries, and territories are typically viewed as having both a geographical area and a temporal extent. Therefore, the description of item-level geographical regions alone may not be sufficient for drawing conclusions about the topical coverage of a collection that contains those items. For example, items that are connected to a geographic area that currently corresponds to the state of Kansas may not in fact have a topical connection to the state, which was admitted to the United States in 1861, but rather to the Kansas Territory or to the earlier Missouri Territory.

Based on this information about the Kansas Territory, we can propose this rule for propagation from the item-level information about geographic and temporal coverage to collection-level information about coverage.

$$\begin{aligned}
&\forall y((Collection(y)\,\&\,\forall x(isGatheredInto(x,y)\ \supset\\
&\quad\exists v_1\exists v_2((SpatialCoverage(x,v_1)\,\&\\
&\quad GeographicWithin(v_1,\mathrm{box}(\mathrm{lat}(37^\circ\mathrm{N},40^\circ\mathrm{N}),\mathrm{long}(94^\circ 35'\mathrm{W},102^\circ 3'\mathrm{W}))))\\
&\quad\&\,(TemporalCoverage(x,v_2)\,\&\,TemporallyWithin(v_2,1854-1861)))))\ \supset\\
&\quad(Coverage(y,\mathrm{KansasTerritory})))
\end{aligned}\tag{5.15}$$

However, this account does not express the information that the spatial area and time period of interest have a connection to the Kansas Territory. As such, it does not seem to directly express the information that is being propagated from the item level to the collection level. In order to model the connection to the Kansas Territory, we can replace the values in the constraint relations $GeographicWithin$ and $TemporalWithin$ with functions that are evaluated for the argument "Kansas Territory". Below is a version of the rule where we have replaced "box(lat(37° N, 40°N),long(94°35'W,102°3'W))" with "$GeographicExtentOf$(KansasTerritory)" and "$1854-1861$" with "$TemporalExtentOf$(Kansas Territory)".

$$\begin{aligned}
&\forall y((Collection(y)\,\&\,\forall x(isGatheredInto(x,y)\ \supset\\
&\quad\exists v_1\exists v_2((SpatialCoverage(x,v_1)\,\&\\
&\qquad GeographicWithin(v_1,GeographicExtentOf(KansasTerritory)))\,\&\\
&\quad(TemporalCoverage(x,v_2)\,\&\\
&\qquad TemporallyWithin(v_2,TemporalExtentOf(KansasTerritory)))))\ \supset\\
&\quad(Coverage(y,\mathrm{KansasTerritory})))
\end{aligned}\tag{5.16}$$

This gives us a general form for propagating information about spatial and geographic

coverage at the item level to a collection-level coverage attribute, by replacing "Kansas Territory" with the individual variable z.

$$\begin{aligned}
&\forall y \forall z((Collection(y) \,\&\, \forall x(isGatheredInto(x,y) \supset \\
&\quad \exists v_1 \exists v_2((SpatialCoverage(x,v_1) \,\& \\
&\qquad GeographicWithin(v_1, GeographicExtentOf(z))) \,\& \\
&\quad (TemporalCoverage(x,v_2) \,\& \\
&\qquad TemporallyWithin(v_2, TemporalExtentOf(z))))) \supset \\
&\quad (Coverage(y,z)))
\end{aligned} \tag{5.17}$$

Generalizing from this example rule gives us this general schema for propagation from items to collection on the basis of two attributes at the item level.

$$\begin{aligned}
&\forall y \forall z[\forall x((Collection(y) \,\&\, isGatheredInto(x,y)) \supset \\
&\quad \exists v_1 \exists v_2((A_1(x,v_1) \,\&\, C_1(v_1, f_1(z))) \,\& \\
&\qquad (A_2(x,v_2) \,\&\, C_2(v_2, f_2(z))))) \supset (B(y,z))]
\end{aligned} \tag{5.18}$$

The coverage example can be generated from this general schema by substituting $GeographicCoverage$ for A_1, $TemporalCoverage$ for A_2, $GeographicWithin$ for C_1, $TemporalWithin$ for C_2, $GeographicExtentOf$ for f_1, $TemporalExtentOf$ for f_2, and $Coverage$ for B.

We can further generalize from the case with two attributes to give a general account of rules that allow us to reason about a single collection-level attribute-value pair from n

item-level attributes.

$$\forall y \forall z[\forall x((Collection(y) \,\&\, isGatheredInto(x,y)) \supset \exists v_1, v_2, \ldots, v_n((A_1(x,v_1) \,\&\, C_1(v_1, f_1(z))) \,\& \\ (A_2(x,v_2) \,\&\, C_2(v_2, f_2(z)))) \;\&\; \ldots \\ (A_n(x,v_n) \,\&\, C_n(v_n, f_n(z)))) \supset (B(y,z))] \tag{5.19}$$

5.4 Value Constraint Relationships

Value constraint relationships stand between collection-level and item-level attribute values. The framework has categories based on whether the constraint between values is identity (value propagation) or some other relationship (value constraint). We can consider whether it is possible to generate further meaningful categories in the framework on the basis of the features of constraint relationships.

Since value constraint relationships are binary relations over a domain we can assess their order properties. The domain over which the constraint relation operates is determined by the attributes that appear in the rule. For example, a rule linking a collection-level *dateItemsCreated* attribute and an item-level *created* attribute will include a $TemporallyWithin$ constraint relation that operates over temporal entities such as years or year ranges.

$$\forall y \forall z[(Collection(y) \,\&\, \forall x(isGatheredInto(x,y) \supset \\ \exists w(created(x,w) \& temporallyWithin(w,z)))) \\ \supset (dateItemsCreated(y,z))] \tag{5.20}$$

Similarly, a rule linking collection-level and item-level geographic information will use a constraint that operates over geographic entities such as regions and places.

The $TemporallyWithin$ relation is reflexive, since any interval temporally contains itself. It is antisymmetric since the only case where $TemporallyWithin(x,y)$ together with

$TemporallyWithin(y, x)$ will be true is when y equals x. In addition, we can see that $TemporallyWithin$ is transitive, since if x is temporally within y and y is temporally within z, then x is temporally within z. These properties together mean that $TemporallyWithin$ is a partial order over a domain of temporal entities.

In some scenarios, it may be desirable to use a rule that reflects a less strict relationship between dates and intervals than temporal within-ness. For a temporal coverage attribute, a collection-to-item rule like the one that follows, which uses the looser association of temporal overlap may provide a better representation of the relationships between collection-level and item-level metadata.

$$\begin{aligned}
&\forall y \forall z((TemporalCoverage(y, z) \,\&\, Collection(y)) \supset \\
&\quad \forall x(isGatheredInto(x, y) \supset \\
&\quad\quad \exists w(TemporalCoverage(x, w) \,\&\, TemporalOverlap(w, z))))
\end{aligned} \tag{5.21}$$

The temporal overlap relation also operates over a domain of temporal entities such as dates and date ranges. Temporal containment is a special case of temporal overlap. If a temporal interval x overlaps with a temporal interval y and none of y occurs outside of x, then x temporally contains y (which we would write in a rule as $TemporallyWithin(y, x)$). $TemporalOverlap$ is reflexive since any temporal entity overlaps itself. The relation is symmetric because whenever x temporally overlaps y, y will also temporally overlap x. $TemporalOverlap$ is not anti-symmetric since it is possible for x to overlap y and y to overlap x without x and y being equal. In addition, it is possible for x to overlap y and y to overlap z without x overlapping z, so $TemporalOverlap$ is not transitive. Therefore, in contrast to $TemporallyWithin$, the $TemporalOverlap$ relation is not a partial order over temporal entities.

