# DATA RELEVANCE: CONCEPTUALIZATION AND THEORIZATION FROM THE PERSPECTIVE OF KNOWLEDGE ORGANIZATION

by

Li Yang

A Dissertation Submitted in

Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

in Information Studies

at

The University of Wisconsin-Milwaukee

August 2025

ABSTRACT

DATA RELEVANCE: CONCEPTUALIZATION AND THEORIZATION FROM THE PERSPECTIVE OF KNOWLEDGE ORGANIZATION

by

Li Yang

The University of Wisconsin-Milwaukee, 2025
Under the Supervision of Professor Margaret E.I. Kipp

This study explores relevance from a data perspective in the context of knowledge organization. Via conceptualizing relevance in knowledge organization, examining relationships in knowledge organization, and theorizing data relevance, this study breaks the dominance of relevance studies from an information retrieval perspective, fills the theoretical gap of data relevance, and supplements the relevance functioning factors.

Nine definitions of relevance in knowledge organization were proposed based on Saracevic's relevance definition framework after examining knowledge organization activities, including subject analysis, indexing, cataloging, and social tagging. These definitions provided a deconstructive understanding of relevance in knowledge organization and laid the groundwork for relationship studies between data, and relationship studies between data and information professionals.

Relationships in knowledge organization were examined based on the meta-typology of relationships, including relationships in the bibliographic universe, the information universe, and the data universe. Relationships in the bibliographic and information universes showed a trend toward transitioning to data relationships. In the data universe, the concept of relationships has been generalized due to the RDF model; thus, the relationships have been

expanded largely with the support of semantic technologies. Relationships explain how data are related and provide a foundation for theorizing data relevance.

The data relevance model was created by exploring how information professionals considered data relevance. The user study included a two-stage task and an interview. The grounded theory based data analysis showed 28 functioning factors of data relevance in three levels. The six top-level factors were features of resource, purpose, users, context, potential use, and prior knowledge. The second-level factors showed more data features, such as enabling linkages, scope of terms, granularity, using single/as few schemas, etc.

Data relevance theory was developed in the end based on the relevance definitions in knowledge organization, the relationships, and the data relevance model.

© Copyright by Li Yang, 2025
All Rights Reserved

To

my parents,

my husband,

and especially my daughter

TABLE OF CONTENTS

LIST OF FIGURES ..... x

LIST OF TABLES ..... xi

LIST OF ABBREVIATIONS ..... xii

LIST OF APPENDICES ..... xiii

ACKNOWLEDGEMENTS ..... xiv

1 Introduction ..... 1

1.1 Background ..... 1

1.1.1 Conceptualization of Relevance ..... 1

1.1.2 Knowledge Organization Context for Relevance Study ..... 3

1.1.3 From Information Relevance to Data Relevance ..... 5

1.1.4 Linked Data and Data Relevance Study ..... 7

1.2 Research Questions ..... 9

1.3 Significance and Contribution ..... 10

1.3.1 For Relevance Research ..... 10

1.3.2 For IR Research ..... 10

1.3.3 For KO Research ..... 11

1.3.4 For LOD Research ..... 12

1.4 Outline of the Study ..... 12

2 Literature Review ..... 15

2.1 Conceptualization of Relevance in Information Science ..... 15

2.1.1 Different Views of Relevance ..... 15

2.1.2 Frameworks to Understand Relevance ..... 25

2.1.3 Paradigms Used in Relevance Research ..... 29

2.1.4 Relevance Assessment ..... 35

2.2 Relationships in Relevance Studies ..... 37
2.3 Relevance Study Beyond IR ..... 40
2.4 Evidence of Data Relevance ..... 42
3 Methodology ..... 45
3.1 Research Paradigm ..... 45
3.2 Research Design ..... 47
3.2.1 Research Method ..... 47
3.2.2 Task Design ..... 49
3.2.3 Interview Design ..... 52
3.2.4 Sampling Method ..... 53
3.3 Participant Recruitment ..... 56
3.4 Data Collection ..... 56
3.5 Data Analysis ..... 58
3.6 Evaluation ..... 59
3.6.1 Reliability ..... 59
3.6.2 Validity ..... 60
4 Relevance in Knowledge Organization ..... 62
4.1 A Conceptual Perspective on Relevance as a Foundation ..... 62
4.2 Subject Analysis ..... 65
4.3 Aboutness ..... 69
4.4 Indexing ..... 76
4.5 Cataloging ..... 80
4.6 Social Tagging ..... 88
4.7 Warrant ..... 96
5 Relationships in Knowledge Organization ..... 103
5.1 Meta-typology of Relationships ..... 104
5.2 Relationships in Bibliographic Universe ..... 106

5.2.1 Bibliographic Relationships ..... 106
5.2.2 Models and Relationships ..... 107
5.3 Relationships in Information Universe ..... 114
5.3.1 Hyperlinks ..... 114
5.3.2 Social Tags and Folksonomy ..... 118
5.4 Relationships in Data Universe ..... 123
5.4.1 Relationships in Metamodel and Encoding Standards ..... 125
5.4.2 Relationships in Value Vocabularies ..... 132
5.4.3 Metadata Schema, Relationships, and Relevance ..... 142
5.5 Relevance as a Relationship ..... 147
6 User Study and Data Relevance Model ..... 149
6.1 Participant Information ..... 150
6.2 Responses in the First Stage ..... 151
6.3 Responses in the Second Stage ..... 153
6.3.1 General Information and Coding ..... 153
6.3.2 Categories in Axial Coding ..... 159
6.3.3 Discussion about the Categories ..... 168
6.4 Interview ..... 169
6.4.1 General Information and Coding ..... 169
6.4.2 Overlapping Categories ..... 174
6.4.3 New Categories ..... 186
6.4.4 Common Questions ..... 194
6.5 Supplement by Discriminant Views ..... 204
6.6 Selective Coding and Modeling ..... 205
7 Data Relevance, Discussion and Future Work ..... 215
7.1 Summary of the study ..... 215
7.2 Highlights of the Findings ..... 217

7.3 Data Relevance Theory and Use Cases ..... 222
7.4 Discussion ..... 224
7.5 Limitations of the Study ..... 227
7.6 Future Work ..... 228
References ..... 231
Appendix A: IRB Approval ..... 253
Appendix B: Informed Consent ..... 254
Appendix D: Structured Information Sample ..... 258
Appendix E: Comparison Table for Stage 2 ..... 260
Appendix F: Interview Questions ..... 261

LIST OF FIGURES


| Figure # | Figure title | Page # |
|---|---|---|
| Figure 1 | Meta-typological model of relationships in knowledge organization | 106 |
| Figure 2 | Relationships in the FRBR model | 108 |
| Figure 3 | Agent relationships in the FRAD model | 109 |
| Figure 4 | Relationships of entities in FRSAD | 110 |
| Figure 5 | Relationships of entities in LRM | 112 |
| Figure 6 | Bibliographic relationships of classes in BIBFRAME | 113 |
| Figure 7 | The framework of KOS in the Semantic Web | 125 |
| Figure 8 | An RDF triple | 126 |
| Figure 9 | A hierarchy of semantic relationships defined in SKOS | 130 |
| Figure 10 | Snippet 1 of in vivo open coding for the second-stage responses | 155 |
| Figure 11 | Snippet 2 of in vivo open coding for the second-stage responses | 156 |
| Figure 12 | Statistics of the in vivo codes for the second-stage responses | 156 |
| Figure 13 | Snippet of axial coding with collapsed in vivo codes for the second-stage responses | 157 |
| Figure 14 | Axial codes for the second stage responses | 158 |
| Figure 15 | Snippet 1 of in vivo open coding for the interview scripts | 170 |
| Figure 16 | Snippet 2 of in vivo open coding for the interview scripts | 171 |
| Figure 17 | Snippet 3 of in vivo open coding for the interview scripts | 171 |
| Figure 18 | Statistics of the in vivo codes for the interview transcripts | 172 |
| Figure 19 | Snippet of axial coding with collapsed in vivo codes for the interview transcripts | 172 |
| Figure 20 | Snippet of axial codes for the interview transcripts | 173 |
| Figure 21 | Screenshot of the Obsidian interface | 208 |
| Figure 22 | Category connections diagram | 209 |
| Figure 23 | Data relevance model in the KO context | 213 |
| Figure 24 | Data relevance factors in the objective-subjective view | 225 |

LIST OF TABLES


| Table # | Table title | Page # |
|---|---|---|
| Table 1 | Different manifestations of relevance in Saracevic's studies | 26 |
| Table 2 | Summary of the main views of relevance in a dichotomous manner | 30 |
| Table 3 | Hjørland's simplified relevance criteria in four psychological paradigms | 34 |
| Table 4 | Hjørland's simplified relevance criteria in four epistemological schools | 34 |
| Table 5 | Research questions and the corresponding data types | 56 |
| Table 6 | Statistics of the duration of data collection at each phase | 58 |
| Table 7 | User tasks as defined in ICP, FRBR, FRAD, and FRSAD | 86 |
| Table 8 | Relevance definitions in KO based on Saracevic's formula | 101 |
| Table 9 | Relationship scopes in KO | 105 |
| Table 10 | Baron's taxonomy of link types | 115 |
| Table 11 | Kopak's functional link type taxonomy | 116 |
| Table 12 | Cantador et al.'s Purpose-oriented social tag categories | 122 |
| Table 13 | First three levels of Park's tag typology in health social media | 122 |
| Table 14 | RDF properties in RDF Schema 1.1 | 127 |
| Table 15 | SKOS properties and domains | 129 |
| Table 16 | Value vocabulary relationships and their types in the conventional view | 135 |
| Table 17 | Properties in UNESCO Thesaurus | 136 |
| Table 18 | Snippet of properties in Getty vocabularies | 138 |
| Table 19 | Properties in AGROVOC Thesaurus | 140 |
| Table 20 | Metadata fields and the corresponding relevance criteria | 146 |
| Table 21 | Background information of the participants | 151 |
| Table 22 | Categories created in selective coding | 206 |
| Table 23 | Categories and their associated categories | 206 |
| Table 24 | Statistics of category connections | 210 |
| Table 25 | Category hierarchy in axial coding | 212 |

LIST OF ABBREVIATIONS

| | |
|---|---|
| AGROVOC | Used for: Food and Agriculture Organization of the United Nations Vocabulary |
| BIBFRAME | Bibliographic Framework |
| DC | Dublin Core |
| DIKW | Data, Information, Knowledge, and Wisdom |
| FRAD | Functional Requirements for Authority Data |
| FRBR | Functional Requirements for Bibliographic Records |
| FRSAD | Functional Requirements for Subject Authority Data |
| IFLA ICP | IFA Statement of International Cataloguing Principles |
| IR | Information Retrieval |
| IRI | Internationalized Resource Identifier |
| KO | Knowledge Organization |
| KOS | Knowledge Organization Systems |
| LOD | Linked Open Data |
| LOD KOS | Used for: Linked Open Value Vocabularies |
| LOV | Linked Open Vocabularies |
| LRM | Library Reference Model |
| OWL | Web Ontology Language |
| RDF | Resource Description Framework |
| RDFS | RDF Schema |
| SKOS | Simple Knowledge Organization System |
| UNESCO | United Nations Educational, Scientific and Cultural Organization |

LIST OF APPENDICES


| Appendix # | Appendix title | Page # |
|---|---|---|
| Appendix A | IRB Approval | 253 |
| Appendix B | Informed Consent | 254 |
| Appendix C | Pre-questionnaire | 257 |
| Appendix D | Structured information sample | 258 |
| Appendix E | Comparison table for Stage 2 | 260 |
| Appendix F | Interview questions | 261 |

ACKNOWLEDGEMENTS

I would like to express my sincere appreciation to my supervisor and chair of my committee, Dr. Margaret E.I. Kipp, for her invaluable guidance, unwavering support, and constructive feedback. She provided me with more than academic guidance; her care extended to various aspects of life. She is like a sister to me, always giving me strength. Her signature saying “it never hurts to ask” gave me great courage at the moments I was about to give up or unsure how to proceed. I would also like to thank the members of my dissertation committee, Dr. Dietmar Wolfram, who has helped me in significant ways during my PhD journey, Dr. Cristina Pattuelli, Dr. Min Sook Park, and Dr. Wan-Chen Lee. Their thoughtful insight, substantive suggestions, and encouragement have been essential to my academic growth.

My heartfelt thanks go to the participants of this study, whose time, knowledge, and willingness to share their experiences made this research meaningful. Without their contributions, this work would not have been possible.

I would like to express my deepest gratitude to my parents for their unconditional love and support. The loss of my mother during my doctoral studies was a tremendous sorrow, and it remains one of my greatest regrets that she could not see me complete this journey. Her strength and love continue to inspire me every day.

Last but not least, I would like to thank my husband, Siran Feng. His unwavering support and belief in me have made my PhD journey possible. I could not have done it without him. Thanks to my daughter Reese Feng. Her company fills my days with joy and reminds me what matters most. Sharing this journey and growing together with her makes everything more meaningful.

# 1 Introduction

## 1.1 Background

### 1.1.1 Conceptualization of Relevance

Relevance is a term that can be easily understood but is difficult to define formally. It is normally used intuitively in all walks of life. As a universal concept, relevance has attracted researchers in the areas of logic, philosophy, communication, information science, and computer science (Saracevic, 2007). Information technologies have affected the way we interact with information. During the interaction process, relevance plays a significant and underlying role in acquiring, organizing, retrieving, and using information (Saracevic, 2016). Information science naturally pays great attention to understanding, interpreting, and applying relevance.

Since the initial study of relevance in information science in the 1950s, intellectual expansion on relevance and the contention of a broad range of thoughts and ideas has never stopped. Many mechanisms, perspectives, and factors affect elaborating the notion of relevance. Relevance is thus a generic and vague concept that can be comprehended and interpreted through many different paradigms, theoretical foundations, and views. Many terms have emerged to explain what relevance is, what dimensions and attributes it has, and by what criteria to judge it, including but not limited to logical relevance, situational relevance, system relevance, topical relevance, cognitive relevance, user-based relevance, subjective relevance, objective relevance, psychological relevance, etc. (Cooper, 1971; T. K. Park, 1994; Saracevic, 1975; Swanson, 1986; Wilson, 1973).

Generally, there are two schools of thought on defining relevance. The first one is the widely recognized definition that relevance is a relation (Mizzaro, 1997). For example, two usages of relevance were to indicate "the relationship of a term or a document to a field of interest" and "a relationship between some system output….and some kind of information requirement" (Cuadra & Katter, 1967, p. 292). In the Cranfield II relevance assessment experiment, relevance was defined as relation between every document and every question (Harter, 1971). Logical relevance (Cooper, 1971) was a relation between a minimal premise set of stored sentences and a component statement of information need. Situational relevance (Wilson, 1973) is a relationship between a piece of information in a personal information base and questions of interest to the person. Relevance is related to the communication effectiveness between a source and a destination (Saracevic, 1975). Relevance depends on the users' judgments of the relationships between information and information need (Schamber et al., 1990). Relevance is a relation between an assumption and a context, whether the assumption is deduced, erased, strengthened or weakened in the context (Harter, 1992; Sperber & Wilson, 1995). This school of thought agrees that relevance represents relationships; the differences between various views are the different relationships and the subjects/objects of the relationships (Cosijn & Ingwersen, 2000; Saracevic, 1996).

The second school of thought refers to the utility of information. Wilson (1968) defined utility as one type of question relevance. His later views such as evidential relevance and situational relevance were partly in line with utility (Wilson, 1973). Green considered relevance to be a property of information that is helpful to meet a user's information need, and they studied relevance relationships to support the ideas (Green, 1995; Green & Bean, 1995).

Borlund (2003) considered subject relevance, with subcategories interpreted as relationships, as the utility of information to achieve users' goals, interests, tasks, or problems. Relevant information increases the possibility of achieving the goal implied by the task, in agreement with situational relevance formulated as relationships (Hjørland, 2010). Though the second thought focuses on the functional aspect, relationships were usually incorporated as a key element to study relevance.

The two schools of thought on defining relevance imply that the nature of relevance is relationships. Both schools of thought suggest that relationship studies are essential and a promising perspective to study relevance and its theories.

### 1.1.2 Knowledge Organization Context for Relevance Study

Traditionally, most relevance studies have occurred in the field of information retrieval (IR) including the information behavior area. Although a prolific research topic, we can barely see relevance studies in other subfields of information science. This is a common orientation in information scientists (from the published research) and practitioners (from my interview that is discussed in the later section). How metadata librarians and catalogers handle relevance is less important than how indexers do; How indexers do subject analysis is less important than how they assign terms to facilitate retrieval (Hutchins, 1977; Mai, 2005). We pay more attention to the front-end relevance because it matters directly whether the end users can find the information they want. However, it is necessary to understand relevance from a more foundational perspective to complement the established epistemological models.

The knowledge organization (KO) field lays the foundation for relevance study. KO researchers have been naturally interested in relationship studies. Bibliographic relationships

have been studied for years. Tillett (1987) created a taxonomy of bibliographic relationships with seven top-level categories, supplemented by Smiraglia's (1992) taxonomy of the derivative relationships. Velucci (1997) identified bibliographic relationships between music resources. Knowledge organization systems (KOS), for resource description, management, and information retrieval, rely on semantic relationships to enable concept associations and interoperations between resources (Hodge, 2000).

Relationships are the intellectual foundation for organizing information and knowledge (Svenonius, 2000). Inquiring about the relationships between resources is an essential step to organizing them. Research on thesauri, subject headings, classification systems, taxonomies, and ontologies shows various degrees of complexity and rich relationships (Glushko, 2013). For example, a consensus has been reached that the equivalence, hierarchical and associative relationships are the basic semantic relationships in KOS such as thesauri and subject headings. Besides the basic types, such as equivalence, hierarchy, and associative, there are numerous relationships that are specific to certain areas. For example, the Unified Medical Language System (UMLS) Semantic Network consists of 54 semantic relationships (*Fact Sheet UMLS Semantic Network*, n.d.). The Relation Ontology (RO) of the Open Biological and Biomedical Ontology (OBO) has 594 semantic relations intended for use across biological ontologies in the OBO Foundry and related libraries (*OBO Relation Ontology*, n.d.).

KOS serves "as a bridge between the user's information need and the material in the collection" (Hodge, 2000, p. 3). On one hand, with rich relationships, KOS can be used to link digital resources and guide the users to identify and discover information objects without awareness and prior knowledge of them (Hodge, 2000); On the other hand, KOS provides terms

to describe the resources and act as intermediate files to represent a user's information need (Hodge, 2000; Svenonius, 2000). From the relation perspective, relevance is a relation between a user's information need and the information object, and the 'relation' is not only an abstract notion but could also be interpreted concretely. Therefore, it is reasonable to study relevance by studying KOS and in turn, study KOS from the perspective of relevance.

Additionally, KO has to do with relevance study naturally not only in leveraging KOS to represent the subject of information and potential information need but also in affecting relevance judgment through the nature and structure of domain knowledge (Bean & Green, 2001). Saracevic (2016) stated that "topical relevance may not necessarily involve an IR system at all" (p. 22), which can be seen as a justification for the rationality of studying relevance beyond the field of information retrieval.

### 1.1.3 From Information Relevance to Data Relevance

The science paradigm has evolved from simulation (computational) studies a few decades ago to data-intensive studies (Hey, Tansley, & Tolle, 2009). Along with the evolution, the context of studying relevance has been shifting from information to data.

In information science, the common understanding of information is based on data. The DIKW (Data, Information, Knowledge, and Wisdom) pyramid (widely credited to Ackoff (1989)) represents the hierarchical relationship, where data are the foundation of information, knowledge, and wisdom. Scholars understand information in terms of data, e.g., information is derived from data assisted by computers to answer who, what, where, when, and how many (Ackoff, 1989), information is well-organized data by certain syntax (rules) to generate semantics (Floridi, 2010), information is a combination of data and meaning (Yu, 2015),

information systems books and knowledge management textbooks tend to define information as organized data (Rowley, 2007).

Data are a general concept, developing in an increasingly broader manner. It is hard to define data without context (Bordignon & Maisonobe, 2022). Data mean different things in different disciplines. In natural and formal sciences, data might be an image of the observation of the galaxies, a measurement of the speed of an object, documentation of a research project, biological sequence like DNA, regression analysis result, binary digits, or logs of human-computer interaction, etc.; In humanities and social sciences, a considerable portion of data are generated by human participants. Bank transactions, maps, surveys, image records of paintings, social media content with tags, interviews, recordings of folk songs, and from a broader perspective, manuscripts, articles, and a set of research results are all data. Data are a set of signs to record observations, e.g., words, numbers, symbols, sounds, and images, representing properties of the observations about things, events, and activities (Ackoff, 1989; Rowley, 2007; Yu, 2015; Zins, 2007). They are heterogeneous raw materials for the generation of information, knowledge, and wisdom.

It is deconstructive thinking that suggests moving from information relevance to data relevance. Information and data are different but related. The information and data contexts bring different but related insights to the concept of relevance. In the information context, relevance deals with the purpose of the IR system. The research keywords include logic, problem, information need, query, information, document, user's cognitive space, user's judgment, criteria, subjectiveness, etc. (Borlund, 2003; Mizzaro, 1997). In the data context, the traditional IR system may not be a factor to consider; certain properties such as purpose of

collection may be required when data sharing and reusing; the users of data and the users of information may be different, consequently, the cognitive and psychological space varies thereupon; the topic of document/information may not be applicable when talking about data; the problems to be solved are distinct, in other words, information need is not equal to data need; some existing types of relevance may not apply and new types may emerge. Data is fact-bearing evidence for reasoning and decision-making while information is used to shape an individual's cognitive state (Liu et al., 2020). All these differences motivate data relevance study.

Many scholars consider information as extracted from data to be meaningful to users; they suggest that data are raw, unprocessed, not contextual, and meaning-neutral or have no meaning (Rowley, 2007; Zins, 2007). However, the preparation and capture of data are intentional with well-designed tools and devices, the organization of data or the way we present them are biased and purposeful with existing schemas and good practice, the data ecosystem and the promotion of data sharing are thoughtfully conceived and meticulously developed. From this perspective, data are not raw but contextual and meaningful (Bordignon & Maisonobe, 2022; Gitelman, 2013), which is the foundation of studying data relevance. However, the value of data "may not be apparent until long after those data are collected, curated, or lost" and may vary "widely over place, time, and context" (Borgman, 2015, p. 4), which makes studying data relevance more complex yet crucial.

### 1.1.4 Linked Data and Data Relevance Study

The Semantic Web provides a framework allowing data publishing and sharing with abundant relationships. As a key implementation step to the Semantic Web, Linked Data or

Linked Open Data (LOD, data without intellectual property restrictions) have been adopted for sharing data by creating links between different resources on the Web (Berners-Lee, 2006). Linked Data has been studied to facilitate resource discovery for innumerable data links and abundant relationships. Concepts and the distances between the concepts in DBpedia, the Linked Data version of Wikipedia, were used to suggest relevant topics in Open Innovation (Stankovic et al., 2011). Huo et al. (2020) proposed ensuring data relevancy across domains for electronic record management systems by leveraging the Linked Data technique and the links between data.

Linked KOS can be a semantic foundation for representing, organizing, classifying, and visualizing big data (Shiri, 2014). The mapped, combined, aligned, and linked KOS has the potential to support and manage interactions with big data. A large number of KOS, e.g., AGROVOC Thesaurus, Dublin Core terms, EU vocabularies, Getty vocabularies, Medical Subject Headings, and Library of Congress vocabularies, have been converted into LOD datasets to facilitate data sharing and resource discovery on the Web. The umbrella term "LOD KOS" emerged during the practice of adapting KOS to the Semantic Web, referring to "all of the value vocabularies and lightweight ontologies within the Semantic Web framework" (Zeng & Mayr, 2019, p. 209), which: (1) Use global identifier HTTP URIs to represent entities or resources for resource location, (2) Are expressed as RDF (Resource Description Framework) triples and encoded in RDF supporting syntax for machine readability, (3) Link to other LOD or LOD KOS datasets to leverage the relationships for resource discovery, (4) Can be accessed through SPARQL endpoints or APIs, and (5) Are licensed under such as Creative Commons licenses or public domain dedication for openness (Berners-Lee, 2006).

LOD KOS enables semantics and interoperability of massive data. Many KOS have been adapted to RDF and converted into LOD KOS, each with numerous links within and between them. With their rich relationships, LOD KOS are suitable candidates for studying data relevance.

## 1.2 Research Questions

Hjørland (2001) claimed that "It is, however, important to realize that there are many different mechanisms at play here. There are many different actors and layers influencing the relevance of the final output." I cannot agree more. Studying data relevance from the perspective of KO is an underexplored topic in information science.

This study seeks to inquire about the concept of relevance from a data perspective. The purpose of this study is to conceptualize relevance in the knowledge organization context, explore data relevance by leveraging the rich semantic relationships in LOD KOS, and develop a data relevance theory. The research questions are:

Question 1. What is relevance in terms of knowledge organization?

Different from the relevance studies in the IR area, this study is set in the KO context and examines the understandings of relevance from the KO perspective using literature thematic analysis.

Question 2. What relationships exist in knowledge organization?

Relationships of resources exist extensively in knowledge organization. Resources can be primary resources like bibliographic resources, online information, etc., and surrogate resources like catalogs, metadata records, etc. They can be full text or information entities in smaller units like concepts and terms. To answer Question 2, an investigation of the subjects of

relationships is necessary before studying relationships further. To study data relevance, relationships in LOD KOS (as a type of data) must be examined.

Question 3. What is data relevance in the KO context?

A user study was conducted to conceptualize data relevance theory, together with the answers to Question 1 and Question 2. Users in this study are information professionals with experience in using KOS. The data relevance model is created to answer Question 3.

## 1.3 Significance and Contribution

The study has theoretical values, methodological values, and practical values for relevance research, IR research, KO research, and LOD research. It supports extensive leverage of the valuable products accumulated in the KO area for deep integration of information science into the data era.

### 1.3.1 For Relevance Research

This study presents a new integrated perspective and a new way of thinking. It breaks the dominance of relevance study from an IR perspective, where relevance is an indicator or a part of an indicator to evaluate the performance of information retrieval systems. It expands to the KO area to bring a supplemental understanding of relevance in the data context.

Data relevance has not been studied systematically. The proposed data relevance theory will fill the theoretical gap and promote the current relevance theories to adapt to the evolving external technological and cognitive environment.

### 1.3.2 For IR Research

The proposed relevance inquiry from a data perspective in the KO and Linked Data context is an unexplored way to study relevance in information science. The result of this

research will supplement the current understanding of relevance criteria and measurements in an IR context. Data relevance theory is expected to extend the boundary of relevance research in IR and deepen the understanding of the common notions shared in the current relevance framework and data relevance theory.

### 1.3.3 For KO Research

Hjørland (2016) wrote "if the purpose of a KOS is to help users to identify relevant documents, then IR is a family of competing approaches when compared to the approaches studied by the KO community." (p. 478). Using the concept of relevance to build connections between IR and KO is optimistic to bring novel views of key concepts in KO. For example, relationship study has always been a crucial topic for structuring knowledge. Relevance study provides motivation to examine the current relationship study in the KO area. The paradigm, theories, and methodologies adopted in the relevance study may promote the expansion of the current generic associative relationship, one of the three major semantic relationships in KOS.

Practically, this study proposes a new application of the rich relationships developed in KO studies. Standard vocabularies for describing metadata elements and values and their relationships are efficient in creating automatic linkages between datasets (Subirats & Zeng, 2020). The study of semantic relationships between LOD KOS will boost the theoretical and practical value of KOS.

Additionally, the study provides a way to address the challenge of determining "how the LOD KOS can be used as more than traditional 'controlled vocabularies' or can function as more than just being 'value vocabularies' in the Semantic Web" (Zeng & Mayr, 2019, p. 19).

Relevance study provides an opportunity to leverage LOD KOS in other ways (e.g., in the sense of judging relevance) and strengthens the fundamentality of KOS in organizing resources.

### 1.3.4 For LOD Research

Vocabulary reuse, rather than reinvention, is one of the best practices of LOD (Lóscio et al., 2017). The proposed data relevance theory provides a fundamental logic for vocabulary reuse. Measuring relevance is a core task of designing algorithms for vocabulary search. Embedding data relevance theory into vocabulary searching or recommendation tools will improve their performance for better interoperability, redundancy and ambiguity reduction, and automatic data processing.

Data relevance will have practical implications for data linking. The current data linkage and vocabulary reuse studies are mainly based on matching, mapping, and similarity. This study makes it possible to leverage relationships for semantic enrichment in data linkage. The abundant LOD KOS semantic relationships will enrich the data linkages beyond simple relationships. It will also contribute to establishing standards for automatic linking between datasets.

## 1.4 Outline of the Study

In the following Chapter 2, theoretical foundations are discussed starting from the conceptualization of relevance in information science. Relevance has been examined from different viewpoints based on various epistemologies, including preliminary views, system view, topical view, and logical view, which represent objective-oriented views on relevance, and situational view, psychological view, user-oriented view, and socio-cognitive view, which represent subjective-oriented views on relevance. Although there is no absolute boundary

between objective- and subjective-oriented views, the downsized dichotomy is helpful to understand the mainstream studies on relevance. Relevance is easy to understand but difficult to define formally. The frameworks and paradigms to define and study relevance are discussed subsequently, followed by relevance assessment from an empirical standpoint. Then relationships, as a core concept, have been examined within the context of relevance. Relevance studies in information science, but beyond the information retrieval area, and data relevance studies have been observed.

In Chapter 3, the research paradigm is discussed first to lay the foundation of the entire study, followed by research design, including research method, task design, interview design, and sampling method. Subsequently, tasks such as participant recruitment, data collection, and data analysis are introduced.

In Chapter 4, relevance is conceptualized in the KO context. Firstly, objectivity and subjectivity have been clarified in Section 4.1 to build up a theoretical foundation for the subsequent discussions. Then, relevance is studied in the tasks of KO, including subject analysis, indexing, cataloging, and social tagging. The related concepts of aboutness and warrant are discussed as well. A set of definitions of relevance along with the tasks has been proposed. Chapter 4 addresses the first research question: What is relevance in terms of knowledge organization?

According to the definitions in Chapter 4, relationships are one of the two central aspects of relevance. In Chapter 5, a meta-typology of relationships is proposed, covering the bibliographic universe, the information universe and the data universe. Following the meta-typology, relationships are examined in each universe. Relevance as a relationship, i.e.,

relevance relationship, is discussed next. Chapter 5 answers the second research question: What relationships exist in knowledge organization?

In Chapter 6, data from the user study are examined alongside three stages. General features of the participants' responses in the first stage are analyzed, covering metadata fields, reused elements, schemas, granularity, and structure. The responses in the second stage are analyzed using open coding and axial coding methods, with many categories emerging from the notes explaining participants' design behavior. The interview transcripts are analyzed using open coding and axial coding methods as well, with more categories emerging. Categories from the second stage and the interview stage are combined for selective coding. The data relevance model with three levels of functioning elements is created after selective coding.

Chapter 7 concludes the study by: 1) summarizing the progression of the study from proposing research questions to answering the questions step by step, showing how the study is addressed overall; 2) highlighting the findings for each research question, showing the main results at each step; 3) concluding data relevance theory and listing some use cases; 4) discussing issues and other findings alongside the research; and 5) looking ahead to the future research to strengthen, extend, and expand this study.

# 2 Literature Review

## 2.1 Conceptualization of Relevance in Information Science

Relevance studies in Information Science have spanned many years, starting with the evidence demonstrated in Vickery's two papers (Vickery, 1959b, 1959a), which aimed to enhance the performance of information retrieval systems. Novel concepts and ideas have been proposed along with the evolution of information science to meet the emerging features of the academic discipline, in the aspects of theoretical foundations and perspectives, definitions, dimensions and manifestations, criteria and assessments, applications, etc.

### 2.1.1 Different Views of Relevance

**Preliminary and system views**

Vickery (1959a) proposed two relevance criteria *literary warrant* and *user relevance*, indicating the forthcoming dichotomous perception of objective relevance and subjective relevance, which dominated relevance studies for a long time. Literary warrant enables "prior discrimination", regarding the relevance defined by the information retrieval systems according to the relationships between documents (Vickery, 1959a, pp. 863–864). Vickery suggested a combination of literary warrant and user relevance for better system performance.

Wilson (1968) specified two senses of question relevance: One was the logical concept of relevance, underlining the logical relationships between statements and indicating objective relevance without considering factors such as person and time; The other one was about utility, the relationships between text, problem, person, and time, focusing on a person's beliefs, preferences, and decisions.

Maron and Kuhns (1960) studied probable relevance from a system view. They used "Probabilistic Indexing" to generate relevance numbers for documents to probabilistically measure a document's relevance (in the form of probability) to a request. The probability was inferred by the statistical measures of semantical closeness between index terms based on cross-referencing.

Bookstein (1979) argued that relevance was the relation between an individual and a document. It was the "degree to which the user senses that the need bringing him to the system is satisfied" (p.270) to serve the goal of information retrieval systems. Generally, the system view focused on the different degrees of match between a query and a document.

**Topical relevance**

Schutz (1970) suggested three manifestations of relevance: topical relevance, interpretative relevance, and motivational relevance. Topical relevance was "the first form of relevance" (p. 26), including intrinsic topical relevance (theme, non-imposed) and imposed topical relevance ("by means of social interaction" (p. 29), e.g., experiences representing "any shift of attention of consciousness" (p. 28)). Interpretative relevance was based on previous knowledge. Motivational relevance was relevance for an individual's future actions. If we look at it from a dichotomous perspective, intrinsic topical relevance is more objective, while imposed topical relevance is more subjective. Harter (1992) considered topical relevance "weak relevance" because on-the-topic information didn't necessarily "cause cognitive change", which meant relevance to him (p. 612).

In further exploration of topical relevance, Huang and Soergel (2006, p. 1) stated that "topical relevance lies at the heart of relevance". They created a three-level taxonomy of

topical relevance with four top categories: direct relevance, indirect relevance, context relevance, and comparison relevance. Unlike direct relevance, indirect relevance could not provide a direct answer but offered an inference helping to answer a question. Context relevance provided a setting to support an argument. Comparison relevance identified similar features for understanding topics. A following study showed that topical relevance was user-oriented and could be determined either by systems or users (Huang & Soergel, 2013).

Topical relevance has been studied in objective, subjective, and mixed viewpoints. Studies associated with these viewpoints created distinct relevance criteria, but some views in common emerged: The first one was that topical relevance was one of the fundamental dimensions of relevance. The second was that the foundation of topical relevance was the relationships between the topic/subject of information and the topic/subject of the users' information needs.

**Logical relevance**

Cooper (1971) defined logical relevance based on logical consequence (entailment). A statement was logically relevant to a question if the statement belonged to "some minimal premise set" (p. 24) of statements from which the answer to the question could be found. Cooper (1971) considered objectifying information need, though an invisible "psychological state" (p. 21), as the information needs representation (i.e., sentences). Logical relevance means stored information entails representations of information needs. Cooper regarded logical relevance as topicality, which meant the information pertained to the topic of an information need, which was closely related to topical relevance. Logical relevance didn't guarantee usefulness, which was insufficient to address utility.

Wilson (1973) argued that Cooper's proposal, as deductive logic, was just a type of logical relevance in a narrow sense since a situation where the premises were true statements, but the conclusions were false might happen. He also explained that logical relevance did not depend on a person's subjective cognition, which was objective and independent of personal understanding and views. He observed evidential relevance, which was inductive and increased "the degree of confirmation, or probability, of conclusions in relation to given premises", or "the premises support the conclusion but do not entail it" (Wilson, 1973, pp. 459–460).

**Situational relevance**

Wilson (1973) proposed the concept of situational relevance based on Cooper's logical relevance, inductive evidential relevance, and personal knowledge background and concerns, which gained great attention in later relevance studies. Situational relevance is relevance directed at an individual's unique concerns, influenced by their background knowledge to interpret the situation. He differentiated situational relevance from psychological relevance, which "has to do with the actual uses and actual effects of information" and was hard to reach a consensus (Wilson, 1973, p. 458). Schamber et al. (1990) proposed a dynamic and situational perspective to study relevance. The relevance of information might change due to different people at different times since relevance judgments were "in relation to their information need situations at particular points in time" (p. 771). Barry (1994) confirmed that situational factors affected users' relevance judgment, e.g., experience, background, knowledge, beliefs, personal preferences, etc.

Borlund (2003) argued that situational relevance "best fulfills the requirements to embrace both the multidimensional and dynamic nature of relevance" (p. 921). Raper (2007)

expanded situational relevance to encompass geographic relevance to meet the mobile information-seeking boom, as a relation between geographic information needs (GINs) and the spatio-temporal expression of the geographic information objects (GIOs). Four geographic relevance domains were recognized, which were the allocentric spatio-temporal domain, the allocentric place-oriented domain, the egocentric spatio-temporal domain, and the egocentric place-oriented domain. Sabbata et al. (2015) described a set of components of relevance, including geographic dimension, including topic, activity, preferences, social, mobility, and context.

Rousi et al. (2018) examined situational relevance in the context of music information seeking. They found 21 types of situational relevance corresponding to various modes of music information, e.g., notated traditions of performance, the gestural language of the concert pieces, and conventions of scientific writing. These types can hardly be applied to domains other than music.

**Psychological relevance**

Wilson (1973) used the term psychological relevance to contrast with logical relevance. Logical relevance was formal, objective and independent of cognition, while psychological relevance was "irrespective of standards of propriety" (p. 458) and related to subjective interest or usefulness. Swanson (1986) categorized psychological relevance as a subjective relevance, related to whether users changed their views after receiving information. Harter (1992) explained that psychological relevance was about usefulness without considering topicality. He interpreted psychological relevance, which Sperber and Wilson proposed in the book Relevance: Communication and Cognition in 1986. Information needs bridge psychological

relevance and information retrieval. According to Harter, information needs were dynamic and were the current cognitive or psychological state of an information seeker. Psychological relevance changed along with the change in cognition. For this reason, Saracevic (1996) argued that psychological relevance was more cognitive relevance and pointed out, “It is only about representation of an information need, and changes in cognitive structure when receiving answers, rendering the original representation inadequate” (p. 208).

**User-oriented relevance**

By stepping away from the system view, topical view and logical view, user-oriented relevance focuses more on user factors in relevance judgments. Cuadra and Katter (1967) did a user study to identify effective variables on relevance judgments to evaluate information retrieval systems. The experiment showed that the criteria of relevance judgment changed according to the different contexts of the tasks informed by the use orientations. Some variables may have a greater impact on results, highlighting the contextual dependence and variability of relevance judgments.

Schamber and Eisenberg (1988) introduced a user-centric approach to define relevance, regarding relevance as a multidimensional concept based on human judgment and cognitive and situational factors. The dimensions of relevance in the user-centric cognitive model were decided by users, not systems, which differed from the user-oriented system method. Schamber et al. (1990) reaffirmed the dynamic, situational user-centric cognitive view of relevance by adding dynamic human judgment attributes of relevance. They mentioned “It is users’ criteria for relevance judgments, not their relevance judgments per se, in which we are primarily interested” (Schamber et al., 1990, p. 771).

Park believes relevance "totally belongs to an individual's situational and psychological state" (T. K. Park, 1993, p. 325) and proposed the concept "user-based relevance" (T. K. Park, 1994). Consistent with psychological relevance, users' cognitive state and context were the vital factors of user-based relevance. Context in this study was constructed psychologically, including: internal context, referring to previous experience, beliefs, expertise, and knowledge background; external context, referring to the judgment of the search goal, search process, and research stage and output; and problem context, how to understand and figure out the problem by referring the same/similar/different problems or information for definitions, backgrounds, methodologies, frameworks, etc.

In Barry's (1994) user study on relevance criteria, users applied factors beyond topicality to evaluate relevance. Chen and Xu (2005) proposed a conceptual model of user-oriented relevance, which included topicality, novelty, reliability, understandability, and scope, but only topicality and novelty were supported by the study.

**Socio-cognitive relevance**

Unlike psychological and user-oriented relevance that depend on self-cognitive status, socio-cognitive relevance studies how the social environment and interactions influence individuals' understanding and judgment of relevance. Cosijn and Ingwersen (2000) added the concept of socio-cognitive relevance as one manifestation instead of "motivational or affective relevance" proposed by Saracevic (1996). Socio-cognitive relevance was "the relation between the situation, work task or problem at hand in a given socio-cultural context and the information objects, as perceived by one or several cognitive agents" (Cosijn & Ingwersen, 2000, p. 541). Hjørland (2002) advocated for a socio-cognitive view in information science

because an individual's cognition was in the framework of society, history, and culture. Thus, the criteria were shaped by a group of associated people (discourse community). Under a socio-cognitive view, relevance was defined as: "Something (A) is relevant to a task (T) if it increases the likelihood of accomplishing the goal (G), which is implied by T" (Hjørland & Christensen, 2002, p. 964).

In line with Cosijn, Ingwersen and Hjørland, Van der Veer Martens and Van Fleet (2012) applied the socio-cognitive view to study dimensions of relevance for information agencies (libraries, archives, intelligence organizations, etc.). They found additional subdimensions besides system, time, task, and assessor, incorporating:

> the system's ethos as determining the information being made accessible to users, the varying time values of the information required and provided, the different assessments being made of the information and its uses, and, finally, the strengths and limitations of the various ways in which information can and cannot be queried on behalf of user tasks (Van der Veer Martens & Van Fleet, 2012, p. 944).

**The objective-subjective dichotomy**

A dichotomy between subjective and objective relevance emerged from the studies from the 1960s to the 1980s. Könige (1971) analyzed three types of relevance: formal relevance, subject/content relevance, and subjective relevance. Formal relevance was the relevance between the document and the request. Subject/content relevance was the relevance between the content of the document and the content of the request. Subjective relevance was pertinency, addressing the utility of documents. The first two types fell into the objective category while the third one was subjective.

Foskett (1972) used the term pertinence for the subjective features and the term relevance for the objective features. Though the two concepts were often treated identically, relevance corresponded with a paradigm, public knowledge, or a consensus while pertinence was about personal preference. Therefore, one document may be pertinent to a request but irrelevant, and vice versa. By expanding Foskett's idea, Kemp (1974) considered relevance and pertinence as paired concepts, corresponding to public knowledge and private knowledge in the philosophy of science respectively. Relevance was public and objective while pertinence was private and subjective.

Swanson (1977) recognized two kinds of relevance: relevance judged by a person subjectively to evaluate whether information need was satisfied and relevance that was topically identical/related but didn't guarantee utility. Documents satisfying the second relevance might not meet the first one. Applying the two kinds of relevance in information retrieval systems separately could get different results. Relevance for new knowledge was subjectively created by the users, not assigned by the systems or the experts, without any bearings on documents or requests. In his opinion, topical relevance was partly subjective as well since it was "a matter of opinion", but "it can make sense to speak of a consensus" within this type of relevance (p. 142). Swanson then explicitly elaborated on the distinctions between objective and subjective relevance in the following study and confirmed that the two kinds of relevance represented subjective and objective relevance respectively (Swanson, 1986). According to him, subjective relevance was mentally related while objective relevance was logically related. He stated system design and testing should be based on objective relevance, "the relationships between requests and documents" independent of the requestor (p. 396).

Pertinence, psychological relevance, utility, and usefulness are subjective while logical relevance and topical relevance (though not pure) are objective.

Harter (1992) further explained that objective relevance was assessed by experts (intermediaries) while subjective relevance was judged by users. Both objective and subjective relevance adopted topicality as one of the relevance criteria. Borlund (2003) considered objective relevance was system-based while subjective relevance was human(user)-based. Huang and Soergel (2013) argued that neither system relevance nor user relevance was the gold standard; they were not opposite but complementary. Ruthven (2021) introduced the concept of resonance as a dimension of relevance and proposed a triple-stratified model called the resonance spectrum encompassing three progressive components: resonance as agreement, resonance as arousal, and resonance as action. Compared to relevance "where (more or less) objective and intellectual criteria dominate…", resonance was a concept space "where both subjective and nonintellectual criteria are more influential, and this is where resonance offers a different perspective" (p.10). Resonance was a key factor in determining subjective relevance (e.g., emotional needs) other than objective relevance (e.g., topical relevance).

Hjørland (2010) argued against the mainstream dichotomy and called for a reconceptualization of the foundations of relevance. He questioned the concept system's relevance and stated, "Relevance is thus never 'a system's', but always 'human' and therefore the dichotomy is wrong" (P. 231).

### 2.1.2 Frameworks to Understand Relevance

Relevance has been interpreted in various ways. To build a big picture of the notion of relevance, several researchers sought to find a way to accommodate and coordinate different viewpoints.

From a communication perspective, Saracevic (1975) constructed a formula in a source-to-destination way with five factors to define relevance: Relevance is the A (measure, degree, dimension, estimate, appraisal, relation…) of a B (correspondence, utility, connection, satisfaction, fit, bearing, matching …) existing between a C (document, article, textual form, reference, information, fact …) and a D (query, a request, information used, point of view, information requirement statement…) as determined by an E (person, judge, user, requester, information specialist…) (p. 328). This formula was comprehensive and laid the foundation for defining relevance in information science.

Saracevic (1975) proposed seven views of relevance based on the formula and the former studies: subject knowledge view, subject literature view, logical view, system's view, destination's view, pertinence view, and pragmatic view. In a following study, he reconsidered relevance and reduced the seven views to five manifestations: system or algorithmic relevance, topical or subject relevance, cognitive relevance or pertinence, situational relevance or utility, and motivational or affective relevance (Saracevic, 1996). The five manifestations were modified by Cosijn and Ingwersen (2000), which was mentioned in the Socio-cognitive relevance section and is discussed in the following part.

A summary of Saracevic's various views of relevance is shown in Table 1.

Table 1: Different manifestations of relevance in Saracevic's studies

| Views of relevance | Source | Destination |
|---|---|---|
| Subject knowledge view | Knowledge about the subject | Topic (question) about the subject |
| Subject literature view | Literature/Subject | Topic (question)/ Literature |
| Logical view | Premises on a topic | Conclusion from a subject/literature |
| System's view of relevance | Contents of file | A subject/literature/question |
| Destination's view | Documents | Topic(question) |
| Pertinence view | The stock of knowledge | Subject knowledge/literature |
| Pragmatic view | Problem | Information |
| System or algorithmic relevance | Query | Information objects |
| Topical or subject relevance | Subject or topic of a query | Subject or topic covered by retrieved texts |
| Cognitive relevance or pertinence | State of knowledge and cognitive information need | Retrieved texts |
| Situational relevance or utility | Situation/task/problem | Retrieved texts |
| Motivational or affective relevance | Intents/goals/motivations | Retrieved texts |

Kitiyadisai (1991) proposed a combination of semiotics and affordance as the framework to study relevance comprehensively and systematically, purporting a universal framework covering kinds of relevance, including relevance in phenomenology, law, logic, communication and cognition, and information science. The four branches of semiotics: pragmatics, semantics, syntactics, and empirics, provided a stratified basis to investigate the logical, contextual, subjective, and utility features of relevance. Relevance as an affordance meant "enabling a person to perform an appropriate action" based on the person's perception and knowledge in certain contexts for certain purposes (p. 129).