The following is an example rule for collection-to-item reasoning about geographic cov-

erage that uses a withinness constraint.

$$\forall y \forall z((GeographicCoverage(y,z)\,\&\,Collection(y)) \supset$$
$$\forall x(isGatheredInto(x,y) \supset$$
$$\exists w(GeographicCoverage(x,w)\,\&\,GeographicWithin(w,z)))) \tag{5.22}$$

This rule states that if a collection has a particular geographic coverage, then all of the items in the collection will have geographic coverage that corresponds to some geographic entity (a place or a region) that is within the region indicated for the collection. The constraint for geographic withinness used here will also be reflexive, since any geographic area contains itself. It will also be antisymmetric because if an area x contains an area y and y also contains x, then y and x must be the same area. Finally, $GeographicWithin$ is transitive since if an area x contains y and y contains z, then x will contain z. Since $GeographicWithin$ is reflexive, anti-symmetric and transitive, it is a partial order over a domain of geographic entities.

Depending on the application, it may be appropriate to select a rule for geographic coverage attributes that uses a different constraint relation to reflect the association between collection-level and item-level values. The following rule uses a geographic overlap constraint instead of geographic containment.

$$\forall y \forall z((GeographicCoverage(y,z)\,\&\,Collection(y)) \supset$$
$$\forall x(isGatheredInto(x,y) \supset$$
$$\exists w(GeographicCoverage(x,w)\,\&\,GeographicOverlap(w,z)))) \tag{5.23}$$

This rule states that if a collection has a particular geographic coverage, then all of the items in the collection will have geographic coverage that corresponds to some geographic entity that overlaps with the region indicated for the collection. Geographic containment is a special case of geographic overlap. If the geographic region x overlaps the region y and none of y occurs outside of x, then the region y is geographically within the region x. Geographic

overlap is reflexive since every geographic region overlaps itself, and symmetric since if the geographic region x overlaps with the region y then y overlaps with x. Geographic overlap is not antisymmetric, since it is possible for the region x to overlap the region y and for y to overlap x without x and y being equal. Similarly, geographic overlap is not transitive since it is possible for the region x to overlap the region y and y to overlap the region z without x overlapping with z. Therefore, *GeographicOverlap* is not a partial order over geographic regions.

The following is a rule for *itemType* that appears in Wickett et al. (2010).

$$\forall y \forall z((itemType(y,z) \,\&\, Collection(y)) \supset \exists x(isGatheredInto(x,y) \,\&\, \exists w(type(x,w) \,\&\, Generalizes(w,z)))) \tag{5.24}$$

This rule states that if a collection has an *itemType*, then there is a member of the collection that has a type that stands in a generalization relationship with the collection-level *itemType*. We say that a value x *Generalizes* a value y when x can be applied to everything that y can be applied to. For example, in the analysis of a metadata repository in Wickett et al. (2010), "images" was considered a generalization of "photographs/slides/negatives" since everything that falls with the category of "photographs/slides/negatives" also falls within the category of "images." Given this definition of the relationship, *Generalizes* will be reflexive, antisymmetric, and transitive and therefore will constitute a partial order over a set of vocabulary terms.

Instead of using a generalization constraint for type values, it may be appropriate to select a rule that states that the type value will be a member of a specified set of values. Consider a family archive collection y that is described with the collection-level *itemType* value "photos, letters, documents." The following rule states that any item x in the collec-

tion has a type value that is a member of the set {photo, letter, document}.

$$\forall y \forall z((itemType(y, \{photo, letter, document\}) \,\&\, Collection(y)) \supset$$
$$\exists x(isGatheredInto(x, y) \,\&$$
$$\exists w(type(x, w) \,\&\, memberOf(w, \{photo, letter, document\}))))) \tag{5.25}$$

This rule uses the set membership relation $memberOf$ to reflect the fact that the item-level type value is a member of the set of values given at the collection level. Generalizing to any set of type values gives the following rule.

$$\forall y \forall z((itemType(y, \{v_1, v_2, \ldots, v_n\}) \,\&\, Collection(y)) \supset$$
$$\exists x(isGatheredInto(x, y) \,\&$$
$$\exists w(type(x, w) \,\&\, memberOf(w, \{v_1, v_2, \ldots, v_n\})))) \tag{5.26}$$

The relation properties of the set membership relation can be assessed by assuming that the relation conforms to the axioms given for ZFC in Fraenkel and Bar-Hillel (1958). Set membership is not reflexive as it is not the case that every set is a member of itself. In fact, in ZFC, no set can be a member of itself, so set membership is irreflexive. Set membership is not symmetric, and is in fact asymmetric since if y is a member of x then x cannot be a member of y. The relation is antisymmetric, but only as a consequence of asymmetry since antisymmetry is defined by a conditional (if x is a member of y and y is a member of x then x and y are equal) that always has a false antecedent due to the asymmetry of set membership. Set membership is not transitive, as it is possible for x to be a member of the set y and y to be a member of the set z without x itself being a member of the set z. Therefore, set membership is not a partial order over a set of vocabulary terms.

Although such categories are not developed here, these examples suggest a further division in the framework, between rules that use constraints that form a partial order over the values for the attributes in a rule, and rules that use constraints that are not partial orders. The order properties of a constraint are not determined solely by the domain of val-

ues that the constraint operates over. As shown in the previous examples, given a domain of values (such as time periods, geographic regions, or vocabulary terms) it is possible to define constraints that either are or are not partial orders.

5.5 Categories

The rules discussed in this chapter can classified into categories on the basis of the logical features of the rules. The most fundamental division between rules for reasoning about collection-level and item-level metadata is between rules that support reasoning about items on the basis of collection descriptions (collection-to-item rules) and rules that support reasoning about collections on the basis of item descriptions (item-to-collection rules). The other categories are based on specializing these top-level categories.

5.5.1 General Rule Schemas

In order to incorporate the rules developed in Section 5.3, which support reasoning on the basis of multiple attributes at either the collection or item levels, the general rule schemas for collection-to-item and item-to-collection reasoning can be expanded as follows.

Item-to-Collection Universal Propagation (ICUP)

$$\begin{aligned}&\forall y \forall z[(\forall x(isGatheredInto(x,y) \supset \\ &\exists v_1,\ldots,v_n((A_1(x,v_1)\,\&\,C_1(v_1,f_1(z)))\,\&\,\ldots\,\&\,(A_n(x,v_n)\,\&\,C_n(v_n,f_n(z))))) \supset \\ &\quad (Collection(y)\,\&\,B(y,z)))]\end{aligned} \tag{5.27}$$

This general rule schema states that for any y, if y is a collection and for any x if x is gathered into y then there exists v_1 through v_n and z such that v_1 through v_n are values for the attributes A_1 through A_n respectively, and for constraints C_1 through C_n and functions f_1 through f_n, for each value v_i $C_i(v_i, f_i(z)))$, then the collection y has the value z for the attribute B.

Categories can be generated from this general schema by applying the specialization

conditions described below. For example, a rule that operates on the basis of a single attribute (SA) at the item level, the index n will be equal to one and the function connecting the result of the value constraint to the collection-level value z will be identity. Therefore, in these rules, the complex clause accounting for multiple attributes will reduce to $(A(x, v) \,\&\, C(v, z))$.

Item-to-Collection Existential Propagation (ICEP)

$$\begin{aligned}
&\forall y \forall z[(\exists x(isGatheredInto(x, y) \,\& \\
&\exists v_1, \ldots, v_n((A_1(x, v_1) \,\& C_1(v_1, f_1(z))) \,\& \ \ldots \ \& (A_n(x, v_n) \,\& C_n(v_n, f_n(z))))) \supset \\
&\quad (Collection(y) \,\& B(y, z)))]
\end{aligned} \tag{5.28}$$

This general rule schema states that for any y, if y is a collection and there exists x such that x is gathered into y and there exists v_1 through v_n and z such that v_1 through v_n are values for x for the attributes A_1 through A_n respectively, and for constraints C_1 through C_n and functions f_1 through f_n, for each value v_i $C_i(v_i, f_i(z)))$, then the collection y has the value z for the attribute B.

Collection-to-Item Universal Propagation (CIUP)

$$\begin{aligned}
&\forall y((Collection(y) \,\& \exists v_1, v_2, \ldots, v_n(A_1(y, v_1) \,\& A_2(y, v_2))) \ \& \ \ldots \ \& A_n(y, v_n)) \supset \\
&\forall x(isGatheredInto(x, y) \supset \\
&\quad \exists z(B(x, z) \,\& (C_1(v_1, f_1(z)) \,\& C_2(v_2, f_2(z)) \,\& \ \ldots \ \& C_n(v_n, f_n(z)))))
\end{aligned} \tag{5.29}$$

This general rule schema states that for any y, if y is a collection and there exists v_1 through v_n such that v_1 through v_n are values for y for the collection-level attributes A_1 through A_n respectively, then for any x, if x is gathered into y then there exists a z such that z is the value for the item-level attribute B for x and for constraints C_1 through C_n and functions f_1 through f_n, for each value v_i $C_i(v_i, f_i(z)))$.

Collection-to-Item Existential Propagation (CIEP)

$$\begin{aligned}
&\forall y((Collection(y)\,\&\\
&\quad \exists v_1, v_2, \ldots, v_n(A_1(y, v_1)\,\&\,A_2(y, v_2)))\ \&\ \ldots\ \&\ A_n(y, v_n))\ \supset\\
&\exists x(isGatheredInto(x, y)\,\&\\
&\quad \exists z(B(x, z)\,\&\,(C_1(v_1, f_1(z))\,\&\,C_2(v_2, f_2(z))\ \&\ \ldots\ \&\,C_n(v_n, f_n(z)))))
\end{aligned} \tag{5.30}$$

This general rule schema states that for any y, if y is a collection and there exists v_1 through v_n such that v_1 through v_n are values for y for the collection-level attributes A_1 through A_n respectively, then there is an x that is gathered into y and there exists a z such that z is the value for the item-level attribute B for x and for constraints C_1 through C_n and functions f_1 through f_n, for each value v_i $C_i(v_i, f_i(z)))$.

5.5.2 Specialization Conditions

Table 5.1 shows the specialization conditions that might be applied to the general rule schemas for universal and existential propagation from collections to items and from items to collections.

Table 5.1: Specialization Conditions

$n = 1$	single attribute propagation (SA)
$n > 1$	multiple attribute propagation (MA)
$A = B$	attribute propagation (AP)
$\neg(A = B)$	attribute differentiation (AD)
$\forall x \forall y(C(x, y) \equiv x = y)$	value propagation (VP)
$\neg\forall x \forall y(C(x, y) \equiv x = y)$	value constraint (VC)

Each of the six specialization conditions in Table 5.1 can be applied directly to item-to-collection universal propagation (ICUP), item-to-collection existential propagation (ICEP), collection-to-item universal propagation (CIUP), and collection-to-item existential propagation (CIEP), which will result in twenty-four specialized rule categories.

The specialization conditions on multiple or single attributes, attribute differentiation

or propagation, and value propagation or constraint can also be combined in the ways shown in Table 5.2. Combinations that are logically contradictory are not considered (e.g. $(A = B) \,\&\, \neg(A = B)$). In addition, multiple-attribute (MA) propagation rules support reasoning about a single attribute at one level on the basis of a combination of attributes at the other level. Therefore cases attribute propagation (AP), which feature the same attribute at the collection and item levels, will only occur when all of A_1 through A_n are equal to each other and to B. In these cases, multiple values for the same attribute will be contributing to a conclusion about an attribute at the other level of description. Similarly, value propagation will occur in multiple-attribute propagation only when all of C_1 through C_n are identity, and v_1 through v_n are all equal. All cases of attribute propagation with value propagation will be cases of single-attribute propagation, since statements of the same value for the same attribute are just repetitions of the same property, so would not appear as distinct statements in a rule.

Table 5.2: Combinations of Conditions

$n = 1 \quad \& \quad (A = B) \quad \& \quad \forall x \forall y (C(x, y) \equiv x = y)$	(SA) AP-VP
$n = 1 \quad \& \quad (A = B) \quad \& \quad \neg\forall x \forall y (C(x, y) \equiv x = y)$	SA-AP-VC
$n > 1$ & $(A_i = B)$ for $i = 1, \ldots, n$ & $\neg\forall x \forall y (C(x, y) \equiv x = y)$	MA-AP-VC
$n = 1 \quad \& \quad \neg(A = B) \quad \& \quad \forall x \forall y (C(x, y) \equiv x = y)$	SA-AD-VP
$n > 1$ & $\quad (A = B) \quad \& \quad \forall x \forall y (C_i(x, y) \equiv x = y)$ for $i = 1, \ldots, n$	MA-AD-VP
$n = 1 \quad \& \quad \neg(A = B) \quad \& \quad \neg\forall x \forall y (C(x, y) \equiv x = y)$	SA-AD-VC
$n > 1 \quad \& \quad \neg(A = B) \quad \& \quad \neg\forall x \forall y (C_i(x, y) \equiv x = y)$ for $i = 1, \ldots, n$	MA-AD-VC

5.5.3 Specialized Rule Categories

Under the constraints mentioned above, each of the six combined specialization conditions in Table 5.2 can also be applied to ICUP, ICEP, CIUP and CIEP, yielding the twenty-eight fully specialized rule categories that follow.

Item-to-Collection Universal Propagation Categories

$$\text{UP/AP/VP/(SA)}: \quad \forall y \forall z[\forall x(isGatheredInto(x,y) \supset A(x,z)) \supset (Collection(y) \,\&\, A(y,z))]$$

$$\text{UP/AP/VC/MA}: \quad \forall y \forall z[\forall x(isGatheredInto(x,y) \supset \exists v_1,\ldots,v_n((A(x,v_1) \,\&\, C_1(v_1,f_1(z))) \,\&\, \ldots \,\&\, (A(x,v_n) \,\&\, C_n(v_n,f_n(z)))) \supset (Collection(y) \,\&\, A(y,z)))]$$

$$\text{UP/AP/VC/SA}: \quad \forall y \forall z[\forall x(isGatheredInto(x,y) \supset \exists w(A(x,w) \,\&\, C(w,z))) \supset (Collection(y) \,\&\, A(y,z))]$$

$$\text{UP/AD/VP/MA}: \quad \forall y \forall z[\forall x(isGatheredInto(x,y) \supset (A_1(x,z) \,\&\, \ldots \,\&\, A_n(x,z))) \supset (Collection(y) \,\&\, B(y,z))]$$

$$\text{UP/AD/VP/SA}: \quad \forall y \forall z[\forall x(isGatheredInto(x,y) \supset A(x,z)) \supset (Collection(y) \,\&\, B(y,z))]$$

$$\text{UP/AD/VC/MA}: \quad \forall y \forall z[\forall x(isGatheredInto(x,y) \supset \exists v_1,\ldots,v_n((A_1(x,v_1) \,\&\, C_1(v_1,f_1(z))) \,\&\, \ldots \,\&\, (A_n(x,v_n) \,\&\, C_n(v_n,f_n(z))))) \supset (Collection(y) \,\&\, B(y,z))]$$

$$\text{UP/AD/VC/SA}: \quad \forall y \forall z[\forall x(isGatheredInto(x,y) \supset \exists w(A(x,w) \,\&\, C(w,z))) \supset (Collection(y) \,\&\, B(y,z))]$$