Mizzaro (1997, 1998) created a framework to embrace plenty of views, methods, types, criteria, etc., of relevance for a comprehensive review. The framework had four dimensions: Dimension 1 represented information resources, e.g., information, document, and surrogate;

Dimension 2 represented the user's problem, information need, request, and query; Dimension 3 included topic, task, context, and combinations of them; Dimension 4 represented time. All kinds of relevance could be seen "as a point in a four-dimensional space" (Mizzaro, 1997, p. 812). The four-dimensional framework was revised later in De Sabbata, Mizzaro and Reichenbacher's (2015) study by adding the geographic related factors to satisfy the new features of the mobile information seeking behavior. The components of relevance included topic, activity, preferences, social context, mobility, and context. The modified framework included resources, user problem, space-time (replacing time), and world (newly added) (De Sabbata et al., 2015).

Cosijn and Ingwersen (2000) studied the five manifestations of relevance proposed by Sarasevic (1996) and argued that motivational/affective relevance was not a category of relevance manifestations but attributes affecting subjective relevance. They proposed socio-cognitive relevance as the fifth type of manifestation. Together with the other four manifestations: algorithmic, topical, cognitive/pertinence, and situational/utility, and the five key attributes, i.e., relation, intention, context, inference, and interaction, they developed a revised version of the original manifestation model. Saracevic (2007) expanded the five attributes to nine by adding internal context, external context, selection, and measurement, and restated the five manifestations as system or algorithmic, topical or subjective, cognitive/pertinence, situational/utility, and affective relevance, and pointed out affective relevance underlined other manifestations.

Kagolovsky and Mohr (2001) developed a process model of information retrieval as a framework to study relevance. The methodology aimed to identify relevance relationships

during the IR process and evaluate the relationships to interpret the notion of relevance. They proposed a list of inquiries about the possible relevance relationships as a guide for comprehensive studies. The relevance relationships, according to the process model of IR, were 1) between the user's state of knowledge and a problem, 2) between information need and user's state of knowledge, 3) between the user's information seeking strategy and explicit information need, implicit information need, or a problem, 4) between a formulated query and user's implicit information need or explicit information need, 5) between retrieved documents and a submitted query, a user's explicit information need, a user's implicit information need, or a user's problem solving.

Borlund (2003) designed a framework to interpret the varied understanding of relevance, which was based on three different IR processes: noninteractive, occurring within a single moment in an IR session, and during a dynamic and interactive IR session.

Huang and Soergel (2013) proposed a unified framework consisting of a conceptual framework (Framework 1) and major variables supporting the conceptual framework (Framework 2). In Framework 1, the notion of relevance was categorized into two groups: Relevance-as-is and Relevance-as-determined. Relevance-as-is was a relationship R between an information object I and an information need N; Relevance-as-determined was the assessment of relevance-as-is by an agent A based on representations I' of I and N' of N. Framework 2 covered the variables of information object I, representation I', information need N, representation N', relationship R between I and N, determining agent A, determination factor/condition, and characteristics of person or system.

Table 2. shows an overview of the views discussed and the frames of relevance. Categorizing different views into objective/subjective/combination does not imply support for the dichotomous cognition of relevance. It was created in the way of reduction and simplification to show a clear outline of the wide range of viewpoints over the years.

### 2.1.3 Paradigms Used in Relevance Research

Sperber and Wilson (1995) explored the notion of relevance in 1986 and 1995 (the first and second editions) within the communication framework. They stressed the importance of context in understanding relevance. They analyzed the irrelevant situations in a context and demonstrated that being relevant must have a contextual effect. They argued relevance was a relative term rather than a precise measurement. Comparative relevance had two conditions: the more contribution to understanding a situation (contextual effect), the more relevant it was; the less effort put into understanding, the more relevant it was. For a long time, the communication paradigm dominated relevance studies in information science (Schamber & Eisenberg, 1988). A source-to-destination definition of relevance is the result of adopting the communication paradigm, e.g., Saracevic's (1975) formula of relevance. One reason for the popularity of the communication paradigm is that relevance has been seen as an assessment of the effectiveness of information retrieval systems.

Schamber and Eisenberg (1988) advocated for a user-centric approach to replace the communication paradigm based on the assumptions that relevance was multidimensional, situational, and cognitive, and relevance was judged and contextualized by users.

Table 2: Summary of the main views of relevance in a dichotomous manner

| Name | Year | Objective/system-based | Subjective/user-based | Both |
|---|---|---|---|---|
| Vickery | 1959 | Literary warrant | User relevance | |
| Maron & Kuhns | 1960 | | Probable relevance | |
| Wilson | 1968 | Logical concept of relevance | Utility | |
| Schutz | 1970 | Topical-non-imposed relevance | Topical-imposed relevance; Interpretative relevance; Motivational relevance | |
| Cooper | 1971 | Logical relevance = topicality/topic-appropriateness | | |
| Königová | 1971 | Formal relevance<br>Subject/content relevance | Subjective relevance=pertinency | |
| Foskett | 1972 | Relevance | Pertinence | |
| Wilson | 1973 | Logical relevance/Evidential relevance | Psychological relevance | Situational relevance |
| Kemp | 1974 | Relevance | Pertinence | |
| Cleverdon and Kidd | 1976 | Redundancy relevance | | |
| Swanson | 1977 | | Relevance judgment as creation of new knowledge | “On the same topic” relevance |
| Bookstein | 1979 | | User sense relevance | |
| Swanson | 1986 | Logical relevance | Pertinence; Psychological relevance;Utility; usefulness | Topical relevance |
| Schamber and Eisenberg | 1988 | | User-centric view of relevance | |
| Schamber et al. | 1990 | | Dynamic, situational user-centric view of relevance | |
| Harter | 1992 | | Psychological relevance | |
| Park | 1993, 1994 | | User-based relevance based on psychological relevance and cognitive state | |

| | | | | |
|---|---|---|---|---|
| Saracevic | 1996 | System or algorithmic relevance, Topical or subject relevance | Cognitive relevance or pertinence; Situational relevance or utility; Motivational or affective relevance | |
| Cosijn and Ingwersen | 2000 | Algorithmic relevance | Topical relevance; Cognitive relevance/pertinence; Situational relevance/utility; Socio-cognitive relevance | |
| Borlund | 2003 | System-based relevance | Human(user)-based relevance; Situational relevance; Psychological relevance | |
| Raper | 2007 | | | Geographic relevance |
| Huang and Soergel | 2013 | | Topical relevance | |
| Sabbata et al. | 2015 | | | Geographic relevance |
| Rousi et al. | 2018 | | | Situational relevance in music |
| Ruthven | 2021 | | Resonance | |

Park (1994) called for naturalism as a new paradigm to study user-based relevance. The naturalistic paradigm had many implications when user-based/subjective relevance accounted for an essential part of the relevance study. Park (1994) summarized 14 implications: natural setting, human instrument, utilization of tacit knowledge, qualitative methods, purposive sampling, inductive data analysis, grounded theory, emergent design, negotiated outcomes, case study reporting mode, idiographic interpretation, tentative application, focus-determined boundaries, and trustworthiness. These implications also helped to address the methodological problems criticized by researchers. For example, a long-existing provocative issue of experimental studies was that relevance judged by assessors other than real users might lead to different conclusions. Naturalists preferred natural settings and purposive sampling to increase the understanding of the phenomenon (T. K. Park, 1994). The reality was multiple, interrelated, interactive, and nonmanipulable in naturalism. These features enabled naturalistic inquiries suited for relevance study.

The constructivist paradigm was applied in two relevant assessment studies (Brooks, 1995; Costello, 2017). In the constructivist paradigm, relevance was subjectively constructed by the users based on psychological status, cognitive status, previous knowledge, etc. (Brooks, 1995). For example, Swanson's (1977) relevance judgment as the "creation of new knowledge", regarding relevance as new knowledge constructed by the subjective experience of the user, and the argument of Schamber, Eisenberg, and Nilan (1990, p. 771) that "relevance criteria are spatially and temporally bound to, and internally constructed by, the user" are in line with constructivism.

Saracevic (1975) claimed the subject knowledge view was the foundation of the other six views (see Section 2.1.2 Frameworks to understand relevance). The subject knowledge view of relevance was "the relation between the knowledge on or about the subject and a topic (question) on or about the subject" (Saracevic, 1975, p.338). It "stresses the nature, structure and extent of the subject knowledge on a topic given by a question" and "…corresponds, to some degree, to Schutz's 'topical relevances'…" (Saracevic, 1975, p. 332). In Saracevic's (1996) reconsideration of relevance, he seemed to drop the notion and substituted it with topical relevance. However, Hjørland (2010) reemphasized the subject knowledge view but interpreted it differently. Other than Saracevic's topical relevance sense, Hjørland (2010) argued that the subject knowledge view of relevance provided a better theoretical framework than system/algorithmic relevance or the users' view and regarded it as a socio-cognitive paradigm of relevance based on a pragmatic theory of knowledge.

Hjørland (2002) listed different relevance criteria under four psychological paradigms: behaviorism, cognitivism, neuroscience, and psychoanalysis (see Table 3), and under four types of epistemology (see Table 4). The two tables show the importance of paradigms for relevance studies. Relevance criteria vary according to the epistemology adopted in the research paradigm.

Hjørland (2010) suggested using other epistemological schools: introducing universalism/objectivism (positivism) for systems or algorithmic relevance, individual subjectivism for user-oriented and cognitive relevance, and pragmatism for the subject knowledge view of relevance. He reinterpreted the objective-subjective dichotomy as two variants of positivist epistemology. He recognized that a fundamental way to study relevance,

"to investigate underlying theoretical assumptions and epistemologies both in the queries and in the documents in the databases", was missing (p. 210).

Table 3: Hjørland's simplified relevance criteria in four psychological paradigms (2002, p. 263)

| Behaviorism | Cognitivism | Neuroscience | Psychoanalysis |
|---|---|---|---|
| Relevant: Information about responses to specific kinds of stimuli. Kind or organism are of minor importance. (High priority to intersubjective controlled data.) | Relevant: Information about mental information mechanisms and processing. Analogies between psychological and computer processes. Measures of channel capacities, etc. | Relevant: Information correlating brain processes or structures with forms of behavior or experience. | Relevant: Information about dreams, symbols, mental associations, personal meanings associated with stimuli, etc. Data collected in therapeutic sessions by trained therapists who can interpret the data (thus giving lower priority to intersubjective controlled information). |

Table 4: Hjørland's simplified relevance criteria in four epistemological schools (2002, p. 269)

| Empiricism | Rationalism | Historicism | Pragmatism |
|---|---|---|---|
| Relevant: Observations, sense-data. Induction from collections of observational data. Intersubjectively controlled data. | Relevant: Pure thinking, logic, mathematical models, computer modeling, systems of axioms, definitions, and theorems. | Relevant: Background knowledge about preunderstanding, theories, conceptions, contexts, historical developments, and evolutionary perspectives. | Relevant: Information about goals and values and consequences both involving the researcher and the object of research (subject and object). |
| Nonrelevant: Speculations, knowledge transmitted from authorities. "Book knowledge" ("reading nature, not books"). Data about the observers' assumptions and preunderstanding. | Low priority is given to empirical data because such data must be organized in accordance with principles that cannot come from experience. | Low priority is given to decontextualized data of which the meanings cannot be interpreted. Intersubjectively controlled data are often seen as trivia. | Low priority (or outright suspicion) is given to claimed value free or neutral information. For example, feminist epistemology is suspicious about the neutrality of information produced in a male dominated society. |

### 2.1.4 Relevance Assessment

Relevance criteria have been studied since the 1960s. Schamber (1991), in her doctoral dissertation, identified 10 top-level and 22 subcategory criteria. The top-level criteria were accuracy, currency, specificity, geographic proximity, reliability, accessibility, verifiability, clarity, dynamism, and presentation qualities. She summarized 80 relevance factors from previous studies (Schamber, 1994).

Barry (1994) identified 23 relevance criteria in seven groups, covering criteria of the information content of documents, the user's previous experience and background, the user's beliefs and preferences, other information and sources with the information environment, the sources of documents, the document as a physical entity, and the user's situation. Schamber (1994) summarized 80 relevance factors over judges, requests, documents, information system, judgement conditions, and choice of scale, mainly from the previous research.

Both Schamber's (1991) and Barry's (1994) empirical studies adopted user assessments but had different groups of users creating relevance judgments in various domains. In their successive research (Barry & Schamber, 1998), they compared the two studies and found shared criteria across domains and users, subsuming depth/scope/specificity, accuracy/validity, clarity, currency, tangibility, quality of sources, accessibility, availability of information/source of information, verification, and affectiveness.

Vakkari and Hakala (2000) conducted an empirical study to show that users' relevance criteria changed during the process of task performance. The criteria had different discrimination power, e.g., topicality was more powerful than references. In line with Vakkari and Hakala, Taylor, Zhang, and Amadio (2009) found that the variation of relevance criteria

along the information retrieval process was statistically significant. Taylor (2013) further found work task and relevance criteria were statistically related, i.e., users adopted different criteria to judge relevant information when they had different work tasks.

Savolainen and Kari (2006) conducted an empirical study and proposed 18 criteria for web searching based on the previous research from other researchers, which were ability to understand, accessibility, affectiveness, clarity, cost, curiosity, currency, familiarity, language, novelty, reliability, security, specificity, time constraints, topicality, usability, validity, and variety. Albassam and Ruthven (2018) identified 28 relevance criteria in 8 groups for video, in which topicality was the leading one. The discrimination ability of each criterion changed in different contexts, e.g., academic, work, or leisure contexts. Habit, the newly emerged criterion, was more related to the leisure context.

Bodoff (2006) studied the differences between relevance for browsing and searching by categorizing relevance criteria. The tripartite categories of relevance criteria are: first, relationships between documents and user needs, subsuming topicality match, topicality non-match, and match but not topicality; second, monotonic document characteristics, including validity, clarity, quality, visibility, accessibility, etc.; third, time constraints as the feature of user's need. The first category may have different effects on browsers and searchers. The result showed that relevance for browsing could predict the relevance for searching but the non-topical factors might cause a negative correlation between them.

The variation of relevance criteria caused by different factors shows that relevance is dynamic, contextual, situational, multidimensional, and complex.

Relevance assessment bridges the conceptualization and application of relevance, enabling the abstract concept to land on practical ground. There are no uniform relevance criteria or a single criterion to judge relevance. The school for the user-based view of relevance regards the users' judgments as the gold standard; the school against it questions the reasonability of generalizing principles from individuality (Hjørland, 2010).

## 2.2 Relationships in Relevance Studies

The nature of relevance is relationships. There are two senses of relationships in relevance studies. The first sense is a general notion, used widely in literature, to present the definition of relevance. For example, relevance is the relationship between information and users' information needs. The second sense is the specific relationships that exist between documents/information. For example, the hierarchical relationship between the subject of a document and the subject of retrieved information. Research on relationships is one of the two major areas of relevance study (the other one is the factors or criteria of relevance) (Saracevic, 1975, 1996).

Relationships have different meanings in different views of relevance. For example, in topical relevance, the two parts of the relationships are the subject in a query and the subject of the information (documents, retrieved text, etc.), where the subjects can be reduced to the terms discussed above; situational relevance is defined as "an item of information and a particular individual's personal view of the world and his situation in it" (Wilson, 1973, p. 458); Motivational relevance referred to information and the intents and motivations of an individual (Saracevic, 1996).

Maron and Kuhns (1960) examined probable relevance by leveraging the semantical relationships (based on meaning, e.g., synonymity, partially implies) and the statistical relationships (relative frequency of occurrence) to measure the closeness between index terms.

Goffman (1964) considered document-query relationships were essential to determine relevance, but not sufficient. Relevance judgment was based on the relationships between documents.

Bodoff (2006) categorized relevance criteria into monotonic and indeterminate. For monotonic relevance, it did not matter what the users' information needs were. For example, "consensus within the field" was monotonic (Bodoff, 2006, p. 71). It was an objective view focusing on the characteristics of information per se and not changing according to a user's information need. The intrinsic connections of information were in the monotonic category, indicating that the relationships are a relevance criterion.

The school of topical relevance study pays more attention to relationships. Traditionally, topical relevance was almost equal to topical matching, which was too narrow to represent the rich relationships existing between documents/information (Bodoff, 2006; Green, 1995; Huang & Soergel, 2006). Green and Bean (1995) applied KOS to represent both information and information needs to study topical relevance relationships. The study was not based on the representation of users' information needs, which was represented by subject headings. The study had two assumptions: users' information needs could be represented by subject headings; users would find information under a subject heading relevant to the need represented by the subject heading. They argued that the development of subject headings was not only on literary warrant but also on user warrant. The end-users, as indexers, judged

relevance in the indexing stage for improving the "chance of obtaining a fully representative sample of topical relevant texts" (Green & Bean, 1995, p. 655). Vocabulary relationships were used to represent the relationships of relevance in this study. The identified 33 relationship types in a 5-level relationship typology fall into three major categories: topical matching, hierarchical relationships, and structural relationships.

Huang and Soergel (2006) studied the relationships between information and users' needs in a reasoning-from-evidence perspective and identified four types of topical relevance: direct, indirect, context, and comparison. Direct relevance was an exact and explicit answer (information) to a question (information need). Indirect relevance was implicit or inferential evidence to a question, subsuming generic indirect relevance, backward inference, forward inference, and inference from cases. Context relevance referred to contextual information to help question understanding, e.g., physical, social, political, and cultural settings, causal and time sequence, etc. Comparison relevance stressed the comparable factors, e.g., persons, places, events, experiences, etc., to help understand the topic. Later in Huang's dissertation (2009), she built up a faceted typology of relevance relationships based on Rhetorical Structure Theory and referring to the semantic relationship framework created by Green and Bean (1995). The typology consisted of three facets (function-based, reasoning-based, and semantic-based) and 227 types of relevance relationships. Though Huang separated the semantic-based facets from the other two facets, the function-based and reasoning-based facets were more or less based on semantic relations.

Huang and Soergel (2013) argued that relationships were the central element of relevance. They categorized relevance into Relevance-as-is and Relevance-as-determined

(discussed in 2.1.2 Frameworks to understand relevance). Relevance-as-is was the relationships and relevance-as-determined was the assessment of the relationships.

The rationality of the transition from the relationships between subject/topic in queries and subject/topic in documents to the relationships between the terms in queries and in documents provides an insight for relevance research: it is possible to study relevance through studying the relationships between vocabularies.

The existing research shows two patterns of relevance relationship studies, consistent with Bean and Green's viewpoint (2001, p. 123) that "relationships between the topics of user needs and the topics of texts previously judged to be relevant to those needs". One pattern is the direct way that draws the relationships by assessing relevance with users. The other pattern is the indirect way that represents the potential information needs using KOS and drawing the conclusions with the descriptors (indexing by information professionals belongs to this type). Bean and Green (2001) argued that "the domain of topical relevance relationships is limited only by the boundaries of human cognitive associations" and "every relationship imaginable can serve as a topical relevance relationship" (pp. 129-130). Huang and Soergel (2013) called for more research on the topical relational nature of relevance because it had not been studied as much as criteria-based studies of users' viewpoints.

## 2.3 Relevance Study Beyond IR

Hillman (1964) studied relevance in the documentation context by converting the study of relevance to the study of the relatedness of concepts. Studying relevance involved two problems: first, defining concepts via properties based on similarity judgments; second,

defining conceptual relatedness based on partial similarity judgments. A document was similar to another if they shared common properties or concepts, e.g., abstract.

Perreault (1966) suggested that KOS (e.g., classifications, subject headings), as structural, systematic conceptual schemes full of hierarchical relationships, could be utilized to find relevant information. The hierarchical relationships were the indicators of relevance.

Wilson (1968) proposed two senses of relevance: logical relevance and relevance enlightening users (he did not name the second type of relevance). The second sense of relevance focused on the helpfulness/usefulness to users. It had four-dimensional relationships between text, problem, person, and time. Wilson claimed both senses of relevance were the basic notions of bibliographic control.

Bade (2007) indicated that subject analysis by librarians was a process of judging relevance. In subject analysis, librarians "attempt to communicate information about topicality and relevance to unknown future users" (p. 842). However, they cannot get in touch with the unknown user and "pertinence can never enter into the act of subject analysis"; thus, this kind of relevance is in fact objective, situational- and context-free (p.842).

Hjørland (2000) suggested analyzing citation be another way to study relevance. The citation selecting and deciding process might imply norms and criteria different from relevance research in the IR area.

White (2011) used relevance theory to explain citation behaviors. Relevance is determined by "cognitive effects and the effort needed to process them" (p. 3345). Citation behaviors follow the power-law distributions because self-citation and recitation of a limited number of references have high cognitive effects and require minimal effort, which implies self-

citation and recitation have higher relevance than others. Following the same line of inquiry, White (2017) used relevance theory to explain why relevance judgment using different scales in IR retrieval experiments presented different distributions. Because vague scale points and strict scale points need more effort to process, users usually avoid using them, implying that they are irrelevant.

Not limited to the IR area, these studies demonstrate that relevance can be studied in broader areas in information science, including the areas of documentation, KOS, bibliographic control, citation, etc. Relevance studies beyond IR, to some extent, demonstrate the feasibility of my study from the perspective of knowledge organization.

## 2.4 Evidence of Data Relevance

Researchers use both relevance and relevancy when it comes to data, i.e., data relevance and data relevancy. In this study, we consistently use data relevance. Data relevance has been studied in data reuse and data retrieval contexts in different domains, usually as a determinant to evaluate such as quality, reusability, and meeting the purpose of data.

There are many challenges of reusing data. Some of them were "scientists' capability of discovering and selecting relevant datasets" and "the difficulty to locate datasets that can be particularly relevant to scientists' topic of study" (Curty, 2015, pp. 38–39). Borgman (2015) argued "Having the right data is usually better than having more data. […] Having no data is all too often the case, whether because no relevant data exist; they exist but cannot be found" (p. 4). Data relevance is one of the criteria to evaluate the "right" data.

Wang and Strong (1996) created a conceptual framework for data quality in the commercial context. According to their study, data relevance was a dimension to evaluate contextual data quality, and data must be relevant to users for tackling tasks.

In pharmacology, data relevance was proposed as an indicator of data quality and their use in hazard and risk assessment (Klimisch et al., 1997). Relevance was defined as "the extent to which data and/or tests are appropriate for a particular hazard identification or risk characterization" (p. 2). Klimisch et al. (1997) claimed that more reliable data might be more relevant and data relevance must be decided by experts. As an attribute of data quality, data relevance was also defined as "The extent to which data are applicable and helpful for users' needs" (Caro et al., 2008, p. 536).

Sedghi et al. (2008) found 27 relevance criteria for medical images through a user (providers and researchers) study, covering technical attributes, contextual information, professional features, availability and copyright, etc. The study also showed that users apply different criteria in different situations to evaluate relevance.

Data relevance was studied as a factor to assess data reusability (Faniel & Jacobsen, 2010). To determine relevant data, the participants assessed how the data met their problems. The relevance criteria varied but focused on whether the model can be validated using the data, including parameters and context information of the model. The participants examined the context regarding how the data were generated, then matched the key parameters to decide whether the data were relevant. However, in a later study, Faniel, Kriesberg, and Yakel (2016) found data relevance didn't show significantly positive relations to data reusers' satisfaction in social sciences.

In a medical white paper, data relevance, as one of the two measurements for data "fit-for-purpose", was defined as whether data were "robust and representative of the population of interest" in a clinical context where data were related to patient health status and care delivery (Duke-Margolis Center for Health Policy, 2018, p. 5). Furthermore, data relevance was measured by data representativeness (of the population of interest, exposures, covariates and outcome), sufficient subject, longitudinality, and patient-level linking.

Liu et al. (2020) identified topicality, accessibility, authority, quality, and usefulness as five data relevance criteria and uncovered the interactive process of users' relevance judgment. The process had three progressive stages: users used topicality to assess relevance first, which could affect the judgment of other criteria; then users used quality, authority, and accessibility to do the assessment; and last, users evaluated usefulness to make an overall judgment.

Zhang et al.(2021) studied scientific research data and found nine data relevance criteria, including topicality, availability, quality, completeness, authority, currency, convenience, usability and standardization. They explored the relationship between metadata and relevance criteria and concluded a many-many relationship, i.e., a metadata field was associated with several criteria and many metadata fields were associated with one relevance criterion.

These studies defined data relevance directly in specific contexts and domains without elaborating on the conceptualization of data relevance. An increasing number of researchers have engaged in data relevance research, demonstrating its substantial potential and future research opportunities, especially in the overwhelming data and AI trend.

# 3 Methodology

The literature review shows a large number of relevance studies for decades, from conceptual studies to empirical studies, from logical empiricism, rationalism, to naturalism, from cognitive approach, psychological approach, user-centered approach to social-cognitive approach, from comparative study to longitudinal study (Hjørland, 2010; Huang & Soergel, 2013; T. K. Park, 1994; Schamber et al., 1990). Unlike most relevance studies in the IR area, this study will examine relevance in the KO context. It is exploratory in nature and will employ a qualitative research design to gain an in-depth understanding of data relevance.

## 3.1 Research Paradigm

This study adopted pragmatism as the research paradigm. The selection of a research methodology under pragmatism should be subject to solving the research problems and achieving the expected results (Kaushik & Walsh, 2019). Specifically, pragmatism understands reality by considering the practical effects. Pragmatists believe a true proposition must have a positive effect on a practical consequence.

The pragmatic paradigm mainly refers to American pragmatism, with the representative figures being Peirce (1839-1914), James (1842-1910), Dewey (1859-1952), and Mead (1863-1931) (Goldkuhl, 2012; Sundin & Johannisson, 2005). Peirce described the pragmatism maxim, the core of pragmatism, as our conception of the practical consequences of the object is the whole of our conception of the object (Houser & Kloesel, 1992, p. 132). There are three general pragmatic views: first, beliefs and mental states are identified as habits; second, concepts and theories are designed for certain purposes and can be seen as instruments to achieve the purposes or change the existence; third, interpretation is goal-oriented and the content of a

sign follows the way we use it (Legg & Hookway, 2021). Pragmatism allows choosing methodology freely to facilitate achieving the research goal. (Goldkuhl, 2012; Kivunja & Kuyini, 2017; Legg & Hookway, 2021; Sundin & Johannisson, 2005).

Pragmatism acknowledges multiple versions of reality and the choice of a certain version depends on the intents and purposes (Kaushik & Walsh, 2019). Various views of relevance may cause cognitive confusion, among which some conflict with each other while others are consistent. Therefore, it is crucial to select one or a combination of views as the foundation to study relevance. In this sense, pragmatism adequately justifies this study.

Knowledge from pragmatic research is called constructive knowledge with different forms, subsuming explanations, understanding, prescriptive knowledge, normative knowledge, and prospective knowledge (Goldkuhl, 2012). The application of all possible methods in pragmatic research enables rich knowledge representation to achieve the purpose. The value of knowledge should not be limited to current research. It should be able to expand to other fields or general practices (Goldkuhl, 2012). This dissertation contends that the study of data relevance rests on the conventional relevance study in IR, and the relationship study in KO and LOD in the Semantic Web. The expected data relevance theory will be applicable not only to information science but extended to data science, computer science, and other domains involving a great deal of data. It is sufficient to do the study in the pragmatic paradigm.

Furthermore, adopting the pragmatic paradigm in information science was advocated by Hjørland (2002, 2010). He claimed that pragmatism, as the basic epistemological theory (the other three were empiricism, rationalism, and historicism), could be the best paradigm for the philosophy of inquiry (Hjørland, 2013). The concept of relevance in pragmatism is related to

"information about goals and values and consequences both involving the researcher and the object of research (subject and object)" (Hjørland, 2002, p. 269). In pragmatists' views, relevance is both subjective and objective; it needs expert assessment but experts may change their opinions according to the goals; different systems and algorithms are needed to solve different perspectives of the problems (Hjørland, 2010). The dynamic, multidimensional, situational, and contextual attributes of relevance call for an inclusive paradigm in a practical and useful manner.

## 3.2 Research Design

### 3.2.1 Research Method

This study seeks to examine relevance in the KO context and with a different subject (shifting from information to data) in an exploratory manner. It is qualitative in nature. Various qualitative methods are used to address the research questions to ensure adequate investigation.

**Qualitative Content Analysis**

Qualitative content analysis is used to answer "Question 1. What is relevance in terms of knowledge organization?" and "Question 2. What relationships exist in knowledge organization?".

Qualitative content analysis uses predefined categories to code qualitative material to generate typologies and patterns (Schreier, 2012). Building a coding scheme is central to conducting qualitative content analysis (Flick, 2018; Mayring, 2014). The coding scheme is usually created in two ways: either from the literature (theory-based or empirically based) or from the research interest (Flick, 2018).

Specifically, a qualitative-content-analysis-based literature analysis is adopted to address Question 1. There are plenty of conceptualizations/definitions of relevance in the IR context, serving as a good reference. Saracevic's (1975) conceptual framework of relevance is used as the coding scheme: Relevance is the A of a B existing between a C and a D as determined by an E (p. 328). To expand beyond the current schools of relevance definitions, several KO tasks and concepts, including subject analysis, indexing, cataloging, social tagging, aboutness, and warrantare examined.

For the second research question, a meta-typology of relationships in KO is created to examine relationships in detail. The meta-typology consists of three universes: the bibliographic universe, the information universe, and the data universe. For the bibliographic universe, bibliographic relationships are investigated using research publications and bibliographic data models. Hyperlinks and social tags (as links in the social network) are examined in the information universe using research publications and existing taxonomies of relationships. The relationships of the conventional KOS, LOD KOS, and metadata schemas are explored in the data universe using research publications, semantic data model and encoding schemas, and LOD KOS.

**Grounded theory**

Grounded theory is used to answer Question 3: What is data relevance in the KO context? Grounded theory aims to generate a theory from analyzing empirical data (Flick, 2018). It is a qualitative research design used for developing "a general explanation (a theory) of a process, an action, or an interaction shaped by the views of a large number of participants" (Creswell & Poth, 2018, p. 133). By providing insights into the phenomena that contribute to

identifying concepts, grounded theory has the potential for conceptualization (Corbin & Strauss, 2015; Hussein et al., 2014). The concepts, not defined beforehand, are derived from data. Research is an iterative two-way (data collection and data analysis) interactive process until saturation.

Grounded theory avoids preconceptions (Hussein et al., 2014). The patterns and concepts are not preformed or pre-developed. This feature makes grounded theory appropriate to answer Question 3. The iterative interaction between data collection and analysis of grounded theory allows rich data to be assessed and generalized during the inductive processes and eventually generates a data relevance theory.

However, there are potential weaknesses in grounded theory research. First, there are a great amount of data during the iterative data collection and analysis processes, which makes the data difficult to manage. The qualitative data analysis tool NVivo is used to make the process transparent.

Second, grounded theory research cannot be exhaustive. Moreover, there are no rules to determine the category saturation or the sufficiency of the theory (Creswell & Poth, 2018). A suggested discriminant sampling (which will be discussed in the Sampling part) may be used to address the problem (Creswell & Poth, 2018).

### 3.2.2 Task Design

A user study was conducted to investigate Question 3. Participants were not asked to judge whether the given data are relevant or not, or choose relevant documents from a collection,  or search relevant data/documents to answer a question, as did in other relevance judgment studies such as Cuadra and Katter (1967), Lesk and Salton (1968), Barry (1994), Bruce

(1994), Park (1997), Choi and Rasmussen (2002), Maglaughlin and Sonnenwald (2002), Chu (2011), Liu et al. (2020), etc. Otherwise, they were asked to complete a 2-phase offline task to disclose data relevance indirectly. The key design philosophy is to examine what they do, not what they say, since actions speak louder than words (Duffy & Feltovich, 2002). A similar design philosophy can be found in Taylor's study (2013), where subjects were asked to do a series of work tasks (other than just searching for information to answer questions), record the document selection to complete these tasks, and choose relevance criteria from a prepared list.

Participants were provided with two resources (as task datasets) and did the tasks in a naturalistic setting. The naturalism paradigm for relevance studies was advocated by Park (1994) to allow the studies to be situated in the multiple, interrelated, interactive, and nonmanipulable conditions. Naturalistic and experimental settings have been proven to act differently on the use of relevance criteria and on the use of types of relevance criteria (Tang & Solomon, 2001). The two task datasets are: 1) A National Audubon Society field note titled *Atlantic Provinces Region* (https://sora.unm.edu/node/115408), which is free for personal and non-commercial use; and 2) Description of a Chinese rubbing named the Hot Spring Inscription, which was mainly adapted from the research paper *On the "Hot-Spring Inscription" Preserved by a Rubbing in the Bibliothèque Nationale at Paris* (https://www.jstor.org/stable/20185481?seq=1). The reason for choosing the task datasets in different cultures and disciplines was to reduce the possibility of bias from the participants. The probability of a participant being familiar with both resources in depth is low. The participants were also provided with a structured version of each resource for reference (Appendix D).

In the first stage, participants were asked to design two metadata schemes for the resources respectively. They were encouraged to use Linked Open Vocabulary (LOV, https://lov.linkeddata.es/dataset/lov/) to choose the metadata elements as much as possible. If they could not find an appropriate term, they could use terms from other vocabularies or create a new one. After they returned the answers to the first stage tasks, they would move to the second stage.

In the second stage, they were provided with two existing standards for the resources and asked to compare them with the metadata schemes they had designed in the first stage. The two standards were Audubon Core (https://ac.tdwg.org/termlist/), a set of vocabularies designed to represent metadata for biodiversity multimedia resources and collections, and the People's Republic of China Cultural Relics Protection Industry Standard: Metadata for rubbings – Cataloguing rules (translated excerpt). The purpose of the comparison was not to assess or judge the quality of the designed schemes but to study data relevance. The comparison tasks included: 1) Deciding whether to substitute any fields. 2) Deciding whether to add/remove any fields. 3) Deciding whether to adjust the structures of their design. Participants were required to record all the changes and the reasons for these changes in the comparison table (Appendix E). They did not need to make their metadata schemes the same, similar, or consistent with the official ones. They were free to decide whether to change any field based on their own experience. When participants returned their answers to the tasks in the second stage, the offline part of the user study was completed. The whole offline task was anticipated to take 3-4 hours to complete. However, some participants spent more time than expected.

A pilot study was conducted to evaluate all the materials. Adjustments and supplements to the materials were made after the pilot study to guarantee feasibility, consistency, and trustworthiness. In the pilot study, the pilot participant found it was time-consuming to do the offline tasks, which might make participants exhausted, affecting the result of the user study, or even worse, causing them to withdraw from the study. Therefore, I added structured information for both resources as a hint to guide participants to read the original resources effectively and quickly. The pilot study also found that participants might provide wildly divergent responses, not covering key elements that were necessary to inform the study. A comparison table was designed to be used in the second stage to reduce the risk.

### 3.2.3 Interview Design

When the offline tasks were completed, semi-structured one-on-one interviews were used to provide useful, detailed, and in-depth information. Participants were asked about the reasons for their decisions to revise or not to revise the metadata schemes. They were also asked to identify the relationships between the modified fields and the original fields they had designed, and between the fields they had designed and the official schemes. Questions were about their previous experience with metadata and the two resources, their general impression of the study, the most difficult thing to complete the tasks, and their previous projects involved locating data, etc. (Appendix F). Probes, as follow-up questions to clarify or elaborate additional information, were used during the interview to explore in-depth and detailed information (Creswell, 2012). Though covering the same scope, the set of questions for each participant varied and was tailored based on their answers to the tasks. The interview took approximately 60 minutes for each participant.

### 3.2.4 Sampling Method

To answer all the research questions, samples in this study consist of research publications//documents/reports, data models, KOS, LOD KOS, and participants.

**Purposive Sampling for Qualitative Content Analysis**

Purposive sampling is a collection of purposeful sampling strategies, subsuming extreme case sampling, critical case sampling, deviant case sampling, homogeneous case sampling, maximal variation sampling, sensitive case sampling, snowball sampling, typical case sampling, etc. (Flick, 2018; Palinkas et al., 2015). The standard for selecting samples is informativeness, which means the identified cases enable researchers to learn most to explore and gain insight (T. K. Park, 1993). The purpose of purposive sampling is not for theory development (compared to theoretical sampling which will be discussed in the next session) but focuses on gaining insight from the cases (Flick, 2018).

In this study, purposive sampling was used to address Question 1 and Question 2. For Question 1, the samples are research publications related to core concepts associated with relevance in KO. For Question 2, the samples are research publications, documents, reports, data models, and LOD KOS.

Theoretically, with newly and continuously created knowledge, the linkages and semantic relationships are endless, rendering a complete collection an impossible task. In the flourishing data universe, typical case sampling as the main sampling strategy and snowball sampling as the triangulation are appropriate for data collection.

Typical case sampling was  used to understand the key points of a phenomenon by leveraging the representativeness of the typical, normal, or average cases (Palinkas et al.,

2015). The typical cases in this study are the typical KOS, including RDF, SKOS, OWL, and LOD KOS with SPARQL endpoints: UNESCO Thesaurus, Getty vocabularies, and AGROVOC Thesaurus. The limitation of typical case sampling is that it ignores the atypical or abnormal cases that may contribute to the insight.

Snowball sampling recruits cases from the existing ones, like rolling a snowball (Creswell & Poth, 2018). In this study, the cases were "recommended" by the typical cases through the linkages between each other, as a supplement to typical case sampling. To reduce bias caused by snowballing, a domain threshold was used to keep a balance among different domains.

**Theoretical Sampling for the Grounded Theory Method**

Theoretical sampling is the default configuration of grounded theory. It is "the process of data collection for generating theory whereby the analyst… decides what data to collect next and where to find them, to develop his theory as it emerges. The process of data collection is controlled by the emerging theory" (Glaser & Strauss, 1967, p. 45). The samples are theoretically selected to further "elaborate and refine categories" to help generate the theory (Charmaz, 2006, p. 96). It means that besides typical cases, extreme cases might be selected to facilitate an understanding of the phenomenon, extract the categories, and identify the properties (Flick, 2018). Theoretical sampling does not emphasize the empirical generalization of categories to represent the population but the sufficiency of the categories and the emergent theory (Flick, 2018; Rapley, 2014).

Theoretical sampling is emergent, progressive, and inductive (Charmaz, 2006; Rapley, 2014). First, select a homogeneous sample of individuals to study. After the initial categories

emerge, select a heterogeneous sample to decide whether the categories hold true or not (Creswell & Poth, 2018).

The samples to answer Question 3 were participants who work with KOS or metadata, usually called metadata specialists, data engineers, or metadata librarians. The sample size is decided by theoretical saturation (Flick, 2018). When new data no longer provides additional theoretical insights, evidence, or properties of the categories, the collection of data can be stopped (Charmaz, 2006; Creswell & Poth, 2018; Flick, 2018). Because this study requires a specific group of participants, it increases the difficulty of recruiting. Convenience sampling and snowball sampling were used in this study as well.

**Discriminant Sampling Triangulation for The Grounded Theory Method**

Discriminant sampling is seen as a follow-up to supplement theoretical sampling in grounded theory. It is usually introduced in the late stage of the study when the theory has been built but not finalized (Corbin & Strauss, 2015). It is used to decide the sampling saturation and verify the emerging theory (Corbin & Strauss, 2015; Creswell & Poth, 2018). If the sampling is not well saturated, discriminant sampling also serves as a new category compensation for the underdeveloped categories (Corbin & Strauss, 2015).

Individuals, documents, or other samples, different from the initial ones, are deliberately selected to exploit the full potential for verification and compensation. In this study, a data engineer who completed an expert talk was used as the discriminant sample. The data engineer was from industry and had dinstinctive working experience and background compared with other participants, which are introduced in Section 6.1.

## 3.3 Participant Recruitment

I utilized my professional network and open information on web pages to recruit candidates. I browsed or used “metadata librarian”, “metadata specialist”, and “data engineer” to search library or data repository directories.

The recruiting email consisted of a brief introduction of the study, Informed Consent (Appendix B) with IRB approval (Appendix A), a digital recruiting flyer, and a pre-questionnaire (Appendix C) with four questions for screening. From November 2022 to April 2024, I sent out 547 recruiting emails and received 35 responses (including professional networking recruitment), among which 26 candidates showed willingness to participate in this study. 17 participants withdrew from the study or didn’t respond after they received the first stage or the second stage tasks. In the last, 9 participants completed all the tasks and interviews. One participant switched to expert talk to contribute to the study. This expert came from the industry and worked in a different domain, contributing as a discriminant sample. Previous relevance studies demonstrated that data reached saturation with fewer than 10 participants (Barry, 1994).

## 3.4 Data Collection

This study requires different sets of data to address three research questions. Table 5 shows the research questions and the corresponding data types needed.

Table 5: Research questions and the corresponding data types

| Research Questions | Data |
|---|---|
| Question 1. What is relevance in terms of knowledge organization? | Research publications |
| Question 2. What relationships exist in knowledge organization? | Research publications, documents, reports, data models, KOS (including LOD KOS) |
| Question 3. What is data relevance in the | Answers to tasks in two stages, interviews |

| | |
|---|---|
| knowledge organization context? | |

Research publications, documents, and reports cover the topics about: 1) philosophical opinions on objectivity and subjectivity; 2) core activities in knowledge organizing process; 3) core concepts in knowledge organizing process and related to relevance; 4) relationships in knowledge organization including bibliographic, information, and data relationships; 5) models at various levels including bibliographic models and data models; ect. The literature was collected through digital libraries, search engines, and websites.

KOS are the data source of data relationship studies. This study used RDF, SKOS, OWL, and LOD KOS including UNESCO Thesaurus, Getty vocabularies, and AGROVOC Thesaurus. Linked Open Vocabularies (LOV), as an aggregation of various types of vocabularies including metadata schemas, value vocabularies, and ontologies, was used as well for its rich relationships. ALL the LOD KOS and LOV used in this study have SPARQL endpoints. The relationships in RDF, SKOS, and OWL were accessed using their official documents. The relationships in LOD KOS and LOV were queried via SPARQL endpoints.

Each response to the tasks was collected via email as soon as it was completed to proceed to the next phase. The responses to the first stage tasks were metadata elements that the participants designed/selected. The responses to the second stage tasks were modifications made by the participants and the reasons why they made or did not make changes. For fresh memories, participants were recommended to complete interviews within two weeks after they submitted the responses in the second phase.

The interviews were conducted via Microsoft Teams and audio recorded for research purposes. The interviews were transcribed automatically using the transcription function of

Teams and then cleaned manually. All the data were collected between April 2023 and August 2024. Table 6 shows the statistics of the durations of data collection at each phase. The number of days of Stage 1 and Stage 2 refers to the duration from sending the tasks to responding. The number of days of Stage 3 refers to the interval between the completion of Stage 2 and the interview.

Table 6: Statistics of the duration of data collection at each phase

| Stage | Mean (days) | Min (days) | Max (days) | Median (days) |
|---|---|---|---|---|
| Stage 1: Metadata design | 58 | 2 | 145 | 30 |
| Stage 2: Metadata modification | 40 | 4 | 89 | 39 |
| Stage 3: Interview | 6 | 1 | 11 | 6 |

## 3.5 Data Analysis

Phase 1: literature-based analysis. Literature is the principal data in this phase. The expected outcomes are a theory-grounded conceptualization of relevance in KO and relationships being studied in KO. Literature-based analysis addresses Question 1 and part of Question 2.

Phase 2: datasets analysis. For RDF, SKOS, OWL, the selected LOD KOS, and LOV data, the key point is analyzing how relationships are expressed and how the terms and concepts in these vocabularies are linked. The expected outcome is a comprehensive understanding of relationships in KO from the conventional view to the Semantic Web context. Dataset analysis addresses Question 2.

Phase 3: Data relevance modeling. Grounded theory is used to code and analyze the recorded reasons in the second stage and the interview transcripts. This study follows three steps of the grounded theory coding method: open coding, axial coding, and selective coding.

In open coding, sentences are coded as in vivo codes for better contextualization. Themes emerge after open coding. In axial coding, themes and relationships between the in vivo codes are analyzed, and many categories are generated separately for the second and the interview stages. Then the categories are combined for further analysis. In selective coding, relationships among the categories are examined carefully based on the in vivo codes. The relationships are organized into a graph for pattern discovery. The graph shows how the categories are related to each other. Many relationship chains are discovered to support modeling. Through analyzing the incoming and outgoing connections of each category, the core categories are identified, and a hierarchy of all the categories is created. The data relevance model is built based on the core categories and the hierarchy.

NVivo 1.7.2 is used to organize interview data. NVivo is a qualitative data analysis software that can accelerate data processing and searching, facilitate data management, comprehensively display data, increase the quality and transparency of research, and further help to validate the analysis (Flick, 2018). Obsidian is used to draw complex relationships into a graph. Obsidian (https://obsidian.md/) is a personal note-taking application that supports users in taking notes, organizing notes through internal links, identifying and linking relationships, and visualizing notes and their relationships as a graph. The dynamic graph view of Obsidian facilitates identifying the core categories and modeling.

## 3.6 Evaluation

### 3.6.1 Reliability

Procedural reliability (Flick, 2018) and test-retest reliability (Creswell, 2012) are used to guarantee the reliability of this study. Procedural reliability is secured by: 1) The process of

data generation is recorded for return checking. 2) For the user study, structured information about the resources in the first stage and a table listing the required information for comparison in the second stage are provided to the participants, for the purpose of coordinating the tone of the study, and increasing the comparability and interpretability of the data. 3) During the interview, the interview guide and common questions are created first, providing the groundwork for customizing the questions for each participant. 4) In data analysis, NVivo 1.7.2 is used to generate vivo codes, organize the codes into categories, and take annotations, making the research process transparent and manageable. The coding process strictly follows the grounded theory coding steps.

Test-retest reliability is secured by: 1) Coding and recoding. The time interval of coding and recoding is four months, which is a sufficient time to lower the risk of remembering the first round result (Polit, 2014). In the first round of coding, the in vivo codes are terms; in the second round of coding, the in vivo codes are sentences for more contextual information. Then, the in vivo codes are compared to verify consistency and ensure that all important data are captured. 2) Iterative coding is used in each coding round to ensure consistency.