Item-to-Collection Existential Propagation Categories

$$\text{EP/AP/VP/(SA)}: \quad \forall y \forall z[\exists x(isGatheredInto(x,y) \,\&\, A(x,z)) \supset (Collection(y) \,\&\, A(y,z))]$$

$$\text{EP/AP/VC/MA}: \quad \forall y \forall z[\exists x(isGatheredInto(x,y) \,\&\, \exists v_1,\ldots,v_n((A(x,v_1) \,\&\, C_1(v_1,f_1(z))) \,\&\, \ldots \,\&\, (A(x,v_n) \,\&\, C_n(v_n,f_n(z))))) \supset (Collection(y) \,\&\, A(y,z))]$$

$$\text{EP/AP/VC/SA}: \quad \forall y \forall z[\exists x(isGatheredInto(x,y) \,\&\, \exists w(A(x,w) \,\&\, C(w,z))) \supset (Collection(y) \,\&\, A(y,z))]$$

$$\text{EP/AD/VP/MA}: \quad \forall y \forall z[\exists x(isGatheredInto(x,y) \,\&\, (A_1(x,z) \,\&\, \ldots \,\&\, A_n(x,z))) \supset (Collection(y) \,\&\, B(y,z))]$$

$$\text{EP/AD/VP/SA}: \quad \forall y \exists z[\exists x(isGatheredInto(x,y) \,\&\, A(x,z)) \supset (Collection(y) \,\&\, B(y,z))]$$

$$\text{EP/AD/VC/MA}: \quad \forall y \forall z[\exists x(isGatheredInto(x,y) \,\&\, \exists v_1,\ldots,v_n((A_1(x,v_1) \,\&\, C_1(v_1,f_1(z))) \,\&\, \ldots \,\&\, (A_n(x,v_n) \,\&\, C_n(v_n,f_n(z))))) \supset (Collection(y) \,\&\, B(y,z))]$$

$$\text{EP/AD/VC/SA}: \quad \forall y \forall z[\exists x(isGatheredInto(x,y) \,\&\, \exists w(A(x,w) \,\&\, C(w,z))) \supset (Collection(y) \,\&\, B(y,z))]$$

Collection-to-Item Universal Propagation Categories

UP/AP/VP/(SA) : $\forall y \forall z((A(y,z)\,\&\,Collection(y)) \supset \forall x(isGatheredInto(x,y) \supset A(x,z)))$

UP/AP/VC/MA : $\forall y \forall z((Collection(y)\,\& \exists v_1, v_2, \ldots, v_n((A(x,v_1)\,\&\,C_1(v_1, f_1(z)))\,\&\,\ldots\,\& (A(x,v_n)\,\&\,C_n(v_n, f_n(z))) \supset \forall x(isGatheredInto(x,y) \supset \exists w(A(x,w)\,\&\,C(w,z)))$

UP/AP/VC/SA : $\forall y \forall z((Collection(y)\,\&\,A(y,z)) \supset \forall x(isGatheredInto(x,y) \supset \exists w(A(x,w)\,\&\,C(w,z))))$

UP/AD/VP/MA : $\forall y \forall z((Collection(y)\,\& \exists z(A_1(x,z)\,\&\,A_2(x,z)\,\&\,\ldots\,\&\,A_n(x,z)) \supset \forall x(isGatheredInto(x,y) \supset B(x,z))$

UP/AD/VP/SA : $\forall y \forall z((A(y,z)\,\&\,Collection(y)) \supset \forall x(isGatheredInto(x,y) \supset B(x,z)))$

UP/AD/VC/MA : $\forall y \forall z((Collection(y)\,\& \exists v_1, v_2, \ldots, v_n((A_1(x,v_1)\,\&\,C_1(v_1, f_1(z)))\,\&\,\ldots\,\& (A_n(x,v_n)\,\&\,C_n(v_n, f_n(z))) \supset \forall x(isGatheredInto(x,y) \supset \exists w(B(x,w)\,\&\,C(w,z)))$

UP/AD/VC/SA : $\forall y \forall z((A(y,z)\,\&\,Collection(y)) \supset \forall x(isGatheredInto(x,y) \supset \exists w(B(x,w)\,\&\,C(w,z))))$

Collection-to-Item Existential Propagation Categories

$$\text{EP/AP/VP(SA)}: \quad \forall y \forall z((A(y,z) \,\&\, Collection(y)) \supset \exists x(isGatheredInto(x,y) \,\&\, A(x,z)))$$

$$\text{EP/AP/VC/MA}: \quad \forall y \forall z((Collection(y) \,\&\, \exists v_1, v_2, \ldots, v_n((A(x,v_1) \,\&\, C_1(v_1, f_1(z))) \,\&\, \ldots \,\&\, (A(x,v_n) \,\&\, C_n(v_n, f_n(z))) \supset \exists x(isGatheredInto(x,y) \,\&\, \exists w(A(x,w) \,\&\, C(w,z)))$$

$$\text{EP/AP/VC/SA}: \quad \forall y \forall z((A(y,z) \,\&\, Collection(y)) \supset \exists x(isGatheredInto(x,y) \,\&\, \exists w(A(x,w) \,\&\, C(w,z))))$$

$$\text{EP/AD/VP/MA}: \quad \forall y \forall z(Collection(y) \,\&\, (A_1(x,z) \,\&\, \ldots \,\&\, A_n(x,z)) \supset \exists x(isGatheredInto(x,y) \,\&\, B(x,z)))$$

$$\text{EP/AD/VP/SA}: \quad \forall y \forall z((A(y,z) \,\&\, Collection(y)) \supset \exists x(isGatheredInto(x,y) \,\&\, B(x,z)))$$

$$\text{EP/AD/VC/MA}: \quad \forall y \forall z((Collection(y) \,\&\, \exists v_1, v_2, \ldots, v_n((A_1(x,v_1) \,\&\, C_1(v_1, f_1(z))) \,\&\, \ldots \,\&\, (A_n(x,v_n) \,\&\, C_n(v_n, f_n(z))) \supset \exists x(isGatheredInto(x,y) \,\&\, \exists w(B(x,w) \,\&\, C(w,z)))$$

$$\text{EP/AD/VC/SA}: \quad \forall y \forall z((A(y,z) \,\&\, Collection(y)) \supset \exists x(isGatheredInto(x,y) \,\&\, \exists w(B(x,w) \,\&\, C(w,z))))$$

Each of the example rules discussed previously is an instance of one of these fully specialized categories. The *itemType* rules for collection-to-item propagation (formula 5.6) and for item-to-collection propagation (formula 5.13) are both cases of existential attribute differentiation with value propagation for a single attribute (EP/AD/VP), the *dateItemsCreated*

rules (formulas 5.2 and 5.9) are cases of universal attribute differentiation with value constraint (UP/AD/VC), and the *ownedBy* (formula 5.7) and *subject* (formula 5.14) rules are cases of universal attribute propagation with value propagation (UP/AP/VP).