### 3.6.2 Validity

In qualitative research, “there has been a shift from validity to validation and also from assessing the individual step or part of the research towards increasing the transparency of the research process as a whole” (Flick, 2018, p. 546). For the validity of this study, it is important to show the process by which the conclusion is reached in qualitative content analysis and how the theory is grounded.

The validity of this study is guaranteed by: 1) the philosophical foundation is clarified in Section 4.1, building up the specific scope for the analysis in Chapter 4; 2) A meta-typology of relationships in KO is created to outline the framework to make sense the following analysis in Chapter 5; 3) A pilot study makes the materials used in the user study trustworthy; 4) Strictly following the open coding, axial coding, and selective coding steps of grounded theory and the detailed analysis, especially at the selective coding step, make the study transparent and easy to understand. 5) Triangulations are used in research methods, sampling, and data collection to reduce bias.

# 4 Relevance in Knowledge Organization

This chapter explores core activities and concepts related to relevance in knowledge organization, via analyzing how relevance functions in these activities and the relationship between relevance and these concepts.

When information professionals consider how KO impacts users' use of information, relevance starts to take effect. The concept of relevance is scattered and implied in various research areas in KO, but instead of referring to it directly, scholars have adopted other specific terms. Hjørland (2001) noticed the isolation of the studies of aboutness, subject analysis, relevance, topicality, etc., in an IR context. I am going to add more by expanding it to the KO context. In this part, I will examine the major related terms, especially the concepts used for knowledge organizing processes, for further exploration.

## 4.1 A Conceptual Perspective on Relevance as a Foundation

Relevance is an elastic concept that can be interpreted in various ways in different paradigms. A clarification of relevance will establish a foundation for this study.

Every piece of information has some inherent features (e.g., topic, subject, and logic) that are independent of an individual's purpose, background, cognition, values, preferences, etc. These intrinsic features are objective or bear the common cognition (as fact or common knowledge) of a certain group or human society. For example, the sun is a star and occupies the center of the solar system. This statement expresses a widely accepted fact regarded as truth; thus, we can say it is objective. However, before the Copernican Revolution, if we said the sun was the only star and the center of the solar system, no one would consider it an objective fact since geocentrism was dominant then. In this study, objectivity is not Platonic. Objectivity is

based on subjective cognition, just the cognition that is at least acknowledged by most people, not individual cognition or minority cognition. This philosophical view resonates with Foskett's (1982) viewpoint[1]:

> If a document is retrieved from a given collection in response to a given request, its relevance to that request may be estimated by a panel of those skilled in the art, but its pertinence can only be judged by the originator of the request. [...] In other words, relevance is a consensus judgment, pertinence an individual judgment. (p. 117)

It aligns with Gaukroger's (2012) statements as well:

> […] we find a move away from a negative understanding of objectivity as freedom from prejudice or bias, towards the positive idea that objectivity consists in accurate representation. […] While the negative conceptions distance objectivity from truth, the positive one brings the two together in an apparent area of overlap, for accurate representation can be thought of both in terms of truth and in terms of objectivity. (p.9)
>
> […] This is the idea that something is objective if it leads to conclusions which are universally accepted. (p. 10)

Objectivity is not immutable or absolute. While the group cognition of a thing shifts, the objectivity of the thing might change. Notwithstanding, the shift is a lengthy process and will not occur in a short time. In this study, I want to draw attention to the relative stability of objectivity with the inherent features, not changes and evolution in the long river of history.

The inherent features can build up connections of information and the connections bear

[1] This view point was originated from Lancaster: Information retrieval systems. Second edition, Chapter 18, with enrichment by Foskett.

relationships of the inherent features. Thus, identifying these features and their relationships correctly is critical to building an information (or knowledge) network. However, the features and relationships are not equal, which leads to various degrees of relevance. From this point, relevance is entailed in the objectivity features of information. The inherent features play a foundational role in further decisions on relevance. This view is in line with the school of objective relevance.

Using information is the ultimate purpose of information searching, collecting, organizing, storing, and communicating. Considering the information per se and how we use it, concerning only objectivity is not sufficient to study relevance. The other key attribute of relevance is subjectivity, which is evident from the research findings of the past few decades. In previous studies, numerous scholars have underlined how contexts, users, backgrounds, expertise, purposes, tasks, and even emotions can affect relevance judgment. All the frameworks to study relevance discussed in the literature review section encompass the subjective elements. Unlike objective relevance which is not affected by an individual's perception, subjective relevance is decided by individuals in a certain context, based on previous accumulation, with personal preference, and for a specific task.

Objective relevance and subjective relevance work together to help users decide what information meets their needs. However, they don't have an absolute correlation or consistent relationship. Taking the previous example, "the sun is a star and occupies the center of the solar system" is relevant to me because I would like to use it to exemplify my argument. In this specific case, this type of relevance is about utility, which has been studied by many scholars, e.g., Wilson's (1968) second kind of power and Boyce's (1982) nontopical relevance, like

serendipity. However, it is not objectively related to my research. The inherent feature (in this case, the topic) of the sentence is unrelated to the subject matter of my dissertation. In another situation, this sentence reflects the cognition shift from Geocentrism to Heliocentrism, representing a major advancement in astronomy. It is an inherent feature related to the history of science. If someone who doesn't know astronomy much is asked to study the history of science, the person may think it is irrelevant to their task. However, whether the person is aware of its relevance or not, the statement is relevant to the task, objectively, not subject to the individual's cognition. Generally, in this study, relevance encompasses both objectivity and subjectivity. They can be used to decide on relevance together, but they do not have an absolute correlation or a consistent relationship.

## 4.2 Subject Analysis

Subject analysis, also called subject description, subject identification, or document analysis, is the first step of cataloging and indexing. The quality of subject analysis is a key determinant of the quality of cataloging and indexing, and furthermore, the effect of KO and IR. Langridge (1992) asserted "Unless we start with a correct subject analysis, it is obviously quite impossible to proceed to a correct class number of index entry" (p. 64).

Subject analysis is a process of identifying the subjects, the overarching ideas, and/or the central themes of an information entity. It is usually studied with indexing or cataloging as a prerequisite process in information science. Subject analysts usually refer to catalogers, classifiers, or indexers, depending on the area, the follow-up work, and the purpose of the analysis. Subject analysis has been investigated within frameworks that feature different stages, e.g., the widely used two-phase approach consisting of conceptual analysis to determine

subject matter and translation into cataloging or indexing terms, where it is also referred to as conceptual analysis (Albrechtsen, 1993; Holley & Joudrey, 2021; Mai, 2005); three-phase approach including 1. analyzing and determining subject content, 2. selecting the key concepts, and 3. using vocabulary to express the concepts (Hjørland, 1997)[2], where phases 2 and 3 are subject analysis; four-phase approach comprising 1. analyzing text, 2. expressing subjects in own words (natural language), 3. translating into cataloging or indexing vocabulary, and 4. expressing subjects in vocabulary (Chu & O'Brien, 1993), where subject analysis is in the first and second phases. Generally, subject analysis starts with analyzing the information entity and then identifying the subject matter. In this study, I will discuss subject analysis first and then follow-up indexing and cataloging.

Many empirical studies have shown that it is hard to reach high or even medium consistency in subject analysis (manifested by indexing or cataloging inconsistencies) while they have claimed consistency is critical in subject analysis and is a criterion to evaluate its quality at the same time (Chan, 1989; Olson & Wolfram, 2008; Saarti, 2002). However, subject analysts neither identify the same subject matter (in the subject analysis phase) nor use the same descriptors for the subject matter (in the cataloging or indexing phase). Many studies seek to improve or change this unsatisfactory situation. What is the main cause of the inconsistency? Let's go back to the question per se: What do we ask for when we pursue consistency? One expectation is objective analysis with as little bias as possible. The objective analysis is to identify the inherent features and their relationships (since in most cases, listing unlimited

[2] Hjørland cited ISO 5963:1985 Documentation — Methods for examining documents, determining their subjects, and selecting indexing terms, which was withdrawn officially in Oct. 2024 and will be replaced by ISO/AWI 5963 which is under development at present.

inherent features is non-realistic, relationships are useful to determine the key subject matter among numerous subject concepts). It is the inherent features that represent the real/correct subject matter. Thus, subject analysts start from the objectivity of an information entity, identifying the inherent features and focusing on the key features to represent subject matters.

Although Hjørland criticized the concepts of objectivity and subjectivity, he admitted that “subject descriptions may be more or less ‘objective’ (intersubjective) or ‘subjective’” (Hjørland, 1997, p. 34). However, according to Hjørland, the objective features are external features to describe the non-intelligent content, thus he claimed “describe the features of a document that most easily can be understood by many people (i.e., the "objective" features) often lead to a focus on the more trivial features of the document that on the more important ones” (Hjørland, 1997, p. 34). Hjørland’s objective features differ from the inherent features I discussed in Section 4.1 which encompasses the substantive intelligent content, the “informative potential” (Hjørland, 2001, p. 777).

An information entity has many inherent features and a single subject analyst cannot perceive them all, due to an individual’s limitations like culture, knowledge background, purpose comprehension, etc. For example, when analyzing the composition of the human body, there are many inherent features that can be seen as subject concepts, e.g., tissues (skin, bones, muscles…), organs (heart, lungs, liver…), organ systems (respiratory system, digestive system, cardiovascular system…). These inherent features are common knowledge of medicine; thus, they are objective and can be perceived by (at least) most subject analysts. In traditional Chinese medicine, acupuncture points are a human body composition as well. They are distributed along the meridian system and stimulation of them can relieve pain and improve

body circulation. However, the cognition of acupuncture points is mainly limited to the Chinese culture. Academic skeptics have questioned and attacked acupuncture points and their existence for years (Langevin & Wayne, 2018). Therefore, acupuncture points are common knowledge (i.e., objective facts according to the discussion in 4.1) within the Chinese culture but are not inherent features of the human body beyond the scope of this culture. A subject analyst without traditional Chinese medicine background knowledge will probably not use acupuncture points as a subject concept of human body composition. Even if he/she sees this term in the information entity, he/she will probably not choose it as subject matter. At this moment, it is something not about "aboutness", but about relevance that influences the decision. In other words, subject analysis is a process of selection; subject analysts choose among the inherent features actively or passively. Active choosing means they are aware of many subject concepts and choose with conscious awareness. Passive choosing means they lack full awareness and don't recognize they are choosing since they assume they have perceived the entirety of existence. Active and passive choosing can coexist in a subject analysis process. This viewpoint explains why inconsistency in subject analysis exists, although subject analysis reflects the objective matter of an information entity. Relevance, therefore, becomes a mechanism of action in subject analysis. Relevance plays a greater role in the active choosing process than the passive choosing process.

Hjørland(1997) stated "A subject analysis cannot be viewed in isolation from the culture, the environment, and the context in which the analysis is performed, including the individual and collective knowledge interests that the analysis is based upon" (p. 41). Now we can identify at which step of subject analysis the culture, the environment, and the context are engaged.

Some researchers discussed the author's intention/view in subject analysis since information entities were created by authors (Hjørland, 1997; Mai, 2001; Šauperl, 2002). An author's subjectivity, such as views, background, and intention, is expressed as information entities. This study doesn't consider an author's subjectivity separate from information entities and regards it objectively.

Hjørland (2001) argued that "Subject was defined as the informative potential of a document, which is close to the relevance of a document" (p. 777). As an abstract concept, subject is not only close to the relevance of a document, but relevance plays an active role in generating subject matter. When examined in detail, the two types of relevance are different. The relevance (the first one) in Hjørland's statement implies using the subject as a criterion to judge the relevance of an information entity. The relevance (the second one) I just discussed means functioning in determining the subject matter of an information entity, which also can be seen as a criterion. The second relevance happens before the first relevance. To further understand the second relevance, I will introduce the concept of aboutness.

Using Saracevic's framework (1975), in subject analysis, relevance can be defined as:

**DEFINITION 1**: Relevance is the relationship between the inherent features and the subject matter of an information entity determined by a subject analyst.

## 4.3 Aboutness

If we see subject analysis as a process, aboutness is the result of this process, which subject analysis is supposed to identify. The notion of aboutness was introduced to information science from logical and linguistic studies as a property of documents (Fairthorne, 1969). Scholars in information science have considered aboutness and subject synonyms, which can be

seen from their use of subject to explain aboutness (Beghtol, 1986a; Hjørland, 2001; Holley & Joudrey, 2021). Aboutness was also defined as a relationship: “the relation between a work and its subject matter” (IFLA Working Group on the Functional Requirements for Subject Authority Records (FRSAR), 2011, p. 10).

Objectivity is the property of aboutness. What a subject analyst is supposed to identify are the inherent features, not personal understanding. The objectivity of aboutness exists because the information entity it represents exists objectively as well. Beghtol (1986a) differentiated aboutness from meaning (a concept serving for utility) in the sense that aboutness was the intrinsic subject that was permanent but with different meanings used by various users, highlighting the objectivity of aboutness itself. She stated “…the document, itself unchanging, is assumed to possess a fundamental aboutness” (p.85). In this sense, aboutness is something that cannot be determined but can be identified and then manifested by subject analysts. Therefore, in this context, it is inappropriate to ask “whose aboutness” or say “the aboutness of subject analysts” or “the aboutness of users”, unlike relevance asking about whose relevance, because subject analysts, at this moment, do not consider the utility issue, e.g. who will use it? How will they use it? In what scenario and context? From this point of view, aboutness is similar to the objective relevance I discussed in Section 4.1, which is based on the inherent features.

However, we must see that the identification and manifestation of aboutness (as a part of the subject analysis process) inevitably incorporates subjectivity. When a subject analyst identifies aboutness from the inherent features whether actively or passively, factors like culture, value, knowledge background, etc. will play an active role. This view corresponds to

Hutchins's (1978) argument:

> the 'presupposition' approach to 'aboutness' (as we might call the concept I have been describing) does not compel indexers to make any assumptions about the general knowledge or cultural background of a 'common' or 'ideal' reader. The indexers' task is to establish and record the topic or topics which the authors of documents themselves assume their readers should be starting from; they are required only to take account of the authors' own assumptions about their ('ideal') readers; they do not have to formulate their own image of readers. In this way, in theory at least, any 'interference' by the indexer between reader and documents should be minimized. (p. 179)

Therefore, the rationale becomes clear that Ingwersen (1992) differentiated author aboutness, indexer aboutness, and user-related aboutness (The position of user's aboutness has also been used in Functional Requirements for Subject Authority Data, FRSAD (IFLA Working Group on the Functional Requirements for Subject Authority Records (FRSAR), 2011)). According to Ingwersen, indexer aboutness provides a unanimous description of an information entity, dominant over author aboutness. He indicated that indexer aboutness is objective and reflects the "truth" (p. 51). However, he failed to specify the theoretical foundation to explain why author and user-related aboutness were subjective while indexer aboutness was objective since authors, indexers, and users are all human beings. Additionally, considering indexers and users are outsiders other than authors of an information entity, why are indexer aboutness and user-related aboutness different in subjective-objective properties? The subjectivity of manifestation of the objective inherent features (aboutness) provides a way to explain the rationale.

The objectivity of aboutness in nature and the unavoidable subjectivity in practice

explains why Hjørland (1997) advocated that “the aim of the [subject] analysis should not be subjective, for instance, influenced by the person who is performing the analysis. Rather, it should be as precise and objective as possible and should conform as much as possible to its object (pragmatic realism)” (pp. 41-42).

Some scholars have noticed the relationship between aboutness and relevance: Aboutness is not relevance but has something related to it. Sometimes, they are intertwined in the explanations. Maron (1977) proposed three types of aboutness: S-about (subjective about), O-about (objective about), and R-about (retrieval about). The study implies that S-about is a psychological understanding of an information entity, O-about is using terms to represent the psychological status, and R-about is to what extent the retrieved information object is relevant to the terms. Schamber et al. (1990) commented S-about “would seem to come closer to a user-oriented definition of relevance based on meaning than to a system-oriented definition based on topicality” (p. 759).

Bruza et al. (2000) modeled aboutness as a binary relation between information objects and represented aboutness using symbols via logical relationship analysis of information containment, composition, and preclusion. According to their analysis, nonrelevance equals nonaboutness while relevance equals aboutness when studying semantic distance. They claimed that “similarity is a form of aboutness” (p. 1100), indicating that aboutness, like relevance, was not binary but rather a range of degrees, which resonates with Maron’s (1977) argument “instead of saying that an indexer must decide whether or not a document $D_j$ is about (the topic named by) $I_j$, we can allow aboutness to admit of degrees” (pp. 39-40). Svenonius (2000) claimed that “A common fallacy in addressing the concept of aboutness is attempting to

define it in isolation and not recognizing that it is a variable" (p. 46). O'Neill et al. (2017) also shared a similar point: words can have strong or weak aboutness (or subject association) (p. 2471) . They proposed an aboutness coefficient to measure the word-subject relationship. Aboutness can be understood as the relationship between an information entity and its subject and between subjects and terms.

Harter (1992) differentiated aboutness from relevance. He pointed out that we used "about" to mean "on the topic" (p. 603). In contrast, we used relevance to describe the relationship between an information entity and its potential impact (e.g. to support or undermine). Then he argued:

> In the former sense, relevance is objective, and is measured by asking experts in the discipline to make relevance judgments. In the latter sense, relevance is subjective and personal, and must be assessed by the user himself. In both cases, but especially in the former, topicality, or "aboutness," is usually required to be the criterion of relevance, rather than "bearing on the matter at hand." (p. 606)

I have just examined relevance in the context of aboutness. Let me examine it from the opposite perspective by exploring aboutness in the context of relevance. In this perspective, aboutness has been studied mostly as an objective criterion/property of relevance. Cooper (1973) argued that relevance "has to do with 'aboutness' (or 'pertinence', or 'topic-relatedness') and is ultimately definable in terms of logical implication" (p. 92) to compare with utility, suggesting that aboutness represented the objectivity of relevance. Boyce (1982) proposed "relevance would appear to involve a two-stage judgement: first of topicality and then of informativeness" (p. 106), where topicality was about a topic, corresponding to aboutness.

Schamber et al. (1990) differentiated topical relevance from aboutness, where topical relevance was a system-oriented term emphasizing objectivity while aboutness was a user-oriented term emphasizing subjectivity. They specified that "topicality in the sense of matching system artifacts like subject terms is valuable primarily for its practicality […] topicality is clearly insufficient" (p. 59). They agreed that "there is a kind of subjective about […] subjective about would seem to come closer to a user-oriented definition of relevance based on meaning than to a system-oriented definition based on topicality" (p. 59). The objectivity and subjectivity of aboutness have been mapped to system-oriented relevance and user-oriented relevance as instances of different kinds of relevance, respectively. Borlund (2003) emphasized the subjectivity of aboutness more: "The subjective class of relevance may, as a generic concept, refer to the aboutness, usefulness, usability, or utility of information objects […] In Saracevic's subjective topical-like relevance category the concept of topic is understood as aboutness, not contents…"

Intuitively, there is a positive correlation between aboutness and relevance, e.g., Bruza and Huibers (1996) claimed, "If the object is deemed to be topically related, or about, the request, the probability of relevance of the object with respect to the information need is assumed to increase" (p. 383). However, this is not always the case. Their correlation can be embodied in a set of expressions for discussion. These expressions are created with the assumption that subject analysts are aware of all inherent features of an information entity. The affection of subjective cognitive bias is not considered. For example, we do not consider the situation that A is about B, but a subject analyst misunderstands A is not about B and consequently affects the judgment of relevance.

I employ inductive reasoning, deductive reasoning, constructive proof, proof by contradiction, etc. to explore the relationships and the factors working on them.

Expression 1: If A is about B, A is relevant to B.

Expression 2: If A is about B, A is irrelevant to B.

Expression 3: If A is not about B, A is relevant to B.

Expression 4: If A is not about B, A is irrelevant to B.

In the expressions, A is an informational unit like sentences, paragraphs, texts, documents, etc., bearing semantics, topics, and meanings. B is a group of terms or concepts. Both aboutness and relevance are not binary but a matter of degree. To make the analysis more intuitive, I simplify it by examining the extreme cases, i.e., the binary cases. Since it is within the subject analysis process, I only discuss them from the subject analyst's perspective. A user's perspective of relevance is studied later in Chapter 6.

In Expression 1: If A is about B, it means A is on the topic of B, and A and B are related objectively. A must be relevant to B (Here, we do not differentiate the linguistic nuances that we do not use aboutness to represent relevance). Thus, Expression 2 is false if we do not consider the mechanism users play. Aboutness, as an absolute relationship, is a kind of relevance; as a matter of degree, it lays the foundation of relevance.

In Expression 3: If A is not about B, it means A is not on the topic of B, but A might be relevant to B, with some remote connections other than the aboutness relationship. For example, deforestation is not about climate change, but is relevant to climate change. Thus, Expression 4 might be false.

From this point of view, aboutness is sufficient but not necessary for Relevance, as Boyce

(1982) stated “Topicality may or may not be necessary for relevance. It is surely insufficient.” (p. 105) … “The material is relevant but not topical.” (p. 106). This is not incompatible with my statement “concerning only objectivity is not sufficient to study relevance” in the previous section, because the discussion here is within the objectivity context.

Aboutness and relevance share some common features: they are both matters of degree and involve objective and subjective attributes. Aboutness is an inherent property of an information entity and is independent of relevance. However, be-about can be seen as a criterion of relevance. Aboutness is one of the properties of objective relevance, where under certain circumstances, the correct aboutness is the condition of relevance. Irrelevance cannot necessarily undermine aboutness. If we see aboutness as an aboutness relationship, relationships with a broader spectrum can serve as another foundation, which is examined in detail in Chapter 5.

## 4.4 Indexing

Subject analysis is the first step of indexing and cataloging, identifying the aboutness/subject of an information entity for further investigation. When subject analysis is seen as an objective process to identify the inherent features, indexing and cataloging are considered more subjective due to the inner criteria applied by indexers and catalogers. Borrowing the terms from Mai (2005), subject analysis can be understood as document-oriented, while indexing and cataloging are user-oriented (similar to requirements-oriented discussed by Albrechtsen (1993)). User-oriented is not identical to subjective; it is a factor to consider when determining index terms, and implying subjectivity when utility, psychological status, etc. (or pertinence used in previous studies) are considered for/by users.

Subjectivity is inevitable in indexing. As a mechanism, relevance plays an important role in the indexing phase when certain situations are involved, e.g., users' situations. Bade (2007) stated that indexing and subject analysis were identical to the process of users' relevance judgment, which was situational and contextual. Hutchins's (1977) view supplements on how relevance is related to indexers and users:

> For the reader a document may be relevant if it contributes in some way to the 'solution' of questions that interest him; for the indexer a document may be relevant if it discusses the topic of interest, whatever the value of its contribution on a particular occasion. For the reader, relevance is a function of his current interests and his personal 'state of knowledge'; for the indexer, relevance is a function of the place of the document in the current 'state of knowledge' as a whole, in Popper's world of autonomous 'objective knowledge'. (pp 34-35)

Indexers use more objective criteria to judge relevance to serve their purpose, which is not answering questions or solving problems as users, but describing information entities in controlled terms to facilitate users' information retrieval. The object of relevance in indexing might be multi-level, which means indexing is not as simple as its definition suggests. Additionally, it is widely recognized and proven by empirical studies that full indexing consistency is the ever-receding horizon (Chan, 1989; Saarti, 2002). Consequently, researchers seek approaches to reduce the subjectivity of indexing, such as creating guidelines for the process (Redigolo et al., 2022) and using standards. Analyzing the mechanism by which relevance operates during the indexing process will be helpful to address this problem.

The main task of indexing is to assign index terms to represent the aboutness/subject

identified in the subject analysis phase by professional indexers. The subjectivity comes from different aspects. The first aspect is the indexers themselves. It is their knowledge background, experience, cognitions, considerations, preferences, etc., that decide, to some extent, issues like what terms and how many terms they will use or whether they will add broader terms or narrow terms. The other aspect is the assumptions that indexers will make. It is generally accepted that indexers, though not communicating with users directly, should not be isolated from users (Albrechtsen, 1993; Bade, 2007; Hutchins, 1978; Mai, 2005). From a pragmatism perspective (Hjørland, 1992), thinking about users is a good practice. Hjørland (2001) argued that "the best subject analysis is the one that makes the best prognosis of the future use of the document" (p. 776). Indexers hold certain positions and make assumptions that are critical in indexing. Indexers will probably ask these questions to help them determine the terms: Who are the potential users of the information entity? What terms will the users use to search for the information entity? What are the users' information needs? What do the users expect from the information entity? How will the users use the information entity? For what purpose? (Albrechtsen, 1993; Bade, 2007; Hutchins, 1978; Mai, 2005). Indexers may assume various contexts and scenarios, possible problems at hand and questions asked to address the problems, the knowledge base and cognition status of the potential users, etc. Could indexers make assumptions based on imaginary users with fabricated needs in illusory scenarios? The answer is yes, but it is limited to an indexer's own experience and levels of cognition. For example, for the book with the title "Basic Generative AI: Beginner's Guide to Artificial Intelligence, ChatGPT and Machine Learning, Practical AI Applications", an indexer with relevant knowledge will probably assume the potential readers are novices not savvy in AI because the

title explicates this, who might not know ChatGPT represents a type of Gen-AI, the relationship between LLM and ChatGPT, machine learning is an approach to AI, who might use this book to get a brief idea of what ChatGPT is and its technology foundations, and further use this book as a starter to enter into the field or see what ChatGPT can do for them. However, there might be different groups of potential users, e.g., instructors who might use it as a textbook, coming with different follow-up assumptions. It might be an endless assumption list that indexers must make choices. Then it runs into a pragmatic defect: "if it is assumed that a given document is to be included in relation to all its possible uses, then this would give rise to all too numerous repetitions or multiple classifications" (Hjørland, 1992, p. 180). At this point, indexers employ relevance unconsciously. Any assumption will come into play when bound to a certain relevance. Indexers must make at least two types of choices: the first one is choosing from the assumption list to decide the limited assumptions they can serve and make sure they do not conflict with each other; the second one is determining what inherent features are central to these assumptions. Correspondingly, there are two types of relevance. Using Saracevic's framework, relevance in indexing can be defined as:

**DEFINITION 2**: Relevance is the degree of correlation between the assumptions of users' information needs as determined by an indexer.

**DEFINITION 3**: Relevance is the measurement of the bearing between inherent features and the selected assumptions as determined by an indexer to identify subject matter.

After indexers make assumptions and match inherent features bearing the possible answers to the assumptions, the remaining task is mapping the inherent features to controlled vocabularies. The mechanism that relevance works at this point is to decide what relationship to

apply to select the terms when the matching relationship doesn't work. For example, when both Gen-AI and ChatGPT do not exist in ERIC (Institute of Education Sciences) Thesaurus, what terms does an indexer use to describe ChatGPT, the broader term Artificial Intelligence, or the related term Natural Language Processing or Robotics? These semantic relationships with different semantic distances serve as a measurement of how relevant a controlled term is to an inherent feature. In essence, mapping inherent features to controlled vocabularies is also a process of determining relevance. This is the third type of relevance in indexing:

**DEFINITION 4**: Relevance is the semantic relationship between the inherent features and controlled vocabularies as determined by an indexer to select index terms.

Automatic indexing is another form of indexing aside from controlled vocabularies used by indexers or author-provided keywords. It relies on automatic techniques to deal with exponentially increased documents and automatically generated index terms, of which the algorithms still need human input. Definitions 2 to 4 are applicable to automatic indexing as well.

## 4.5 Cataloging

Cataloging is normally considered a sequential combination of descriptive cataloging and subject cataloging, followed by classifying (Joudrey et al., 2015). Referring to Wilson's (1968) two kinds of power - knowledge and bibliographic control, catalogers have power over information entities. How catalogers describe resources, assign terms to the resources, determine relationships of the resource, and classify them into classifications, to some extent, shape knowledge understanding and cognition of users, and can affect relevance judgment as well. Although based on subject analysis, catalogers have a wide scope of discretion when they

apply different values and principles. Relevance, thus, plays an active role in the cataloging process in determining the terms and classifiers.

Cataloging and indexing are the subsequent tasks of subject analysis. Compared to indexing, a metadata schema and a classification scheme must be used in cataloging, which means assigning terms/values/classifiers to metadata fields. Indexing is a much closer concept to subject cataloging. They are similar in several ways, such as subject analysis is the required precondition, terms are used to represent the aboutness of an information entity, they are both carried out by information professionals, etc. The intrinsic motivations of cataloging and indexing are identical, helping users find relevant information to address their information needs and solve their problems. In this section, I will analyze the aspects of cataloging that differ from indexing.

The process of cataloging is the process of decisions and selections. Catalogers must make many decisions. Some decisions involve no relevance judgment, such as descriptive cataloging, where some answers are right there, shown directly in the information entity. Other decisions need relevance judgment using various criteria, especially in subject cataloging with classifying, such as whether a subject term is appropriate (relevant) for targeted users, what terms to use if no subject headings are found, etc. However, a thorough inquiry of these criteria will not be helpful in this study since it gets entangled easily in an analytical spiral, occupied with numerous trivial criteria while losing sight of high-level principles. Exemplification of relevance along the cataloging process is a sound way to study the interaction between relevance and cataloging.

IFLA (International Federation of Library Associations and Institutions) established

thirteen cataloging principles, which were about convenience of the user, common usage, representation, accuracy, sufficiency and necessity, significance, economy, consistency and standardization, integration, interoperability, openness, accessibility, and rationality (Galeffi et al., 2017). These principles describe the "ideal" cataloging results, which can be seen as the problems catalogers target to address. To reach the ideal status, catalogers will use these principles as criteria during the cataloging process, and these principles are utilized to judge which terms to choose, etc. Thus, they can be seen as relevance criteria[3]. However, some principles are more impactful, while others are less significant. These principles suggest that the combined effort of objectivity and subjectivity affects the cataloging process. Taking this further, relevance in cataloging can be interpreted as practitioner relevance with a specific consideration of the user's relevance that a cataloger assumes.

**Practitioner relevance**

Practitioner relevance is based on a combination of inherent features of information entities and the best practice of cataloging, including the recommended standards, principles, data models, and frameworks. It supports the decisions that catalogers make during the cataloging process. Some inherent features can be used directly to answer the cataloging questions, e.g., what is the title? Who is the creator? When is the publishing date? For these examples, the relationship between these inherent features and the information needs of catalogers (represented as the attributes or descriptors) is of an extreme degree of relevance. When we say cataloging, we usually mean cataloging an entire resource, but the assumption

[3] In the same way, other guidelines like the practical cataloging frameworks, standards, or ethical codes can serve as relevance criteria as well.

doesn't exist in indexing. For example, it doesn't make much sense to catalog a paragraph of a book. Thus, some features of an information entity are important for cataloging but may not apply to indexing. For catalogers, relevance can be reflected by the structural information of a resource, i.e., certain fields are naturally of a higher degree of relevance for catalogers. An empirical study conducted by Šauperl (2002) found that although slight variations existed for each individual, catalogers used the fields of title, table of contents, abstract (including summary on the title page), cover page and title page, and pictures and illustrations with captions, which were most frequently cited to do subject cataloging. My viewpoints resonate with Bade's (2007) argument: "In a library catalogue every element in a MARC record represents the interpretation and judgement of a human being, the subject headings being in fact the cataloguer's assertions of relevance according to a defined taxonomy" (p. 836).

Other inherent features, other than the ones with "obvious" relevance, have the potential for relevance judgment. From a communication perspective, once a genre emerges from the text and context relationship, it serves as a shared media among authors, catalogers and users to guide the understanding of the resources and thus can be used to judge relevance (Rauen, 2009).

Besides the direct functioning on relevance, inherent features can serve as connections to transmit relevance. Šauperl (2002) reported catalogers used associated bibliographic records with the same subject headings as a reference to confirm/evaluate the candidate subject terms and determine which subject headings were relevant to their tasks. In the same study, the property author was used by catalogers as a connection of different works to help determine subject headings and classifiers since the publications of the same author were probably in the

same domain. When assigning a classifier, catalogers referred to the classifier assigned to the resources with the same subject headings. These shared features provide important information to help catalogers determine the most relevant terms and classifiers.

The best practice is the agreement reached in the information professional community about what to do and how to get the best result. Generally, best practices cover ethical codes, values, objectives and functions, ways to consider, approaches and techniques, models including attributes and relationships, etc. As I argued, best practices could be seen as the tasks that catalogers would like to complete. Accordingly, the cataloging process can be reinterpreted as a process of determining whether cataloging behavior is relevant to the best practice, i.e., whether the determination of terms and classifiers is relevant to the best practice. For example, in IFLA Statement of International Cataloguing Principles (ICP), there are two principles: "common usage" means terms used in cataloging "should be in accordance with that of the majority of users"; "representation" means terms used "should represent a resource as it appears. […] If this is not feasible, the form commonly used in reference sources should be used" (Galeffi et al., 2017, p. 5). The two principles, "common usage" and "representation," can be seen as the objective relevance criteria (according to my explanation in Section 4.1, objectivity is not Platonic but based on the common cognition) for catalogers to determine the terms to represent the inherent features. "Consistent description" is another relevance criterion for catalogers, but users do not care about or are not aware of it.

Practitioner relevance is the relevance of cataloging behavior itself. Using Saracevic's framework, it can be defined as:

**DEFINITION 5**: Relevance is the relationship between the inherent features of an information entity and the descriptions, subject terms, and classifiers assigned by a cataloger to fulfill the cataloging tasks.

**User's relevance**

User's (equals to end-user in this section) relevance is not a parallel type of relevance of practitioner relevance but a critical specific consideration that can be seen as a subtype of it. User needs are a critical factor in several cataloging principles, guidelines and models, e.g., IFLA ICP (Statement of International Cataloguing Principles), IFLA Functional Requirements for Bibliographic Records (FRBR), IFLA Functional Requirements for Subject Authority Data (FRSAD), IFLA Functional Requirements for Authority Data (FRAD), and IFLA Library Reference Model (LRM). In FRBR, "Intended Audience" is an attribute of a work (IFLA Study Group on the Functional Requirements for Bibliographic Records, 2009), suggesting the relevance of a resource to users. In ICP, the most important cataloging principle is "convenience of the user", which means "Decisions taken in the making of descriptions and controlled forms of names for access should be made with the user in mind" (Galeffi et al., 2017, p. 5). It means that the more relevant terms are the ones that are more "comprehensible and suitable for the users" (p. 5). These are user-oriented principles and lead to user-oriented relevance judgment in cataloging.

Hoffman (2009) asked "If users' needs cannot be met in cataloging practice and users are not the focus of cataloging research and standards, then how should cataloging help users?" (p. 636). However, users' relevance in cataloging is not decided by users but by the assumptions of catalogers, which makes it a challenging task. Catalogers cannot focus on the needs of individuals since their needs are diversified (Hoffman, 2009). Rather, it is feasible to address the

needs of groups of users if catalogers focus on the common features. The connotation of users' relevance differs from the relevance in information retrieval, although the purposes are the same: to meet users' information needs.

The term user tasks is used to define users' needs in cataloging. User tasks reflect "functions of the catalog" and the purpose of cataloging, stressing how users can benefit from catalogs (Joudrey et al., 2015, p. 998). The generic description of user tasks in the guidelines and models is shown in Table 7 (IFLA Working Group on the Functional Requirements for Subject Authority Records (FRSAR), 2011).

Table 7: User tasks as defined in ICP, FRBR, FRAD, and FRSAD[4]

| ICP | FRBR | FRAD | FRSAD |
|---|---|---|---|
| Find | Find | Find | Find |
| Identify | Identify | Identify | Identify |
| Select | Select | | Select |
| Obtain | Obtain | | |
| Navigate | | | Explore |
| | | Contextualize | |
| | | Justify | |

In the four frameworks, user tasks are mapped to certain attributes and relationships to support fulfilling the tasks. Selection, defined as "to determine the suitability of the resources found, and to be enabled to either accept or reject specific resource" (Riva et al., 2017, p. 16), is a task that is closely related to relevance and relevance judgment. The corresponding attributes and relationships can provide sufficient information to support relevance judgments of users, e.g., subject, language, technical requirement, and access and use restrictions (IFLA Study

[4] Table 7 is adapted from Figure 6.2 in IFLA Working Group on the Functional Requirements for Subject Authority Records, (FRSAR), Zeng, M. L., Žumer, M., & Salaba, A. (2011). Functional requirements for subject authority data (FRSAD) a conceptual model. Berlin; New York : De Gruyter Saur, with an added part of ICP.

Group on the Functional Requirements for Bibliographic Records, 2009; Riva et al., 2017). In FRAD, contextualization, interpreted as allowing users to place an entity in context and clarify the relationships between them, is a preliminary step to the following relevance judgment (IFLA Working Group on Functional Requirements and Numbering of Authority Records (FRANAR), 2013).

In practice, studies found that certain cataloging properties were more important than others. Noh (2011) reported title, subject, description, and creator were more important than other properties for users when searching for web-based reference resources. Lee et al. (2015) found price, platform, genre, series, and style were the top five properties according to users when seeking new video games. Hu et al. (2018) observed work type, title, and creation date were the most important properties for users to search and browse murals, a type of cultural heritage product. Pauman Budanović et al. (2021) reported that the importance and necessity of different properties for cataloging from catalogers' and users' views had common features but were not identical. They found that users didn't know the meanings of certain properties, e.g., ISBN, edition, which were important in practitioner relevance judgment. Some properties are necessary for users to judge relevance other than the fields about aboutness. For certain groups of users, properties such as binding, print size, content list, etc., were important to evaluate potential use (Pauman Budanović et al., 2021). Format is a critical field for visually impaired users to assess the accessibility (Chapman, 2007). Descriptions "such as whether it is illustrated, what equipment is needed to use it, or that its mode of access is through the web" are helpful for some users to judge relevance (Joudrey et al., 2015, p. 15).

Some concepts must be clarified before defining the user's relevance in cataloging.

User's real relevance is: the relationship between the properties and attributes (as values of properties) and their information needs determined by users. User's relevance assumed by catalogers is: Whether users use certain properties and attributes assumed by catalogers to judge their real relevance. Therefore, user's relevance in cataloging can be further explained as whether the properties and attributes assumed by catalogers match the properties and attributes used by users to meet their information needs. Using Saracevic's framework, user's relevance in cataloging can be defined as:

**DEFINITION 6**: Relevance is the relationship between the properties and attributes assumed by catalogers and used by users to meet their information needs.

## 4.6 Social Tagging

Tagging is similar to indexing when we see it as assigning terms to an information entity. However, unlike indexing, tagging neither requires using controlled vocabulary nor being carried out by information professionals. Tags can be created by information professionals, content creators, and users (Mathes, 2004). Correspondingly, tags are specified as intermediary descriptors, author keywords, and user tags (Kipp, 2005; Lu & Kipp, 2014). Similarly, Van Setten et al. (2006) distinguished tags created by creators, catalogers, social annotation (by users), and systems. Rafferty and Hidderley (2007) examined three types of indexing: expert-led indexing, author-generated indexing, and user-oriented indexing. Tagging is usually associated with social media, which is called social tagging. Social tags are considered specifically user tags (generated by users) in many studies, users refer to end-users and creators inclusively since users contribute to the content in the social tagging context.

Trant (2009) proposed a framework to study social tagging, including tagging,

folksonomy,  and social tagging systems. In this framework, tagging stresses the process of assigning terms to information entities of both content users and creators; folksonomy focuses on the collective tagging result: the collection of tags; social tagging systems address the relationship between users, tags, and resources, e.g., users use tags to retrieve information to solve their problems. I use this framework to analyze relevance at three levels: tagging as a behavior, folksonomy as a result, and social tagging systems as an implementation.

**Tagging as a behavior**

Tagging is a process extending beyond a subject analysis of what an information entity is about. For example, the time and task management tags such as “todo” and “new” found by Kipp (2005) are not about aboutness. Users exhibit distinct characteristics via tagging. Diversified tags with such as commentary and descriptive qualities provide valuable potential for resource discovery (Rafferty, 2018).

Tagging behavior has been studied as a relevant indicator to show a user’s interactivity level, information organization level, and interest in a content for user modeling for an adaptive system (Carmagnola et al., 2007). In other words, tagging shows the relevance between tags and their profiling. Kehoe and Gee (2011) studied social tagging from a corpus linguistic perspective on aboutness. They compared the keywords showing in the text (seen as the aboutness of the text) and the tags assigned by users but not showing in the text (seen as the “about indicators” in their study, but relevance indicators in essence). They grouped the not-showing-in-text tags into nine categories, such as evaluative words, attempts to categorize, attempts to build topic hierarchy, metadata, etc., indicating that the tagging activities were more than subject analysis. Kipp et al. (2017) identified 11 motivations of user tagging activities,

encompassing affective reactions, opinion on photos, interest in subjects, contribution to description, knowledge sharing, improving findability, social network, appreciation, personal use, personal relationship, and archive or preserving purpose, in which affective reactions, opinion on photos, social network, appreciation, personal use, personal relationship, and archive or preserving purpose are probably not about subject analysis. Each motivation provides a justification as to why the tags are relevant/assigned to an information entity. These intentions or tasks driven by the intentions provide the context for users to decide the words they use to tag. For example, the tag “read later”, which is in the personal use context, shows the information about the following action to the information entity. From the perspective of relevance, “read later” is relevant to personal use; each motivation leads to a type of relevance. Tags reflect the bias caused by intentions and the cognition, knowledge background, and culture behind intentions. These types of relevance may be further categorized into topical relevance, situational relevance, psychological relevance, user-oriented relevance, etc.

The source of tagging (the subject the tags are created by) can be embedded into this level to study relevance. Studies show that authors assigned different keywords from tags generated by users, and author keywords and user tags have different impacts on recall but similar performance on precision, indicating that tags from creators and users imply different relevance. (Kipp, 2005; Lu & Kipp, 2014). For a content creator, more latent features of an information entity might be disclosed according to the creator’s intention. In social media, the scope of tags is unlimited. Creators can use any terms to describe the work they create. For example, tags like fisheye, rokinon 8mm, and pentax might be used to describe a photo, showing the intention that the creator targets special groups by listing specific photography

techniques. It is difficult to identify these latent features until end users see them in text. Users always assess an information entity from what they observe directly and make assumptions according to their situations. Users might assign a tag “example of unnatural expression” to the same photo. Users’ information needs (intention) do not necessarily match the creator’s intention. In this example, the user’s motivation is personal use, while the creator’s motivation is social networking in a professional group. In this sense, the creator’s tags are irrelevant to the user’s need. This belongs to the interaction scope, and I discuss it in the social tagging systems section. According to the previous discussion, when creators and users have different intentions, they are in different contexts, thus, they are associated with different types of relevance and use different criteria to judge relevance. Motivation plays a critical role in assigning relevant tags. Relevance in the tagging process can be defined as:

**DEFINITION 7:** Relevance is the relationship existing between the tags and the tagging intentions as determined by users (including content creators) in the tagging process.

From the linguistic view, tagging serves three functions: topic-marking function, tagging processes (verbs), participants (nouns), and circumstances (prepositional or adverbial phrases); interpersonal function, tagging users’ interactions and attitudes; and textual function, organizing the text using tags, according to the systemic functional linguistics (SFL) theory (Halliday, 1978) and its interpretation by Zappavigna (2015). In the context of knowledge organization, the topic-marking function is not limited to tagging topics (or aboutness) but includes motivations, as interpreted as processes, participants, and circumstances by Zappavigna. The interpersonal function can be interpreted as social interactions, including attitudes, enabled by tags. If attitude is seen as motivation, it can be categorized as a topic-

marking function. Therefore, the interpersonal function mainly focuses on social interaction, which is addressed in the social tagging system section. The textual function represents how tags can be organized and structured (in KO as folksonomy), "the relevance function of language, without which the other two do not become actualized" (Halliday, 1978, p. 123), which is discussed in the following part.

**Folksonomy as a result**

Tags are the result of tagging activities. With various motivations for tagging behavior, the range of tag typologies is extensive. Gupta et al. (2011) summarized 11 types: content-based tags, context-based tags, attribute tags, ownership tags, subjective tags, organizational tags, purpose tags, factual tags, personal tags, self-referential tags, and tag bundles. These types suggest that the relationships embedded in a collection of tags are novel compared to controlled vocabularies that have a limited scope of relationships (Kipp & Campbell, 2010).

Folksonomy is an aggregation of tags. Tagging the same resource by different people using their own terms is called broad folksonomy; narrow folksonomy means tags assigned for a resource by one person, usually the creator(s) (Vander Wal, 2005). In both types, folksonomy represents "power to the people" since "the term-significance relationship emerges by means of an implicit contract between the users" (Quintarelli, 2005), which can be seen as a variety of one of the two kinds of power proposed by Wilson (1968). It is widely believed that folksonomy is flat, without strict hierarchical structures and explicit semantic relationships as taxonomies, classifications, thesauri, etc. (Mathes, 2004; Quintarelli, 2005). This does not mean folksonomy is a scattered mess; conversely, it implies rich relationships and encompasses various possibilities to interpret and apply. The numerous motivations and the types of tags confirm this

as well. For example, tag bundles, one of the tag types, create hierarchies in folksonomy (Gupta et al., 2011). Furthermore, some scholars consider folksonomies as social or user-generated or grassroots community classifications, and the hints can also be found at the origin of the term (Mathes, 2004; Quintarelli, 2005; Vander Wal, 2007). The classification view suggests that there are relationships and structures in folksonomy yet to be discovered.

Empirical studies seek to discover hierarchical features, relationships, and categories in folksonomies. For example, studies on hierarchical categorization schemes (taxonomy) for a folksonomy to understand the structural feature of tags (Cantador et al., 2011; Heymann & Garcia-Molina, 2006; Luo et al., 2018), tag-tag co-occurrence and similarity (Cattuto et al., 2007; Kim et al., 2012; Kipp & Campbell, 2006), and tag-tag relationships (Kome, 2005). In this sense, relevance in a folksonomy can be defined as:

**DEFINITION 8:** Relevance is the relationships between tags in a folksonomy.

**Social tagging systems as application**

Social tagging systems emphasize the interaction and relationships between tags, resources, and users. They target problems in the application scope, such as how to use tags to facilitate users in finding relevant resources and how to help taggers assign appropriate tags to resources. Wetzker et al. (2010) build up the links among them as “the semantics of tags based on who applied them and by looking at the items they have been assigned to” (p. 72). Tags per se “can be viewed as setting up an attributive relationship between the tweet as a tagged token and the label as its type” (Zappavigna, 2015, p. 281). Using RDF, tags can be interpreted as the predicate to link resources and users (Marlow et al., 2006). In the resources-tags-users triple, tags are surrogates or attributes of resources, the intention and needs expression of users, and

the relationships between resources and users. In this sense, regardless of what functions, motivations, or intentions tags represent, they can be seen as the values of properties that users use to describe a resource. Resource meanings and user needs converge at one point. In this context, the concept of tag relevance has been proposed to address the application issue.