5.5.4 Logical Relationships Between Categories

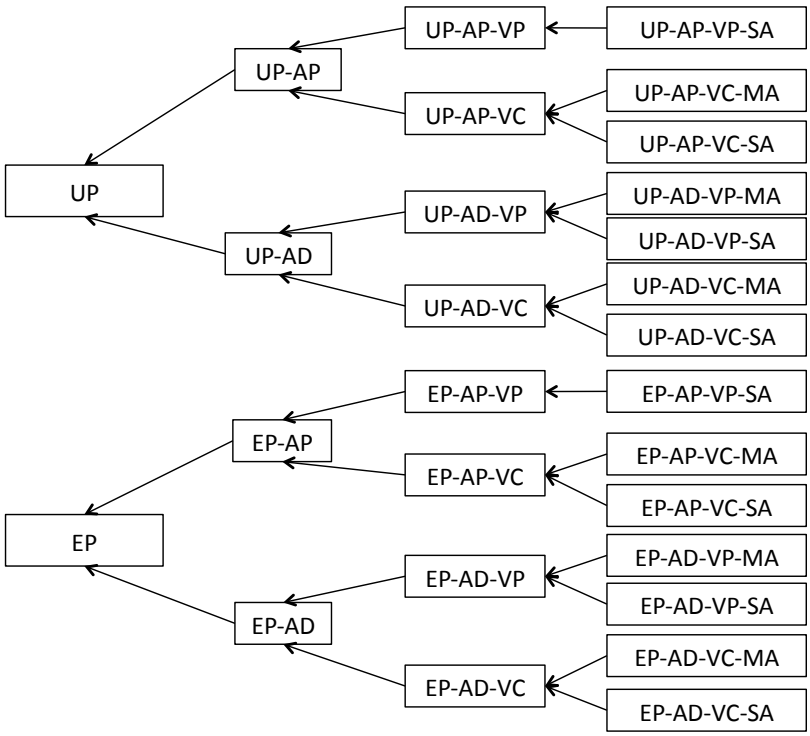

Figure 5.1: Collection-to-item propagation categories and entailments based on the specialization hierarchy.

The frameworks of rules for collection-to-item propagation and for item-to-collection propagation have the same structure, since they are both constructed by the adding the specialization conditions from Table 5.1. The structure of the collection-to-item categories is shown in Figure 5.1 and the structure of the item-to-collection categories is shown in 5.2[3]. The logical relationships between the categories have two sources: the specialization/generalization structure of the framework and the relationship between universal and

[3]The technique seen here and in the following figures for the visual presentation of the rule categories and their relationships was developed by Richard Urban.

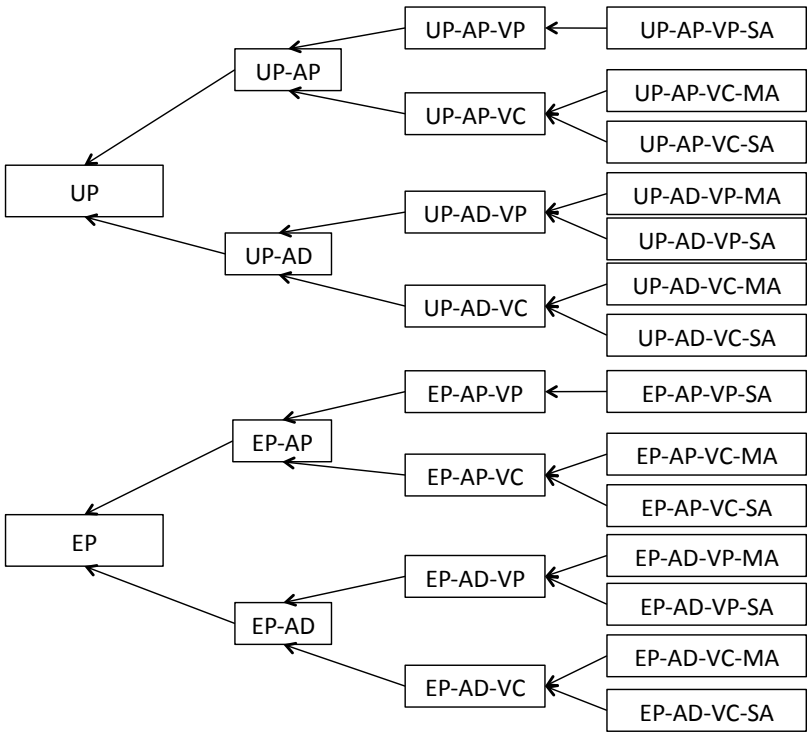

Figure 5.2: Item-to-collection propagation categories and entailments based on the specialization hierarchy.

existential quantification.

Because the framework is generated by conjunctively adding specialization conditions to UP and EP it has the logical structure of a specialization/generalization hierarchy. For example, a rule like the *itemType* rule (formula 5.4), which is in the universal attribute differentiation with value constraint (UP-AD-VC) category, is also in the universal attribute differentiation (UP-AD) category as well as in the universal propagation (UP) category. This means if the framework is being used to classify attributes according to their collection-to-item propagation features, any attributes that fall into to the UP-AD-VC rule will also fall into the UP-AD and UP categories. These implications are shown in Figures 5.1 and 5.2.

In addition, according to the standard semantics for the universal and existential quantifiers any collection-to-item universal rule logically implies the corresponding existential rule by universal instantiation and existential generalization (assuming there are no empty

collections). Roughly, if a collection level attribute implies that every item in a collection has some attribute (UP), then it implies that at least one item in the collection has that attribute (EP). Figure 5.3 shows these relationships for the top three levels of the hierarchy of collection-to-item propagation rules.

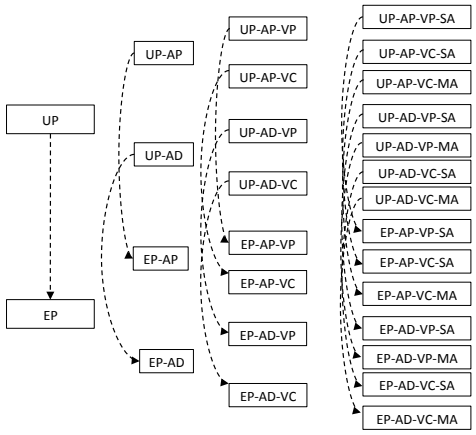

Figure 5.3: Collection-to-item propagation categories with implications based on the relationship between universal and existential propagation

For any universal collection-to-item propagation rule, we can consider the existential version. For example the existential version of a *dateItemsCreated* rule will look like this:

$$\begin{aligned} &\forall y \forall z((dateItemsCreated(y,z)\,\&\,Collection(y)) \supset \\ &\quad \exists x(isGatheredInto(x,y)\,\& \\ &\quad\quad \exists w(created(x,w)\,\&\,temporallyWithin(w,z)))) \end{aligned} \tag{5.31}$$

In any case where the universal version of the rule (formula 5.2) that was discussed above is true, this existential version will also be true.

The logical implication for item-to-collection rules is similar, but operates in the other direction. It can be shown that any existential item-to-collection rule implies the corresponding universal rule, by taking the existential version as a premise and deriving the universal version using standard rules of inference for first order predicate logic. Roughly, if having one item in a collection with some property is enough to infer something about the

collection as a whole, then every item in the collection having that property will certainly be enough to infer the fact about the collection. Figure 5.4 shows these relationships for the top three levels of the hierarchy of item-to-collection propagation rules.