Vig et al. (2009) explained tag relevance as "the degree to which a tag describes a given item" and tag preference as a measure of "the user's sentiment to the given tag" (p. 48). The definitions separate the triple into two parts: resources-tags relates to tag relevance while tags-users relates to tag preference. They further explained, "tag relevance and tag preference are orthogonal to one another: the former is item-specific and the latter is user-specific" (p. 48). Li et al. (2016) argued that "tag relevance […] estimating the relevance of a tag with respect to the visual content of a given image and its social context" for image resources (p. 3). To make the definition generalized, tag relevance is the relationship between a tag and the content of a resource and its social context. Social context here is an inclusive concept, referring to any context related to users and creators. When using tags to help users find relevant information, the context information can help users to determine the relevance of the tags, such as "who provided the annotation (the person and his background) and the intent of the person who provided the annotation." (Van Setten et al., 2006, p. 8). Kotkov et al. (2021) differentiated object-centered tag relevance and personalized tag relevance. Object-centered tag relevance is the relationship between a tag and an information entity. This type of tag relevance is similar to the relevance in the tagging process and Definition 1 of relevance in Section 4.2. It also represents the first half of the triple: resources-tags. Personalized tag relevance considers the user as a factor and is the relationship between a tag and an information entity for a specific

user.

Zappavigna (2015) analyzed the effect of the general-specific spectrum of tags on relevance to end users. Too specific tags can hardly reach a broad audience, while too general tags have difficulty returning effective results. Taggers need to find an appropriate position in the spectrum to help users find resources.

Tags cluster resources with the same attributes. One resource can be "classified" into different categories for the assigned tags. Thus, the relevance of tags transitions to the relevance of resources. The evidence of the empirical studies using tags to calculate the similarity between resources also supports this view (Wetzker et al., 2010). In the resources-tags-users interactions, tags serve as the intermediary, providing search points and categories for searching and browsing the resources. Leveraging the relevance of tags (referring to Definition 8), users are provided a list of related tags to expand their search and browsing, where they can easily reach a huge resource pool that is relevant to their interests. After users click to expand to another category (tag), a new list of relevant tags will be provided for their further consideration. Users' interests, thus, will be strengthened or weakened, which is also a criterion for them to proceed, go back, or stay in a category to browse. In this repeated interaction, relevance is the unifying thread and forms a relevance chain to move the interaction forward until users find satisfying resources. This is similar to the principles of Linked Data: tags are the predicate (attribute, value of a property) to link resources (subjects) and users (objects). Relevance in this context can be defined as:

**DEFINITION 9:** Relevance is the relationship of tags existing between an information entity and information needs as determined by users.

## 4.7 Warrant

Warrants are "logical reasons for believing and accepting premises" (Martínez-Ávila & Budd, 2017, p. 708). In the KO area, they are the rational justification, limits, and verification of decisions about inclusion or exclusion of terms and classes and how they are structured in a KOS (Beghtol, 1986a; Tennis, 2005). From a high-level perspective, warrant is a methodological component of constructing KOS, paralleling design, data collection method, sampling, etc. (Martínez-Ávila & Budd, 2017). Specifically, a warrant is the core factor to be considered when developing a new KOS and using KOS to represent and classify information entities. Warrants accompany the processes of indexing and cataloging, where warrants are the justification to select specific terms to represent particular concepts to describe information entities and their subjects (Smiraglia, 2016). In the constraint theory proposed by Fraser (1978), selecting index terms was a process constrained by literary warrant, user warrant, and logical warrant.

### Types of Warrant

Since the first concept of literary warrant was proposed by E. Wyndham Hulme (1859-1954) more than a century ago, various types of warrants have been studied (Barité, 2018; Beghtol, 1986b; Bullard, 2017). Barité (2019) summarized a typology of warrants containing over 20 types. I will analyze the representatives in this study to show the connections between warrant and relevance.

Literary warrant refers to the view that terms used in KOSs originate from the literature in related subject areas, and thus, classifying these terms and the information entities

represented by them is based on the logical relationships shown in the literature (Barité, 2018; Beghtol, 1986b). Semantic warrant, a comprehensive notion of literary warrant, scientific/philosophical warrant, educational warrant, and cultural warrant, is a semantic rationale, providing "the principal authorization for supposing that some class or concept or notational device will be helpful and meaningful to classifiers and ultimately to the users of documents" (Beghtol, 1986b, p. 111). User warrant means that choosing the terms and classifiers which are used by users, not from authorities or existing literature (ANSI/NISO, 2010). Folksonomy can be used as a source of user warrant in this sense (Merholz, P. "Metadata for the masses" cited in (Rafferty & Hidderley, 2007)). Martínez-Ávila and Kipp (2014) proposed the concept of market warrant, "the new criterion for the inclusion and arrangement of classes based on current market demands and interests" (p. 385), via analyzing the categories of the Book Industry Standards and Communications Subject Headings List (BISAC). Epistemic warrant "is a means that can assist information seekers in locating content that closely matches the desired purpose of searches" (Martínez-Ávila & Budd, 2017, p. 708). Dobreski (2020) identified conceptual warrant and terminological warrant. Conceptual warrant justifies the "choice of relevant concepts" while terminological warrant justifies "the process of choosing a suitable term to name an entity" (p. 63). The two types of warrants are applied in the subject analysis phase and the indexing/cataloging phase, respectively.

Generally, the logic of these types is similar to the types of relevance. From the perspective of various orientations, the logic-oriented approach leads to literary warrant and scientific/philosophical warrant and correspondingly logical/topical relevance, the cultural-oriented approach leads to cultural warrant and correspondingly situational relevance and

socio-cognitive relevance, and the user-oriented approach leads to user warrant and correspondingly user-oriented relevance. The orientations decide the epistemology and values of warrants and relevance.

**Functions of Warrant**

In the KO area, warrants are commonly used for KOS development, knowledge classification, and knowledge representation, including indexing and cataloging. Warrants function as the principles, guidelines, or instructions that information professionals bear in mind when making decisions.

On a macro level, warrants can be seen as the atmosphere in which information professionals are immersed when they make decisions, like the "invisible support" Barité (2019) mentioned. For example, author-based indexing is a literary warrant approach, while the warrant of democratic indexing "comes from the constructive interpretation of its users." (Rafferty & Hidderley, 2007, p. 408). Objectively and explicitly, warrants provide the rationale for decision-making. Subjectively and implicitly, they subtly shape the mindset of a decision maker. Warrants establish the context and lay the groundwork of relevance. In other words, contexts/situations of relevance may be constrained by warrants. The concept of situational relevance is an exemplary illustration, in which warrants can be a boundary of the situations. My view is in line with Begthol's (1986a) argument:

> "In the case of classifying documents, these elements [elements of the text which are found important by the readers] of contextual relevance would be contributed by the cultural and classificatory warrant of the classification system and by the particular library for which the classifying is done." (p. 104)

And with Barité's (2019) point as well:

> "In the territory of knowledge organization, the warrant strengthens in the first place the relevance and adequacy of a term to represent a concept, so that this term can be used for the classification or indexing of data, documents or information." (p. 648)

Examining warrants as a collective concept, they are mechanisms/tools/means for representation, construction, evaluation, and revision (Barité, 2019; Kwasnik, 2010). Seeing warrants with various types, they are also principles/rationale/constraints and they can be used either alone or in combination. For example, literary warrant was used as authority to generate reliable records and support functional requirements in electronic record management (Duff, 1998). Literary warrant can be used to restrict, disambiguate, and structure terms in vocabulary control (Barité, 2018). Warrant can be used as the rationale for critiques of classification theory and practice (Tennis, 2005). Ethical warrant consists of ethical principles and the study implies the principles, and in turn, ensures the implementability of ethical warrant (Beghtol, 2002).

When warrants are seen as principles, according to my analysis of cataloging where principles can be seen as relevance criteria, warrants are closely related to relevance. Additionally, Martínez-Ávila and Budd (2017) provided other evidence by explaining epistemic warrant as aboutness, which represents the inherent features of observations or thoughts. Kwasnik (2010) argued that

> "…warrant is not a feature of a classification as, say, the classes are, but rather it acts as a lens through which we can assess the extent to which a classification has meaning and utility for the purposes it was created. In fact, an articulation of warrant can be thought of as a conceptual framework for assessing any knowledge-representation system." (p.

107)

From Kwasnik's view, warrants act as a tool to assess the relationship between utility (actual status) and purpose (ideal status), where the assessment itself involves relevance judgment. Warrants can be reasonably seen as a (relevance) criterion, like the "disguised criteria for the selection of terms" stated by Barité (2019).

**Warrant, KOS Lifecycle, and Relevance**

To explore warrants further, I will demonstrate them in the context of the lifecycle of KOS. Drawing on the lifecycle of information system development (A. Lee & Brewer, 2004), the lifecycle of KOS has several phases: initiation, development, evaluation, implementation, and maintenance. In the initiation phase, the planning strategies include determining epistemological views, values, context, and scope, such as defining objectives, identifying stakeholders, establishing boundaries and requirements, etc. At this stage, warrants function more as the "atmosphere" to determine the key tone. In the development phase, tasks are analyzing and designing the terms, relationships, and structures of the KOS, deciding the reuse of the existing KOS, and developing the KOS. Warrants play a tangible role in concept formation/representation and term selection. Relevance judgment is the functional concretization of warrants at this stage. In the evaluation phase, tasks include user and performance testing, verification, etc. Beghtol (1986b) implied that the verification ensured by warrant followed the first justification function, which was explained by Barité (2019) as "at the time of the evaluation of the quality and relevance of a term" (p. 649). Relevance judgment is a component to perform evaluation. In the implementation phase, tasks include preparing user guidance, user training, promotion, and bringing the KOS into operation. In the maintenance

phase, tasks are reevaluation, updating, revision, etc. (A. Lee & Brewer, 2004; Soergel, 2009). Warrants come into play again in the lifecycle, and so does relevance.

In Kwasnik's (2010) perception,

> "In fact, an articulation of warrant can be thought of as a conceptual framework for assessing any knowledge-representation system. In this way warrant, as a concept, joins the ranks of relevance as a pivotal notion, open for many interpretations and applications but rich in its ability to capture essential qualities." (p. 107)

Resonating with Kwasnik and based on my analysis, I assert that warrant bears the same relationship to KO as relevance (the mainstream understanding) does to IR. Even if defining warrant using Saracevic's relevance formula (See Section 2.1.2), its meaning would remain appropriate and harmonious. For example, literary warrant can be defined as the justification of choices between terms and information entities the terms represent as determined by an information professional.

I investigated the knowledge organization processes of subject analysis, indexing, cataloging, and social tagging, and the accompanying key concepts of aboutness and warrant to conceptualize relevance in KO. Table 8 is a compilation of the relevance definitions in KO.

Table 8: Relevance definitions in KO based on Saracevic's formula (1975)

| Tasks | Definitions |
|---|---|
| Subject analysis | Definition 1: Relevance is the relationship between the inherent features and the subject matter of an information entity determined by a subject analyst. |
| Indexing | Definition 2: Relevance is the degree of correlation between the assumptions of users' information needs as determined by an indexer. |
| Indexing | Definition 3: Relevance is the measurement of the bearing between inherent features and the selected assumptions as determined by an indexer to identify subject matter. |
| Indexing | Definition 4: Relevance is the semantic relationship between the inherent features and controlled vocabularies as determined by an indexer to select |

| | |
|---|---|
| | index terms. |
| Cataloging | Definition 5: Relevance is the relationship between the inherent features of an information entity and the descriptions, subject terms, and classifiers assigned by a cataloger to fulfill the cataloging tasks. |
| Cataloging | Definition 6: Relevance is the relationship between the properties and attributes assumed by catalogers and used by users to meet their information needs. |
| Social Tagging | Definition 7: Relevance is the relationship existing between the tags and the tagging intentions as determined by users (including content creators) in the tagging process. |
| Social Tagging | Definition 8: Relevance is the relationships between tags in a folksonomy. |
| Social Tagging | Definition 9: Relevance is the relationship of tags existing between an information entity and information needs as determined by users. |

These definitions represent the deconstruction view for understanding relevance. Two orientations of the studies emerged from the decomposition to explore relevance in KO: relationships study leveraging KOS, from Definitions 1, 4, 5, and 8, and user (information professional) study, from Definitions 2, 3, 6, 7, and 9. The unbalanced distribution of assigned definitions to each orientation doesn't mean the difference in importance.

We are unable to accurately predict the needs of users (including both information professionals and end-users), as we cannot identify the use scenarios since there are always numerous possibilities and unpredictable situations. User studies are a promising method and many have been conducted to investigate how users approach information to solve their problems. The user study to explore relevance in KO is discussed in Chapter 6. If we see user needs as a pool or a scope, not specific ones, relationships in KOS undoubtedly wave a net as a backbone to support the pool. In the next chapter, I explored relationships as an embodiment of relevance.

# 5 Relationships in Knowledge Organization

To examine relationships in KO, a meta-typology of relationships is created in this chapter. Relationships are investigated, followed by the levels of the meta model.

A key aspect of knowledge organization is to establish the inherent connections between information entities (encompassing data, information, and knowledge in this study) and differentiate among them. KOS, themselves as a type of structured data and an approach to represent information entities, are the concentration or distillation of what information entities convey. Van Dijk (1979) argued in the discourse view that "Textual relevance is defined in terms of textual structures, such that certain structures are assigned a higher degree of relevance than others on general structural grounds" (p. 113), where sentence-level relevance contributes to global relevance (text as a whole with sentences and passages). Similarly, the relationships entailed in the KOS structures inform the connections between information entities, based on the assumption and the practice that information entities are represented by terms.

Relevance studies investigate whether information entities, their context, and users' information needs are related and the factors that work on them. Relationships answer the questions about how they are related and in what concrete form they manifest. Therefore, as a core concept in relevance studies, relationships are not a type of relevance but a sufficient term for relevance. In other words, at least some kind of relationship must exist to enable relevance. For the concept of relationship, there is an attempt to universalize it, which makes it easy to expand. This study does not put relationships in the purely social, interpersonal, or personal context but focuses on the relationships between information entities (chapter 5) and between information entities and information professionals (chapter 6).

## 5.1 Meta-typology of Relationships

To understand relationships in KO, it does not make much sense to simply differentiate them as data relationships, information relationships, or knowledge relationships according to the DIKW pyramid. Typology of relationships in KO has been studied in different ways and relationships in KO exhibit an array of typological variations. Tillett (2001) summarized four types of bibliographic relationships: primary relationships, content relationships, whole-part and part-to-part relationships, and shared characteristic relationships, plus two other types, responsibility relationships (between information and people/organizations) and subject relationships. Green (2001) grouped relationships in KO into bibliographic relationships, intratextual and intertextual relationships, subject relationships, and relevance relationships. She rebuilt the typology with a hierarchy in the follow-up study (Green, 2008). The top level is non-subject bibliographic relationships (including document-to-document relationships, responsibility relationships) and conceptual content relationships (including subject relationships and relevance relationships).

In the context of KO and KOS, relationships in this study mainly focus on semantic relationships, specifically the meaning associations between entities. The relationships focusing on non-meaning relationships, e.g., structural (organizational), syntactic, and co-occurrence relationships, or on partial-meaning relationships, e.g., pragmatic relationships, are only referred to when necessary. Semantic relationships that occur at the word level are called lexical relationships, at the concept level are called conceptual relationships, and at a higher level, e.g., the document level, might be called bibliographic relationships if in the bibliographic scope (Khoo & Na, 2006). In the meantime, inspired by the DIKW pyramid and borrowing the common

terms in the KO area, a meta-typological model of relationships in KO was created for this dissertation study to facilitate understanding the landscape of relationships. Bibliographic universe, information universe, and data universe are the three scopes in which relationships are located.

Table 9: Relationship scopes in KO

| Scopes | Relationship examples |
|---|---|
| Bibliographic universe | Bibliographic relationships |
| Information universe | Hyperlinks, social network connections |
| Data universe | Relationships in classifications, thesauri, taxonomies, ontologies, etc. |

Bibliographic universe means that the subject in this universe is bibliographic resources. The relationships between bibliographic resources have been studied for years, represented by bibliographic relationships and the relationships in bibliographic models. Information universe means that the subject in this universe is information. To differentiate information from bibliographic resources, this study focuses on online information such as hypertext and social networks. Data universe means that the subject in this universe is data. To differentiate data from information and bibliographic resources, and set in the KO context, this study uses KOS as the data source. In a broad sense, data covers bibliographic resources and information. Rather than being understood broadly, the three scopes are approached with a focus on differentiation. Given this premise, the meta-typological model of relationships is demonstrated in Figure 1.

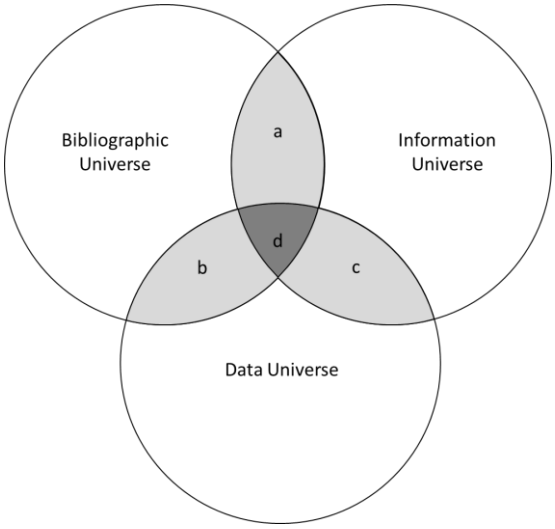



Figure 1: Meta-typological model of relationships in knowledge organization

The three universes intersect with each other, and some relationships are shared between them. For example, hyperlink relationships are part of Intersection a, thesauri relationships are part of Intersection b, and ontology relationships and metadata relationships can be part of Intersection d.

## 5.2 Relationships in Bibliographic Universe

### 5.2.1 Bibliographic Relationships

Bibliographic relationships are the relationships and associations existing between bibliographic works (Tillett, 1987). With the expansion of library collections, bibliography covers works of multimedia format, such as “textual, cartographic, audio-visual, graphic, and three-dimensional materials; to paper, film, magnetic tape, and optical media; and to acoustic, electric, digital, and optical recording modes” (Harmon, 1989; IFLA Study Group on the Functional Requirements for Bibliographic Records, 2009, p. 4).

Tillett (1987) created a taxonomy of bibliographic relationships, including equivalence relationships, derivative relationships, descriptive relationships, whole-part relationships,

accompanying relationships, sequential relationships, and shared characteristics relationships. Smiraglia (1992) supplemented the derivative relationships, covering simultaneous derivations, successive derivations, translations, amplifications, extractions, adaptations and performances. However, Smiraglia's derivative relationship taxonomy does not entirely agree with Tillett's taxonomy. For example, amplification is a subtype of descriptive relationships in Tillett's taxonomy rather than a subtype of derivative relationships in Smiraglia's taxonomy. Extracts/excerpts and the whole work are a whole-part relationship in Tillett's taxonomy, whereas in Smiraglia's taxonomy, they are an extraction derivative relationship. Vellucci (1997) studied bibliographic relationships in the music domain and found some new categories, mostly in whole-part, derivative, and accompanying relationships. Some new categories are domain-specific, e.g., forms of musical presentation and notational transcripts. Green (2008) considered the citation relationship as a type of bibliographic relationship in the categories of document-to-document relationships and content relationships. Furner (2002) proposed co-citation, co-relevance, and co-indexing relationships as subtypes of shared-characteristic relationships, where two information entities were cited for the same work, were relevant to the same information need, and were indexed about the same topic.

### 5.2.2 Models and Relationships

**FRBR**

The Functional Requirements for Bibliographic Records (FRBR) is a conceptual model providing a common understanding of bibliographic records, their relationships, and user tasks. (IFLA Study Group on the Functional Requirements for Bibliographic Records, 2009). The core elements in this model are entities, attributes and relationships. It differentiates entities in three

groups: Group 1 representing bibliographic records: work, expression, manifestation, and item; Group 2 representing agents responsible for the bibliographic records: person and corporate body; Group 3 representing subjects of Group 1: concept, object, event and place. The FRBR model shows the relationships between the three entity groups vertically and horizontally. Vertical relationships are the relationships among the three groups and the hierarchical relationship in Group 1. Horizontal relationships are the relationships within each type of entity in Group 1. Figure 2 shows the vertical and horizontal relationships in FRBR. The figure ignores the horizontal relationships of work-expression, expression-expression for different works, and manifestation-item because of the duplicated relationships and the relationships of expression-expression for different works originate from the work-work relationships.

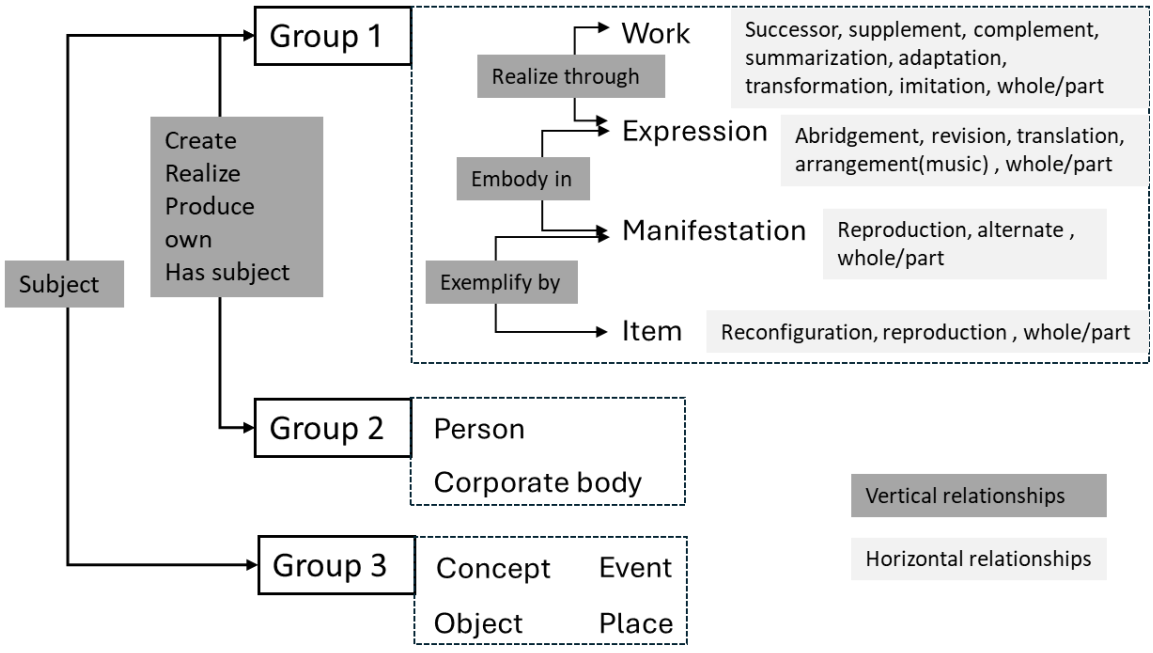



Figure 2: Relationships in the FRBR model

Empirical studies observed that supplement, successor, transformation, and adaptation were the most frequent work-to-work relationships in AMICUS, the Canadian national catalog

(Arsenault & Noruzi, 2012), and work-work relationships and manifestation-manifestation relationships were the main relationships, and whole-part relationships in the work-work level, derivative relationships in the expression-expression level, and equivalent relationship in the manifestation level were dominant relationships in the OCLC Connexion database (T. K. Park & Morrison, 2017).

The FRBR model doesn't consider horizontal relationships in Group 2 and Group 3. IFLA's Functional Requirements for Authority Data (FRAD) demonstrates suitable relationships for Group 2, and IFLA's Functional Requirements for Subject Authority Data (FRSAD) demonstrates suitable relationships for Group 3.

**FRAD**

The FRAD model is an extension and expansion of the FRBR model (IFLA Working Group on Functional Requirements and Numbering of Authority Records (FRANAR), 2013). The FRAD model added the family entity to Group 2. The relationships of agents in FRAD are shown in Figure 3.

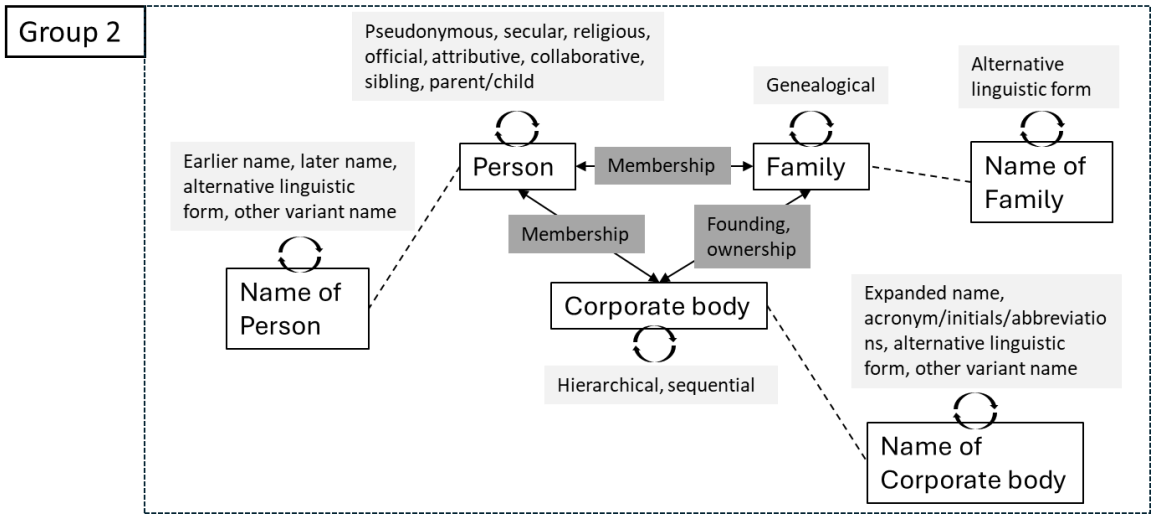



Figure 3: Agent relationships in the FRAD model

Person, family, and corporate body demonstrate relationships existing within their own groups and in relation to one another. The FRAD model used Tillett's (1987) bibliographic relationships for the work-work relationship. Name of work relationships include alternative linguistic form, conventional name, and other variant name.

**FRSAD**

The FRSAD model was developed to address subject authority data issues and their use (IFLA Working Group on the Functional Requirements for Subject Authority Records (FRSAR), 2011). It defines the relationships between work, thema, and nomen. Thema is "any entity used as a subject of a work" (p.15). Nomen is "any sign or sequence of signs (alphanumeric characters, symbols, sound, etc.) that a thema is known by, referred to, or addressed as" (p.15). Nomen is like the name entity in FRAD, but nomen includes name, identifier, and controlled access point. The relationships in FRSAD are shown in Figure 4.

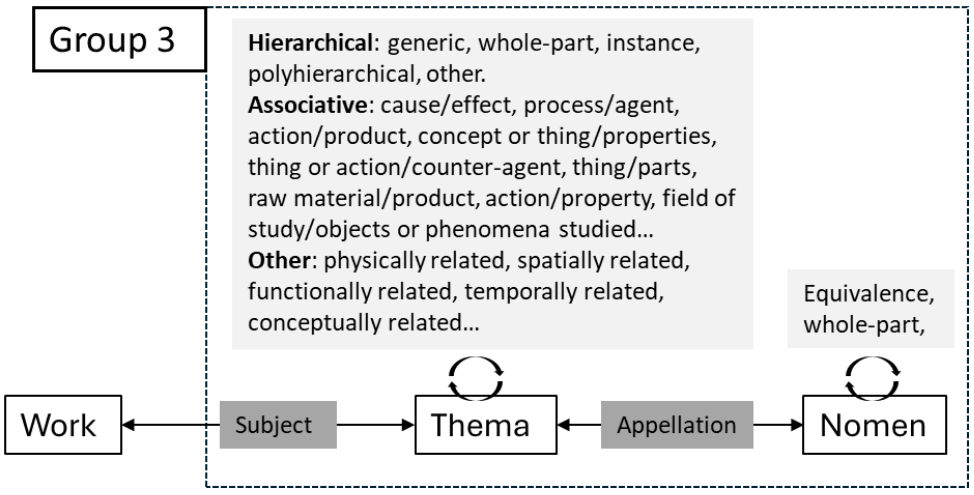



Figure 4: Relationships of entities in FRSAD

**LRM**

IFLA Library Reference Model (IFLA LRM) was created to combine the FR family FRBR, FRAD, and FRSAD into one model for better understanding and utilization (Riva et al., 2017). In

LRM, relationships are approached abstractly at a high level. LRM defines relationships of entities including resources (or things), works, expressions, manifestations, items, agents (including person and collective agents), nomens, places, and time-span. LRM doesn't categorize entities into three groups as FRBR, but into three levels. The only top-level entity is "res". Res is any entity in the bibliographic universe, fictional or real, e.g., work, person, creature, character in a novel, figure, substance, geographical area, etc. It can be an object, a work, a concept, an event, or a family. The concept of res is much closer to data than bibliographical resources. The relationships between res and other entities evidence this point as well: the relationship between res and nomen is appellation, and the relationship between res and work is subject. This datafication inclination makes LRM more compatible with data-oriented technologies and the data universe. LRM identified relationships and inverse relationships, which makes the LRM model clearer, enables reasoning links between resources, and facilitates greater interoperability.

Same as the FR family, the relationships in LRM exist between entities and within each entity. Based on the understanding of res as a data entity, the relationships of LRM entities are shown in Figure 5. The relationships and inverse relationships were merged for easy understanding. The relationships are clustered into three groups similar to FRBR but from a data perspective. Two groups are with data attribute and one with bibliographic attribute. From this grouping, we can see that data relationships are different from bibliographical relationships. In the case of IFLA model, several bibliographic relationships, e.g., accompany/complement, inspiration, alternate, reproduction, don't show in data relationships.

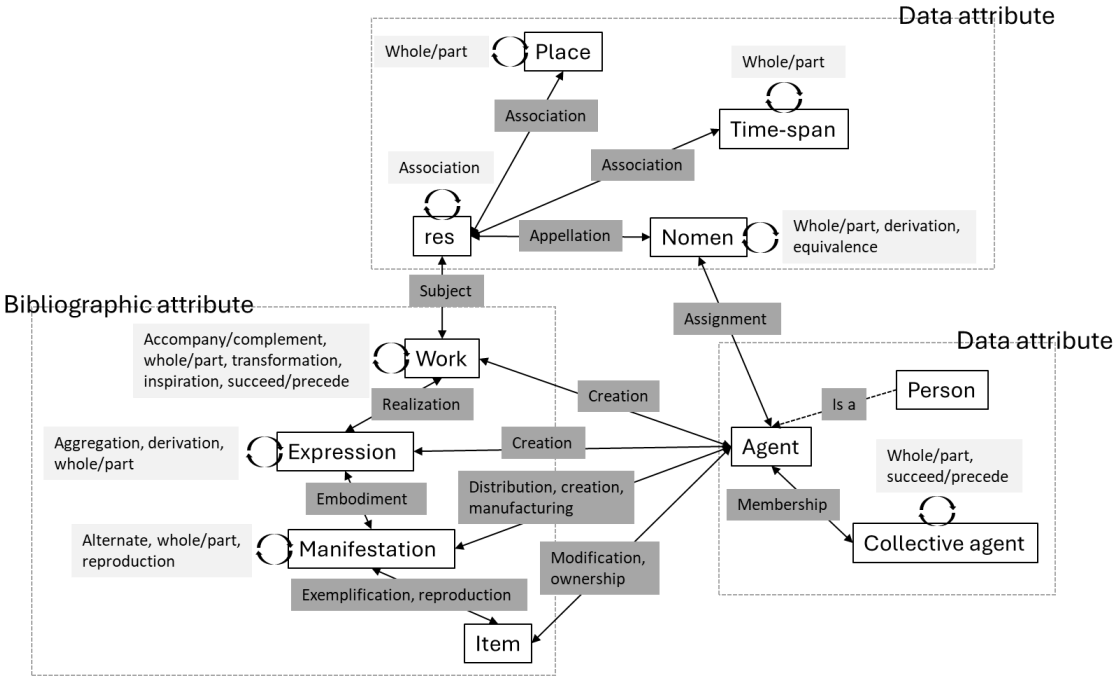



Figure 5: Relationships of entities in LRM

**BIBFRAME**

Bibliographic Framework (BIBFRAME) is a Linked Data model for bibliographic description to facilitate bibliographic resource sharing on the web (*Overview of the BIBFRAME 2.0 Model*, 2016). It can be seen both in the bibliographic universe and the data universe, which is discussed in the following section. BIBFRAME organizes bibliographic resources in three levels: work, instance, and item, with agents, subjects, and events that are closely related to them. Work is defined as the conceptual essence of the cataloged resource, which includes author, language, and subject. Work in BIBFRAME maps to work and expression in FRBR. The instance is defined as the material embodiment of a work, mapping to manifestation in FRBR. The item is a copy of an instance, same as the item in FRBR.

Because BIBFRAME is an RDF (Resource Description Framework) Linked Data model, each property can be used to represent a relationship. In this section, only the bibliographic

relationships are discussed, specifically, all the subproperties (relationships) under the top-level property relatedTo. The bibliographic relationships are two-level hierarchical. Most of the relationships are paired, e.g., hasInstance and instanceOf, hasDerivative and derivativeOf. For easy understanding, only one relationship is used to represent the paired relationships, e.g., using instance instead of hasInstance and instanceOf, derivation instead of hasDerivative and derivativeOf. The bibliographic relationships in BIBFRAME are shown in Figure 6.

In Figure 6, the level of the relationships is shown as UpperCategory: relationship. For example, whole/part: series, subseries means the entities have the relationships series and subseries that are both subcategories of whole/part. Same as LRM, entities with data attribute have fewer relationships identified than entities with bibliographic attribute. It is partly attributed to their bibliographic model nature.

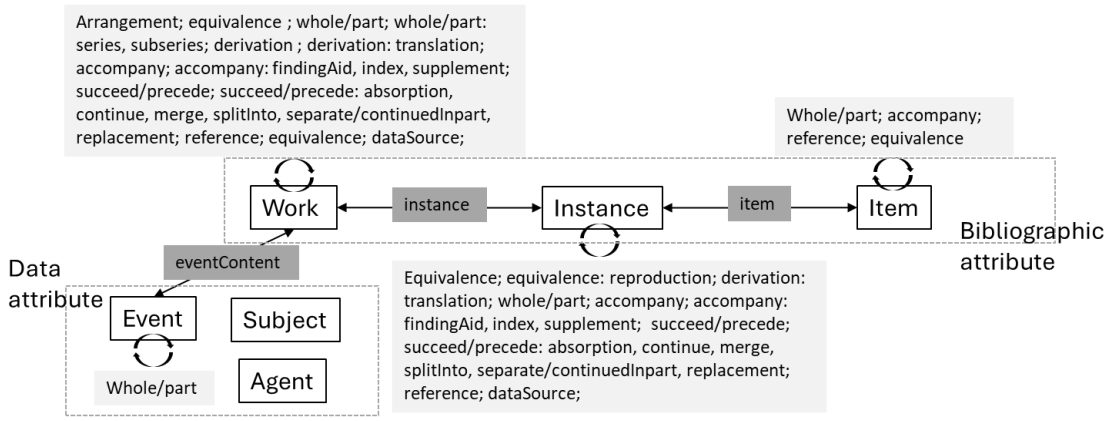



Figure 6: Bibliographic relationships of classes in BIBFRAME

The above analysis suggests that there is a continuous expansion from bibliographical relationships toward data relationships in the context of Linked Data and the Semantic Web, to make bibliographic resources more discoverable on the web. More discussion is in the data universe section.

## 5.3 Relationships in Information Universe

One focus of the information universe is the relationships between pieces of information. Online information, as a vast source of information with innumerable links and connections in various forms such as hyperlinks, labels, and tags, was used to explore relationships. Bibliographic resources can be a type of online information, but in this section, they are treated as general online information without considering their bibliographic features.

### 5.3.1 Hyperlinks

Hyperlinks are compositions with anchor text in the source information, the link, and the anchor text in the target information. Hyperlinks represent relationships between information nodes, making online information linked together and helping users navigate and explore. Hyperlinks present related information in a linear or non-linear way by positioning the links in different areas of information nodes, which can be understood as a form of organizing and structuring. Links and anchor text have been successfully used to help users search for relevant information, e.g., Google was designed to use hyperlinks and anchor text to search the web (Brin & Page, 1998). Ma et al. (2021) proposed an anchor-document relevance idea and argued that anchor-document relations were similar to relevance between queries and documents.

The types of hyperlinks have been studied for decades. Trigg (1983) created a link typology of textnet, which was interwoven by scientific works in the units of concepts, sentences, paragraphs, passages, etc., as information nodes. The links specify semantic relationships between these information nodes. The typology has two top-level categories: normal link and commentary link. Normal link types include explanation, solution, support, generalization/specification, summarization/detail, abstraction/example, alternate-view,

formalization/application, background, methodology, data, update, citation with subcategories c-source, c-credit, c-leads, etc., argument with subcategories a-deduction, a-induction, a-analogy, and a-intuition, etc. The commentary link types include comment, points, environment, problem posing, arguments, style, and data, each having its own set of subcategories. For example, data has d-comment, d-inadequate, d-dubious, d-ignored, d-irrelevant, d-inapplicable, and d-misinterpreted. These links provide the semantic foundations for relevance without considering users' tasks, which is objective and similar to topical/logical relevance. Trigg's link typology only covers scientific works, which is just a part of online information.

Baron (1994) expanded the link typology study to hypertext and included non-semantic links referred to as organizational links. Baron's typology has two top categories: organizational links describing the structure, e.g., next page, and content-based links describing the relationships of meaning, including semantic links, rhetorical links, and pragmatic links. Baron explained that semantic links were about words, not the user's task; the rhetorical links, with the purpose of persuasion, were used to help users get sufficient information to support their goals; the pragmatic links led users to the information to address their tasks directly. The content-based links are functional links. The taxonomy is shown in Table 10.

Table 10: Baron's taxonomy of link types (Baron, 1994, p. 43)

| Semantic | Rhetorical | pragmatic |
|---|---|---|
| Similar<br>Contrast<br>Part/kind of | Definition<br>Explanation<br>Continuation<br>Illustration<br>Summary | Warning<br>Pre-requisite<br>Usage<br>Example |

Baron's definition of semantic, rhetorical, and pragmatic links, representing three types

of relationships, implies understanding relevance from three perspectives. The semantic links lay the foundation of the relevance of information nodes semantically: relevance comes from semantic relationships. At this level, being relevant can be claimed to be equal to having semantic links. The rhetorical links represent how information is discoursed to help users understand. The rhetorical links are extensions of semantic relationships. From the authors' perspective, rhetorical links are semantic relations for clarifying an information node; from the users' perspective, they bridge an information node and how they understand it and get prepared to determine relevance in the next step. Although belonging to the scope of semantic links, rhetorical links show a transitional function from semantic/logical/topical relevance to utility/situational/user-oriented relevance. The pragmatic links define the boundary, helping users understand how they use an information node via the linked nodes. The pragmatic links belong to the semantic scope and show the transitional function as well, but they are much closer to the utility/situational/user-oriented end.

Kopak (2000) created a hierarchical structure of 26 link types representing functional relationships between information nodes on the Web. The study also reported that some relationships were more relevant than others to users via a user test, implying link types contributed differently to relevance judgment. The taxonomy is shown in Table 11. The underlined terms are the types, while the other terms are the categories of the types.

Table 11: Kopak's functional link type taxonomy

| First level | Second level | Third level | Four level |
|---|---|---|---|
| Antecedent | support | <u>backgound</u> | |
| | | pre-condition | <u>Argument</br>Pre-requisite</u> |
| | | Proof | <u>Evidence</br>Condition</u> |

| | | | |
|---|---|---|---|
| | | Provenance | Citation<br>Application |
| | Reason | Causation<br>Purpose | |
| Instruction | Usage<br>Procedure | | |
| Outcome | Attention | Warning<br>Future | |
| | Resolution | Solution<br>Addendum | Correction<br>Update |
| Abstraction | Generalization<br>Summary<br>Content | | |
| Exposition | Perspective | Comparison<br>Alternate view | |
| | Specification | Clarification | Definition<br>Explanation<br>Comment |
| | | Instantiation | Example<br>illustration |

Link typology studies provide an understanding of objective relevance in a different way from bibliographic relationships. The typologies of hyperlinks built up by Trigg (1983), Baron (1994), and Kopak (2000) vary but show some common features:

1) The three typologies are link types representing semantic relationships at various levels (word level, concept level, document level, etc.). Although Baron discussed organizational links (a type of non-semantic links), she did not include them in the link type taxonomy.

2) Link types represent relationships between the linked information nodes, which serve as justifications or the foundation of justifications for why online information is related, and further support relevance judgment.

3) A link-type taxonomy can be used as a quick tool to help users identify which links are

more relevant to their current tasks with the information at hand.

### 5.3.2 Social Tags and Folksonomy

Social tags can be seen as a type of hyperlink. The difference is that a hyperlink leads users to a related resource, while a tag leads users to a list of resources labeled with the same tag. As the hyperlinks to hypertext, so are social tags to social networks.

We cannot conclude that no relationships exist between tags because folksonomies are flat or tag clouds are piles of terms without links. Chapter 4 discussed social tagging and folksonomy with a focus on the concept of relevance, and defined relevance as the relationships between tags in a folksonomy (Definition 8). In this section, the relationships between social tags are examined, with a primary focus on their semantic relationships.

Lee et al. (2009) extracted semantic relations between tags in del.icio.us using the Wikipedia corpus. They predefined three types of relationships: equivalent, subsumption, and similar relationships (defined as sharing common characteristics, different from the computational linguistics view). Wu and Zhou (2009) demonstrated that numerous semantic relationships existed in del.icio.us folksonomies as well via studying the semantic relatedness of tags, but without any specific semantic relationships being identified. Trabelsi et al. (2010) developed an algorithm to extract non-taxonomic relationships (semantic relationships excluding the hierarchical ones) from folksonomy, aiming at converting folksonomies to ontologies. They listed some example relationships: *has_been*, *mentions*, *is_being_developed_within*, *added*, *revealed*, and *provides*, without giving a typology of the relationships they discovered.

Hierarchical relationships were evidenced in tag sets in the del.icio.us folksonomy

(Kome, 2005). Heymann and Garcia-Molina (2006) reached the same conclusion using a computational method and calculating similarity scores. These studies demonstrate the two kinds of folksonomy hierarchies as stated by Kiu and Tsui (2011): user-powered and automatic derivation by statistical algorithms.

From the statistical and computational view, semantic relationships are understood as semantic relatedness and similarity to facilitate using arithmetic methods to address how tags are related to each other. This view is distinct from the KO view on the study of relationships in the information universe. Compared to typology and enumeration methods in KO, semantic relationships are usually studied as a whole. If we see the two lines of studies in a grid, the relationships in the KO view extend horizontally by adding new relationships, while the relationships in the statistical and linguistic view extend vertically by adding numbers representing different levels of similarities using algorithms. The statistical and computational view cannot address how the tags are related, but to what degree they are related. In the KO view, the relationships can be studied from tag-tag relationships and tag-post relationships.

**Tag-tag relationships**

Kipp (2005) examined tags assigned by users, authors, and intermediary indexers for academic journal articles and found patterns and relationships of the tags from different views. Seven types of tag relationships emerged, which were thesaural relations including same, synonym, broader term, narrower term, and related term, and non-thesaural relations including related and not related. The study found that thesaural relations between tags were less common than non-thesaural relations. Related relationships, constituting nearly half of all the relationships, were related but could not fit into the thesaural relationships, e.g., relationships

between an object and its field of study. Not related relationships could be further divided into time and task management tags, geographic descriptors, specific details and qualifiers, generalities, emergent vocabulary and other. Dix et al. (2006) proposed the concept of Semantic Halo to provide more semantic and contextual information of tags in flat and no-link folksonomies. A Semantic Halo, with a 4A feature, aggregates related terms (Aggregation), differentiates them in generalization or specialization groups (Abstraction), identifies co-occurring tags in a specific context as contextual tags (Ambience), and orders the ambience tags over time to see the semantic evolution (Age). The Semantic Halo shows a framework to understand how tags are related in folksonomies. Chen et al. (2008) identified eight types of relationships between tag pairs (co-occurring tags) in a small Flickr data sample, which were *is-a-measure-of*, *is-located-in*, *induces*, *is-induced-by*, *is-style-of*, *is-of*, *is-for*, and *is-a-method-of*, to train on large-scale data. The concept of Linked Tag was proposed by Im and Park (2015), leveraging the semantic relationships of DBPedia to link tags for image resources. For example, *isDesignedBy* links the designer tag of the subject tag of an image; *islocatedIn* links the geo-name tag and the possible creator tag, or the subject tag. Tags are terms. The essence of tag-tag relationships is term relationships. Kipp (2011) examined and compared results from previous studies that studied thesaural relationships between tags, title keywords, author keywords, and controlled vocabularies and found exact match, partial and related match relationships across domains. Tag-tag relationships exhibit a strong presence of data attribute, which is further examined in Section 5.4: Relationships in Data Universe.

**Tag-post relationships**

As discussed above, the relationships of information nodes are manifested in link types

in the KO view. Accordingly, types of social tags represent the relationships between tags and posts in parallel. Cantador et al. (2011, p. 13) claimed that "a categorisation of tags based on the users' tagging purposes can help to discard non-relevant tags, and thus improve content retrieval processes", which aligns closely with my perspective.

Tag typologies have usually been studied generically in the user studies or studies of tag suggestions (Carmagnola et al., 2007; Gupta et al., 2011; Kehoe & Gee, 2011; Xu et al., 2006). These flat typologies provide frameworks to understand, consisting of content-based tags, contextual tags, subjective tags, attribute tags, purpose tags, word-level semantics, ownership tags, organizational and self-referential tags, etc. For the content-based tags, it can be subdivided into: claim-specific, topic-specific, source-specific, sensitivity, manipulated media, and path or virality (Morrow et al., 2022).

Beaudoin (2007) identified 18 categories of image tags in Flickr, consisting of adjective, compound, emotion, event, humor, language, living thing, number, person, photographic, place-general, place-name, poetic, rating, thing, time, unknown, verb. Although the categories show the relationships between tags and the images, some of them lack semantics, e.g., adjective and verb.

Bischoff et al. (2008) developed a classification taxonomy of tags from sample tags in Del.icio.us, Last.fm, and Flickr, representing webpages, music, and image social media platforms. The categories are topic, time, location, type, author/owner, opinions/qualities, usage context, and self-reference. The study reveals that tag types contribute differently to various domains, with some types being more beneficial to users. For example, usage context is important in the music domain.

Cantador et al. (2011) created a hierarchical typology of social tags by leveraging the semantics of WordNet, Dbpedia and Yago. Table 12 shows the categories of the tags. The typology reflects users' intention, implying that tags in these categories are more relevant to users than those that are not.

Table 12: Cantador et al.'s Purpose-oriented social tag categories

| Top-level | Second-level | Third-level | Fourth-level |
|---|---|---|---|
| Content-based | Physical entity | Artefact | |
| | | Living entity | Animal, Person, Plant |
| | Non-physical entity | Organization | |
| | Entity | | |
| Context-based | Location<br>Time | | |
| Subjective | Opinion<br>Quality | | |
| Organizational | Self-reference<br>Task<br>Action | | |

Park (2016) studied social tags in the health domain and identified an 8-level hierarchical typology of 126 semantic types of tags in Tumblr using the UMLS semantic types. The first three levels are shown in Table 13. The types in the higher levels are more domain-independent, while the types in the lower levels (not shown in the table) are more domain-specific. Similarly, Wang et al. (2018) identified 34 semantic types of tags from BibSonomy using UMLS and FrameNet semantic types.

Table 13: First three levels of Park's tag typology in health social media

| Top-level | Second-level | Third-level | | |
|---|---|---|---|---|
| Entity | Conceptual Entity | Idea or concept | organization | finding |
| | | Occupation or discipline | Intellectual product | Group |
| | | Organism attribute | Group attribute | Language |
| | Physical Object | Manufactured object | Substance | |

| | | |
|---|---|---|
| | | Anatomical structure Organism |
| Event | Phenomenon or process | Natural phenomenon or process Injury or poisoning<br>Human-caused phenomenon or process |
| | Activity | Occupational activity Behavior<br>Daily or recreational activity Machine activity |

In the information universe, relationships have been studied via links and types of links for hypertext and tags for social media. Typologies of social tags are domain specific. In hierarchical typology, the lower the level, the more domain-specific the types become. However, there are some high-level common types across domains: content-based, contextual-based, and task-based, corresponding to the dimensions of relevance. Some types of links/tags have more semantic embodiments, while others don't. For example, the categories of "specific tags", "organizational links", "typos", "non-English", and "invented tags" can hardly be used to inform relevance studies. The typologies of tags reflect users' perspectives and support relevance studies from multiple dimensions.

Another feature appearing in the research process is that most typology studies have used external tools such as WordNet, DBpedia, UMLS, Yago, etc., by leveraging their abundant relationships. Relationship study transitions from the context of the information universe to the data universe.

## 5.4 Relationships in Data Universe

In this section, knowledge organization systems are used as the subject to study relationships in the data universe. Compared to the bibliographic universe and the information universe, KOS in the data universe are more structured, and the relationships are at a more

foundational level.

KOS are all types of schemes to organize information, such as term lists, classification schemes, taxonomies, ontologies, etc. (Hodge, 2000). They are different combinations of terms and their relationships for various purposes to support information description, structuring, arrangement, searching, and access. KOS can be knowledge structures of domain knowledge at a high level as well. Thinking about classifications, taxonomies, and ontologies, they present users with a framework of terms and their relationships for domain knowledge in an efficient and concise way. KOS, even without formal semantics, can still "provide a convenient and intuitive map of some subject domain" (Miles & Bechhofer, 2009).

KOS can be grouped into four categories from the structure and function perspectives: term lists, metadata-like models, classification and categorization, and relationship models, with progressively greater complexity (Hodge, 2000; Zeng, 2008). Term Lists: Lists, Dictionaries, Glossaries, and Synonym Rings; Metadata-like Models: Authority Files, Directories, and Gazetteers; Classification and Categorization: Subject Headings, Categorization Schemes, Taxonomies, and Classification Schemes; and Relationship Models: Thesauri, Semantic Networks, and Ontologies.

In the context of Linked Data, Semantic Web technologies, including metamodels and encoding standards such as RDF, SKOS, and OWL, are utilized to enable KOS to be adapted and integrated into the Semantic Web, broadening the scope of conventional KOS. To study relationships in the data universe, I proposed a framework to organize KOS and their supporting technologies in the Semantic Web, inspired by the report from the Library Linked Data Incubator Group (Isaac et al., 2011), the LOD KOS concept (Zeng & Mayr, 2019) and the

Taxonomy of KOS (Zeng, 2008). The framework is shown in Figure 7. It shows that relationships are a foundational element across all levels to support data description and knowledge representation. Relationships are analyzed at each level in the following part.

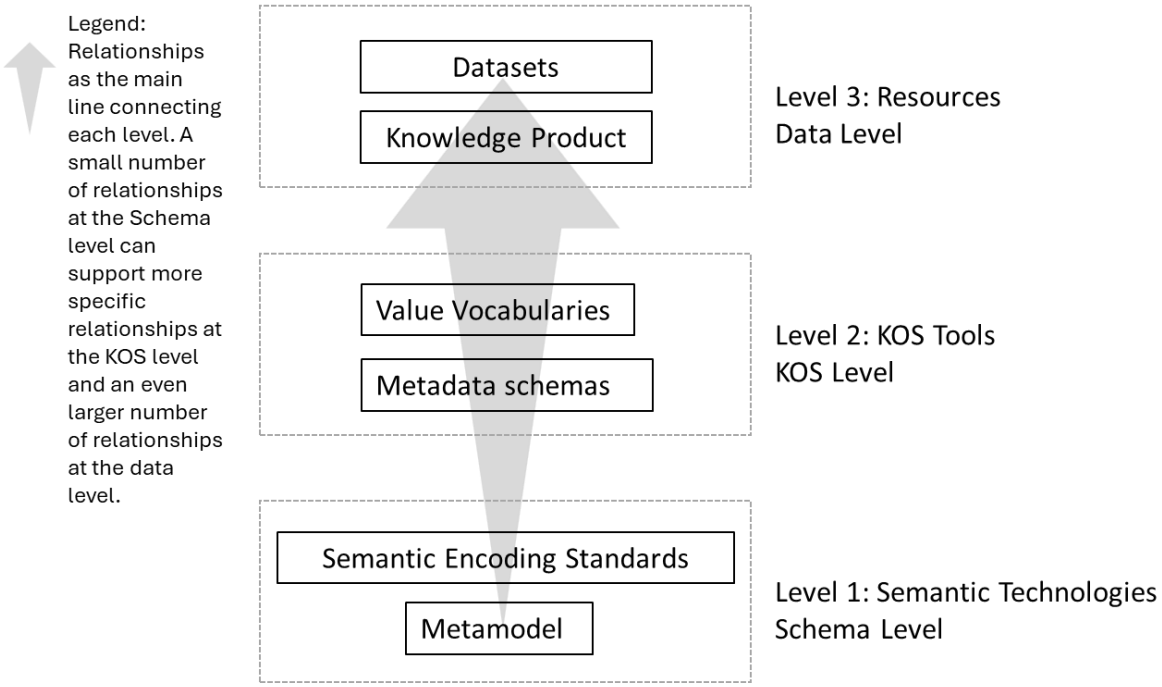



Figure 7: The framework of KOS in the Semantic Web

### 5.4.1 Relationships in Metamodel and Encoding Standards

**RDF and RDF Schema**

RDF is a data model developed by W3C to describe and interchange data on the Web (Cyganiak et al., 2014). The data model has two structures: RDF graphs and RDF datasets. RDF graphs are represented by a set of abstract 3-tuple syntax: subject-predicate-object, which are called RDF triples. An RDF dataset is a collection of RDF graphs.

An RDF triple is shown in Figure 8. Subjects and objects are resources at various levels, from strings, concepts, to documents. Predicates link subjects and objects, showing the relationships between them. A subject can have many predicates linking to many objects. An

object in a triple can be the subject of another triple.

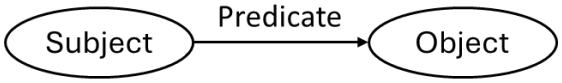



Figure 8: An RDF triple

In an RDF graph, the predicate is the relationship. All the terms used as predicates can be seen as various relationships, e.g., *isCreatorOf*, *isPartOf*, and *isAdaptedFrom*. However, RDF is a metamodel that does not define these relationships but only provides a framework for expressing these relationships.

Although RDF does not define specific relationships between subjects and objects (resources, Level 3 in the Fig. 7 framework), it defines relationships between RDF graphs and RDF datasets. At the graph level, there are *entailment*, *equivalence*, and *inconsistency* relationships (Cyganiak et al., 2014). *isomorphic RDF-term mapping* [5] represent the relationship that two graphs have the same data structure, the same relationship (predicate), and same subjects and objects in IRIs (Internationalized Resource Identifier, W3C proposed IRI to permit a wider range of Unicode characters of URIs) or literals (if blank nodes, they can be different). At the dataset level, the *dataset-isomorphic* relationship means that two datasets have *isomorphic RDF-term mapping* default graphs and named graphs[6]. *isomorphic RDF-term mapping* and *dataset-isomorphic* are relationships from the structure perspective.

RDF Schema (RDFS) provides a set of vocabulary, extending the RDF vocabulary, to define the RDF model, including subjects, predicates and objects to enable semantics among

[5] This relationship is not defined in the current version of W3C Recommendataion RDF 1.1 Concepts and Abstract Syntax (Cyganiak et al. 2014). It is defined in the draft version RDF 1.2 Concepts and Abstract Syntax (March, 2025), https://www.w3.org/TR/rdf12-concepts/#dfn-isomorphic-rdf-term-mapping

[6] The original definition is in RDF 1.1 Concepts and Abstract Syntax (Cyganiak et al. 2014). This explanation borrowed the concept *isomorphic RDF-term mapping* in the draft version.

them (Brickley & Guha, 2014). The RDFS vocabulary provides a way to describe classes and properties, not to provide classes or properties. RDF defines the syntax (the triples) while RDFS defines semantics for knowledge inference.

RDFS has class vocabulary and property vocabulary. Defining a resource using RDFS means defining the classes/subclasses and properties/subproperties of it. The RDF/RDFS property vocabulary is shown in Table 14. The properties with domain in *rdfs:Resource* can be used as relationships of RDF data (resources at the data or KOS levels in Figure 7), i.e., as predicates. For example: *rdfs:label*, *rdfs:comment*, *rdfs:seeAlso*, and *rdfs:isDefinedBy*. The reason is that resource is defined broadly in RDF: "All things described by RDF are called resources" (Brickley & Guha, 2014). Other properties are used at the schema level, e.g., *rdfs:subClassOf* and *rdfs:subPropertyOf* enable the hierarchical relationships of classes and properties. A specific example of an RDF triple "*rdfs:isDefinedBy*(S) - *rdfs:subClassOf* (P) - *rdfs:seeAlso* (O)" shows that *isDefinedBy* is a subclass of *seeAlso*.

Table 14: RDF properties in RDF Schema 1.1 (Brickley & Guha, 2014)

| Property name | domain | range |
|---|---|---|
| rdf:type | rdfs:Resource | rdfs:Class |
| rdfs:subClassOf | rdfs:Class | rdfs:Class |
| rdfs:subPropertyOf | rdf:Property | rdf:Property |
| rdfs:domain | rdf:Property | rdfs:Class |
| rdfs:range | rdf:Property | rdfs:Class |
| rdfs:label | rdfs:Resource | rdfs:Literal |
| rdfs:comment | rdfs:Resource | rdfs:Literal |
| rdfs:member | rdfs:Resource | rdfs:Resource |
| rdf:first | rdf:List | rdfs:Resource |
| rdf:rest | rdf:List | rdf:List |
| rdfs:seeAlso | rdfs:Resource | rdfs:Resource |
| rdfs:isDefinedBy | rdfs:Resource | rdfs:Resource |
| rdf:value | rdfs:Resource | rdfs:Resource |
| rdf:subject | rdf:Statement | rdfs:Resource |
| rdf:predicate | rdf:Statement | rdfs:Resource |

| | | |
|---|---|---|
| rdf:object | rdf:Statement | rdfs:Resource |

*rdfs:type* is a special property with domain in *rdfs:Resource* and range in *rdfs:Class*. It can be used to connect the data/KOS level and the schema level (in Figure 7). For example, *ex:blender* (as an instance) is *rdfs:type* of *ex:kitchenAppliance* (as a class), where instance is at the data/KOS level and class is at the schema level.

**SKOS**

SKOS (Simple Knowledge Organization System) is an RDF/RDFS-based datamodel to share value vocabularies such as thesauri, taxonomies, classification schemes, subject heading systems, and other controlled vocabularies in Linked Data (Miles & Bechhofer, 2009). SKOS itself is LOV (Linked Open Vocabulary). It reuses RDF and OWL (Web Ontology Language, discussed next) elements to define its classes and properties, leveraging the foundational classes and properties in RDF and the more expressive semantics of OWL. SKOS has four classes: *skos:Collection*, *skos:Concept*, *skos:ConceptScheme*, and *skos:OrderedCollection*. The properties and their domains in SKOS Simple Knowledge Organization System Reference (Miles & Bechhofer, 2009) are shown in Table 15.

SKOS uses *skos:Concept* to represent resources such as ideas, meaning, objects, events, specifically, items in KOS (Isaac & Summers, 2009). The scope of *skos:Concept* is a specific type of resource that exists at the data level. Thus, SKOS properties with domains in *rdf:Resource* or *skos:Concept* can be used to represent the relationships at the data level. The properties with domains in *skos:ConceptScheme* and *skos:Collection*/*OrderedCollection* not only show the schema-level (e.g., *skos:member* shows a part/whole relationship) but also the data-level relationships, because KOS are expressed as collections of concepts. It is hard to differentiate

the schema level from the data level since KOS, as data to SKOS, can be schemas as well.

Table 15: SKOS properties and domains

| Properties | Domains | Properties | Domains |
|---|---|---|---|
| skos:inScheme | - | skos:broader | skos:Concept |
| skos:hasTopConcept | skos:ConceptScheme | skos:broaderTransitive | skos:Concept |
| skos:topConceptOf | skos:ConceptScheme | skos:narrower | skos:Concept |
| skos:altLabel | rdf:Resource | skos:narrowerTransitive | skos:Concept |
| skos:hiddenLabel | rdf:Resource | skos:related | skos:Concept |
| skos:prefLabel | rdf:Resource | skos:semanticRelation | skos:Concept |
| skos:notation | rdf:Resource | skos:broadMatch | skos:Concept |
| skos:changeNote | rdf:Resource | skos:closeMatch | skos:Concept |
| skos:definition | rdf:Resource | skos:exactMatch | skos:Concept |
| skos:editorialNote | rdf:Resource | skos:mappingRelation | skos:Concept |
| skos:example | rdf:Resource | skos:narrowMatch | skos:Concept |
| skos:historyNote | rdf:Resource | skos:relatedMatch | skos:Concept |
| skos:note | rdf:Resource | skos:member | skos:Collection |
| skos:scopeNote | rdf:Resource | skos:memberList | skos:OrderedCollection |

SKOS differentiates relationships within one knowledge organization system and between KOS by using specific terms. A set of mapping relationships is used for relationships between KOS: *skos:broadMatch*, *skos:closeMatch*, *skos:exactMatch*, *skos:mappingRelation*, *skos:narrowMatch*, and *skos:relatedMatch*.

SKOS defines hierarchical and associative relationships between concepts within one knowledge organization system: *skos:broader*, *skos:broaderTransitive*, *skos:narrower*, *skos:narrowerTransitive*, *skos:related*, and *skos:semanticRelation*. These properties can also be used to link concepts between KOS without emphasizing the mapping relationship, which is confusing, but SKOS allows this to exist[7]. A hierarchy of the semantic relationships defined in SKOS is shown in Figure 9.

[7] More explanation can be found at https://www.w3.org/TR/skos-reference/#overview Section 10.6.1.

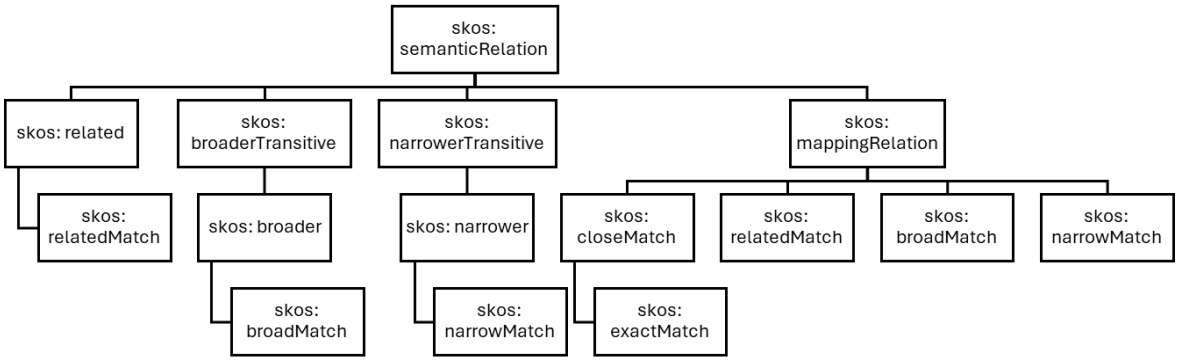



Figure 9: A hierarchy of semantic relationships defined in SKOS

**OWL**

OWL (Web Ontology Language) builds formal semantics based on RDF and is seen as a "vocabulary extension" of RDF for expressing ontologies (Bechhofer et al., 2004). OWL uses RDF classes and properties in its vocabulary. The latest version is OWL2 (https://www.w3.org/TR/owl2-overview/). Classes in OWL have hierarchical (*rdfs:subClassOf*) and equivalent (*owl:equivalentClass*) relationships.

OWL has three types of properties: object properties linking individuals to individuals, datatype properties linking individuals to data values, and annotation properties linking annotations to ontologies, axioms, or IRIs (Internationalized Resource Identifier) (Motic et al., 2012). Same as classes, these properties have inherent hierarchical (*rdfs:subPropertyOf*) and equivalent (*owl:equivalentProperty*) relationships. Object properties and data properties are not determined by OWL but by specific ontologies. Each ontology has different object properties and data properties depending on the domains and the subjects. Using the examples created by Motic et al. (2012), *parentOf* is an object property and *hasAge* is a data property. OWL2 used some annotation properties including *rdfs:label, rdfs:comment, rdfs:seeAlso, rdfs:isDefinedBy, owl:backwardCompatibleWith, owl:incompatibleWith, owl:versionInfo,* and *owl:priorVersion*.

Most of the annotation properties describe the relationships at the KOS and data levels. There are other properties in OWL functioning for bottom logic and axioms for reasoning, which are far from the properties discussed as relationships. Besides property and class relationships, OWL defines relationships between individuals: *owl:sameAs*, *owl:differentFrom*, and *owl:AllDifferent*, which can be used at the KOS and data levels.

Because of the RDF data model, the concept of relationship has been generalized: any concept linked between two resources is a relationship. Additionally, the concept of resource has been generalized as well, far beyond the common understanding in the library and information science domain. Resources can be as small as a string and as big as a dataset aggregation, causing the self-referential statement such as "*rdfs:Class* is an instance of *rdfs:Class*" (Brickley & Guha, 2014), making relationships more difficult to understand.

As discussed above, the common relationships at this level are hierarchical and equivalent relationships. However, these relationships are not used for (at least not specifically designed for) resource description as in the bibliographic or information universes, but for the structures of the schemas, i.e., they express the relationships between the elements of the schemas.

The generalization of resources brings about a positive effect that the properties in these schemas can be seen as the relationships not only for the schemas but also for the resources that the schemas support. However, not all the relationships reflected by the properties fall in the scope of relevance research. Many relationships can hardly have a direct impact on relevance judgment, since they mainly provide a mechanism to describe the relationships between resources, not the relationships per se, which can be called the relationships of

relationships.

### 5.4.2 Relationships in Value Vocabularies

Value vocabularies mean the vocabulary in which the terms are used as values (also called attributes) for properties (metadata elements), such as taxonomies, classification systems, thesauri, subject heading systems, authority files, term lists, etc. (Isaac et al., 2011; Zeng & Mayr, 2019). It is a type of controlled vocabulary, different from the property vocabularies – elements in metadata schemas. Relationships in value vocabularies have been studied for years, and some common understandings have been reached.

A value vocabulary is a collection of terms and their relationships in a specific domain. Three main types of value vocabulary relationships[8] have been identified in previous studies and widely accepted: equivalence relationships, hierarchical relationships, and associative relationships (Clarke, 2001; Hjørland, 2015; Zeng, 2008; Hodge, 2000; ANSI/NISO, 2010; Harpring, 2010). It does not mean that each value vocabulary must have all three types. For example, term lists do not have hierarchical relationships.

**Equivalence relationships**

Corresponding semantic relationships at the lexical level to the vocabulary (a collection of terms) level, equivalence relationships can be subdivided into synonyms, lexical variants, near synonyms, generic posting, and cross-reference to elements of compound terms (ANSI/NISO, 2010). In each sub-equivalence relationship, there are many types. For example, synonym types include Common/scientific names, non-proprietary/trade names, standard names/slang,

[8] The term of thesaural relationships can be used here since a thesaurus expresses all three types of relationships. The term of controlled vocabulary relationships applies too.

abbreviations/acronyms, terms from different cultures, terms of different linguistic origin, current/outdated terms, irregular plurals, etc. (ANSI/NISO, 2010; Clarke, 2001).

Generic posting is not a relationship but a technique to treat broader terms as preferred terms. Preferred term is an expression of equivalence relationships. It represents that a term is granted priority over other terms by information professionals. Designed for addressing the duplication, variants, and polysemy problems in indexing, it is not about semantic priority but involves certain types of warrants. Not all value vocabularies have preferred terms, e.g., taxonomies usually do not have preferred terms. Different expressions of preferred terms are used in various vocabularies. For example, in the Art and Architecture Thesaurus (AAT), the notation “preferred” is used following the preferred term; in the Faceted Application of Subject Terminology (FAST), the preferred terms are listed on the top without any notation, followed by “USED FOR” to indicate the unpreferred terms.

Antonyms are also included in equivalence relationships as near-synonyms in ANSI/NISO Z39.19-2005 (R2010) (ANSI/NISO, 2010). Antonyms provide an opposite view to help users determine the correct terms, although it can be confusing to include antonyms in equivalence relationships because antonyms cannot have equivalent meanings.

In multilingual vocabularies, equivalence means semantic equivalence, cultural equivalence and/or structural equivalence, and can be subdivided into exact equivalence (semantic and culturally equivalent, synonymy), near equivalence (semantic same but culturally different, quasi-synonymy), partial equivalence (broader/narrower semantic, quasi-synonymy), one-to-many equivalence, and non-equivalence (IFLA Working Group on Guidelines for Multilingual Thesauri, 2009).

**Hierarchical relationships**

Hierarchical relationships link the superordinate terms and subordinate terms (ANSI/NISO, 2010). In thesauri, hierarchical relationships can be expressed by BT (broader term) and NT (narrower term), e.g., Library of Congress Subject Headings (LCSH), or by a hierarchy tree, e.g., AAT, where terms are at the upper level, the lower level, or the parallel level.

Hierarchical relationships have three sub-relationships: the generic relationships, the instance relationships, and the whole-part relationships (ANSI/NISO, 2010). The generic relationship can be expressed as *is-a*, *is-a-kind-of*, *is-a-type-of*, *is-a-subtype-of*, *is-a-subset-of*, *manifestation-of*, etc. The instance relationship is expressed as *is-a* and *is-an-instance-of*. The whole-part relationship can be expressed as *is-a*, *is-a-part-of*, *is-in*, etc. (Harpring, 2010; Zeng, 2008).

**Associative relationships**

Associative relationships are a cluster of relationships that are neither equivalent nor hierarchical. Although ANSI/NISO Z39.19-2005 (R2010) claimed " it is important to make explicit the nature of the relationship between terms linked in this way and to avoid subjective judgments as much as possible" (ANSI/NISO, 2010, p. 51). The definition leaves a space for subjectivity to play a role in deciding which terms are related and to what extent. Defining associative relationships using the logic of elimination makes the concept unclear. Previous studies have enumerated common types listed in Table 16. Table 16 is adapted from previous studies to show the conventional relationships and their type of value vocabularies (ANSI/NISO, 2010; Harpring, 2010; IFLA Working Group on Guidelines for Multilingual Thesauri, 2009; Iyer, 1995; Zeng, 2008).

Table 16: Value vocabulary relationships and their types in the conventional view

| Relationship | Sub-relationships | Types |
|---|---|---|
| Equivalence | Synonyms | common/scientific names, non-proprietary/trade names, standard names/slang, terms from different cultures, terms of different linguistic origin, current/outdated terms, cross-reference to elements of compound terms, inter-language synonym |
| | Lexical variants | abbreviations/acronyms, irregular plurals, inverted order, orthographic variants, stem variants, verb/noun, |
| | Near-synonyms | inter-language quasi-synonym with cultural difference, inter-language quasi-synonym with lightly broader/narrower meaning, antonym, points on a continuum |
| Hierarchical | Generic relationship | genus/species, parent/child, specific/generic |
| | Instance relationship | things, events, concepts |
| | Whole-part relationship | member/collection, systems and organs/body, geographic location, department/corporate |
| Associative | | Cause/effect, thing/property, thing/function, process/product, process/agent, Process/device, Process/counteragent, Action/property, Action/target, Concept of object/property, raw materials/product…… |

The table shows the conventional view on value vocabulary relationships. In the Linked Data environment, due to the metamodel RDF and semantic technology, the concept of relationships in value vocabularies has been widely expanded.

**LOD KOS**

The concept LOD KOS emerged during the practice of adapting KOS to the Semantic Web, referring to "all of the value vocabularies and lightweight ontologies within the Semantic Web framework" (Zeng & Mayr, 2019, p. 209). I investigated some LOD KOS with SPARQL endpoints and queried the properties using the SPARQL query language. As discussed in Section 5.4.1, properties are relationships in LOD KOS. These LOD KOS are UNESCO Thesaurus, Getty vocabularies, and AGROVOC. I also queried LOV (Linked Open Vocabularies) to get insights from

more comprehensive data. LOV is not a single LOD KOS but a collection of vocabularies, including value vocabularies, data model and encoding schemas, metadata schemas, etc.

UNESCO (United Nations Educational, Scientific and Cultural Organization) Thesaurus is a value vocabulary used in subject analysis and information retrieval in education, culture, natural sciences, social and human sciences, communication and information domains (*About UNESCO Thesaurus*, n.d.). The Linked Data version of UNESCO Thesaurus is built on SKOS. A list of 43 properties is shown in Table 17. In order to improve the readability of the table, I used labels to represent the IRIs (Internationalized Resource Identifier) for the properties. For example, the label *skos:narrower* is used for the property "http://www.w3.org/2004/02/skos/core#narrower" queried from the thesaurus.

Table 17: Properties in UNESCO Thesaurus

| | | | |
|---|---|---|---|
| dc:source | dcterms:publisher | rdf:type | skos:scopeNote |
| dc:title | dcterms:rights | skos:altLabel | skos:topConceptOf |
| dcterms:conformsTo | dcterms:rightsHolder | skos:broader | skos-thes:microThesaurusOf |
| dcterms:created | dcterms:source | skos:closeMatch | skos-thes:subGroup |
| dcterms:creator | dcterms:title | skos:hasTopConcept | skos-thes:superGroup |
| dcterms:description | dcterms:type | skos:inScheme | umu:uneskos#hasMicroThesaurus |
| dcterms:identifier | geo:wgs84_pos#lat | skos:member | uneskos:contains |
| dcterms:issued | geo:wgs84_pos#long | skos:narrower | uneskos:hasMainConcept |
| dcterms:language | owl:sameAs | skos:notation | uneskos:mainConceptOf |
| dcterms:license | rdf:comment | skos:prefLabel | uneskos:memberOf |
| dcterms:modified | rdf:label | skos:related | |

Getty vocabularies are a set of vocabularies developed by the Getty Research Institute, including Art & Architecture Thesaurus (AAT), Getty Thesaurus of Geographic Names (TGN), Cultural

Objects Name Authority (CONA), Union List of Artist Names (ULAN), Getty Iconography Authority (IA), and Categories for the Description of Works of Art (CDWA). CDWA is a metadata schema with cataloging rules. All the other five vocabularies are value vocabularies. AAT, TGN, and ULAN are now available in LOD format and for SPARQL queries (*The Getty Vocabularies*, n.d.). The SPARQL endpoint returned a total of 435 properties for AAT, TGN, and ULAN. These properties can be categorized into the common group, the AAT group, the TGN group, and the ULAN group. Snippets of each group are shown in Table 18. Labels are used to represent the IRIs (Internationalized Resource Identifier) retrieved from the SPARQL endpoint. For example: the label *aat:produce* is used for the property "http://vocab.getty.edu/ontology#aat2211_produce".

Table 18: Snippet of properties in Getty vocabularies

| | | | |
|---|---|---|---|
| dc:title | getty:parentStringAbbrev | getty:broaderNonConcept | geo:long |
| dcterms:created | skos:broaderTransitive | getty:broaderPreferred | rdfs:subClassOf |
| schema:deathPlace | rdfs:subPropertyOf | owl:imports | skos:broadMatch |
| schema:elevation | owl:disjointWith | getty:displayOrder | prov:used |
| vann:preferredNamespacePrefix | getty:contributorNonPreferred | getty:estEnd | prov:wasGeneratedBy |
| vann:preferredNamespaceUri | owl:inverseOf | getty:estStart | foaf:focus |
| getty:broaderInstantialExtended | geo:lat | getty:eventNonPreferred | foaf:homepage |
| aat:locus-setting_for | aat:perform | aat:locus-setting_for | aat:produced-processed_by |
| aat:used-located_in | aat:used_by | aat:takes_place_in | aat:reflect-produced_by |
| aat:produce | aat:use | aat:involved_in | aat:reflected_in-produces |
| aat:produced_by | aat:involved_in | aat:involves | aat:reflects |
| aat:required_for | aat:involves | aat:required_for | aat:reflected_in |
| aat:require | aat:used_by | aat:requires | aat:reflects |
| aat:act_upon | aat:create | aat:locus-setting_for | aat:located_in |
| aat:have_counterpart | aat:involved_with | aat:takes_place_in | aat:possessed_by-existing_in |
| tgn:distinguished_from | tgn:coextensive_with | tgn:capital_of | tgn:member_is |
| tgn:possibly_identified_as | tgn:located_on | tgn:capital_is | tgn:moved_from |
| tgn:adjacent_to | tgn:is_location_of | tgn:ally_of | tgn:moved_to |
| tgn:capital_is | tgn:meaning-usage_overlaps_with | tgn:member_of | tgn:successor_of |
| ulan:partner_of | ulan:worker_was | ulan:romantic_partner_of | ulan:step-child_of |
| ulan:partner_was | ulan:predecessor_of | ulan:relative_by_marriage | ulan:guardian_of |
| ulan:principal_in | ulan:administration_overlaps_with | ulan:sibling_by_marriage_of | ulan:ward_of |
| ulan:member_was | ulan:founded | ulan:adoptive_parent_of | ulan:descendant_of |
| ulan:school_of | ulan:related_to | ulan:adopted_child_of | ulan:ancestor_of |
| ulan:school_was | ulan:sibling_of | ulan:half-sibling_of | ulan:possibly_related_to |
| ulan:worked_with | ulan:child_of | ulan:step-sibling_of | ulan:friend_of |

AGROVOC Thesaurus is a multilingual collection of concepts, terms, and their relationships in the scope of food and agriculture developed and coordinated by the Food and Agriculture Organization of the United Nations, with over 41,300 concepts and 1,198,000 terms (*AGROVOC*, n.d.). A total of 133 properties have been retrieved from its SPARQL endpoint. Labels are used to represent the IRIs (Internationalized Resource Identifier) retrieved from the SPARQL endpoint. For example: the label *agr:affects* is used for the property "http://aims.fao.org/aos/agrontology#affects".

Table 19: Properties in AGROVOC Thesaurus

| | | | |
|---|---|---|---|
| arg:hasCommonNameConcept | arg:isIncludedIn | arg:typeOf | skos:changeNote |
| arg:hasComponent | arg:isMadeFrom | arg:usesProcess | skos:closeMatch |
| arg:hasDisease | arg:isMeasuredBy | vocbench:hasSource | skos:definition |
| arg:hasGoalOrProcess | arg:isMemberOf | dc:creator | skos:editorialNote |
| arg:hasHost | arg:isObjectOfActivity | dc:rights | skos:exactMatch |
| arg:hasMember | arg:isOldNameOf | dcterms:created | skos:example |
| arg:hasObjectOfActivity | arg:isPartOf | dcterms:creator | skos:hasTopConcept |
| arg:hasOldName | arg:isPartOfSubvocabulary | dcterms:isReplacedBy | skos:hiddenLabel |
| arg:hasPart | arg:isPathogenOf | dcterms:licence | skos:historyNote |
| arg:hasPathogen | arg:isPreventedBy | dcterms:modified | skos:inScheme |
| arg:hasPest | arg:isProcessFor | dcterms:source | skos:member |
| arg:hasProduct | arg:isProducedBy | voaf:classNumber | skos:narrower |
| arg:hasProperty | arg:isPropertyOf | voaf:dataset | skos:narrowMatch |
| arg:hasRelatedTerm | arg:isSourceOf | voaf:propertyNumber | skos:notation |
| arg:hasScientificName | arg:isSpatiallyIncludedIn | void:inDataset | skos:note |
| arg:hasSymbol | arg:isStudiedBy | rdf:first | skos:prefLabel |
| arg:hasSymptom | arg:isSymbolFor | rdf:rest | skos:related |
| arg:hasSynonym | arg:isTaxonomicRankOf | rdf:type | skos:relatedMatch |
| arg:hasTaxonomicConcept | arg:isUsedAs | rdf:value | skos:scopeNote |
| arg:hasTaxonomicRank | arg:isUsedIn | rdfs:comment | skos:topConceptOf |
| arg:hasTermType | arg:isUsedToMake | rdfs:domain | skos-xl:altLabel |
| arg:hasType | arg:isUseOf | rdfs:label | skos-xl:hiddenLabel |
| arg:includes | arg:m49Code | rdfs:range | skos-xl:literalForm |
| arg:indicates | arg:makeUseOf | rdfs:subPropertyOf | skos-xl:prefLabel |
| arg:isAchievedByMeansOf | arg:measures | owl:complementOf | foaf:depiction |
| arg:isAcronymOf | arg:pestOf | owl:deprecated | |
| arg:isAffectedBy | arg:prevents | owl:equivalentClass | |
| arg:isAfflictedBy | arg:produces | owl:imports | |

When value vocabularies are converted into Linked Data, more relationships than the conventional view have been identified and used to link the concepts and terms in value vocabularies. UNESCO Thesaurus, Getty Vocabularies, and AGROVOC Thesaurus exhibit many commonalities of relationships in LOD KOS.

1. Because LOD KOS are built based on RDF, the relationships between the terms are expressed in predicates, i.e., properties in the encoding schemas. The expression is a strong tool for building up relationships naturally and diversely, not only limited to RT (related term), NT (narrower term), BT (broader term), USE, or USED FOR in conventional vocabularies, but also in a more specific way such as *arg:isProducedBy*, *tgn:moved_to*, *ulan:ward_of, etc*.
2. The relationships can be classified into three levels. The first level is the encoding schema level, including properties in the RDF, SKOS, or OWL namespaces, e.g., *rdf:label*, *rdfs:subPropertyOf*, *skos:altLabel*, *skos:broadMatch*, *owl:sameAs*, *owl:imports*, *owl:equivalentClass*, etc. The properties not only show the relationships but also the structural functions required by the encoding systems. The equivalence and hierarchical relationships are expressed at this level. The second level is the vocabulary level, including ontologies. Many vocabulary-specific properties are at this level, e.g., *uneskos:mainConceptOf*, *aat:produce*, *tgn:capital_of*, *ulan:step-sibling_of*, *getty:contributorNonPreferred*. The third level is the metadata level. At this level, metadata elements are used as properties, usually to describe the value vocabularies per se. Elements in DC (Dublin Core), DCTerms, and Schema.org are used in the three

LOD KOS. From this view, metadata elements are relationships, which are discussed in Section 5.4.3.

3. The three types of relationships, which are equivalence relationships, hierarchical relationships, and associative relationships, are still useful and have been used in LOD KOS. These relationships might be expressed differently in different LOD KOS. Numerous properties at the vocabulary level are instances of associative relationships, making associative relationships explicit and clarified. Additionally, the same/similar relationships are expressed differently across LOD KOS, e.g., a whole-part relationship can expressed by *owl:members*, *arg:hasMember*, *skos:member*, *uneskos:memberOf*, *tgn:member_is*, *tgn:member_of*, *ulan:member_of*, etc.
4. Numerous relationships are expressed as properties, but some are for the structural requirement of encoding schemas, or are from the perspective of formal semantics, e.g., *owl:inverseOf* shows the relationship between two terms, such as *use* and *is_used_by*. Therefore, some relationships are not as helpful as others in terms of relevance studies because of the generalization of relationships caused by RDF.

I SPARQL queried LOV and got over 1800 distinct properties showing the same commonality. The difference is that LOV contains many metadata schemas in LOD format, strengthening metadata elements as relationships.

### 5.4.3 Metadata Schema, Relationships, and Relevance

Metadata schemas are crucial for identifying relevant data resources (Darby et al., 2012). Relevance judgment has been recognized as one of the functions that metadata supports (Balatsoukas et al., 2009). Prior to the widespread adoption of the term metadata, researchers

used terms such as catalog information, document representations, document information elements, etc., to study how the informational function elements of a document are related to relevance judgment (Barry, 1998; Marcus et al., 1978, p. wang; P. Wang & Soergel, 1998).

**Metadata fields are relationships**

In an RDF view, metadata fields are the relationships (denoted as the predicate) between the attributes and the information entity they describe (denoted as the subject and object). BIBFRAME, as an example, uses properties to represent the relationships between resources. Previous studies show that certain metadata fields are more successful than others in helping users determine relevance, which means that certain relationships are more important than others in relevance judgment. These studies are:

Marcus et al. (1978) found abstract, subjects, and title (in descending order) were indicative of relevance judgment. Janes's findings (1991) were a little different, showing abstract and title were significant while bibliographic and indexing (subjects) could have an impact, but not much on relevance judgment. Barry (1998) found source traits, abstract, title, and index terms had higher co-occurrence with various relevance criteria, indicating that these fields were used more during relevance judgment. The experiment conducted by Crystal and Greenberg (2006) showed title, abstract, date, URL, and format were used to judge relevant online health information. Zavalina (2011) found in aggregated digital collections, collection-level metadata fields, including description, subjects, title, URL, copyright and IP rights, object, and geographic coverage, contributed more to determining relevance (the match to user search queries in the paper). Wildemuth et al. (2019) observed that metadata fields title, keywords, year of production, and run time were used by users to judge relevance of video resources.

**Metadata fields are indicators of relevance criteria**

Metadata elements are associated with relevance criteria and the types of relevance. Wang and Soergel (1998) built up the Document Selection Model, implying the relationships between metadata elements (they called document information elements, DIE), criteria (including relevance criteria), and their values (they identified epistemic, functional, conditional, social, and emotional values which can be seen as types of relevance). For example, the title (a metadata element) is used to decide topicality (a relevance criterion) and contributes to epistemic relevance; the author is used to decide authority and contributes to social relevance; the publication date is an indicator of recency and contributes to functional relevance. A metadata element can correspond to multiple criteria and values.

Barry (1998) found an association between metadata fields and relevance criteria based on co-occurrence in their experiment. The abstract was associated with depth/scope, accuracy/validity, content novelty, tangibility, ability to understand, recency, external verification, effectiveness, and access. The title was associated with all the criteria of the abstract except access. Document/source traits were associated with access and clarity, which did not appear in the abstract and the title, plus depth/scope, accuracy/validity, content novelty, tangibility, ability to understand, and recency. Indexing terms were associated with depth/scope, accuracy/validity, content novelty, and tangibility.

Conversely, relevance criteria guide users in determining which metadata fields to use for effective relevance judgment. Relevance criteria and metadata elements have a corresponding relationship since relevance criteria can be embodied by metadata elements (Crystal & Greenberg, 2006). Balatsoukas et al. (2009) built up an evaluation model of user

interaction with metadata focusing on relevance judgment. In the use-content (metadata surrogates) interaction, users use relevance criteria to rank the importance of metadata elements.

**Metadata fields are indicators of data relevance**

Gregory et al. (2020) found many fields to inform users to judge data relevance, including author, source(repository, data journal, etc.), documentation and metadata, data structure, data format, data size, and collection period, but varied among different domains.

Liu et al. (2020) considered information attributes (i.e., metadata) were clues for relevance criteria studied the relationships between and them. For example, title, keywords, description, and time scope correspond to topicality; data producer, affiliation, and data platform correspond to authority; data acquisition channel, sharing level, download, and data size correspond to accessibility.

Zhang et al. (2021) considered metadata fields as stimuli for relevance judgment and studied the relationships between metadata fields and relevance criteria in the scientific research data domain. Their findings were in line with Wang and Soergel's (1998), indicating metadata fields were bound to relevance criteria. For example, keyword and title indicate topicality; data time indicates topicality, currency and comprehensiveness; cost indicates convenience; data format indicates usability and standardization.

Li et al. (2024) argued that metadata of research data provided systematic and heuristic cues, as a bridge to the data understanding gap, for users to judge relevance. The metadata elements for the systematic cue include data title, keywords, abstract, and data type, while for the heuristic cue, they include data source, authors, publication date, identifier, and citation

frequency.

A synthesis of the metadata fields, relevance criteria and their relationships, created from the previous studies of Barry (1998), Wang and Soergel (1998), Zhang et al. (2021), and Liu et al. (2020), is shown below. The purpose of Table 20 is to demonstrate the relationships between the common metadata fields and relevance criteria. It is not a full list.

Table 20: Metadata fields and the corresponding relevance criteria

| Metadata fields | Relevance criteria |
|---|---|
| abstract, author's affiliation, data sharing level, data size, data time (subject), document/data type, journal | Accessibility and availability |
| abstract, author, author affiliation, data description, data platform, journal, title | Authority |
| abstract, author, author affiliation, data description, data time, document type, geographical location, journal, source, subject, title | Coverage |
| abstract, author, journal, source, subject, title | Novelty |
| abstract, author affiliation, document/data type, experiment method, journal, source, subject, title | Quality |
| abstract, document/data type, source, title | Readability |
| abstract, data time (subject), publish time, source, title, update time | Recency |
| author, author affiliation | Relation/origin |
| data description, data format, experiment method | Standardization |
| abstract, author, author affiliation, data time (subject), data variables, description, document/data type, geographic location, instrument type, journal, keywords, publication date, related publication, subject, title | Topicality |
| abstract, format, resolution, title | Usefulness and usability |

This table is not a simple combination of all the metadata fields and relevance criteria that appeared in the previous studies, but includes semantic analysis, deduplication, aggregation, and categorization. The relevance criteria in Table 20 are from the listed references. There is no unified relevance criteria standard. The domain-specific fields have been

dropped or made domain-independent if applicable. The synthesis doesn't consider the weight of a metadata field indicating relevance. For example, the title and abstract are more important than document type and time scope to judge topicality. They are listed together without an order. Not all the relevance criteria can be embodied in metadata fields. For example, user background and experience, affectiveness, etc., cannot be represented using a metadata field describing an information entity. Not all the features of an information entity are described via metadata. For example, rank order and visit volume are not metadata elements, the downloadability of a dataset usually appears at the downloading part, and the fee for accessing a dataset is listed in the license.

## 5.5 Relevance as a Relationship

As a concept, relevance itself is a relationship, i.e., relevance relationship. When we say A is relevant to B, it implies that we use the relevance relationship. To avoid the self-referential contradiction, relevance is seen in a holistic view in this section.

Green (1995) proposed the concept of topical relevance relationships in response to the limitations of topical matching in information retrieval (especially recall), citation analysis, and knowledge synthesis. Further, Green and Bean studied topical relevance relationships empirically and proposed a typology of them (Bean & Green, 2001; Green & Bean, 1995). They identified 33 content-based relationships and categorized them into three top-level categories: topic matching relationships, hierarchical relationships, and structural relationships. Reference, definition, and attributes are topic-matching relationships; taxonomy and partonomy are hierarchical relationships; structural relationships are more complex and have subcategories such as structural components (object, result, purpose/goal, source, etc.), complex structures

(force, agent, need, etc.), etc. More than half of the relationships in their study fell into structural relationship types. They argued “every relationship imaginable can serve as a topical relevance relationship” (Bean & Green, 2001, p. 130). According to them, relevance is a spectrum from matching to related. Matching is the most relevant relationship in the spectrum. All types of relations specify how information is related. Therefore, the types of relationships identified in the three universes are all relevance relationships.

Some studies sought to categorize relevance relationships into a larger framework. Furner (2002) suggested categorizing relevance relationships, in the scope of bibliographic relationships, as a subtype of Shared-Characteristic relationships. Khoo and Na (2006) argued relevance relationship is a type of semantic relationship and defined it as “the relevance of a document to a query or to the information need of a user” (p. 174). They used the definition of relevance to define relevance relationships.

Relevance relationships cannot be investigated without a deep exploration of relationships, or we will be lost in the self-referential loop. It is similar to the associative relationships to some extent. The focus of relevance relationships is on “relationships”. The more subtypes that are studied, the better the disclosure of what relevance relationships are.

Relationship study has been a focus in the KO area. Chapter 5 explored the relationships in the bibliographic, information, and data universes. Based on the conventional view of relationships, the concept of relationship has been generalized due to the RDF model. In the data universe, the relationships of KOS have been expanded largely with the support of semantic technologies. These relationships have explained how data are related and provided a foundation to explore data relevance from the perspective of information professionals.

# 6 User Study and Data Relevance Model

This chapter first reviews the nine definitions of relevance from Chapter 4 and the relationship study in Chapter 5, following the introduction of the user study. Then the responses to the tasks in the first and second stages and the interview transcripts are analyzed to reach a data relevance model.

The concept of relevance has been examined and disclosed via subtasks and core concepts in KO. In some circumstances, relationships are primary factors in determining relevance. They are: Definition 1 in subject analysis: Relevance is the relationship between the inherent features and the subject matters of an information entity determined by a subject analyst; Definition 4 in Indexing: Relevance is the semantic relationship between the inherent features and controlled vocabularies as determined by an indexer to select index terms; Definition 5 in cataloging: Relevance is the relationship between the inherent features of an information entity and the descriptions, subject terms, and classifiers assigned by a cataloger to fulfill the cataloging tasks; and Definition 8 in social tagging: Relevance is the relationships between tags in a folksonomy. To describe it more directly, relevance is represented by relationships in specific contexts. Relationships have been investigated at three levels, from the bibliographic universe to the information universe and to the data universe. Relevance embodied in these relationships shifts from bibliographic relevance, information relevance to data relevance accordingly.