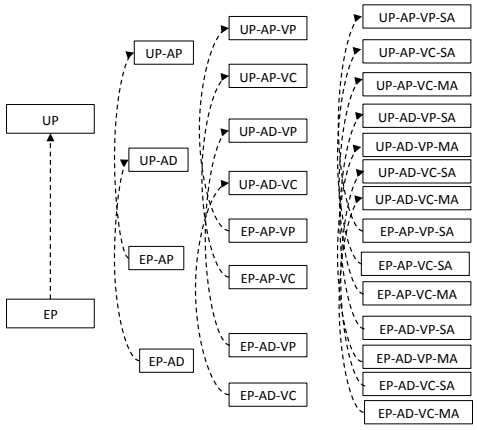

Figure 5.4: Item-to-collection propagation categories with implications based on the relationship between universal and existential propagation

5.6 Conclusion

This chapter presents a complete framework of propagation rules for collection/item metadata relationships. The framework classifies rules according to the following distinctions:

- propagation of information from the collection level to the item level or from the item level to the collection level,
- universal or existential claims at the item level,
- drawing a conclusion on the basis of a single attribute or multiple attributes,
- attribute behavior – whether attributes are the same or different at the item and collection levels, and
- value behavior – whether values for the attributes are the same at the collection and item levels, or related by a constraint.

Rules have been classified according to their logical form into a framework that gives a comprehensive view of how collection-level and item-level description can be systematically related. This system may be directly applied to formalize metadata vocabulary development, enhance retrieval systems, and validate metadata assignments.

Chapter 6

Concluding Remarks

Metadata that describes collections and metadata that describes the items in those collections are related in the sense that we can infer facts about items from descriptions of collections, and facts about collections from descriptions of items. This kind of reasoning is guided by inference rules that are closely connected to our understanding of collection-level and item-level metadata, but these rules are often left implicit and not put to direct use in information organization systems.

Collection descriptions support the use of and access to collections by providing a range of information, such as the purpose, audience, format or type of items, and temporal or geographic coverage of items. The representation of these attributes allows potential users to find collections and to determine both the usefulness of the collection as a whole and the possibility that the collection contains individual items relevant for some purpose. Collection description can also provide an important kind of context that can be used interpret the significance of individual items. When the rules that inform reasoning between collection-level and item-level metadata are represented explicitly and made computationally available for processing by software, they can be used to help improve access to collections and item by supporting the propagation of information between levels of description. These rules can also be used to supplement metadata vocabularies and to aid in the creation, maintenance, and validation of metadata.

In this dissertation I have explored the nature of these rules, developed a logic-based framework of categories for collection/item metadata rules, and provided an analysis of the logical and semantic nature of such rules. In addition, I have taken up foundational questions related to the specification of collection/item metadata relationships, including the

ontological nature of collections, the logic of the collection membership relationship, and the challenges for the translation of metadata records into logic-based knowledge representation languages.

6.1 Summary of Results

The principal results of the preceding chapters can now be summarized.

A number of plausible axioms that might govern the *isGatheredInto* relationship are considered in Chapter 2. For the most part, *isGatheredInto* appears to have the same relation properties as set membership: it is irreflexive, asymmetric, and non-transitive (although arguments in favor of transitivity may be made).

The questions that arise from collection identity and set identity that are discussed in Chapter 2 are more troubling. Since sets are completely determined by set membership, identifying collections as sets will imply that a collection cannot maintain its identity if items are added or removed from the collection, which seems to contradict a commonly held intuition that collections may grow or shrink over time. We might take this as a decisive counterexample against collections being sets, or we might insist that collections really are sets and give an alternative account of what is going on when collections grow or shrink. For example, when a curator states that an item has been added to a collection, this can be understood to mean that the curator has selected a different set (a set that is larger by one member) to serve the curatorial purpose at hand. Although these issues remain unresolved, it does not appear that they present problems for the development of collection/item rule categories.

Facts about collections and items are expressed by metadata records for those entities. The analysis of metadata records presented in Chapter 3 attempts to identify the logical form of the propositional content that is expressed by typical metadata records for items or collections. Three potential formalizations of the record are evaluated with respect to whether they formally entailed the intuitive entailments of the record. It is natural to expect that the logical form of a metadata record is that of a existential claim in which a

resource with certain properties is said to exist. However, there is no formal derivation from this general existential claim to the intuitive entailments, which use the identifier from the record in a way similar to a proper name.

Three approaches to managing this problem are identified: 1) use definite descriptions instead of proper names throughout the formalization process, 2) add a naming clause to the existential claim, and 3) use the identifier directly as a name in binary predications. Although this final option aligns with the logical form of the Resource Description Framework (RDF) and was used in a previous project for exploring rules for collection/item metadata relationships, it relies on treating identifier elements in metadata records as simultaneously describing a property of a resource and giving a name for the resource. This shows that while identifier attributes in metadata records have the syntactic appearance that is common across descriptive attributes, they have the potential to play a distinctive semantic role that involves both the use and mention of individual constants. Again, while challenging issues remain, no obstacles to the development of a collection/item metadata rule framework are apparent.

Inferencing rules for collection/item metadata relationships are identified in Chapter 4 as propagation rules where the grounding relationship is *isGatheredInto*. Although the logical conditionals used to express propagation rules can be troubled by "paradoxes" of material and strict implication, the use of "preemptive modal exclusion" can block problematic consequences. Collection/item propagation rules can be re-written as Horn clauses, and so are computationally tractable. However, the occurrence of variables representing "unknown individuals" in rules means that these collection/item inferencing rules cannot be formulated in OWL.

The new generalized framework is presented in Chapter 5. It accommodates item-to-collection inferencing as well as collection-to-item inferencing, and includes inferencing based on multiple attributes as well as inferencing based on single attributes. The framework consists of 28 categories related by two kinds of logical relationships. This system of rule categories may be directly applied to formalize metadata vocabulary development, enhance

retrieval systems, and validate metadata assignments.

The approach to illuminating the underlying semantics of metadata that has been used in this dissertation can be seen as a variation on axiomatic formal methods. When certain rules are endorsed as normative for inferences between items and collections, we come to a better understanding of both the subtleties of relevant metadata attributes and what it means to be a collection. The rules concerning description that can be built via the rule framework act as axioms that contribute to analysis of collections and of descriptions, and contribute to a developing conceptual scheme for these entities within information organization.

6.2 Next Steps

One exciting area for further work based on the research in this dissertation is in the development of practical techniques for articulating rules and connecting them to metadata schemas. As discussed in Chapter 1, the rule framework presented here can be used to annotate the definition of attributes in metadata schemas or application profiles with rules that express the expected relationships between properties at the collection and item levels. In order to make this kind of supplementation a reality, it will be necessary to develop practical guidelines and tools for metadata developers and repository managers. For example, a tool that allows a repository manager to import a schema and view potential rules generated from the framework would support annotation with rules without requiring training in the techniques used to develop the framework in this dissertation.

The framework itself was developed at a high level of abstraction, which means that the rule categories are very general and not dependent on the practices of particular communities or domains. However, any implementation of rules in metadata vocabularies or to assist in search or navigation should be closely connected to the practices and expectations of particular communities, which will inform the relationships between levels of description. Therefore the development of practical guidelines for using propagation rules will require collaboration and engagement with the communities likely to make use of rules.

Another area for future work on metadata relationships is the exploration of different kinds of rules. As argued in Chapter 4, the propagation rules explored here correspond to Horn clauses. These rules that follow the basic form of "if each of these things is true, then some fact is true," which means that the antecedents of rules are all conjunctions of positive assertions and the consequent is a positive assertion. These were the rules that seemed most likely be useful in current information system, but rules that include disjunctions or negations may also be useful in some scenarios. Rules with disjunctive antecedents could allow the propagation of information on the basis of one or more possible conditions being met, and rules with negations in the consequent could support ruling out certain search results, given particular facts about a search or user.

In addition, although the rules developed in this dissertation all express relationships between collection-level and item-level properties, it would also be possible to develop rules that express relationships at a single level of description. These rules might be used to link types of coverage attributes to a single subject attribute, or to constrain creation dates of resources to the time period during which the creator was known to be alive. As is the case with rules that express collection/item metadata relationships, these rules would provide a formal explicit expression of the background assumptions that are implicit in many descriptive systems.

Several substantial conceptual questions that have been considered in this dissertation remain open and invite continued investigation. As mentioned above, the analysis of colloquial metadata records presented in Chapter 3 showed that while identifier attributes in records have the syntactic appearance of other descriptive attributes, they have the potential to act as simultaneously describing a resource and giving the resource a name within the context of an information system. Further analysis of this topic could connect the identification practices in information systems with theories of naming and identity from the philosophy of language. Open questions about naming and context have also appeared in research on using semantic technologies to link descriptions on the web, which suggests that empirical investigations of how tool builders in these areas go about assigning identifiers

may be worthwhile. This kind of investigation could shed light on whether identifiers from metadata records are used like proper names in systems that bring together RDF descriptions, and what the impact of such an approach is on the usefulness of those systems.

Although the analysis of the *isGatheredInto* relationships presented in Chapter 2 provided valuable insights into this critical relationship for collection/item metadata relationships, there are several areas for further analysis. One point of interest is on the question of transitivity for collection membership. The potential transitivity of collection membership could be explored in large aggregations that have collections and sub-collections. By implementing transitivity of collection membership in such a repository, researchers could develop a clear picture of the benefits and drawbacks of this interpretation of *isGatheredInto*. In addition, although the analysis in this dissertation was intended to be general for any type of information system, it was largely grounded in library practice. Further research could investigate whether there are distinctive practices from archival collections, natural history collections, collections of scientific datasets, or other areas that imply meaningful differences in the expected semantics of a collection membership relationship.

Finally, there is room for additional analysis on the ontological nature of collections themselves. As argued in Chapter 2 discussed above, identifying collections as mathematical sets presents a problem if we want to be able to add or remove items from collections. Developing an ontological account of collections that fully supports curatorial practices like updating collection membership may be possible by introducing a temporal element to the identity of collections. Further investigation of collection practices from archival and museum communities may also provide a more comprehensive account and suggest other directions for a general ontological account of collections.

These and many other challenging problems await us. It is no great surprise that a concept as fundamental as *collection* provides so many challenges for analysis. The work presented in this dissertation has set the stage for continued research on collections and the semantics of metadata records and vocabularies.

References

Atkinson, R. (1998). Managing traditional materials in an online environment: Some definitions and distinctions for a future collection management. *Library Resources & Technical Services*, 42(1):7–20.

Bowen, J. (2010). Moving library metadata toward linked data: opportunities provided by the eXtensible Catalog. In *Proceedings of the 2010 International Conference on Dublin Core and Metadata Applications*, pages 44–59. Dublin Core Metadata Initiative.

Brachman, R. and Levesque, H. (2004). *Knowledge Representation and Reasoning*. Morgan Kaufmann Pub.

Brachman, R., McGuinness, D., Patel-Schneider, P., Resnick, L., Resnick, L., and Borgida, A. (1991). Living with CLASSIC: When and how to use a KL-ONE-like language. In *Principles of Semantic Networks*.

Brack, E., Palmer, D., and Robinson, B. (2000). Collection level description – the RIDING and Agora experience. *D-lib Magazine*, 6(9).

Brockman, W., Neumann, L., Palmer, C., and Tidline, T. (2001). *Scholarly work in the humanities and the evolving information environment*. Digital Library Federation.

Buckland, M. (1995). What will collection developers do? *Information Technology and Libraries*, 14(3):155–159.

Casserly, M. (2002). Developing a concept of collection for the digital age. *portal: Libraries and the Academy*, 2(4):577–587.

Christenson, H. and Tennant, R. (2005). *Integrating information resources: Principles, technologies, and approaches*. California Digital Library. Oakland, CA.

Covi, L. M. and Cragin, M. H. (2004). Reconfiguring control in library collection development: A conceptual framework for assessing the shift toward electronic collections. *Journal of the American Society for Information Science and Technology*, 55(4):312–325.

Currall, J., Moss, M., and Stuart, S. (2004). What is a collection? *Archivaria*, 58:131–146.

Doerr, M. (2003). The CIDOC conceptual reference module: an ontological approach to semantic interoperability of metadata. *AI Magazine*, 24(3):75.

Dubin, D., Renear, A., Sperberg-McQueen, C., and Huitfeldt, C. (2003). A logic programming environment for document semantics and inference. *Literary and Linguistic Computing*, 18(1):39–47.

Dubin, D., Wickett, K., and Sacchi, S. (2011). Content, format, and interpretation. In *Proceedings of Balisage: The Markup Conference*, volume 7.

Dublin Core Collection Description Task Group (2007). *Dublin Core Collections Application Profile*. Dublin Core Metadata Initiative.

Foulonneau, M., Cole, T., Habing, T. G., and Shreeves, S. L. (2005). Using collection descriptions to enhance an aggregation of harvested item-level metadata. In *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 32–41. ACM Press.

Fraenkel, A. and Bar-Hillel, Y. (1958). *Foundations of Set Theory*. North-Holland Publication Company.

Galton, A. (2010). How is a collection related to its members? In *Proceedings of the Third Interdisciplinary Ontology Meetings*, volume 3, pages 9–17.

Golbreich, C. and Wallace, E. (2009). OWL 2 Web Ontology Language: new features and rationale. W3C recommendation.

Gonçalves, M., Fox, E., and Watson, L. (2008). Towards a digital library theory: a formal digital library ontology. *International Journal on Digital Libraries*, 8:91–114.

Gonçalves, M. A., Fox, E. A., Watson, L. T., and Kipp, N. A. (2004). Streams, structures, spaces, scenarios, societies (5S): A formal model for digital libraries. *ACM Transactions on Information Systems*, 22:270–312.

Grau, B., Horrocks, I., Motik, B., Parsia, B., Patel-Schneider, P., and Sattler, U. (2008). OWL 2: The next step for OWL. *Web Semantics: Science, Services and Agents on the World Wide Web*, 6(4):309–322.

Greenberg, J. (2009). Theoretical considerations of lifecycle modeling: an analysis of the Dryad repository demonstrating automatic metadata propagation, inheritance, and value system adoption. *Cataloging & Classification Quarterly*, 47(3):380–402.

Halpin, H. (2011). Sense and reference on the web. *Minds and Machines*, pages 1–26.

Haslhofer, B. and Isaac, A. (2011). data.europeana.eu - The Europeana Linked Open Data Pilot. In *Proceedings of the 2011 International Conference on Dublin Core and Metadata Applications*, pages 94–104. Dublin Core Metadata Initiative.

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

Heaney, M. (2000). An analytic model of collections and their catalogues. *UK Office for Library and Information Science*.

Hill, L., Janee, G., Dolin, R., Frew, J., and Larsgaard, M. (1999). Collection metadata solutions for digital library applications. *Journal of the American Society for Information Science*, 50(13):1169–1181.

Hillmann, D. (2003). *Using Dublin Core*. Dublin Core Metadata Initiative.

Horrocks, I., Patel-Schneider, P., Boley, H., Tabet, S., Grosof, B., Dean, M., et al. (2004). SWRL: A semantic web rule language combining OWL and RuleML. *W3C Member submission*, 21:79.