For other definitions, information professionals play a greater role in determining relevance. They are Definition 2 in indexing: Relevance is the degree of correlation between the assumptions of users' information needs as determined by an indexer; Definition 3 in indexing:

Relevance is the measurement of the bearing between inherent features and the selected assumptions as determined by an indexer to identify subject matter; Definition 6 in cataloging: Relevance is the relationship between the properties and attributes assumed by catalogers and used by users to meet their information needs; Definition 7 in social tagging: Relevance is the bearing existing between the tags and the tagging intentions as determined by users (including content creators) in the tagging process; and Definition 9 in social tagging: Relevance is the relationship of tags existing between an information entity and information needs as determined by users (including information professionals). In these definitions, information professionals determine relevance based on their intentions and assumptions. More subjective factors besides relationships affect relevance judgment.

In this chapter, data generated from the user study are analyzed alongside the stages of the tasks, and the data relevance model is created in the end.

## 6.1 Participant Information

A total of 9 participants have completed all the tasks and experienced all the processes. One participant switched to expert talk during the process. This expert used domain-specific resources rather than the resources I provided to complete the tasks. The expert's data are not analyzed along with the data from the 9 participants, but as a discriminant case to support or undermine the results drawn from them. Table 21 shows the background information of the participants.

Related work experience means experience working with metadata schemas, metadata records, value vocabularies, ontologies, and taxonomies. Although all participants have related experience, the levels vary from newly hired to decades of experience.

Table 21: Background information of the participants

| Subject | 9 participants | 1 expert |
|---|---|---|
| Language | English | English |
| Related work experience | Yes | Yes |
| Position type | Metadata librarian | Data engineer |
| Affiliation type | Academic libraries | Industry |

The positions of nine participants are generically referred to as metadata librarians, though titled differently, such as metadata cataloging librarian, reference, metadata & cataloging librarian, metadata analyst, etc.

Regarding familiarity with the two resources for the tasks (the Audubon field note and the Chinese rubbing), none of the participants had direct experience with the Audubon field note or similar resources, while two out of nine participants had experience with rubbings. Most participants who had no experience with both resources reported that they were more familiar with the Audubon field note because it was a textual resource similar to a journal article. Unless otherwise specified, the participants mentioned from Section 6.2 to 6.4 refer to the 9 participants, excluding the expert.

## 6.2 Responses in the First Stage

Participants were asked to design two metadata schemes for the resources, respectively. They were encouraged to use LOV (Linked Open Vocabulary) to choose metadata elements. There was no limit to the number of elements. Each participant provided a unique response. These responses reveal some commonalities and special considerations from the participants.

1) Fields

For the Audubon field note, the numbers of metadata fields vary from 7 to 37. Four responses fall within the range of 10-20, and three within the range of 20-30. For the Chinese

rubbing, the numbers vary from 10 to 57. Seven responses fall within the range of 10-20.

2) Elements reused from existing schemas

Eight designs reused terms from existing metadata schemas, while one participant did not specify whether to reuse or not, which was regarded as using self-developed terms. For the Audubon field note, Dublin Core terms were most frequently reused by 6 participants, followed by the Darwin Core standard being reused by 3 participants. For the Chinese rubbing, Dublin Core terms were reused by 6 participants as well, followed by VRA Core (a Visual Resources Association data model) being reused by 3 participants.

3) Mixed schemas

Three participants reused terms sourced from a single metadata schema for the Audubon field note, while four reused terms from a single metadata schema for the Chinese rubbing. Other reuses originated from between 2 and 4 vocabularies. A typical case is that one participant reused 20 terms from 11 vocabularies for the Audubon resource and 18 terms from 5 vocabularies for the rubbing resource.

4) Granularity

The designs differ in their level of granularity. DC terms are general, without the ability to provide detailed descriptions. One participant created fields based on DC terms to support a fine-grained description, expanding the reused 8 terms from DC to 23 fields. For example, *dc:description* was extended to *description.weathreport*, *description.obvseason*, *description.activity*, *description.evidence*, etc.; *dc:contributor* was extended to *contributor.observer*, *contributor.photographer*, and *contributor.editor*. Another participant designed the fields using primary terms from DC and listed several more granular substitute

terms for the primary terms. For example, for the primary term *dc:subject*, the substitute terms were *gbif:keywordSet:keyword*, *dwc:HumanObservation*, *gbif:taxonomicCoverage:generalTaxonomicCoverage*. Granularity and structure are closely linked but do not necessarily have a cause-effect relation.

5) Structures

Over half of the designs are flat for both resources, without hierarchies. However, the participants tended to group the fields according to the subject they describe by arranging them closer. For example, the fields describing Season, Weather, and Geographic location were arranged following Observation (the field describing observation is not hierarchically above the other fields); the fields describing Inscription information, Transcription, Mounting, Measurement of Inscription, etc. were grouped under Inscription (no field to describe Inscription but as a wrapper category).

The designs show that the original structures are inherited when reusing several terms from a single hierarchical vocabulary. For example, one design used VRA Core's 7-level hierarchy, and another design used some terms from PREMIS as well as the 2-level hierarchy of these terms.

## 6.3 Responses in the Second Stage

### 6.3.1 General Information and Coding

Participants were asked to compare the metadata schemes they designed in the first stage with existing standards to decide whether to change, add, remove fields, or adjust the structure of their original designs. All participants recorded the reasons why they made changes or did not make changes.

One participant did not make any changes. One participant substituted all the self-created terms with terms in existing schemas. For the Audubon field note, three participants added some elements, three removed some, and two changed elements without altering the total number. Five participants added some elements for the Chinese rubbing and three removed some elements. For both resources, two participants added hierarchies to their original designs.

Some NVivo screenshots show the open coding and axial coding processes. Figures 10 and 11 show the in vivo coding snippets of the second-stage responses from different participants. The highlighted yellow background texts are codes, and the blue background texts are for annotations. Figure 12 shows the statistics of the in vivo codes.

For the second resource

| Stage-1 Schema: Element | Stage-2(record your changes) Schema: Element | What I wanted to represent (you may use the text from the resources) | Reasons for your changes/no-changes |
|---|---|---|---|
| dcterms: source | dbpedia-owl:picture<br>bf:count<br>dcterms:source | Image information: three images are shown on the first page of Resource 2. The source of the images is Chen, Tsu-Lung (1958). You may check the reference section for more detail. | The embedded structure makes the description clear and easy to follow.<br>I missed the count information. It is important to describe image information. |
| | workType: NSCHclassification | | Classification information should be included if generally used for this type of resource. |
| | workType: localclassification | | Same as above |
| dcterms: title | "firstTitle" - "headTitle" | Title of work | Prefer the ability to embed where the title was taken from in the metadata element |
| dcterms: alternative | - | Other known titles | very similar to "otherTitle" |
| Mrel: artist | rubbingCreator | Person who created the rubbing | "rubbingCreator" is more specific and would also accommodate rubbings that could include other artists' contributions |
| Vra: idCurrentAccession | generalRegistrationNumber | identification from the repository | If "registration number" is their preferred nomenclature, use that. |
| | accessionDate | date item was acquired by repository | Great information, if known |
| Vra: place of repository | - | where the item is held | Prefer "place of repository" to the suggested "geographicLocation" because place of repository would include a corporate name, and if information was moved outside of the repository it could still be traced back. |
| Dcterms: medium | materials | what was used in the rubbing | materials is more clear than the DC term |
| Dcterms: format | mountingFormat ; dimensions | the physical extent of the resource | recommended terms are more clear and will provide more consistent metadata |
| Vra: technique | - | techniques used in resource | no change - term nearly equivalent to recommended term |
| Vra: style | scriptForm | Broad description of the style of the rubbing | This, and the following two fields, better capture the type of information important to this art style |
| | lineAndCharacterCount | | |
| | number of characters | | |
| Vra: date type discovered | | | |
| | original Object Description - all fields | | I hadn't considered this – but an incredible addition and valuable information |
| Dcterms: language | - | | equivalent to language |
| Dcndl: transcription | - | | equivalent to fullText; transcription makes it feel more clear that it is the words from the item, to me. |
| Dcterms: description | Adjust to include description subcategories | A free text description of the resource | The subcategories will provide more granularity and easier adjustment of scope for researchers. |
| Dcterms: isReferenceBy | relatedKnowledge | works that have described the item | relatedKnowledge sounds more elegant |
| Mrel: fmo | provenance | who owned the item | provenance takes the sticky "ownership" question away and instead focuses on stewardship; I prefer that language. |
| Bibo: reproducedIn | [in part] | relationship links to places where the | relatedDigitalResources + |

Figure 10: Snippet 1 of in vivo open coding for the second-stage responses

| Stage-1<br>Schema: Element | Stage-2(record your changes)<br>Schema: Element | What I wanted to represent (you may use the text from the resources) | Reasons for your changes/no-changes |
|---|---|---|---|
| dcterms:coverage@temporal | dwc:Event | time period of each observation chronicled within the resource, e.g. June 29-30, 1998 | I decided that the Darwin Core element and related elements below were more precise for detailing what is being covered in the resource. Potentially readers of this resource will be amateur specialists or professionals and, while dcterms:coverage@temporal will get the job done, the specificity of dwc:Event and related elements would better describe the resource for those users |
| | dwc:samplingProtocol | the method of sampling, e.g. area count | As we're ultimately talking about linked open data, the concern here is ensuring that those interested in 'area counts' per se would be able to locate this resource against other types of sampling used, or against more generic cataloging that may lack this specific element. |
| | dwc:eventDate | the date the sightings happened, e.g. June 29-30, 1998 | As this is a child element within a repeatable parent element, it allows for very specific narrowing of details down to the date of an event rather than something more vague like dcterms:coverage@temporal, which, while possibly sufficient, still leaves something to be desired conceptually. |
| dcterms:coverage@geographic | | the geographic area described, e.g. Atlantic Provinces region | I decided to keep this particular element in spite of also using the dc:terms location container element, because I feel that nothing under dcterms:location or within Darwin Core location metadata (unless I missed it) allowed for generalities like the title of the article uses, i.e. Atlantic Provinces region, and I wanted to make sure that was also captured. |
| | dcterms:location | container element meant to provide very specific data about an event or events described in the resource (event is referenced above) | I decided that the Darwin Core element and related elements below were more precise for detailing what is being covered in the resource. Potentially readers of this resource will be amateur specialists or professionals and, while dcterms:coverage@geographic gives the generality I want for the whole of the article, the specificity of dwc:Location and related elements would better describe the encounters listed in the resource for those users |
| | dwc:country | describes the country where an event happened | see dcterms:location above |
| | dwc:countryCode | country code for event location | see dcterms:location above |

Figure 11: Snippet 2 of in vivo open coding for the second-stage responses

**stage 2**

| Name | Codes | References |
|---|---|---|
| ML05 | 53 | 53 |
| ML08 | 2 | 2 |
| ML15 | 32 | 32 |
| ML16 | 11 | 11 |
| ML21 | 8 | 8 |
| ML25 | 14 | 14 |
| ML10 | 20 | 20 |
| ML12 | 2 | 2 |
| ML18 | 6 | 6 |

Figure 12: Statistics of the in vivo codes for the second-stage responses

Figures 13 and 14 show the axial coding results. The in vivo codes are categorized into groups. The groups are not pre-defined but emerge from the in vivo codes.

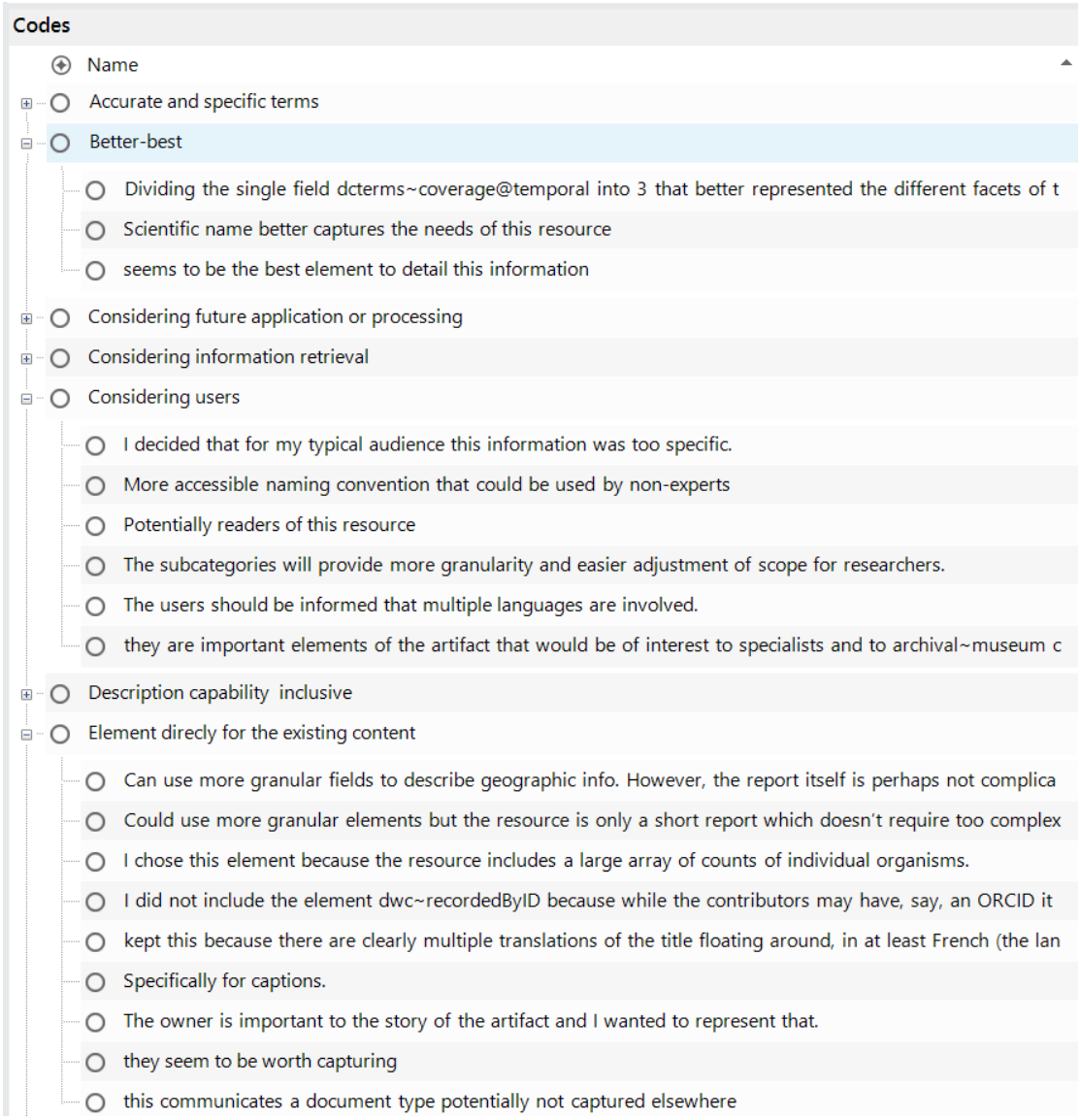



Figure 13: Snippet of axial coding with collapsed in vivo codes for the second-stage responses

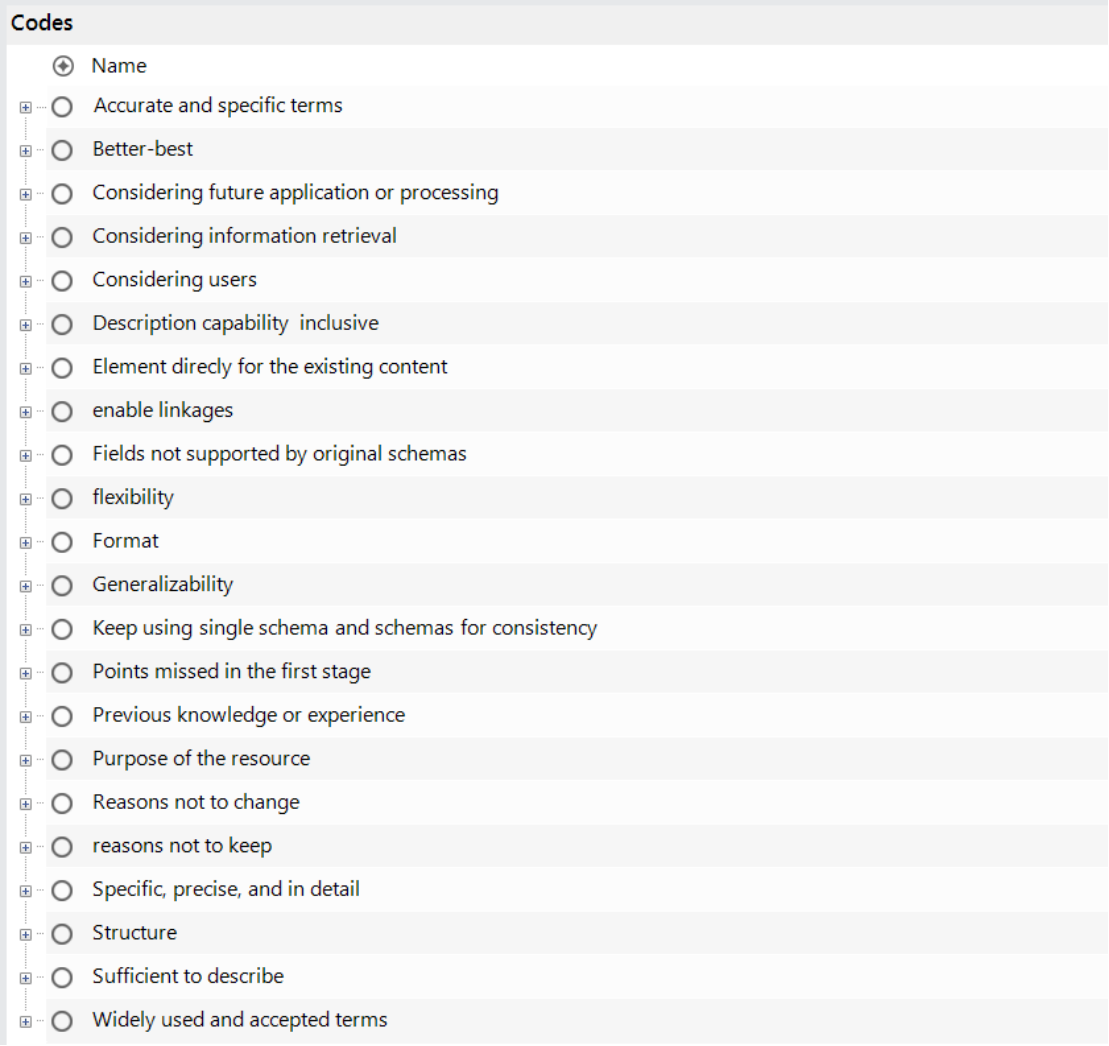



Figure 14: Axial codes for the second stage responses

Here are some reasons for the participants' design behavior in the second stage.

Question 1: Reasons to remove terms

Question 1 explains the change from being relevant to being irrelevant. Examples of the original notes from the participant to demonstrate this category are:

[ML10] "No longer necessary", "Covered by the previous value, removed."

[ML25] "Not needed. Too specific.", "Not needed too much detail."

The general reason why they removed the terms was that they no longer needed them. Except for ML25 who gave specific reasons, the notes didn't address why they did not need

these terms after all.

Question 2: Reasons to keep terms

Question 2 explains why participants considered the terms still relevant to their tasks after comparison. For example: [ML05] “This seems to be the best element to detail this information.”, “No other element would do the job better.” [ML08] “The original Dublin Core schema suffices for Resources 1 and no significant changes were determined.” [ML10] “Unchanged because the current fields provide suitable coverage”, “Unchanged because there's not a really good alternative.” [ML15] “No change. No equivalent element in Audubon Core.” [ML16] “No change. This element accurately describes the data.”, “A free-text field would be appropriate here.” [ML21] “Don’t see a more granular or appropriate element for activity evidence in the provided schema.”

Questions 1 and 2 show general reasons for participants’ actions. The notes were analyzed using the grounded theory coding method. Themes emerged after in vivo open coding. In the axial coding phase, 21 categories of codes emerged for the responses in the second stage.

### 6.3.2 Categories in Axial Coding

Category 1: Accuracy

Accuracy means that the understanding of a data point remains the same, but participants use a more semantically accurate term to describe it. Participants also used synonyms of quasi-synonyms of accurate, such as clear, right, elegant, less confusing, etc. Accuracy is related to specificity and might be a reason for a choice of better terms. For example:

[ML05] “I think this was the right field for representing what I wanted. I had a hard time

distinguishing between the different varieties of 'related' fields, to be honest." [ML10] "Less confusing as a wrapper." [ML15] "transcription makes it feel clearer that it is the words from the item to me.", "relatedKnowledge sounds more elegant", "provenance takes the sticky 'ownership' question away and instead focuses on stewardship; I prefer that language." [ML16] "This seemed like a more accurate representation of the data.", "Based on the description of this element, it seems more accurately reflective of the data." [ML21] "fullTextforinscription might be clearer."

Category 2: Better terms

Better terms refers to a term that can better describe the property of the resources than other terms. Better terms can be further explained as more specific terms, or more accurate terms For example, [ML05] "Dividing the single field dcterms:coverage@temporal into 3 that better represented the different facets of this rubbing, and related materials, was an easy choice." Better terms show the scope of a metadata element, e.g., [ML15] "Scientific name better captures the needs of this resource". However, if seen as a relevance criterion, being better is too vague to show the real scene.

Category 3: Considering users

When participants design metadata schemas, they consider the possible needs of potential users. The users can be grouped according to their information needs, such as researchers, experts and non-experts. The information needs of different groups of users may vary significantly. The users can be end-users who will use the KO product to search for information and information agencies who will create the KO product.

For example: [ML05] "they are important elements of the artifact that would be of

interest to specialists and to archival/museum curators and catalogers.", "Potentially readers of this resource will be amateur specialists or professionals and, while dcterms:coverage@temporal will get the job done, the specificity of dwc:Event and related elements would better describe the resource for those users." [ML15] "The subcategories will provide more granularity and easier adjustment of scope for researchers.", "More accessible naming convention could be used by non-experts." [ML25] "The users should be informed that multiple languages are involved.", "I decided that for my typical audience this information was too specific."

Category 4: Description capability

Participants would like to include terms with a stronger capability to handle different situations. This could be using more generic terms for inclusion, using more specific terms or be more granular to enable detailed information to be recorded, depending on the situations.

For example: [ML05] "it allows for cataloging each of the different birds that were sighted and identified as described in the resource.", "I kept this despite the more specific elements in the schema […] for any significant contributor that would be left out from the others.", "This allows for the multiple ways identifiers might be applied to this item that has had an international life." [ML15] "Prefer 'place of repository' to the suggested 'geographicLocation' because place of repository would include a corporate name, and if information was moved outside of the repository it could still be traced back."

Category 5: Enabling linkages

Participants used terms to enable linkages between resources, making sure resources are linked via certain properties. The purpose of linking data is intended to support information

retrieval and knowledge discovery.. For example:

[ML05] “They allow the data in the resource to link beyond the resource itself.”

[ML21] “this may be useful for connecting researchers of similar content/topics, etc.”, “It also allows for linking between dwc:Occurrence and dwc:Event.”

Category 6: Features of resource

It refers to the elements that are selected to describe the inherent features of the resources, staying true to the resources. In Chapter 4, the concept of inherent features was discussed as a foundation for analyzing relevance in KO. All the KO processes and key terms discussed in Chapter 4 rely on this root. The features might be considered at a detailed level, focusing on certain properties, or focusing on the whole at a general level. Examples in the second stage are:

[ML05] “I did not include the element dwc:recordedByID because while the contributors may have, say, an ORCID, it is not represented in the resource.”, “The owner is important to the story of the artifact and I wanted to represent that.” [ML18] “This communicates a document type potentially not captured elsewhere.” [ML21] “Could use more granular elements but the resource is only a short report which doesn’t require too complex schema.”, “the report itself is perhaps not complicated enough to fit for too granular descriptions.”

Category 7: Fields not supported by original schemas

It means that participants cannot find an appropriate term in the schemas used in the first stage, so they drop the points, but when new schemas are introduced, they add new appropriate ones, e.g., [ML10] “Did not previously have a space to record this.” This category is not typical since reusing terms from different schemas or creating new terms are allowed in this

study.

Category 8: Flexibility

Participants prefer terms that can hold flexible values. Flexible terms could be general terms but could also be related to the formats of the values, e.g., [ML05] “Locality offers more flexibility than Geographic Region.”

Category 9: Format

Participants consider that the format of the value affects the relevance of a property (term). Format is associated with flexibility and certain formats contribute to flexibility. For example: [ML05] “Because of the naming type in the resource, I decided against adding dwc:Taxon and its subelements.” [ML16] “it seems more useful to make this element a free-text form rather than a date format.”, “a free-text field would be appropriate.”

Category 10: Potential use

When determining the fields to describe the resources, participants consider beyond the immediate needs and take possible applications and future processing into account, including how the data will be displayed, how the data will be possibly processed in the future, the information is not at hand but may appear in the future, etc.

For example: [ML15] “Allows for more fluid reporting of time of day” [ML25] “This is needed especially if metadata is going to a larger collection like DPLA.”, “Needed if the information is available”, “I don’t believe this information was supplied but when available it should be added.”

Category 11: Generality

In some circumstances, participants prefer generic terms rather than specific terms. As

discussed above, generality provides flexibility. There are other reasons shown in the data for using general terms, such as a term designed to be used as wrapper element. For example: [ML05] “I feel like it still can act as either a summary field or a catch-all for anything that may not fit into the otherwise comprehensive and specific fields from the schema.”, “I felt it couched generalizations”. [ML10] “Unchanged because it is a suitable general field to use as a part wrapper.”

Category 12: Information retrieval

The purpose of knowledge organization is helping users find information that meets their needs. When designing a metadata schema and resource description, it is natural for participants to consider a step further and think about how users will use a metadata schema or other KOS to search for information. For example: [ML05] “be able to locate this resource against other types of sampling used.” [ML15] “Helpful in narrowing scope of searches”, “Temporal allows for more specificity; particularly for folks looking for a smaller date range.”

Category 13: Points missed

It means that participants missed some properties of the resources in the first stage mainly because they were not aware of a property or were unaware of the importance of a property. So, they added them in the second stage. For example:

[ML05] “Represents classification standards that I didn’t consider the first time.”

[ML15] “I missed the count information. It is important to describe image information.”, “I hadn’t considered this – but an incredible addition and valuable information.”

Category 14: Prior knowledge

Previous knowledge and experience affect participants’ judgments. For example:

[ML05] “I am not familiar enough with rubbings and historic Chinese literary practices to feel like I could say what should be included and what shouldn’t be. I don’t understand the differences between the types of titles, to put it bluntly.” [ML12] “As someone who has had to deal with a lot of data migrations, I want this schema to be as flat as possible.”

Category 15: Purpose

The purpose of a resource is a high-level factor that can affect many other categories. It addresses the why problem of resource description, e.g., [ML16] “Given that this has to do with species identification, this element seemed more appropriate.” However, not much evidence was shown in the first stage.

Category 16: Reducing duplication

Participants would like to avoid using duplicated terms to represent the same property in different positions in a hierarchy, e.g., [ML10] “I want to reduce the duplication of elements in the scheme, and there are geographic elements from the Darwin Core/Audubon Core schemas that I want to use to encode the geographic information of the observations themselves.” Reducing duplication in a metadata schema has a positive impact on accuracy.

Category 17: Specificity

Participants mentioned specificity a lot in the second stage. They prefer to choose terms that can specifically or precisely describe data points (properties) of resources. Specificity contributes to accuracy but is not identical. Specificity has a richer meaning, including the specificity of a term (closest to accuracy) and the specificity of the scope of a term. Evidence shows that specificity affects information retrieval as well and is affected/decided by categories such as considering users, granularity, and scope of terms.

However, as discussed in the category of generality, participants prefer generality than specificity in certain circumstances.

For example: [ML05] “allows for a more detailed parsing of the important information in the resource,”, “against more generic cataloging that may lack this specific element”, “precise for detailing what is being covered in the resource”, “I could immediately appreciate the more specific fields in the schema that I would not have necessarily considered as more of a generalist.”, “Maybe that’s too fine of a way to split it, but it feels right.” [ML10] “Unchanged to keep the specificity of ‘editor’ as the type of contributor”, “More specifically identifies the context of the values”, “More specific to biological data.” [ML15] “The subcategories will provide more granularity and easier adjustment of scope for researchers.”, “relatedDigitalResources + relatedKnowledge cover the same information as reproducedIn but have more defined scopes”, “DWC provides more granular data and additional guidance.” [ML16] “This seems to address the element with more specificity.” [ML18] “this needs some kind of note that it is an image contribution.” [ML21] “Both agent and provenance are not specific enough to describe people who found it. The rubbing schema has elements to describe more details of the work.” [ML25] “To use a more specific date field.”, “Makes sure public display labels make it clear who this is of rubbing or of original.”, “I included coverage without being specific about which kind. We should include both.”

Category 18: Structure

Participants consider adding sub-elements or wrapper elements to make the elements structured, organized and grouped for clarity. Structure is useful when representing complex relationships between the elements. For example: [ML05] “However, I no longer think this

element is sufficient to capture the ‘observers’ and as such, while they can and should still be listed in dcterms:contributor they would also be represented individually in the container element dwc:Occurrence and subelement.” [ML15] “The subcategories will provide more granularity and easier adjustment of scope for researchers.”, “The embedded structure makes the description clear and easy to follow.”

Category 19: Sufficiency

Sufficiency means when describing a resource, the schema covers the necessary properties to represent the inherent features. Sufficiency has two levels of meaning: sufficient terms and sufficient schemas. It implies a term/schema is acceptable to describe the resource although there might be more specific/granular terms/schemas. This is the reason why some participants kept terms without any change. It suggests that participants are aware of more specific terms, but they keep using the original ones. For example: [ML08] “the original Dublin Core schema suffices for Resources 1 and no significant changes was determined.” [ML10] “Unchanged because the current fields provide suitable coverage.” [ML16] “This element is sufficient to describe the data.” [ML21] “The original proposed element(s) seems to be sufficient.”

Category 20: Using single/as few schemas

Some participants prefer a single schema or as few as possible for consistency and clarity. For example: [ML10] “I tried to keep elements in the same schema for consistency.” [ML12] “prefer use of BIBFRAME to record and disambiguate metadata about the bibliographic resource”. [ML15] “recommended terms are clearer and will provide more consistent metadata” [means switching to the single metadata standard provided in the second stage].

Category 21: Widely used

The reason why participants decided to make changes or not is either a widely used term or a widely used schema. For example, [ML05] “I decided to keep to Dublin Core as it is more widely used and accepted.” [ML10] “using Dublin Core because it is so broadly recognized and used.” [ML15] “No change: The most frequently used element for author”, “If ‘registration number’ is their preferred nomenclature, use that.” [ML25] “Commonly used field.”

### 6.3.3 Discussion about the Categories

Though participants did not use the term context much, some categories show context information is an important factor, e.g., Category 2: Better terms and Category 15: Purpose. Creating a hierarchy implies putting a term in a scope, which could be seen as the context of a term, like mentioned in Category 18: Structure. The scope of terms makes using the terms more accurately. The evidence can be found in [ML05]: “I no longer think this element is sufficient to capture the ‘observers’ and as such, while they can and should still be listed in dcterms:contributor they would also be represented individually in the container element dwc:Occurrence and subelement dwc:recordedBy, detailed below.”

When some participants could not find an appropriate term, they preferred to reuse a general term instead of creating a new one. For example, [ML10]: “Unchanged because there's not really a good container for specific weather reports, but this is general enough to be used flexibly and when nested as a child element of the temporal, can be contextualized to be describing that temporal period.” and [ML05]: “I hate this element because it is so ambiguous. […] It should have subelements. However, within the LOV site and within the Audubon A/V core terms I do not see another element that will take pieces of this metadata (vol, issue, etc.)”.

When a schema has more information about the terms and the application of the terms, participants tend to use it because detailed information helps them determine relevance. For example: [ML15] “DWC provides more granular data and additional guidance”.

## 6.4 Interview

### 6.4.1 General Information and Coding

All interview data were collected within 2 weeks upon the completion of the 2-stage tasks. The interview questions consisted of two parts: the common questions and the customized questions according to the responses of each participant. The common questions covered: 1. previous experience with metadata-related work; 2. Metadata schemas they were familiar / worked frequently with; 3: whether they were familiar with the two resources for the tasks; 4. General impression of the tasks; 5. Whether they used the structured information of both resources; 6. Whether they used LOV to search for fields; 7. Their strategies for creating the schemas; 8. Hierarchy and structure issues; 9. The general thoughts on modification; 10. Previous locating data experience; 11. How they defined/thought relevant data. The customized questions were about their designs, modifications, and reasons they listed. Appendix F shows common questions and examples of individualized questions.

Some NVivo screenshots show the open coding and axial coding processes. Figures 15, 16 and 17 show the in vivo coding snippets of the interview scripts from different participants. The highlighted yellow background texts are codes, and the green background texts are for annotations. Figure 18 shows the statistics of the in vivo codes.

I think going back to what I said before, where when in like my everyday work when I'm creating a schema which isn't often umm, but I try to.

I try to keep it as less as less complex as possible because I'm not the only one who's going to be using that schema to create descriptions, so I'll try and limit it, limit the the fields I'm using to 1 schema if possible, and only branch out if it's really necessary.

So that's sort of how I approach this as I was looking for one thing, maybe you know one schema, maybe two that could cover what I wanted.

So for the rubbing, I knew that the VRA core had pretty much everything I needed.

It was. I had to finagle some stuff.

Umm, but that was relatively straightforward for the Audubon one, I think I mentioned previously that I was really struggling to find a thing that would encompass all of that, which is why I ended up going with something way more broad.

Uwm, which then was just really frustrating to work with. So yeah, I think. Yeah, I. But I think I really wanted to stick with that. Just using one schema and because schema.org is relatively flexible, it worked.

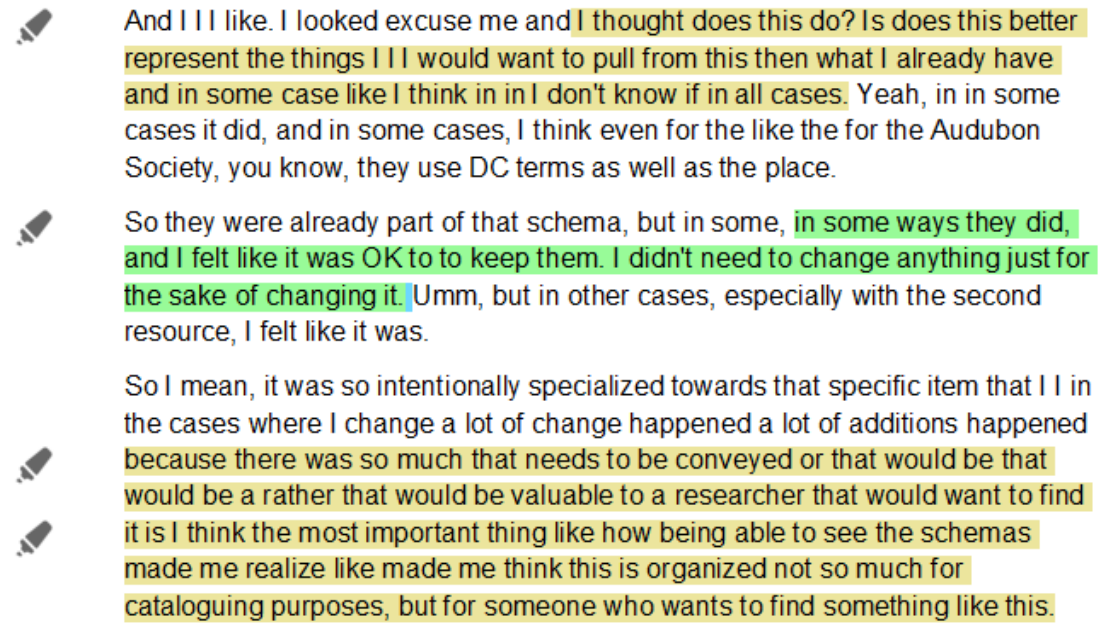
I could, you know, kind of make things make fields function in a way that I needed them to, but if I were to actually use that in a uh, like in a. You know, in a digital collection or something, I would have to be really like careful to actually to describe how each field is being used because it's not necessarily being used in the way it was intended to.

Like I think about the one that for some reason lives in my head is the the like age.

So it's there is the the sort of information that's talking about the age of the birds.

And I think I used. I use something in schema that is generally used to describe like the intended age of like people interacting with this. So specifically talking about humans generally. So in my head, I'm like, yeah, we're describing not humans with this. So it's not ideal, but like you know, if I describe how I'm using it, you'll make sense.

Figure 15: Snippet 1 of in vivo open coding for the interview scripts

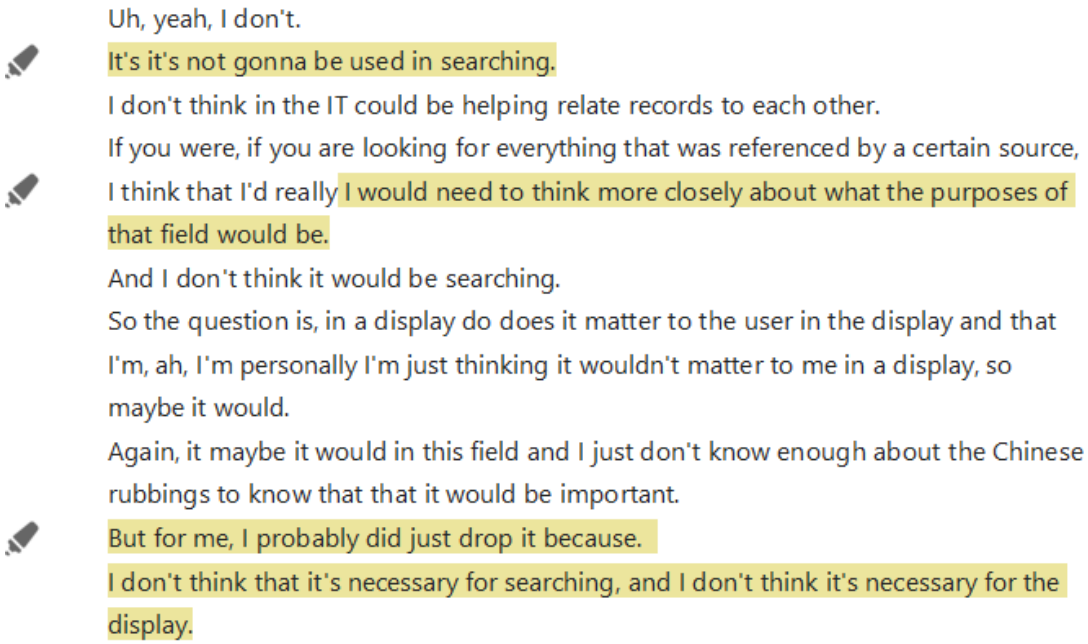


And I I I like. I looked excuse me and I thought does this do? Is does this better represent the things I I I would want to pull from this then what I already have and in some case like I think in in I don't know if in all cases. Yeah, in in some cases it did, and in some cases, I think even for the like the for the Audubon Society, you know, they use DC terms as well as the place.

So they were already part of that schema, but in some, in some ways they did, and I felt like it was OK to to keep them. I didn't need to change anything just for the sake of changing it. Umm, but in other cases, especially with the second resource, I felt like it was.

So I mean, it was so intentionally specialized towards that specific item that I I in the cases where I change a lot of change happened a lot of additions happened because there was so much that needs to be conveyed or that would be that would be a rather that would be valuable to a researcher that would want to find it is I think the most important thing like how being able to see the schemas made me realize like made me think this is organized not so much for cataloguing purposes, but for someone who wants to find something like this.

Figure 16: Snippet 2 of in vivo open coding for the interview scripts



Uh, yeah, I don't.

It's it's not gonna be used in searching.

I don't think in the IT could be helping relate records to each other.

If you were, if you are looking for everything that was referenced by a certain source, I think that I'd really I would need to think more closely about what the purposes of that field would be.

And I don't think it would be searching.

So the question is, in a display do does it matter to the user in the display and that I'm, ah, I'm personally I'm just thinking it wouldn't matter to me in a display, so maybe it would.

Again, it maybe it would in this field and I just don't know enough about the Chinese rubbings to know that that it would be important.

But for me, I probably did just drop it because.

I don't think that it's necessary for searching, and I don't think it's necessary for the display.

Figure 17: Snippet 3 of in vivo open coding for the interview scripts

| Interview | | |
|---|---|---|
| Name | Codes | References |
| ML05 | 48 | 48 |
| ML08 | 47 | 47 |
| ML10 | 70 | 70 |
| ML15 | 38 | 38 |
| ML16 | 31 | 31 |
| ML25 | 44 | 44 |
| ML12 | 59 | 59 |
| ML18 | 36 | 36 |
| ML21 | 43 | 43 |

Figure 18: Statistics of the in vivo codes for the interview transcripts

Figures 19 and 20 show the snippets of axial coding results. The in vivo codes are categorized into groups. The groups are not pre-defined but emerge from the in vivo codes.

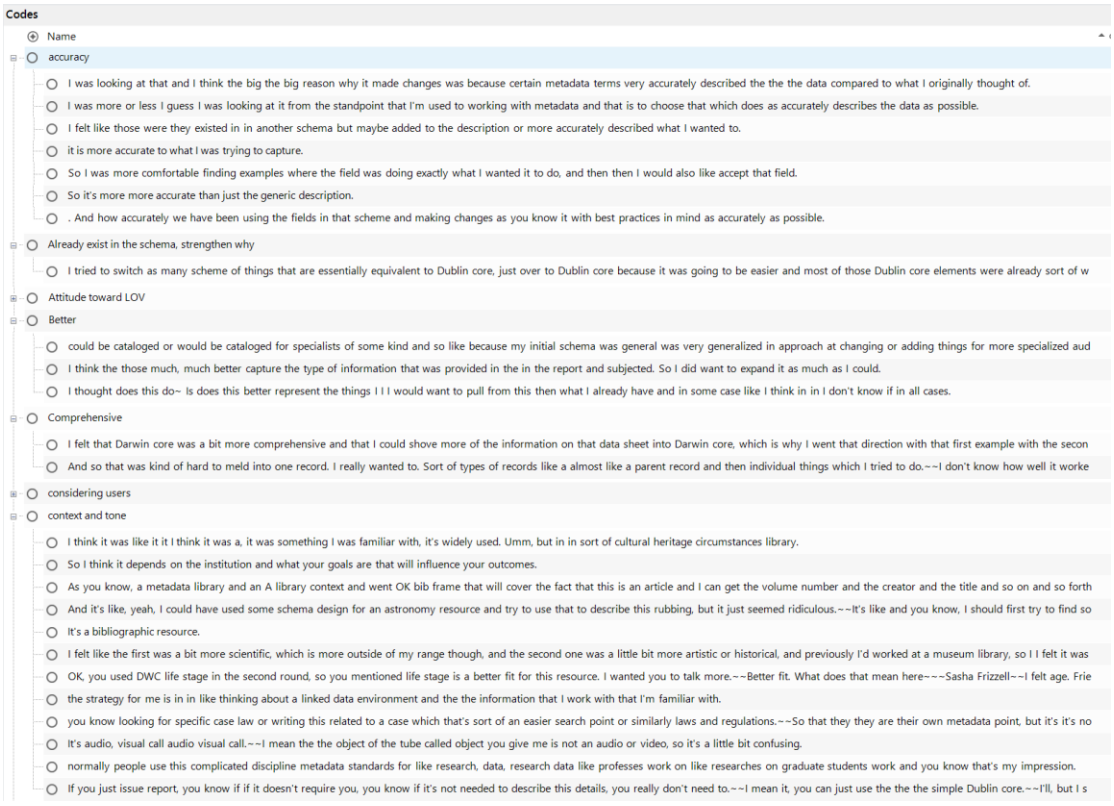



Figure 19: Snippet of axial coding with collapsed in vivo codes for the interview transcripts

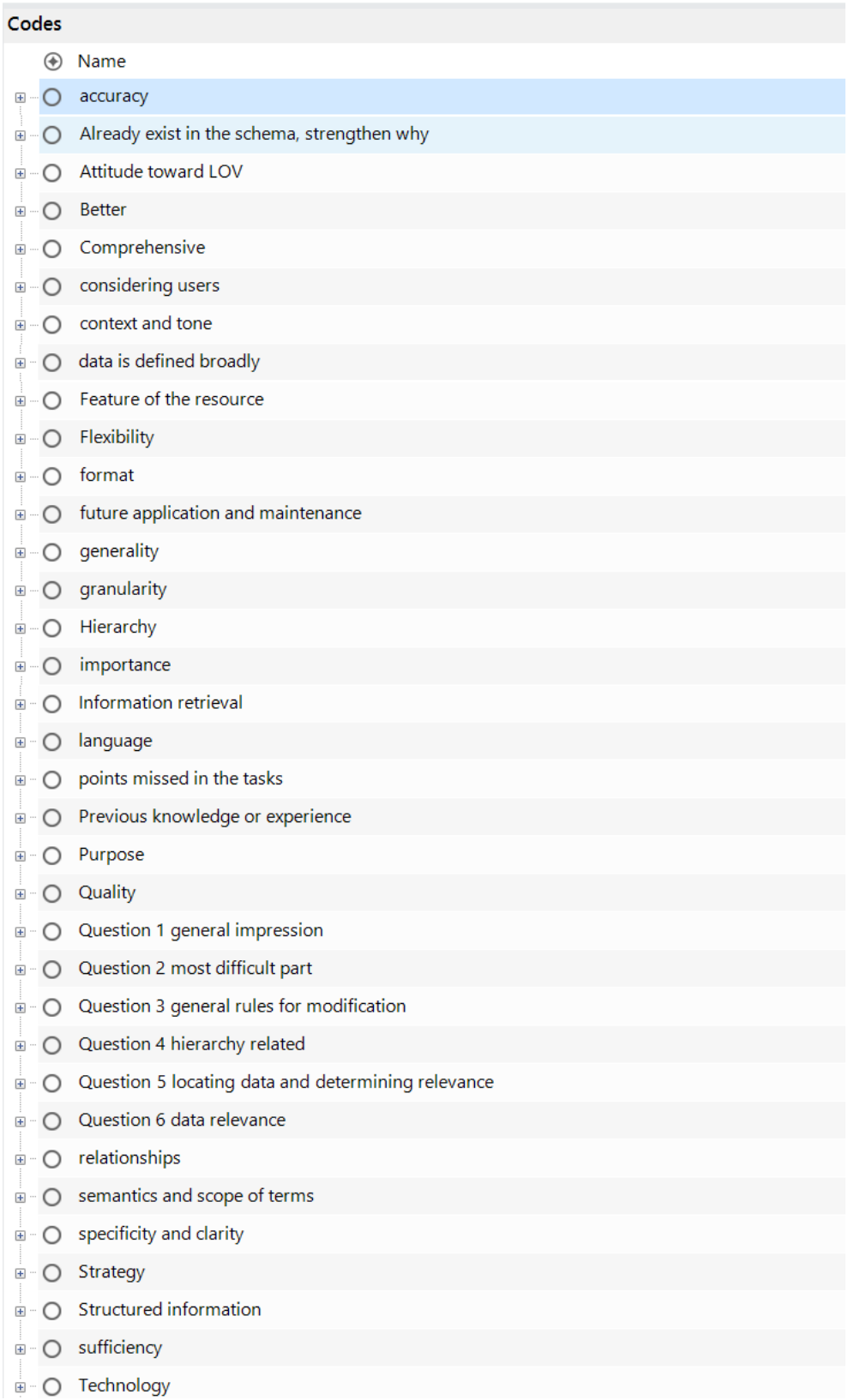



Figure 20: Snippet of axial codes for the interview transcripts

Themes emerged after in vivo open coding. In the axial coding phase, many categories emerged after iterative analysis of the open coding result. These categories overlap with the findings in the second stage because the interviews are partly based on the second stage tasks. Many new categories have been found due to the in-depth communications.

### 6.4.2 Overlapping Categories

There are 15 overlapping categories appearing both in the second stage and the interview stage. For all the overlapping categories in this section, I do not repeat the same explanation in Section 6.3.2, but list new evidence examples, new findings and explanations from the interview.

Category 1: Accuracy

Accuracy can be further explained as when a term accurately represents a property, or a term matches a property semantically. In other words, it could be focused on the term or on the representation. For some participants, being specific equals being accurate. Specificity falls in the spectrum of accuracy because being more specific means being closer to matching.

For example: [ML16] “I was looking at it from the standpoint that I'm used to working with metadata and that is to choose that which does as accurately describe the data as possible.”, “I think the big reason why it made changes was because certain metadata terms very accurately described the data compared to what I originally thought of”, “making changes as you know it with best practices in mind as accurately as possible.” [ML18] “It is more accurate to what I was trying to capture.” [ML21] “So it's more accurate than just the generic description.” [ML25] “So I was more comfortable finding examples where the field was doing exactly what I wanted it to do, and then I would also like accept that field.”

Category 2: Considering users

The data shows two general types of users: end-users using metadata schema to search information, and information professionals using metadata schema to describe resources. Considering users further leads to other decisions such as future application or processing, having sufficient/more information, using single/as few as possible schemas, being general or specific, background knowledge, information retrieval.

**End-users:**

[ML05] “they would think differently in approaching it then say someone just like a general library or museum visitor would I think”, “there was so much that needs to be conveyed […] would be valuable to a researcher that would want to find it is.” [ML08] “I provided so it would have just been more detailed and with the hopes that those details would have benefited users.”, “So basically the fields that I know that our users use all the time to find content.”, “It doesn't require us to be too detailed because they're not doing in depth research”, “Considering that it's a resource in multiple languages, you're never sure how users are gonna search. So, ensuring that whatever you have, there's both languages to suffice for that.”, “We create our own records and try to ensure that they have certain fields that we know users look for.” [ML10] “I think it's very interesting to think about people who don't have a metadata background and how they're approaching.” [ML12] “Dublin core and I think part of that too is our audience for our digital collections is our local undergraduate users.”, “if it's assuming some sort of, like domain expertise like this is somebody who works with this kind of material all the time […]”, “So I'm not describing this stuff for, you know, an expert audience, whether they're, you know, art scholars or collection curators or, you know, any of that.”, “I recognize that you know the this

information has value to somebody who's interested in making use of these resources." [ML15] "if you have a better term and can actually capture that in a way that is helpful to researchers." [ML18] "I wanted to break this up […] so that they could find it without having to dig through a bunch of extra information", "I feel like the type is actually a really relevant data point for people who might be doing research" "for example thinking about archaeologists you know it's important to know if somebody discovered a whole bunch of things or what areas they were active in but I couldn't find a good way to really describe this.", "As a user of the data, how the connections would be made and so I think those would potentially be their own separate objects that would just be linked to this object as opposed to being part of the description of the object." [ML25] "I had sort of is who the audience is that you're describing this for, right. Because as a university Cataloging person, generally it's undergraduates and the needs of undergraduates would be very different from like a scholar who's trying to get information out of this."

**Information professionals:**

[ML10] "I try to keep it as less complex as possible because I'm not the only one who's going to be using that schema to create descriptions, so I'll try and limit the fields I'm using to one schema if possible and only branch out if it's really necessary."

Category 3: Features of resource

Elements must be able to describe the important inherent features of the resources. For example: [ML10] "still use schema specifically to define like part source relationships because there wasn't really a good way to do that with Audubon Core." [ML21] "It's just like a short report, so I don't feel like it needs that level of description", "The nature of the typical object".

Category 4: Flexibility

There was not much discussion on flexibility in the second stage. Flexibility is an ability of a term to be fit for various situations and be inclusive. The generality of a term/schema contributes to flexibility. The format of a value affects flexibility, e.g., a free text field is more flexible than a fixed-type value that allows less possible input. Flexibility can be rooted in the features of resources, i.e., some properties can be naturally more flexible than others, e.g., the field 'subject' is more flexible than the field 'date'.

Examples in the interview stage are: [ML08] "because it was such an open-ended task or I read it as such an open-ended task, I felt that for flexibility purposes, if you were to change it, DC is very amendable to those changes." [ML10] "what is so great about DC is that it is so broad and flexible, but it also is extremely limited in what it can describe." [ML15] "So geographic region, I felt locality would give them a little bit more room to say I was at the base of this mountain in this place.", "so I wanna give the flexibility to capture the data. So if it needs to be more free form, I want to leave space for that, but I understand the importance of using a specific term when you're looking at really granular data.", "Sunset on kind of springtime, so I wanted to make sure while that is more flexibility, I think that better captures the data than if they had to put a date in there, especially if that date wasn't provided and with locality."

Category 5: Format

Format refers to the format of the value field to a property. When designing a metadata schema, participants consider the value first in some cases. Starting from the value or possible value and thinking about the format of the values, participants determine which term is more appropriate to represent the property. The anticipated format of the value of a field (term)

affects the relevance of the term. For example: [ML15] “So I felt with event date that needed like a numerical, a numeral and a lot of those event dates.”

Category 6: Potential use

Participants considered possible applications by information professionals such as cataloging, interoperability of the schemas and terms, and maintenance. Information retrieval (another category) is a subcategory of potential use by end users. Potential use is closely associated with users, since users are defined as information professionals (agencies) and end-users. When considering users, participants consider how users interact with data. Potential use is as well as a factor that can affect/determine other categories, such as granularity, hierarchy, sufficiency, and widely used.

Evidences are: [ML10] “my thought with adding these three as sort of essentially in my mind were you use whichever is relevant to what information you have.” [ML12] “the fewer linked data vocabularies I'm having to call out to and query, the easier the process is and the less maintenance I have to do.”, “So you know, we are looking again at how do we flatten this to be able to more easily port it into a new environment.” [ML21] “if you have a decent rubbing collection, that might be good, but then you need to map it to the more generic standard so that it can display in other digital repository.”, “it needs to be mapped for the greater discovery or greater interoperability.” [ML18] “And I feel like capturing that is probably that's a fairly unique thing to this sort of resource and then I mean the rest of it felt pretty consistent like you could find a lot of similarities with other publications.”, “I think trying to find a way to adequately describe in wiki data system it would be easier to display the language equivalencies whereas depending on your system.” [ML25] “Because if we have to share the data and go

somewhere else with it, Dublin core is a good is very shareable, right?", "But for me, I probably did just drop it because. I don't think that it's necessary for searching, and I don't think it's necessary for the display.", "I was also assuming that I'm creating this metadata set for a collection, not just for one. […] So I decided just to include it now and if we don't have any data in it, you don't have to use it.", "I was thinking that if you're going to have students doing the data input, I want to sort of minimize the opportunity for typos and other problems that students might have."

Category 7: Generality

As discussed previously, generality implies flexibility and inclusion, and is good for wrapper elements. More reasons why participants were using a general term are found in the interview stage, which are: when participants consider a group of users who do not need in depth data, when the features of a resource does not require specific representation, or when a working context or job responsibility prefers general terms. When participants cannot determine an accurate term, they prefer a generic term instead. However, participants have different attitudes on generality, which can be perceived in the category of Granularity as well.

Examples of arguments for generality: [ML05] "And so I think Creator is a much more generous interpretation in this case that I find valuable. […] I do value specificity highly, but I guess in this case authors a little Debbie" [ML10] "I had no idea what to do with that weather report. […] so I just left it in a schema description.", "I don't have a lot of knowledge about metadata schemes for more scientific topics, which is why I went with a more gentle with I used schema which is like broad and general". [ML12] "it's like I want to use a schema that is generic as possible to be able to accommodate all these different types of resources in this aggregated

environment.", "the rubbing number is that equivalent to an accession number and an archival context that I'm familiar with, or is it not, in which case I should use the more generic identifier."

Examples of arguments against generality: [ML05] "I don't actually like Dublin Core, they're kind of what I think of general.", "I think some of the terms are ambiguous in a way that I find frustrating."

ML05 voted for and against generality both. It seems that ML05 was in a dilemma, self-contradictory. In general, ML05 was against generality, but considering other factors, e.g., inclusion, ML05 preferred generic terms in certain cases.

Category 8: Information retrieval

More evidence of information retrieval is shown in the interview stage. From previous analysis and new evidence, information retrieval is affected/supported/decided/included by many other categories, either negatively or positively, such as potential use, enabling linkages, granularity, specificity, widely used, etc. For example:

[ML05] "put in metadata the more specific values […] that someone might appreciate or might be more helpful for them to find what they're looking for." [ML10] "I always kind of joke that all of our users just do keyword […], if you put more information in there, there will be more things that a keyword search or a natural language search is going to pull up." [ML21] "It has so many limitations, you know and, but if you can make the specific, more granular fields local fields, but it will limit your search capability." [ML25] "Because that's gonna help you in the keyword searching and if you use the right vocabulary it can get more specific and really be helpful to people", "So I constantly just try to think about how the user is going to be searching for it and what the user needs to see in order to find it", "So I don't think from a searching point

of view that it was pretty specific for anybody to need that kind of information."

Category 9: Points missed

When the participants acquired new knowledge (like the standards provided in the second stage), their cognitive status changed and the knowledge was added to their prior knowledge. Some points they miss in the first stage due to the lack of knowledge/experience were considered in the second stage. Points missed were affected/caused by prior knowledge. A supplemental example is [ML10] I didn't include that, and that's interesting that I didn't.

Category 10: Prior knowledge

Previous knowledge or experience includes domain knowledge about the resources, knowledge about metadata schemas and resource descriptions, and related working experience. Lack of knowledge may lead to missed points because participants don't know what properties are important that must be included. Participants tend to connect the task to the field they are familiar with, which further affects their judgment.

For example: [ML05] "hierarchy doesn't matter, because in the previous job that I had to look at large data sets of records and I would use a tool to extract.", "I think it was something I was familiar with". [ML10] "Sort of like building on what I had done already. So I had sort of the groundwork with everything in schema and then making the changes to the Audubon core and not making other changes.", "For sort of two reasons, I'm a little more familiar with it than CDWA.", "Sort of integrating as well with Dublin Core, which is very familiar for me.", "More general description areas that I wanted to switch them over to Dublin core because I'm a little more familiar with". [ML12] "I selected Darwin Core and again it was a vocabulary that was known to me that is used for this type of data.", "So with that kind of background knowledge, I

knew exactly where to go." [ML15] "So that was a little bit more challenging, especially looking at the scope and the same with the rubbings, some things I'm more familiar with from working with them like, you know, provenance metadata, things like art styles." [ML16] "if I had more familiarity with the resource too, more regular exposure [she would give it more structure]", "I probably just choose one that seems most familiar, probably schema.org/author.", "I'm thinking I'm really trying to relate everything to Mark 21 because that's what I'm familiar with in March 21.", "I've used schema.org before, so, there is no real reason. I just didn't know any better." [ML18] "I did is because the vocabulary of Fabio was more familiar to me", "I didn't know that the significance of. I didn't know if that was something unusual or if that's like something that is a relevant data point for. For all of the everything like this that would be captured", "I think I had like a wiki data model in mind with those where you know they are they sort of warrant their own layer of individual attention" [ML21] "I have been working with metadata for over 10 years, so I know this standard.", "I'm familiar with the standards in general, especially Dublin Core so that's why I chose the standards." [ML25] "for me it was easier for the Audubon example because I was much more familiar with, you know, with birds and bird sighting.", "I always started I think with both of them with Dublin Core just to use the what I was familiar with", "it was harder to find something when you don't know when you're less familiar with the topic that you're trying to describe.", "And that is because that is what I am most familiar with."

Category 11: Purpose

More evidence of purpose was found in the interview stage. Purpose has an expanded scope in this study. It includes the purpose of the resources, the purpose of resource description, the views/perspectives of participants, etc. Purpose determines the context, and

affects the sufficiency and granularity of resource description.

For example: [ML05] “I think the most important thing like how being able to see the schemas made me think this is organized not so much for cataloguing purposes, but for someone who wants to find something like this.” [ML08] “if it was more of like a class assignment, I was in school trying to do something, I felt that I would have made a few changes.”, “So I think it depends on […] what your goals are that will influence your outcomes.” [ML10] “I have sort of cherry picked a couple of fields for specific kind of administrative needs.”, “I knew I wanted to switch over all of the more scientific aspects to the Audubon. Core.”, “It's like I don't know how I'm gonna do this and I really wrestled with how I wanted to describe that.” [ML18] “I think generally a lot of the data that I work with is more about the schema is built upon the views that we want to build from it.”, “if I was doing for work, you know I would have to dig into the materials more.” [ML21] “it really depends on how you like, how detailed, how granular, how much time you want to spend.” [ML25] “I would need to think more closely about what the purposes of that field would be.”, “when I first got them I just said, you know, I'm gonna create the scheme that I would use at my specific institution.”

Category 12: Specificity

More meanings of specificity are identified in the interview stage. Besides the specificity of terms, and the specificity of the scope of terms (schema), there is specificity of a structure (close to hierarchy and granularity). It is affected by more categories including hierarchy and working context and responsibility.

For example: [ML05] “I thought the two other schemas were just too perfect and too well suited to skip them […] you know is very specific to this kind of artifact.”, “I really

appreciated how specific and exact those were and that I could allow me to work better.", "A very specific schemas that take into account a lot of things that you wouldn't find in a generic schema. [ML10] "This is perfectly created specifically for this. Of course I would use this if I knew that this schema was out there.", "I decided that I was just going to switch everything over to the schema that you had provided because it was so specifically designed for rubbings. […] That makes a lot of sense." [ML12] "I decided I maybe I should look for something that's a bit more specific.", "I can't comfortably just shove it all into a Dublin core description field. […] description but that's not really helpful.", "I'm not able to find a good enough match and you know to provide. It's just too general, so I just was like, I don't know how to do." [ML15] "I normally prefer to use as specific as I can get if I think I can capture the right data." [ML16] "you know that one it was because Dublin Core is kind of one size fits all metadata schema. You just try to find the one that fits best, and that's what takes most of the time." [ML21] "it's more specific and it's more targeted and more narrowly defined, so it's easy to find the element because you can find the accurate one. So it's more relevant."

Category 13: Sufficiency

According to the participants, certain metadata fields are sufficient to describe a resource, representing the core elements that affect relevance judgment. Sufficiency is a spectrum, depending on factors including purpose, future application and processing, features of the resource, and time/resource constraints. Many categories affect sufficiency, such as time and resource constraints, features of the resource, purpose, potential use, and users.

For example: [ML08] "For me specifically with this project, I would say sufficient and just having the basics, which included title, description, if there was an author date and. Subjects.

Those were what I would consider to be suffice for this project", "It's hard to say what is sufficient because it's very much based on the project you're doing. Who's doing it? What does it entail and what are your constraints?" [ML10] "I tried to include as many as possible. If you have the information there and you can logically structure it into a record that you should try and do that."

Category 14: Using single/as few schemas

More evidence for this category was found in the interview stage. This category is closely related to the scope of terms And this decision can be affected by working context and responsibilities and when considering users.

For example: [ML05] "it's kind of strange to me to mix fields across like to make a sort of a mixed-use schema, so that was that I probably could have spent a lot longer going through." [ML10] "So that's sort of how I approach this as I was looking for one thing, maybe you know one schema, maybe two that could cover what I wanted.", "I like to try and use as few schemas as possible to keep it streamlined." [ML12] "I wanted to use as few ontologies as possible." [ML15] "I prefer to keep it to a single vocabulary as possible or as few as possible." [ML18] "I would start with the name spaces that are the schemas that I've already used because it feels like if you have a wide variety of schemas to describe one object, it becomes nearly impossible to do that consistently throughout." [ML21] "Normally I don't like just to mix like 3-4 schemas together."

Category 15: Widely used

Systems might strengthen this phenomenon if many systems support only a few widely used schemas. This is also reflected in using LOV to search terms. When terms are semantically

identical, participants tend to choose the highly reused one. Wide use affects information retrieval and potential use. For example: [ML05] “it was something I was familiar with, it's widely used.” [ML10] “I know it's pretty broadly used and is a little bit easier to access.” [ML15] “Looking to see, particularly from the open vocabulary site, how many times it's been used. If it's been used a whole bunch of times, I feel a little bit more comfortable.” [ML16] “I would probably stick with. Dublin core, just again out of a sense of time and the fact that Dublin core is a widely used term list.” [ML25] “And is widely used in many of the places [means systems] they limit you to Dublin core.”

### 6.4.3 New Categories

Category 16: Context

Context means the circumstances or settings in which the participants make their decisions, which is determined by purpose and features of the resource. Context consists of working context and affects the scope of terms. Context in this study can be understood as the context of the resources, the domains, etc.

For example: [ML05] “it's widely used, but in in sort of cultural heritage circumstances library.” [ML12] “A library context and BIBFRAME will cover the fact that this is an article.” [ML15] “I felt like the first was a bit more scientific, which is more outside of my range though, and the second one was a little bit more artistic or historical, and previously I'd worked at a museum library, so I felt it was a bit more in my wheelhouse.” [ML18] “thinking about a linked data environment and the information that I work with that I'm familiar with.” [ML21] “normally people use this complicated discipline metadata standards for like research, data, research data like professes work on like research on graduate students work and you know that's my

impression.", "if it's like a real project, like a granted project […] maybe you can give more details."

Category 17: Granularity

Granularity means that participants decide to what extent they describe the resources. It is a choice affected by factors such as purpose, time/resource constraints, and features of the resource. It affects the hierarchy of a metadata schema and can affect accuracy and specificity.

For example: [ML10] "I was like doing this as a regular journal article isn't going to work because there's so much specific like report details." [ML15] "Things that are deep in a hierarchical structure are probably more granular metadata", "I'd want to capture all of those elements as granularly as possible." [ML18] "if I could break those up into the smallest parts, whatever the actual schema that made sense could make use of those parts instead of forcing them all into one field." [ML21] "I personally like a more granular", "I'd like to describe things in the more granular ways it if it has a more granular, more accurate element to describe the content, I would prefer that so more granularity, more relevance and yeah, it would be good.", "you know VRA Core is kind of sufficient but the rubbing schema provided in a bit more granular elements." [ML25] "Because I think it depends what level of detail you want to get put in the metadata record."

Category 18: Hierarchy

In the second stage, participants did not use the term hierarchy, instead, they use wrapper and subcategory. Hierarchy is a more specific term than structure in the second stage codes. Hierarchy means metadata elements are organized into levels according to their semantic relationships. Compared to hierarchy, the other choice is flat, where all the elements

are gathered without levels. Some participants consider hierarchy to be good, but do not use it, which is affected by factors such as granularity, relationships of the resource, future application or processing, time/resource constraints. Hierarchy supports the expression of relationships between fields and provides scope for the lower-level terms. It contributes to specificity and clarity but can lead to complexity and negatively affect interoperability. Some participants vote for hierarchy if with limited levels. Some participants find that internal linkages can be a substitute for hierarchy, like ML05 said “the external linking I think is the thing that would kind of reduce the need for the hierarchy to some degree.” The dilemma of hierarchy exists, like ML25 said “I like the flat one better for like simple things, but for more difficult things you've just got to acknowledge that you actually need to have the hierarchy.”

Examples of arguments for hierarchy: [ML08] “So I find applying a hierarchical structure and providing as much information as possible would be ideal. How much you get on that realistically depends on a lot of things.” [ML10] “It's the way it is structured. It's a little less hierarchical than CDWA, which I like.”, “It's kind of funny because I like having hierarchy in metadata, but I find that it can get really complicated and making records essentially out of it when it gets too deep like I'm thinking of MODS right now.”. [ML12] “But I think it is important to and have some structure to provide that clarity and accuracy of description.” [ML15] “I think if you need to start drilling down more into specific information, it makes sense to have hierarchy.” [ML18] “you know that there were specific areas where it felt like it needed a hierarchy and then the rest felt kind of flat.” [ML21] “If everything's flat, maybe it'll be confusing.”, “If you don't have the system back up, you just want to describe the information, then this just local fields, and I think I need to be hierarchical or else is not right.”, “if you want

to it to be coded correctly, you need to put it in the like a hierarchical way."

Examples of arguments against hierarchy/complex hierarchy: [ML10] "If it's like MODS is very complicated, it's very good at what it does, but it's also like there's you really have to know what you're doing.", "I don't think that the hierarchies between those two schemas could have made them work together." [ML12] "there's a lot of structural metadata. There are a lot of layers of hierarchy, and it has been very difficult to manage.".

[ML18] "I think as far as data entry and data retrieval I find flat to be just easier to work with.", "I think hierarchical is easier for like a human to interpret, but a flat one is easier for a machine and most of the time.", "And I struggled to it makes sense to make those into a hierarchy, but then it was like well, would these be their own item that I could just link to instead of associating it directly." [ML25] "it's gonna be easier to create high quality metadata with for me with my experience from the flat structure that anytime you start adding more fields to it, you're opening up opportunity for error."

Category 19: Importance

Participants consider certain properties to be important in resource description. It is a more subjective category. This category is affected by features of the resource and a participant's prior knowledge.

For example: [ML05] "The Who made the discovery is arguably less important to the creation of any of the early creators and they get their own." [ML08] "I didn't think there were significant enough to [make change]." [ML25] "I think that description is an important thing that needs to be communicated."

Category 20: Language

The languages of the resources and schemas affect participants' judgment and may also cause difficulties in understanding. Language reflects the inherent feature of a resource and is affected by the category of considering users.

For example: [ML05] "I think relevance changes depending on the language that you're using." [ML08] "it's helpful to do that [means using the Chinese standard] because the more information you have the better, because it's not as there are different nuances than they are to English within the language.", "would be doing more of the transliteration, so providing the Chinese version and the English version. […] So ensuring that whatever you have, there are both languages to suffice for that." [ML10] "VRA core is a little more user-friendly, also has recently had a full translation into Chinese. So I thought it would be a good opportunity to use that for a Chinese language material." [ML18] "I think when it started mixing like Japanese researchers and Chinese researchers […] OK, so now I'm dealing with two languages that I can't recognize at all. So, and I'm in real trouble."

Category 21: Quality

Quality refers to the quality of metadata records. Quality is affected by sufficiency. The participants did not address the specific factors affecting quality but talked about it generally.

For example: [ML08] "I know for a fact that one of the top five reasons why users can't find stuff is either their lack of understanding of that metadata or insufficient metadata. So it does play out a huge part in our access and discoverability of resources.", "If you're doing like research, then that resource could be hidden because it doesn't have sufficient metadata", "physical items tend to have much better metadata than the electronic resources, […] it does

affect how our resources are discoverable." [ML15] "I hope so. Hopefully we're able to as metadata quality increases and as folks become more aware of the fact that data and research and items can be used in more ways than we may have anticipated and start integrating more of this."

Category 22: Relationship

Participants consider relationships an important property of resources and something they must consider when describing a resource. They prefer to choose schemas that can support the representation of complex relationships. Relationships are determined by the features of the resource. It can be represented via hierarchies, which is one of the reasons why participants opt for hierarchy. Relationships are closely related to the category enabling linkages identified in the second stage.

For example: [ML10] "I didn't think Dublin core was going to be useful for the rubbings because that's really complicated and it just can't support the relationship.", "It's non-hierarchical, so it just can't support the relationship stuff, so I completely ruled it out there.", "I think I was so focused on trying to get the relationships right that I just glossed over the reproduction." [ML12] "The rubbings were a real challenge because there's so many different layers".

Category 23: Scope of terms

It means the range of terms. The scope of terms can be provided by both schemas and hierarchy and is also determined by context. Specifying the scope of terms can increase specificity and accuracy.

For example: [ML10] "I use something in schema that is generally used to describe the

intended age of, like people interacting with this. So specifically talking about humans generally. So in my head, yeah, we're describing not humans with this." [ML12] "I could have used some schema design for an astronomy resource and try to use that to describe this rubbing, but it just seemed ridiculous. It's like and you know, I should first try to find something that is resource specific or discipline specific and sort of try to describe it in context." [ML15] "So I wanted to make sure the scope of the term, even if the term says one thing, if the scope says something else, I wanna make sure that we're fitting within the scope of that term as well." "I felt age wanted numbers in a lot of cases and life stage you could say something, especially in this case like hatching or laying. Or you could give kind of the stage of the animal and that just was overall better than trying to guess an age range for the birds, I think." [ML21] "It's Audiovisual Core. I mean the resource you gave me is not an audio or video, so it's a little bit confusing."

Category 24: Technology

Two participants mentioned technology, but they were referencing different things. One referred to the encoding system and the other referred to the metadata management system. Therefore, technology here refers to the technologies and systems used to operate metadata. Technology supports the category of widely used.

For example: [ML05] "I use Oxygen XML editor and I would get everything I wanted in and then I might rearrange it just purely aesthetically I think." [to argue hierarchy of metadata elements doesn't matter], "I think if you're going to display them in a system later, then that system can make the decision for how it wants to organize things for visual." [ML21] "And because it really depends on the system. I mean which system you have? I mean which one it can be accommodated?", "But if you don't have a system who can accommodate it, then the

only thing you can do is to use DC."

Category 25: Time and resource constraints

This is a category mentioned by most participants. It refers to the limits that affect relevance judgments of the participants. It shows that participants make different decisions when they face time and resource constraints. Participants tend to choose general and sufficient terms/schemas when having limited time and resources. Time limit is the major constraint according to the participants.

For example: [ML08] "But because of the circumstances, the timing and everything, it was the best within those constraints.", "So usually you're not going to be too involved because you don't have the resources to do that.", "Ideally I would say hierarchical would be the better option. If you do have the time and resources to focus on." [ML10] "I'm going to just not use the data because it's not going to be worth my time." [ML12] "I'm gonna try to do this, you know, just as one stop shop as possible. And it's really about time savings", "It's like I only have so much time, I'm not going to look at three different ontologies." [ML15] "it's challenging, it's complicated to find a balance between adding in high quality metadata and also completing projects right." [ML16] "So if I had the time and if I had the need to do it, I would sit down and go through it and learn it." [ML21] "it's time-consuming action to code the details." [ML25] "I think my reason for changing or not changing was feeling a little time pressed because I wanted it to get it done in like a day.", "I think another aspect is just the time that it takes to create the record and we are trying to create as many records as we can in a short period of time."

Category 26: Working context and responsibility

Working context and responsibility is a subcategory of context. It provides a specific

context affecting judgmentCompared to prior knowledge, this category focuses more on the outside, not the “inside” of the participants.

For example: [ML05] “I do have to do a lot of like quality check in terms of the catalog metadata that we have. So I use date a lot for that to identify.” [ML08] “I would say as a new resources librarian, I do work with metadata, but I work with metadata not the same way as a catalog or would work with metadata.”, “So I think it depends on the institution […] will influence your outcomes.” [ML12] “I'm working within a professional context and how much my thinking has sort of been shaped by that”, “And I went to IPTC just because that's what I use on a fairly, regular basis.”, “I think so much of it has to do with my role in managing information and performing authority control within the space of our digital asset management system.” [ML15] “I'm working from a single set and trying to conceptualize what may be useful across multiple different resources. So this was a little bit different, especially since they had two very different focuses.”, “there were pieces of both that were easier for me to find terms for because there are things I work for more frequently.” [ML16] “The librarians and our digital repository itself do use Dublin core, so I refer to that to see how”.

### 6.4.4 Common Questions

This section discusses some common questions in the interviews.

Question 1: Most difficult part

Participants were asked what they thought were the most difficult parts of the tasks. Their responses varied. To sum up, there are several points: first, finding accurate schemas to match the resources; second, identifying the important properties of the resources; third, finding the appropriate terms to represent the properties; fourth, finding a good way to merge

all the elements in one schema; fifth, balancing the tasks and the time constraints.

For example: [ML05] “I think the most difficult part was trying to figure to think about how to best represent the more nuanced pieces of each of the artifacts.” [ML08] “it was difficult because as I was doing it I kept thinking if this was the best that I could […] within those constraints.” [ML10] “And so that was kind of hard to meld into one record.”

[ML15] “Definitely finding terms for places where I personally have knowledge gaps.”, “I'm not familiar with and vocabularies, those were the most challenging.” [ML16] “I think the primarily just trying to find a vocabulary to most accurately match it.” [ML25] “Getting your mind around what the identifying the important pieces of information that needed to get put into the metadata set.”

Question 2: Attitude toward LOV

All the participants visited LOV but only a few used it to search successfully. LOV provides over 50 English metadata schemas, covering various domains. Only one out of nine participants retrieved all the terms from LOV. Some participants used LOV to search for schemas. Some participants used LOV to search for terms, not schemas. Once they got something they were familiar with and sufficient to describe the resources, like they found an anchor, then they stopped searching and turned to the schema directly. The participants showed different paths to using LOV, with various levels of involvement. Path 1 is: search terms/schemas → stop using; Path 2 is: search schemas → looking for the familiar ones in the result → confirm and stop using; Path 3 can be summarized as: search terms → determine a possible term → check the schema of the term → determine the schema → stop using and turn to the schema; Path 4 is: search terms → check terms, related schemas, and scopes →

determine terms (with schemas) → create a schema for each task resource → stop using.

The reasons why they did not engage in LOV were: 1) they didn't get the anticipated result, 2) LOV didn't support subject search (for schemas), 3) they got numerous terms which were hard to handle, 4) the LOV site was difficult to understand, 5) the terms were from different schemas, 6) they decided to rely on previous experiences at the beginning.

An example of comments for LOV: [ML18] "the site was very user friendly as far as coming through and so usually what I did is I would search the term that I was looking for and then look at the results and the different schemas that came up…"

Examples of comments against LOV: [ML05] "I found that difficult to use", "This like the overwhelming utility of the resources", "they're [the terms] not part of the same schema, so that would have been kind of strange to me." [ML10] "I found it very hard to navigate […] it's kind of tricky to search through." [ML15] "if you type something in, you can get 50 or 70 or 500 terms for the same thing that you're looking for. So being able to narrow those down and make sure that I was, I was actually finding the terms that I was looking for was a little challenging." [ML16] "It was just the sheer overwhelming amount of vocabularies that are out there", "I had difficulty understanding the structure of the website […] so I think it's just pure ignorance.", "I think the challenge […] was primarily because I couldn't search by subject." [ML21] "It has so many elements and it's not easy to choose one because it's not clear if you don't study the standard very carefully"

Question 3: Attitude toward structured information

The structured information of each resource was provided to participants in the first stage for their reference. Participants showed their appreciation of the structured information

and thought it was very helpful and timesaving. Some participants even didn't examine the original texts and relied fully on it to complete the tasks. This implies that a well-structured and described piece of surrogate information is more relevant to users' needs than the original text. Structures contribute to users' judgments.

For example, [ML05] "Informational aspects from the artifacts that I think that was very helpful", "it was very helpful, supportive work to figure out what were the most probably the most important aspect." [ML10] "I think if I personally am looking, you know, at data and I can tell it's not structured well, it's going to be very frustrating." [ML12] "There was some text passage and then there were data values that had been selected from that text and I really focused on that abbreviated set of information." [ML15] "So I tried to take what was kind of highlighted in that structuralized information and tried to make sure that I had it some type of element or some place within the schema that that information could be captured." [ML16] "It just saved me a lot of time. I could probably if I had the time, go and pick out that information and the resources themselves, but it's nice to have it right there in front of me." [ML18] "used the structuralized information as sort of the primary data points of how I organized it." [ML25] "I believe that I used it more for the rubbing where it seemed like there were very complicated layers of what you were describing."

Question 4: Strategies

Participants used different strategies to approach the tasks. These strategies are partly associated with the attitude toward LOV, since it was the tool that they used. The features are summarized as below.

1) Keeping the original design if there is no obvious inconsiderateness.

Unless something was obviously lacking consideration or erroroneous, some participants preferred to keep their original design. The original design may not be the best but was sufficient. We can assume that there is a threshold. Within the threshold, any better options are not worth the change. However, participants have different thresholds, which can be seen from the examples that one participant did not make any changes in task 2 while another one changed all the fields.

Some examples of evidence are: [ML08] “I didn’t think the changes would have made that much of a difference based on what I did in the first phase.” [ML25]: “I'm already using Dublin core, I'll just use the Dublin core subject because I don't need to add a new one in there.” [ML15] “I just kept my term. That was probably rude of me, but I could have used either, but I just kept buying because that was in the original and I was doing a revision. So if I met the same scope and I was capturing the same thing, I just kept my term.”

2) Selecting schemas first, then terms within the schemas.

Schemas provide scope and context for the terms. Some participants prefer to have a scope first to help them choose terms within the scope. They believe once they find an accurate schema, it is easy to find accurate terms.

For example: [ML10] “So I did schema first and then worked from there out” [ML12] “I chose the ontologies and schemas first.” [ML25] I found the vocabularies first and then then looked at the terms and then chose the terms that I wanted to use.”

3) Selecting terms first, then examining the vocabulary, and then going to the second strategy if needed.

Using this strategy, participants usually start with their familiar terms. Their familiar terms,

acting like an anchor, provide schema information for further decisions.

For example: [ML08] “I start off with that to trying to find like those like 5 or 6 fields that I know users look for and then go from there.” [ML15] “I focus more on the terms. […] But if I didn't have a property off the top of my head, then I've worked more with and scopes and expanded to different vocabularies.”, “things that I knew that I could capture in vocabularies I already knew, like title or author, those I just checked to make sure that the scope was correct.” [ML16] “I think it would have been terms 1st and when I first started work on this. That's what I did was search by term. I think I built up to trying to search by vocabulary gradually after that.”

Question 5: Locating data and determining relevance

Participants talked about this topic on the broad concept of data, including various types of resources, even physical resources. They talked about where to find data, such as Google and Google Scholar. Several participants mentioned that when looking for data, they would like to start with broader topics and then narrow them down. Using Google was one of the approaches to start with broader topics.

The participants mentioned some properties they used to locate data. Four participants mentioned subject, four mentioned date, three mentioned citation, two mentioned format, two mentioned title, two mentioned author, and one mentioned abstract, description, geographic range, keyword, and context. One participant argued if data were without subject information, it would be frustrating. The result is consistent with the analysis in Section 5.4.3, where metadata fields are analyzed from the perspective of relevance judgments.

For example: [ML05] “I often use subjects to search and sometimes title or author or whatever the representative field. At that point I like because that's as wide as I would gather in

things I've not seen before and that so I usually leave it pretty wide", "to spread my net super wide" [ML08] "I start off with subjects and then I get more granular. So then the description will play a part.", "I start from like the broader topic and then I slowly narrow it down." [ML10] "But broadly, usually just using subject information is oftentimes how all assess whether it's relevant or not.", "If I'm looking for something I know is in a specific date range, we'll do dates first. If I know I'm looking for something, particularly in a geographic range, I'll look in the geographic range first." [ML16] "one of the biggest challenges is defining subjects, and sometimes that takes exploration of the resource itself to determine its subject.", "What does this resource about? Essentially, what is the format of this resource?" [ML18] "I think that for the most part it was based on keyword searches", "I think the only other way would be as I mentioned, like the looking for citations and the crossover of citations as the other search point." [ML25] "Depending on what I'm looking for […] Google Scholar […] I'll go through the citations to see who has cited what.", "I tend to have a broad search and then if it's too broad I'll limit it by date because I don't want all the old stuff and look at what's been published" [ML21] "if you don't know which place to looking for and Google is not bad, and then once you have more specific like you look at a few articles and you can look at the references and find more articles"

Question 6: Relevant data

The responses show participants' views on relevant data or data relevance and are analyzed case by case.

[ML08] "So relevant data depends on the like what you're researching. Like for science research, the date plays an important role and where it's published, who wrote the article. Like the author, the publisher, those play an important role when it becomes something like

historical, then other factors come into place. So I always think it's important to know what type of thing you're researching, where what type of area, science, social science history, and then trying to figure out what's relevant to those fields. And then going from that, so maybe you need to start by doing like a refresher and then trying to figure out what type of resources would work for you based on your research topic."

**Summary:** 1) The importance of data relevance criteria varies across domains; 2) Data must be put in a context to be relevant.

[ML10] "I think relevant data is very context specific, so depending on who is looking for what different data is going to be important, which again I think is why it's important to include as much as possible because it can be relevant in some situation. […] I'm thinking of comics for some reason, like a pencil or something, you know, not including that is going to be a disservice to somebody who is doing research on, like, pencillers and comics. So I think relevance is contextual and as metadata like we should be trying to encompass like as broader range of information as we can to, you know to provide as much information as possible. I'd rather have more information than less, or think that I know better than users what is relevant to them."

**Summary:** 1) Data relevance is contextual, purposeful and situational. 2) By providing full description of resources, information professionals can increase the possibilities of data to be relevant in different situations.

[ML15] "I think it depends on what I believe the use case for the item or the object will be. So I want to make sure that any information that I'm putting on there is useful to the highest amount of people who will be accessing the resource. I recognize now, especially as we're starting to get into more linked data environments, that we can never anticipate all the needs of

every user, every time. That would take me 100 years to catalog the book and I don't have that right time. So I think it's a handoff between trying to understand the information seeking needs of the user and the amount of time and resources you have to dedicate to a project.

**Summary:** 1) relevance is partially determined by the usefulness to most people (users), not a single person. 2) relevance takes time and resources at hand into account.

[ML16] “I guess the way we judge whether it's relevant or not is does IT support discoverability? In other words, does it help someone find our resources? And that's not always the case. Uh, sometimes we keep a vocabulary in there because it represents the ethical mission of the institution, for example.”, “We came across a vocabulary which expresses the LGBTQ community, and normally it's a vocabulary we would not include just that's not relevant to what we're expressing. But because it's talking about this community and the mission of our institution is partially about social justice and inclusion, we keep that term in there to be expressed in our records”.

**Summary:** Relevance has two meanings: 1) On one hand, there is relevance for resource discovery, which means relevance is about the relationships between resources, the properties used to describe resources, and how you describe the resources. It is more focused on the resources per se. 2) On the other hand, we have relevance for the mission, which means relevance is determined by purposes and demands other than resource discovery, mostly affected by factors of human society.

[ML18] “So I think that for me the relevant data is really about the initial connections and the ability to link objects together with the hope that items with more with more links or you know have will have a higher connection to each other, or at least that when you are

looking for something obscure that you know the additional linking will connect you to a resource that you might not otherwise find based on keywords or indexing or something like that. So I mean I think that as noted that you know a lot of it has to do with breaking the whatever metadata or topics involves in an object into its smallest parts so that they can be assembled in a way that is relevant for whatever they use of the researcher."

**Summary:** 1) Relevant data refers to data linked based on relationships. 2) One possible path to make data relevant is breaking down information into granular data units, linking them via relationships, and assembling them case by case to serve a user's purpose.

[ML21] "I think to the relevance it should be relevant to your research goal. Like what you are trying to look for? It could provide you with a, you know, research either context or content. Object to be started on provide you with some sorts to expand your research. Relevant […] it's not that narrow. Actually, if you expanded relevance and then you can actually find more stuff. So and yeah, it will depend on the subject, and depend on your research goal, and also your research process. Probably it can be changed and expanded and it depends on whether it's a static thing you're researching on or it's an adaptive or it can be a growing project."

**Summary:** 1) Purpose provides a context that which relevance is embedded; 2) Relevance is a matter of degree. It can be adjusted to get the anticipated results, e.g., participants mentioned they would like to go broadly first and then narrow down. The threshold of relevance changes alongside the changing scope. 3) Relevance is not static but evolves with changing conditions. Time is a dimension of relevance, which specifically means the changing situations along with time.

[ML25] “Well, and that's sort of interesting too, because I have come across like sort of irrelevant data that is still interesting to me, right? Well, I'll make a side note and I'll say I need to research that later, right? So sometimes when you run into information that's not specifically what you're looking for, it still can be meaningful to you. So I'll just say that about relevance, that sometimes the irrelevance stuff is also meaningful in a different way.”, “I sort of also like to go outside of my field and look for what the other fields are about. […] So it's not necessarily within the field of library science, but it's the same. You've got to look at how that concept is being applied in other areas. So I think that's that I think is 100% relevant.”

**Summary:** 1) Topicality is just one type of relevance. There are many ways that data can be relevant to you. 2) The purpose of information seeking behavior is important to determine relevance. 3) Domains (context) affect data relevance, but it does not mean data need to be in the same context or domain to be relevant. 4) Relevance can be for the present or for the future.

## 6.5 Supplement by Discriminant Views

The expert (DE01) (discriminant sampling) provided the views as a supplement. The key points are:

1) Time is an important dimension of data relevance, since relevance can change over time.
2) The accuracy of values (accurately and precisely describing the value of a property) is important, which affects the potential application of the values.
3) Domain is the factor determining the schema to choose.

4) Data must describe the key features of tasks and support tasks and various stages of the tasks.
5) How the users use the data must be considered. For a metadata schema, it is about how the users can interact with it.
6) Fields enable data to be linked to outside resources.

Most points are covered by the main data, except the second one. Accuracy is a category in the second stage and the interview stage. However, in the main data, accuracy means using a semantically accurate term to represent the property, where the participants don't consider the future application. In the discriminant data, it emphasizes the accurate value that affects potential use.

## 6.6 Selective Coding and Modeling

There are six categories identified in the second stage but not in the interview stage, which are: Category 2: Better terms, Category 4: Description capability, Category 5: Enabling linkages, Category 7: Fields not supported by original schemas, Category 16: Reducing duplication, and Category 18: Structure.

The data show that better terms are specifically explained as specific terms and terms with appropriate scopes, which are duplicated with other categories. Description capability can be affected either by using generic or specific terms, duplicated with the categories of generality and specificity. Category 7 is not representative, since self-created terms are allowed. Therefore, these categories were dropped and were not included in further analysis. Structure is merged with hierarchy. Enabling linkages and reducing duplication are added to the interview stage for selective coding. A total of 28 categories are shown in Table 22.

Table 22: Categories created in axial coding

| | | | |
|---|---|---|---|
| Accuracy | Generality | Potential use | Specificity |
| Considering users | Granularity | Prior knowledge | Sufficiency |
| Context | Hierarchy | Purpose | Technology |
| Enabling linkages | Importance | Quality | Time and resource constraints |
| Features of resource | Information retrieval | Reducing duplication | Using single/as few schemas |
| Flexibility | Language | Relationship | Widely used |
| Format | Points missed | Scope of terms | Working context and responsibility |

These categories are associated with each other according to the previous analysis. Table 23 shows the categories and their associated categories.

Table 23: Categories and their associated categories

| Category | Associated categories |
|---|---|
| Accuracy | Granularity, Hierarchy, Potential use, Reducing duplication, Scope of terms, Specificity |
| Considering users | Generality, Information retrieval, Language, Potential use, Prior knowledge, Specificity, Sufficiency, Using single or as few schemas |
| Context | Features of resource, Purpose, Scope of terms, Working context and responsibility |
| Enabling linkages | Information retrieval, Relationships |
| Features of resource | Context, Flexibility, Generality, Granularity, Importance, Language, Relationship, Specificity |
| Flexibility | Features of resource, Format, Generality |
| Format | Flexibility |
| Generality | Considering users, Features of resource, Flexibility, Working context and responsibility |
| Granularity | Accuracy, Features of resource, Hierarchy, Information retrieval, Potential use, Prior knowledge, Purpose, Specificity, Time and resource constraints, Working context and responsibility |
| Hierarchy | Accuracy, Granularity, Potential use, Prior knowledge, Relationships, Scope of terms, Specificity, Time and resource constraints |
| Importance | Features of resource, Prior knowledge, |
| Information retrieval | Considering users, Enabling linkages, Granularity, Potential use, Specificity, Widely used |
| Language | Considering users, Features of resource, |
| Points missed | Prior knowledge |

| | |
|---|---|
| Potential use | Accuracy, Considering users, Granularity, Hierarchy, Information retrieval, Sufficiency, Widely used |
| Prior knowledge | Considering users, Granularity, Hierarchy, Importance, Points missed |
| Purpose | Context, Granularity, Sufficiency |
| Quality | Sufficiency |
| Reducing duplication | Accuracy |
| Relationship | Enabling linkages, Features of resource, Hierarchy |
| Scope of terms | Accuracy, Context, Hierarchy, Specificity, Using single or as few schemas |
| Specificity | Accuracy, Considering users, Granularity, Hierarchy, Information retrieval, Scope of terms, Working context and responsibility |
| Sufficiency | Considering users, Features of resource, Potential use, Purpose, Quality, Time and resource constraints |
| Technology | Widely used |
| Time and resource constraints | Granularity, Hierarchy, Sufficiency |
| Using single/as few schemas | Considering users, Scope of terms, Working context and responsibility |
| Widely used | Information retrieval, Potential use, Technology |
| Working context and responsibility | Context, Generality, Granularity, Specificity, Using single/as few schemas |

The connection diagram is created using Obsidian to show the links between the categories and the discriminant data. First, all the categories were created as titles of notes. Second, under each title, the associated categories were input as the content of the note via the "Add link" function. To make the association clear and consistent, the active party of a connection was input as the title, while the passive party was input as the content. Figure 21 shows the split screens of Obsidian. The left column shows the categories generated in axial coding. The middle column shows the relationships between the categories. The right column shows the graph of the linked categories. Figure 21 shows the category *Features of resource*, its outgoing linked categories, and its position in the diagram.