Horrocks, I., Patel-Schneider, P., and 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.

Horrocks, I., Patel-Schneider, P. F., Bechhofer, S., and Tsarkov, D. (2005). OWL rules: A proposal and prototype implementation. *Web Semantics: Science, Services and Agents on the World Wide Web*, 3(1):23 – 40.

IFLA (2009). Functional requirements for bibliographic records: Final report. Technical report, International Federation of Library Associations and Institutions.

Kaczmarek, J. (2006). The complexities of digital resources: Collection boundaries and management responsibilities. *Journal of Archival Organization*, 4(1):215–227.

Klyne, G. and Carroll, J. (2004). Resource description framework (RDF): concepts and abstract syntax.

Lagoze, C. and Fielding, D. (1998). Defining collections in distributed digital libraries. *D-Lib magazine*, 4(11).

Lagoze, C., Krafft, D., Cornwell, T., Dushay, N., Eckstrom, D., and Saylor, J. (2006). Metadata aggregation and automated digital libraries: A retrospective on the nsdl experience. In *Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital libraries*, pages 230–239. ACM.

Lalmas, M. (1998). Logical models in information retrieval: Introduction and overview. *Information Processing & Management*, 34(1):19–33.

Le Boeuf, P., Doerr, M., Ore, C. E., and Stead, S. (2012). *Definition of the CIDOC Conceptual Reference Model*. ICOM/CIDOC.

Lee, H. (2005). The concept of collection from the user's perspective. *The Library Quarterly*, 75(1):pp. 67–85.

Lee, H.-L. (2000). What is a collection? *Journal of the American Society for Information Science*, 51(12):1106–1113.

Lewis, C. (1918). *A survey of symbolic logic*. University of California Press.

Liu, J. (2007). *Metadata and its applications in the digital library*. Libraries Unlimited.

Lourdi, I., Papatheodorou, C., and Doerr, M. (2009). Semantic integration of collection description. *D-Lib Magazine*, 15(7/8):1082–9873.

Lynch, C. (2002). Digital collections, digital libraries and the digitization of cultural heritage information. *First Monday*, 7(5-6).

Marcoux, Y., Sperberg-McQueen, C., and Huitfeldt, C. (2009). Formal and informal meaning from documents through skeleton sentences: Complementing the formal tag-set documentation with intertextual semantics and vice-versa. In *Proceedings of Balisage: The Markup Conference*, volume 3.

Meghini, C. and Spyratos, N. (2010). Unifying the concept of collection in digital libraries. *Advances in Intelligent Information Systems*, pages 197–224.

Meghini, C., Spyratos, N., and Sugibuchi, T. (2010). Modelling digital libraries based on logic. In *Proceedings of the 14th European Conference on Research and Advanced Technology for Digital Libraries*, pages 2–13. Springer-Verlag.

Miller, S. J. (2010). The One-to-One principle: Challenges in current practice. In *Proceedings of the International Conference on Dublin Core and Metadata Applications*.

M'kadem, A. and Nieuwenhuysen, P. (2010). Digital access to cultural heritage material: case of the moroccan manuscripts. *Collection Building*, 29(4):137–141.

Moss, M. and Currall, J. (2004). Digitisation: Taking stock 1. *Journal of the Society of Archivists*, 25(2):123–137.

Niles, I. and Pease, A. (2001). Towards a standard upper ontology. In *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems*.

Palmer, C. (2004). Thematic research collections. In *A companion to digital humanities*. Blackwell, Oxford.

Palmer, C. L., Knutson, E., Twidale, M., and Zavalina, O. (2006). Collection definition in federated digital resource development. In *Proceedings of the 69th ASIS&T Annual Meeting (Austin, TX)*.

Powell, A., Heaney, M., and Dempsey, L. (2000). RSLP collection description. *D-lib Magazine*, 6(9):1082–9873.

Powell, A., Nilsson, M., Naeve, A., Johnston, P., and Baker, T. (2007). *DCMI Abstract Model*. Dublin Core Metadata Initiative.

Renear, A., Dubin, D., and Sperberg-McQueen, C. (2002). Towards a semantics for XML markup. In *Proceedings of the 2002 ACM symposium on Document engineering*, pages 119–126.

Renear, A., Wickett, K., Urban, R., and Dubin, D. (2008a). The return of the trivial: Problems formalizing collection-level/item-level metadata relationships. In *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries*.

Renear, A. H., Wickett, K. M., Urban, R. J., Dubin, D., and Shreeves, S. (2008b). Collection/Item metadata relationships. In *Proceedings of the International Conference on Dublin Core and Metadata Applications*.

Rescher, N. (2007). *Conditionals*. MIT Press, Cambridge, MA.

Robinson, J. A. (1965). A machine-oriented logic based on the resolution principle. *Journal of the ACM*, 12(1):23–41.

Russell, B. (1903). *The principles of mathematics*. Cambridge University Press.

Russell, B. (1905). On denoting. *Mind*, 14(4):479–493.

Sacchi, S., Wickett, K. M., Renear, A. H., and Dubin, D. (2011). A framework for applying the concept of significant properties to datasets. In *Proceedings of the 74th ASIS&T Annual Meeting (Pittsburgh, PA)*.

Sebastiani, F. (1998). On the role of logic in information retrieval. *Information Processing & Management*, 34(1):1–18.

Seidenberg, J. and Rector, A. (2006). Representing transitive propagation in OWL. *Conceptual Modeling-ER 2006*, pages 255–266.

Shreeves, S. and Cole, T. (2003). Developing a collection registry for IMLS NLG digital collections. In *Proceedings of the International DCMI Metadata Conference and Workshop*.

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

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

Sweet, M. and Thomas, D. (2000). Archives described at collection level. *D-Lib Magazine*, 6(9).

Urban, R. J. (2012). *Principle Paradigms: Revisiting the Dublin Core 1:1 Principle*. PhD dissertation, University of Illinois at Urbana-Champaign.

Van Rijsbergen, C. (1986). A non-classical logic for information retrieval. *The Computer Journal*, 29(6):481.

Warner, S., Bekaert, J., Lagoze, C., Liu, X., Payette, S., and Van de Warner, H. (2007). Pathways: Augmenting interoperability across scholarly repositories. *International Journal on Digital Libraries*, 7(1):35–52.

Wendler, R. (2004). The eye of the beholder: Challenges of image description and access at harvard. In Hillmann, D. and Westbrooks, E., editors, *Metadata in Practice*, pages 51–6. American Library Association, Chicago, IL.

Wickett, K. (2010). Discourse situations and markup interoperability. In *Proceedings of Balisage: The Markup Conference*, volume 5.

Wickett, K., Urban, R., Zheng, W., and Renear, A. (2009). A testbed approach for metadata inference rule development. In *Workshop on Integrating Digital Library Content with Computational Tools and Services. ACM/IEEE Joint Conference on Digital Libraries (JCDL).*

Wickett, K. M. (2009). Logical expressiveness of semantic web languages for bibliographic modeling. In *Proceedings of the iConference (Chapel Hill, NC).*

Wickett, K. M. (2011). Expressiveness requirements for reasoning about collection/item metadata relationships. In *Proceedings of the iConference.*

Wickett, K. M., Renear, A. H., and Urban, R. J. (2010). Rule categories for collection/item metadata relationships. In *Proceedings of the 73rd ASIS&T Annual Meeting (Pittsburgh, PA).*

Yeo, G. (2012). The conceptual fonds and the physical collection. *Archivaria*, 73.

Zeng, M. and Qin, J. (2008). *Metadata.* Neal-Schuman Publishers, New York.