Figure 22 shows the linkages between the categories. The arrows from A pointing to B in

the diagram indicate that A produces/causes/affects/supports/contains/decides B, where A is the active party and B is the passive party.

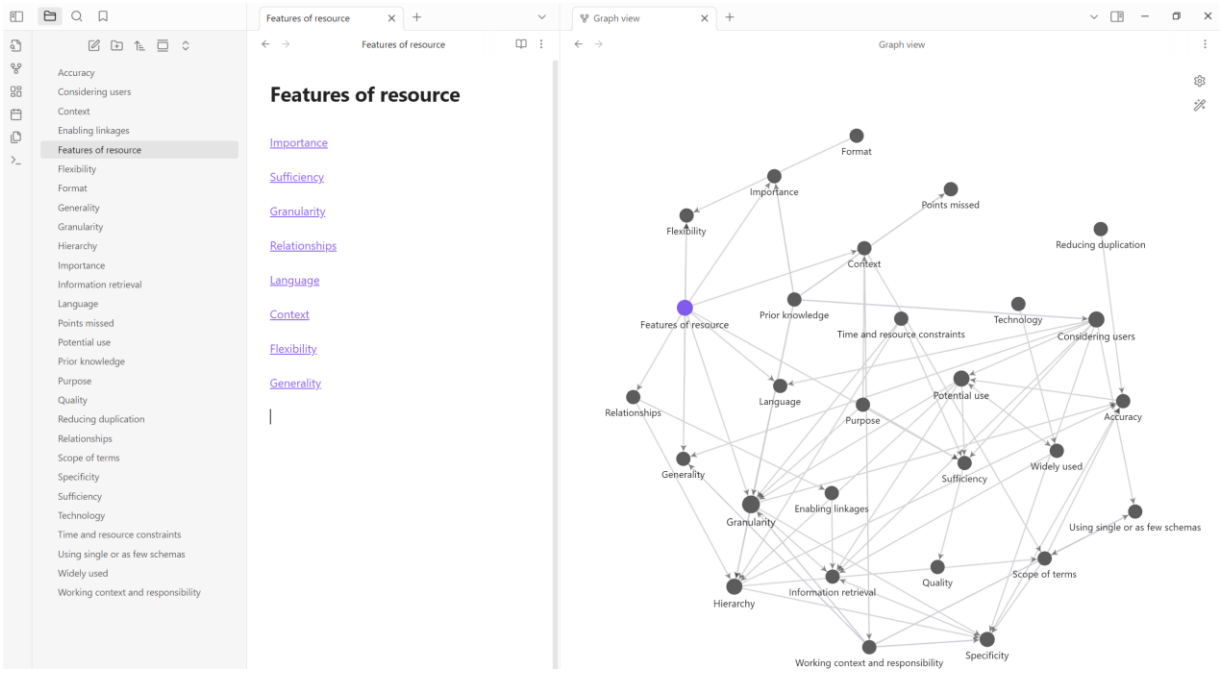



Figure 21: Screenshot of the Obsidian interface

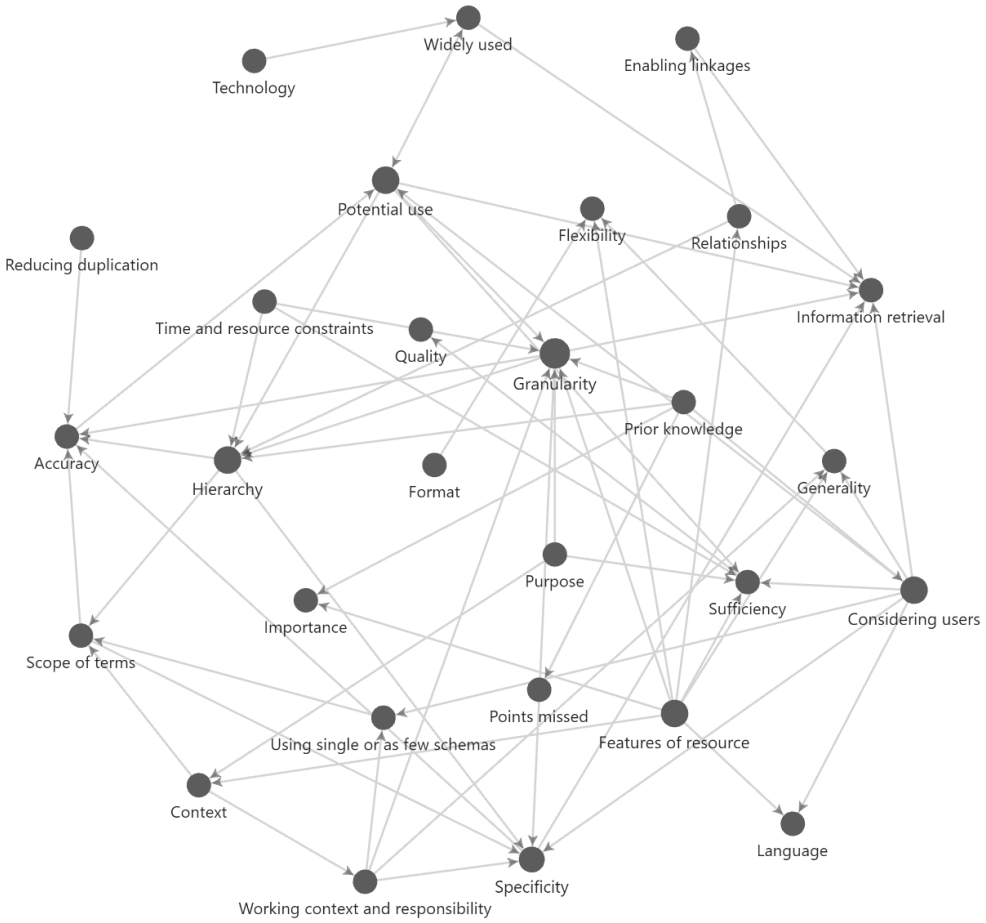


Figure 22: Category connections diagram

Some categories have more connections (both in and out) than others, which means they are the more core phenomena in the data relevance study. However, core phenomena cannot be determined by the number of connections only; incoming-relationship (an arrow pointing toward a node) and outgoing-relationship (an arrow directing away from a node) must be considered as well. Incoming relationships indicate that a node is produced/caused/affected/supported/contained by other nodes. More incoming connections mean the category has more upper-level elements. Outgoing relationships indicate a node produces/causes/affects/supports/contains/decides other nodes. More outgoing connections imply the node has more lower-level elements and has a larger influence, which can be seen as the core category of this study. Table 23 shows the statistics of the categories with a total

number of connections, incoming connections and outgoing connections.

Table 24: Statistics of category connections

| Categories | Number of connections | Incoming connections | Outgoing connections |
|---|---|---|---|
| Features of resource | 8 | 0 | 8 |
| Considering users | 8 | 1 | 7 |
| Potential use | 7 | 2 | 5 |
| Prior knowledge | 5 | 0 | 5 |
| Working context and responsibility | 5 | 1 | 4 |
| Granularity | 10 | 6 | 4 |
| Hierarchy | 8 | 5 | 3 |
| Purpose | 3 | 0 | 3 |
| Time and resource constraints | 3 | 0 | 3 |
| Context | 4 | 2 | 2 |
| Relationship | 3 | 1 | 2 |
| Scope of terms | 5 | 3 | 2 |
| Specificity | 7 | 5 | 2 |
| Accuracy | 6 | 5 | 1 |
| Enabling linkages | 2 | 1 | 1 |
| Format | 1 | 0 | 1 |
| Generality | 4 | 3 | 1 |
| Reducing duplication | 1 | 0 | 1 |
| Sufficiency | 6 | 5 | 1 |
| Technology | 1 | 0 | 1 |
| Using single/as few schemas | 3 | 2 | 1 |
| Widely used | 3 | 2 | 1 |
| Flexibility | 3 | 3 | 0 |
| Importance | 2 | 2 | 0 |
| Information retrieval | 6 | 6 | 0 |
| Language | 2 | 2 | 0 |
| Points missed | 1 | 1 | 0 |
| Quality | 1 | 1 | 0 |

* Categories are sorted by outgoing connections.

When determining the core categories, the three types of numbers are used synthetically. The core phenomena cannot be grounded on a single indicator. The main principles to determine the levels (which can be seen as having different impacts on relevance judgment) of the categories are: 1) The number of connections is the foundational criterion. The

larger the number, the more connections to other categories, and the more impact the category probably has. 2) The numbers of outgoing connections and incoming connections. More outgoing connections mean a category affects more other categories, and more incoming categories mean a category is affected by more other categories. In other words, when two categories have the same number of total connections, the category with more outgoing connections will probably be at a higher level than the category with more incoming connections. Thus, the number of outgoing connections is a key indicator to assess the level of a category. 3) The impact of a category. When a category affects other high-level categories (has outgoing connections to high-level categories), the category is at a high level.

Based on the principles, I observed: 1) Categories with a large number of total connections but a low number of outgoing connections are at a lower level. For example, Accuracy and Sufficient (6 total, 5 incoming, 1 outgoing), and Information retrieval (6 total, 6 incoming, 0 outgoing) are not at the top level. 2) The number of total connections is not an absolute indicator. A category with a small number of connections may be at a higher level, depending on the categories it affects. For example, one of the outgoing connections of Purpose (3 total, 0 incoming, 3 outgoing) points to Context (4 total, 2 incoming, 2 outgoing), and the only incoming connection of Working context and responsibility (5 total, 1 incoming, 4 outgoing) is Context. There is an influence chain: Purpose → Context → Working context and responsibility. When Working context and responsibility is not at a low level (because of the number of outgoing connections), Purpose and Context, although they do not have a significant number of total connections and outgoing connections, are the core categories. Another example is that Granularity has the most total connections, but only 4 outgoing connections that are not at a

high level; thus, Granularity is not a core category. 3) The number of outgoing connections is not an absolute indicator. For example, the Working context and responsibility category has 5 outgoing connections, while the Context category has 2. It cannot be concluded that Working context and responsibility is more important than Context, since Context has an outgoing connection pointing to Working context and responsibility. 4) The same number of total connections, incoming connections, or outgoing connections does not guarantee the same level of the categories.

Based on the principles, an adjusted hierarchy of the categories is shown in Table 24.

Table 25: Category hierarchy in selective coding

| Top-level | Second-level | Third-level |
|---|---|---|
| Considering users | Accuracy | Flexibility |
| Context | Enabling linkages | Importance |
| Features of resource | Format | Information retrieval |
| Potential use | Generality | Language |
| Prior knowledge | Granularity | Points missed |
| Purpose | Hierarchy | Quality |
| | Reducing duplication | |
| | Relationship | |
| | Scope of terms | |
| | Specificity | |
| | Sufficiency | |
| | Technology | |
| | Time and resource constraints | |
| | Using single/as few schemas | |
| | Widely used | |
| | Working context and responsibility | |

There are seven categories without incoming connections, which means in the KO context, they are the source categories or independent categories. Their outgoing connections vary from 1 to 8, and they are scattered across different levels, indicating their different levels of importance in data relevance theory. The importance of the category refers to the influence of

the category on other categories, which does not necessarily mean the weight of the factors of data relevance.

The core elements of data relevance are the top-level categories, which are features of the resource, considering users, potential use, Prior knowledge, purpose, and context. Upon analysis based on Table 23, Table 24, and Figure 22, and after a holistic assessment, a data relevance model is created as in Figure 23.

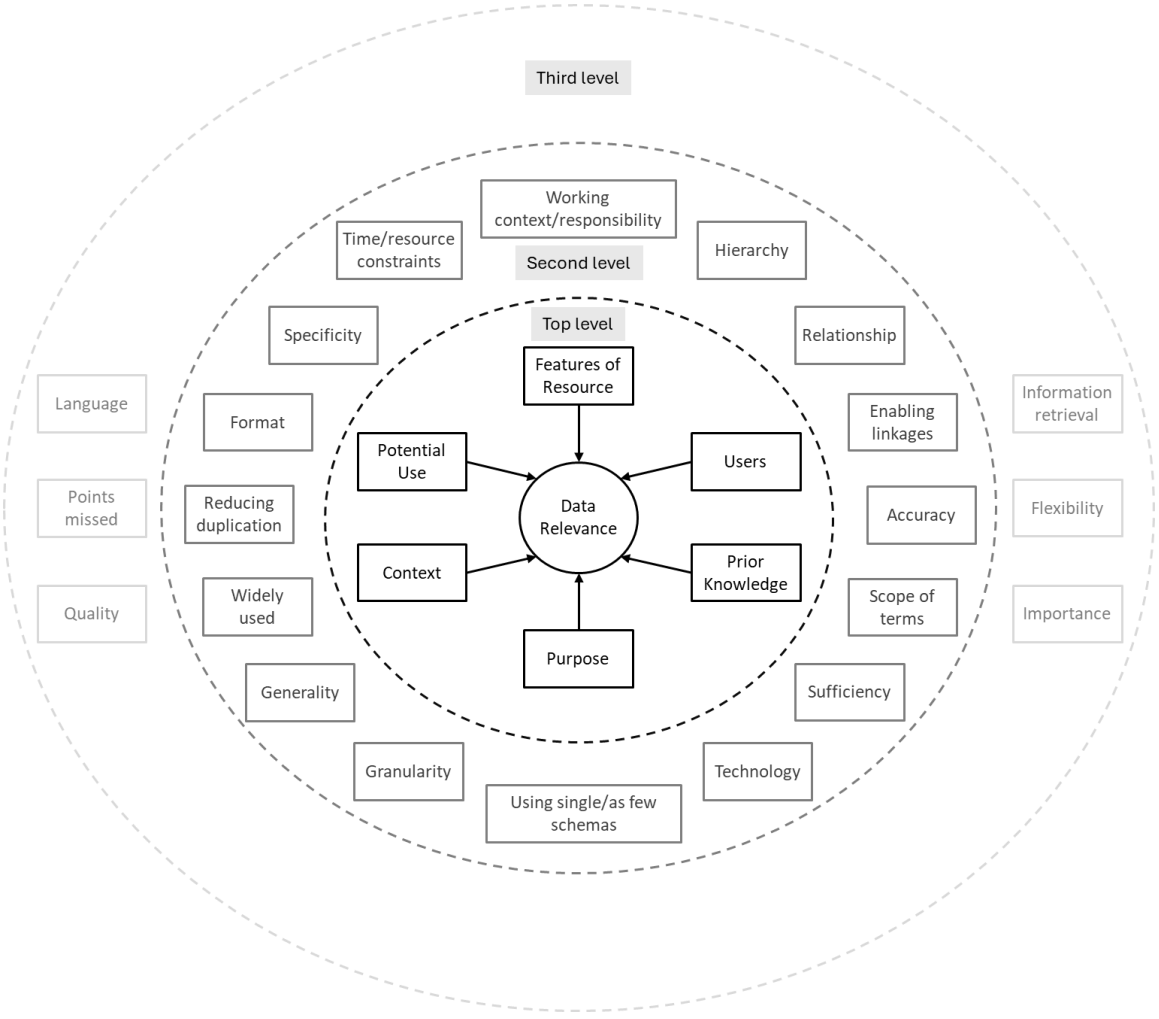



Figure 23: Data relevance model in the KO context

The model has three levels of elements functioning on data relevance. The higher the level, the more impact on other elements in the model, the possible larger influence on data

relevance, since higher level elements can trigger a longer impact chain. For example, the element Purpose can trigger Purpose → Context → Scope of terms → Specificity → Accuracy → Potential use… (shown in figure 10). Purpose may have a wide impact along the chain. As a functioning element, it will have a significant effect on data relevance if many elements on the chain are triggered.

The top-level elements are not data-specific, which can be used for resources other than data. The second level elements show more data features, such as scope of terms, granularity, enabling linkages, using single/as few schemas, etc. It is interesting to see that information retrieval is an end node (without outgoing connections) in the KO context.

In the next section, conclusions will be drawn about the data relevance theory based on Chapter 4: Relevance in Knowledge Organization, Chapter 5: Relationships in Knowledge Organization, and Chapter 6: User Study and Data Relevance Model.

# 7 Data Relevance, Discussion and Future Work

## 7.1 Summary of the study

The goal of this study was to conceptualize data relevance in the knowledge organization context and build a data relevance theory accordingly. To accomplish the goal, the studies answered three research questions: Question 1: What is relevance in terms of knowledge organization? Question 2: What relationships exist in knowledge organization? And Question 3: What is data relevance in the knowledge organization context?

The answers to Question 1 in Chapter 4 laid the foundation of the study and examined whether studying the concept of relevance in the KO context was appropriate and reasonable. Question 1 facilitates researchers and information professionals to jump out of the box of IR relevance studies to approach relevance studies in a new context. Chapter 4 shows that many core concepts in KO are associated with relevance, and it is reasonable to study relevance in the KO context. It demonstrates that the subject of relevance is more about data and naturally prompts data relevance in the KO context. The result shows that relevance in KO is about relationships of information, data, and information professionals. It calls for further study on what relationships exist between information and data (Question 2) and between information/data and information professionals (Question 3).

How should information and data be studied to understand relevance in the KO context? What can KO provide to facilitate study relevance? Based on the conceptualization of relevance and the emerging data relevance in the KO context, and the previous relevance studies in the IR context, these issues converged on one concept: relationships. The relationship was the reason why Question 2 was proposed. Relationships, as a core concept extensively embedded in

knowledge organization systems, and a term frequently seen in interpretations of relevance, are the way to explore relevance. Relationships demonstrate how KO is related to relevance study. The relationship study is not just a conceptual discussion on relevance but moves to ground it in practice. Another important consideration is that relationship studies can leverage numerous KOS and research results to explore relevance. Relationships, especially in the data universe, solidify the study of data relevance. The relationship study in Chapter 5 shows one aspect of data relevance. Users are another vital and indispensable aspect, which is the reason for proposing Question 3.

This dissertation conducted a user study to address the following related questions in Chapter 6: what is data relevance? What processes can be involved in informing relevance? What behaviors/activities need relevance judgment on data? Another noticeable point is that users are not end users but information professionals as an intermediary between data and end users in the KO context. Based on these questions and the analysis in Chapter 4, a user study of information professionals was designed. The user study investigated how the participants designed metadata schemas and how they made their decisions on changing or not changing their designs after extra information was provided. For completing the tasks, participants were involved in subject analysis, term selection, considering future cataloging, etc. Relevance functioned on their decisions throughout the process, especially in the second stage. A data relevance model was created at the end of the user study based on the tasks and the interviews.

In summary, to conceptualize and theorize data relevance in the KO context, this study first examined the processes and concepts relevant to the concept of relevance in Chapter 4,

including subject analysis, aboutness, indexing, cataloging, social tagging, and warrant, to get informed about how relevance functioned in KO. Based on the conceptual exploration, this study further examined relationships as an embodiment of relevance in KO in Chapters 5 and 6. That is to say, there must exist some relationships to enable relevance. Relationships were used to answer the questions about how information and users' information needs were related and in what concrete form they are manifested. This study analyzed the relationships between resources at the bibliographical level, the information level, and the data level, and disclosed how data were linked/related in Chapter 5. In Chapter 6, information professionals, a special group of data users whose routine responsibilities were about resource description, including subject analysis, cataloging, or metadata-related work, were involved to explore the relationships between data and users. The empirical study investigated how participants judged relevant data, what factors they considered, and led to a data relevance model.

## 7.2 Highlights of the Findings

Conceptualization of relevance in knowledge organization, the relationship study, and the user study of data relevance have revealed several significant aspects.

1) Relevance has two foundational attributes: objective relevance and subjective relevance. They work together to determine the relevance of information (for the sake of conciseness, use information to cover information, data, and other information resources) to a user's information needs. Objective relevance is rooted in the inherent features of information, where knowledge organization has systematic ways to deal with these features. Subjective relevance is determined by individuals in a certain context, based on previous accumulation, with personal preferences, and for specific tasks.

Objective relevance and subjective relevance do not have an absolute logical correlation or a consistent relationship. Objective relevance might be subjectively irrelevant; subjective relevance might be objectively irrelevant.

2) Relevance is conceptualized via analyzing knowledge organization activities, including subject analysis, indexing, cataloging, and social tagging. These processes involve many judgments along the way, which are affected by many objective and subjective factors. Objective relevance and subjective relevance have unequal functions in these processes. For example, in social tagging, subjective relevance has a significantly larger influence than objective relevance. The user study in Chapter 6 adds another KO activity: creating metadata schemas, with empirical data to support the concept of relevance in knowledge organization.
3) Aboutness and warrant are KO concepts closely related to relevance. Aboutness and relevance share some common features; they are matters of degree and involve objective and subjective attributes. Aboutness is an inherent property of information and is independent of relevance. Aboutness guarantees a relationship between information (as sentences, documents, etc.) and data (as terms/concepts), and this relationship is an embodiment of objective relevance. Warrants function as the "atmosphere" to determine the key tone, establish the context, and the boundary of situations, in which relevance is determined.
4) Relevance in knowledge organization is understood via nine definitions alongside the KO processes. The definitions can be categorized into two groups. The first group includes the relationships between information/data, falling within the scope of objective

relevance. This group includes Definition 1: Relevance in subjective analysis is the relationship between the inherent features and the subject matter of an information entity determined by a subject analyst; Definition 4: Relevance in Indexing is the semantic relationship between the inherent features and controlled vocabularies as determined by an indexer to select index terms; Definition 5: Relevance in cataloging is the relationship between the inherent features of an information entity and the descriptions, subject terms, and classifiers assigned by a cataloger to fulfill the cataloging tasks; and Definition 8: Relevance in social tagging is the relationships between tags in a folksonomy. The second group is the relationships between information/data and users (including information professionals), falling within the scope of subjective relevance. This group includes Definition 2: Relevance in indexing is the degree of correlation between the assumptions of users' information needs as determined by an indexer; Definition 3: Relevance in indexing is the measurement of the bearing between inherent features and the selected assumptions as determined by an indexer to identify subject matter; Definition 6: Relevance in cataloging is the relationship between the properties and attributes assumed by catalogers and used by users to meet their information needs; Definition 7: Relevance in social tagging is the relationship existing between the tags and the tagging intentions as determined by users (including content creators) in the tagging process; Definition 9: Relevance in social tagging is the relationship of tags existing between an information entity and information needs as determined by users.

5) Relationships are seen as an embodiment of objective relevance. There are numerous relationships at different levels in KO, leading to an understanding challenge. A meta-

typological model of relationships in KO was created for this dissertation study to facilitate understanding the landscape of relationships. The model consists of three types of relationships: relationships in the bibliographic universe, the information universe, and the data universe. The model provides a way to understand relationships in KO systematically. It can be reused in other relationship studies in information science.

6) After examining bibliographic relationships and models including FRBR, FRAD, FRSAD, LRM, and BIBFRAME in the bibliographic universe, the data shows that there is a continuous expansion from bibliographical relationships toward data relationships in the context of Linked Data and the Semantic Web, to make bibliographic resources more discoverable on the web. In the information universe, relationships have been studied via links and types of links for hypertext and tags for social media. The data shows that link types represent relationships between the linked nodes, which serve as justifications for why online information is related, and further support relevance judgment. Tag-tag relationships exhibit a strong feature of the data attribute. Typologies of social tags are domain-specific, with some high-level common types across domains: content-based, contextual-based, and task-based, corresponding to the dimensions of relevance. Most typology studies in the information universe have used external tools such as WordNet, DBpedia, UMLS, Yago, etc., for leveraging their abundant relationships. Relationship study transitions from the context of the information universe to the data universe. Relationship studies in the bibliographic universe and the information universe show a trajectory toward datafication.

7) Knowledge organization systems are used as the subject to study relationships in the data universe. A framework was created to organize KOS and their supporting technologies in the Semantic Web, including the schema level, the KOS level, and the data level. Through examining the relationships in RDF, SKOS, and OWL at the schema level and examining LOD KOS including UNESCO Thesaurus, Getty Vocabularies, and ARGOVOC Thesaurus at the KOS level, some findings are: the RDF model is a strong tool for building up relationships naturally and diversely, not only limited to RT (related term), NT (narrower term), BT (broader term), USE, or USED FOR in conventional vocabularies, but also in more specific way such as *arg:isProducedBy*, *tgn:moved_to*, *ulan:ward_of*, etc.; numerous properties at the vocabulary level are instances of associative relationships, making associative relationships explicit and clarified; the large amount of relationships show how data are related and objectively relevant.
8) In the data universe, metadata fields are relationships. They are the indicators of relevance criteria and the indicators of data relevance. Many common metadata fields correspond to different data relevance criteria, including accessibility and availability, authority, coverage, novelty, quality, readability, recency, relation/origin, standardization, topicality, and usefulness and usability. For example, the field data time is used to judge accessibility and availability, coverage, recency, and topicality.
9) A data relevance model was created in the user study. The model consists of 28 elements that affect data relevance. All the elements are categorized into three levels according to their influence on other categories in the model. The core elements of data relevance are the top-level categories, which are features of the resource, considering

users, potential use, prior knowledge, purpose, and context. The core elements have a probable larger influence on data relevance. The top-level elements are not data-specific, which can be used for resources other than data. The second level elements show more data features, such as scope of terms, granularity, enabling linkages, using single/as few schemas, etc.

## 7.3 Data Relevance Theory and Use Cases

This study aims to address the conceptualization and theorization of data relevance in the KO context. The three research questions form the framework to understand data relevance. Nine definitions of KO relevance have been proposed. Many relationships existing in the bibliographic universe, the information universe, and the data universe have been examined. The relationships between data and information professionals have been explored in the user study. Let's reach back to the core question: what is data relevance from the perspective of knowledge organization?

Data relevance has two levels of meaning: the first, relevance is the relationships between data and between data and information, and *features of resource* are the underlying foundation of the relationships; the second, relevance is the relationship between data and people, and *purpose* is the origin of the relationship. The two levels are determined by both objective and subjective factors, while the first level is affected by more objective factors and the second level is affected by more subjective factors. In this sense, relevant data can be data that are related via inherent properties and features or can be data that serve a user's purpose. Either way is sufficient to identify relevant data.

Data relevance functions in knowledge organization processes. Looking through a

magnifying glass, the two levels of meaning vary practically according to the activities in specific contexts. However, the core of the nine definitions is consistent with the two levels of relationships

Data relevance is affected by many factors. *Features of resource* with their manifestation *relationships* (the first level relevance) serve as a core factor of the second level relevance. The core factors to determine data relevance are considering users, context, features of resource, potential use, prior knowledge, and purpose. The other main factors are accuracy, enabling linkages, format, generality, granularity, hierarchy, reducing duplication, relationship, scope of terms, specificity, sufficiency, technology, time and resource constraints, using single/as few schemas, widely used, and working context and job responsibility. These factors explain: 1) why a user considers some data to be more relevant than others; 2) why some data are more relevant to each other than others, not subject to human will.

Data relevance can be further explained by some use cases. In practice, there might be different situations when applying the two levels of data relevance. First, the two levels are consistent, i.e., both levels stand. Second, the two levels are inconsistent. This could be that the first level stands, but the second level fails, or the first level fails, but the second level stands. Taking some examples from the user study, there are two terms to represent a data point of the rubbing resource: Creator (design in the first round) and originalObjectCreator (standard in the second round). The relationship between them is a subtype, representing that one term is more specific than the other. The participant's goal is to better represent the degree of specificity. The subtype relation between the two terms meets the goal, and the participant considers the term originalObjectCreator to be more relevant than creator to represent the data point. This is a

consistent situation. The other participant prefers a general description after considering the end users' comprehension ability. The relationship subtype doesn't meet the intention, so originalObjectCreator is irrelevant. This is an inconsistent situation. In these cases, there is a close relationship between data A and data B. The essence at the second level is whether the relationship between A and B meets the intention. The two levels can be explained in another way: Data A is relevant to a user. There is a relationship between data A and data B. If the relationship between A and B meets the user's intention, then B is relevant to the user, i.e., relevance is transitive if the relationship meets a user's need. Taking another example from the user study, there are two terms: *subject* and *scientificName*, and the relationship between them is *nameOf*, implying one is more granular than the other. The participant chooses *subject* first. Considering the features of the resource and the relationship between the two terms, the participant decides to treat them as identical and adds *scientificName* in the design as well. Relevance is transmitted from *subject* to *scientificName* due to the relationship.

## 7.4 Discussion

**Objective-subjective Dichotomy**

Objective relevance and subjective relevance have been discussed for decades. Relevance is a complex concept that can be comprehended and interpreted through many different paradigms, theoretical foundations, and views. The derived concepts, e.g., logical relevance, situational relevance, system relevance, topical relevance, cognitive relevance, user-based relevance, psychological relevance, etc., are strong evidence of its complexity. The dichotomous view cannot cover all views concerning relevance; however, it is the downsized manner that focusing on the two key attributes of relevance provides a solid foundation to

study the phenomenon, extended for data and expanded to a different context. Alongside the study, many considerations and factors show both objective and subjective features, which demonstrate the rationality of using this approach. The data relevance model can be adapted to the dichotomous view. Figure 24 shows the remodeled data relevance factors. It provides a different understanding of the factors that inform future research.

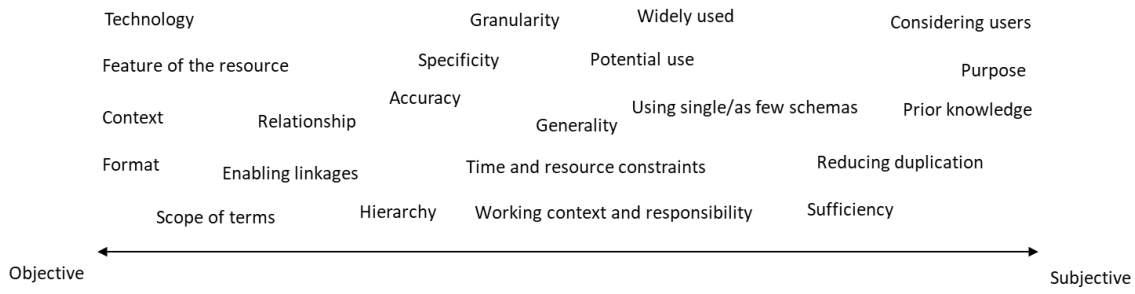



Figure 24. Data relevance factors in the objective-subjective view

**Philosophical foundation of objectivity**

Since the dichotomous view runs throughout the study, the philosophical discussion and a definition of objectivity in Section 4.1 are critical. Clarifying the notion of objectivity underpins the study. Many concepts, views, decisions, and arguments do not have a clear-cut line between subjective and objective. They are in a "vague" status in most cases, which can be seen from Figure 24. Many factors are a combination of objectivity and subjectivity.

**Deconstructive thinking**

Deconstructive thinking is an important approach in this study. The proposal from information relevance to data relevance follows a deconstructive thinking. Exploring relevance in several KO activities applies a deconstructive approach. Designing the user tasks using two stages demonstrates deconstructive thinking as well.

Deconstruction, pursuing details and in-depth understanding, is a construction method.

Based on the connections of data with numerous relationships, users actively determine/construct which relationships are more important to their current situation (van Dijk, 1979). The process of judging relevance is a process of construction. When exploring relevance in the breakdown processes, it helps to examine the construction process (judging relevance) in detail and gain insights, and the construction for a new understanding/interpretation.

**Self-contradiction and dilemmas of participants in the user study**

In the user study, participants sometimes fell into self-contradiction or dilemmas, which usually happened in making decisions on general/specific, hierarchical/flat, sufficient/granular, true to resource/future possibility, using single schema/multiple schemas. For example, a participant preferred a flat schema for interoperability but considered a hierarchy very important and helpful; A participant considered the types of resource at hand when determining a field, but consider future possible types for another; A participant preferred a specific term for one field but a general term for another; A participant dislike a general schema but used and only used it; etc. They used opposite criteria in one task. These are not errors but show the ambivalence of participants in decision-making. One reason for these phenomena is that when they were making decisions, they considered many factors. For example, a participant prefers to simplify the metadata as much as possible to improve their work efficiency, while on the other hand the participant would like to see more information provided; however, if providing users with too much information, they will become confused. Although the user study was set in a natural setting with the same tasks, they had various mindsets and assumptions on these factors. When either choice of a factor was acceptable and justifiable, they randomly chose between them. Furthermore, several participants mentioned that if these

tasks were a job responsibility, they would make a different decision. The more factors they considered, the higher the possibility of having inconsistent decisions, and the higher the possibility of falling into a dilemma.

**Grounded theory**

Grounded theory was used to code and analyze the participant-generated data and create the data relevance model. For reliability, I did two rounds of coding. For open coding, in the first round, I coded terms and phrases (usually called categories) as in vivo codes, without contextual information; in the second round, I coded sentences as in vivo codes for contextual information. In the axial coding, I found contextual information to be critical to understanding the codes and correctly clustering them into categories. Coding terms and phrases without contextual information is not sufficient to determine the axial categories for this study.

In the selective coding, I used graphing to identify the relationships between the axial codes. Graphing is a strong method to support grounded theory in the selective coding (theory generating) phase. It organizes the complex relationship easily and clearly and supports data-driven decisions. In this study, the graph view of Obsidian visualized the relationships between the categories. The graph view allows interaction with the categories, selection of the category of interest for further exploration, examination of the linkages (both incoming and outgoing), and performing the statistical analysis. It allows the study to be firmly grounded in empirical evidence rather than arbitrary conclusions.

## 7.5 Limitations of the Study

Some limitations must be acknowledged in this study:

1) When investigating the concept of relevance in knowledge organization in Chapter 4, Saracevic's formula to define relevance was used. This formula has been widely used as a foundation to understand the concept of relevance. However, it was proposed in the information retrieval context following a source-to-destination principle. This formula limits the scope to explore relevance in knowledge organization. Although the analysis shows that it is sufficient to define relevance in KO, it remains possible that there are better ways to define relevance, moving beyond the constraint of this framework.
2) The major data used to generate the data relevance model came from nine metadata librarians. While the quantity of participants is sufficient to reach the conclusion, more participants with diverse backgrounds will probably optimize the model. The discriminant sample shows some complementary viewpoints, demonstrating that diverse participants contribute to a rich understanding of data relevance.
3) Though knowledge organization provides a seedbed to study data relevance, data relevance should not be limited to the knowledge organization context. Finding data and reusing data are common questions across the subareas of information science. They are critical issues in other domains that are driven by data as well. Exploring data relevance in diverse contexts and across domains will be essential to improve data relevance theory.

## 7.6 Future Work

This dissertation is an attempt to conceptualize and theorize data relevance from the perspective of knowledge organization. Based on the findings, there are several promising directions to expand and extend data relevance theory.

1) Continuing research on the current data relevance theory.

Time is an important dimension of data relevance. This study discussed time from the perspective of time constraints without considering time as a factor. However, the two-stage tasks show that resources at hand, prior knowledge and experience change over time, indicating that time might be an important factor to study data relevance. The current data relevance model is static; a time-dependent model with changing conditions will complement and enrich the current model.

It would also be interesting to learn the end-users' attitude toward the data relevance model. An empirical study to examine the data relevance model from the end users' perspective would provide evaluation and more insight to refine the model.

2) Expanding the domains of data to explore data relevance

The relationship study in Chapter 5 shows that data and their relationships vary significantly across domains. It implies that disparities might exist in data relevance theory in different domains. Expanding the current data relevance study to different domains is a promising way to examine whether data relevance theory is domain dependent.

3) Expanding the context of data relevance study.

Data relevance has been studied within the scope of KO. During the user study, I found that data retrieval is a necessary step to design a metadata schema. For example, participants searched LOV to look for schemas and terms, and they browsed schemas to examine terms. Studying data relevance back to the IR context offers a critical complementary perspective. An extended related problem would be a comparison study of data relevance in KO and in IR.

4) Relationship studies and data models.

There are numerous relationships based on the RDF model on the Semantic Web. Exploring relationships would be a valuable inquiry. A future possibility would be how relationships and data relevance contribute to data modeling, data description, and data organization.

References

*About UNESCO Thesaurus*. (n.d.). UNESCO. Retrieved April 13, 2025, from https://vocabularies.unesco.org/browser/en/about

Ackoff, R. (1989). From Data to Wisdom. *Journal of Applied Systems Analysis*, *16*(3), 9. https://softwarezen.me/wp-content/uploads/2018/01/datawisdom.pdf

*AGROVOC*. (n.d.). Food and Agriculture Organization of the United States. Retrieved April 13, 2025, from https://www.fao.org/agrovoc/about

Albassam, S. A. A., & Ruthven, I. (2018). Users' relevance criteria for video in leisure contexts. *Journal of Documentation*, *74*(1), 62–79. https://doi.org/10.1108/JD-06-2017-0081

Albrechtsen, H. (1993). Subject analysis and indexing: From automated indexing to domain analysis. *The Indexer*, *18*(4), 219–224. https://doi.org/10.3828/indexer.1993.18.4.3

ANSI/NISO. (2010). *ANSI/NISO Z39.19-2005 (R2010): Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies*. National Information Standards Organization. https://doi.org/10.3789/ansi.niso.z39.19-2005R2010

Arsenault, C., & Noruzi, A. (2012). Analysis of Work-to-Work Bibliographic Relationships through FRBR: A Canadian Perspective. *Cataloging & Classification Quarterly*, *50*(5–7), 641–652. https://doi.org/10.1080/01639374.2012.679580

Bade, D. (2007). Relevance ranking is not relevance ranking or, when the user is not the user, the search results are not search results. *Online Information Review*, *31*(6), 831–844. https://doi.org/10.1108/14684520710841793

Balatsoukas, P., Morris, A., & O'Brien, A. (2009). An evaluation framework of user interaction with metadata surrogates. *Journal of Information Science*, *35*(3), 321–339. https://doi.org/10.1177/0165551508099090

Barité, M. (2018). Literary warrant. *Knowledge Organization*, *45*(6), 517–536. https://doi.org/10.5771/0943-7444-2018-6-517

Barité, M. (2019). Towards a General Conception of Warrants: First Notes. *Knowledge Organization*, *46*(8), 647–655. Computers & Applied Sciences Complete. https://doi.org/10.5771/0943-7444-2019-8-647

Baron, L. (1994). *The Effectiveness Of Labelled, Typed Links As Cues In Hypertext Systems* [The University of Western Ontario]. https://ir.lib.uwo.ca/digitizedtheses/2333/

Barry, C. L. (1994). User-defined relevance criteria: An exploratory study. *Journal of the American Society for Information Science*, *45*(3), 149–159. https://doi.org/10.1002/(SICI)1097-4571(199404)45:3<149::AID-ASI5>3.0.CO;2-J

Barry, C. L. (1998). Document representations and clues to document relevance. *Journal of the American Society for Information Science*, *49*(14), 1293–1303. https://doi.org/10.1002/(SICI)1097-4571(1998)49:14<1293::AID-ASI7>3.0.CO;2-E

Barry, C. L., & Schamber, L. (1998). Users' criteria for relevance evaluation: A cross-situational comparison. *Information Processing & Management*, *34*(2), 219–236. https://doi.org/10.1016/S0306-4573(97)00078-2

Bean, C. A., & Green, R. (2001). Relevance Relationships. In C. A. Bean & R. Green (Eds.), *Relationships in the Organization of Knowledge* (pp. 115–132). Springer Netherlands. https://doi.org/10.1007/978-94-015-9696-1_8

Beaudoin, J. (2007). Folksonomies: Flickr image tagging: Patterns made visible. *Bulletin of the American Society for Information Science and Technology*, *34*(1), 26–29. https://doi.org/10.1002/bult.2007.1720340108

Bechhofer, S., Van Harmelen, F., Hendler, J., Horrocks, I., McGuinness, D. L., Patel-Schneider, P. F., & Stein, L. A. (2004, February 10). *OWL Web Ontology Language Reference* (M. Dean & G. Schreiber, Eds.). W3C. https://www.w3.org/TR/owl-ref/

Beghtol, C. (1986a). Bibliographic Classification Theory And Text Linguistics: Aboutness Analysis, Intertextuality And The Cognitive Act Of Classifying Documents. *Journal of Documentation*, *42*(2), 84–113. https://doi.org/10.1108/eb026788

Beghtol, C. (1986b). Semantic validity: Concepts of warrant in bibliographic classification systems. *Library Resources & Technical Services*, *30*(2), 109–125.

Beghtol, C. (2002). A proposed ethical warrant for global knowledge representation and organization systems. *Journal of Documentation*, *58*(5), 507–532. https://doi.org/10.1108/00220410210441

Berners-Lee, T. (2006). *Linked Data*. https://www.w3.org/DesignIssues/LinkedData.html

Bischoff, K., Firan, C. S., Nejdl, W., & Paiu, R. (2008). Can all tags be used for search? *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, 193–202. https://doi.org/10.1145/1458082.1458112

Bodoff, D. (2006). Relevance for browsing, relevance for searching. *Journal of the American Society for Information Science and Technology*, *57*(1), 69–86.

Bookstein, A. (1979). Relevance. *Journal of the American Society for Information Science (Pre-1986)*, *30*(5), 269. Business Premium Collection; ProQuest One Business. https://doi.org/10.1002/asi.4630300505

Bordignon, F., & Maisonobe, M. (2022). Researchers and their data: A study based on the use of the word data in scholarly articles. *Quantitative Science Studies*, *3*(4), 1156–1178.

https://doi.org/10.1162/qss_a_00220

Borgman, C. L. (2015). Big data, little data, no data: Scholarship in the networked world / Christine L. Borgman. In *Big data, little data, no data: Scholarship in the networked world*. The MIT Press.

Borlund, P. (2003). The concept of relevance in IR. *Journal of the American Society for Information Science and Technology*, *54*(10), 913–925. https://doi.org/10.1002/asi.10286

Boyce, B. (1982). Beyond topicality: A two stage view of relevance and the retrieval process. *Information Processing & Management*, *18*(3), 105–109. https://doi.org/10.1016/0306-4573(82)90033-4

Brickley, D., & Guha, R. V. (2014, February 25). *RDF Schema 1.1*. W3C. https://www.w3.org/TR/rdf-schema/

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, *30*(1–7), 107–117. https://doi.org/10.1016/S0169-7552(98)00110-X

Brooks, T. A. (1995). Topical subject expertise and the semantic distance model of relevance assessment. *Journal of Documentation*, *51*(4), 370–387. https://doi.org/10.1108/eb026956

Bruce, H. W. (1994). A cognitive view of the situational dynamism of user-centered relevance estimation. *Journal of the American Society for Information Science*, *45*(3), 142–148. https://doi.org/10.1002/(SICI)1097-4571(199404)45:3<142::AID-ASI4>3.0.CO;2-6

Bruza, P. D., & Huibers, T. W. C. (1996). A study of aboutness in information retrieval. *Artificial Intelligence Review*, *10*(5), 381–407. https://doi.org/10.1007/BF00130692

Bruza, P. D., Song, D. W., & Wong, K. F. (2000). Aboutness from a commonsense perspective. *Journal of the American Society for Information Science*, *51*(12), 1090–1105. https://doi.org/10.1002/1097-4571(2000)9999:9999<::AID-ASI1026>3.0.CO;2-Y

Bullard, J. (2017). Warrant as a means to study classification system design. *Journal of Documentation*, *73*(1), 75–90. https://doi.org/10.1108/JD-06-2016-0074

Cantador, I., Konstas, I., & Jose, J. M. (2011). Categorising social tags to improve folksonomy-based recommendations. *Journal of Web Semantics*, *9*(1), 1–15. https://doi.org/10.1016/j.websem.2010.10.001

Carmagnola, F., Cena, F., Cortassa, O., Gena, C., & Torre, I. (2007). Towards a Tag-Based User Model: How Can User Model Benefit from Tags? In C. Conati, K. McCoy, & G. Paliouras (Eds.), *User Modeling 2007* (pp. 445–449). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-73078-1_62

Caro, A., Calero, C., Caballero, I., & Piattini, M. (2008). A proposal for a set of attributes relevant for Web portal data quality. *Software Quality Journal*, *16*(4), 513–542. https://doi.org/10.1007/s11219-008-9046-7

Cattuto, C., Schmitz, C., Baldassarri, A., Servedio, V. D. P., Loreto, V., Hotho, A., Grahl, M., & Stumme, G. (2007). Network properties of folksonomies. *AI Communications*, *20*(4), 245–262.

Chan, L. M. (1989). Inter-Indexer Consistency in Subject Cataloging. *Information Technology and Libraries*, *8*(4), 349–358. eLibrary; ProQuest One Business.

Chapman, A. (2007). Resource discovery: Catalogs, cataloging, and the user. *Library Trends*, *55*(4), 917–931. https://doi.org/10.1353/lib.2007.0038

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

Chen, M., Liu, X., & Qin, J. (2008). Semantic Relation Extraction from Socially-Generated Tags. *Proceedings of the International Conference on Dublin Core and Metadata Applications: Metadata for Semantic and Social Applications*. International Conference on Dublin Core and Metadata Applications. https://doi.org/10.18452/1257

Choi, Y., & Rasmussen, E. M. (2002). Users' relevance criteria in image retrieval in American history. *Information Processing & Management*, *38*(5), 695–726. https://doi.org/10.1016/S0306-4573(01)00059-0

Chu, C. M., & O'Brien, A. (1993). Subject analysis: The critical first stage in indexing. *Journal of Information Science*, *19*(6), 439–454. https://doi.org/10.1177/016555159301900603

Chu, H. (2011). Factors affecting relevance judgment: A report from TREC Legal track. *Journal of Documentation*, *67*(2), 264–278. https://doi.org/10.1108/00220411111109467

Clarke, S. G. D. (2001). Thesaural Relationships. In C. A. Bean & R. Green (Eds.), *Relationships in the Organization of Knowledge* (pp. 37–52). Springer Netherlands. https://doi.org/10.1007/978-94-015-9696-1_3

Cooper, W. S. (1971). A definition of relevance for information retrieval. *Information Storage and Retrieval*, *7*(1), 19–37. https://doi.org/10.1016/0020-0271(71)90024-6

Cooper, W. S. (1973). On selecting a measure of retrieval effectiveness part 1: The "subjective" philosophy of evaluation. *Journal of the American Society for Information Science*, *24*(2), 87–100. https://doi.org/10.1002/asi.4630240204

Corbin, J. M., & Strauss, A. L. (2015). *Basics of qualitative research: Techniques and procedures for developing grounded theory* (Fourth edition). SAGE Publications.

Cosijn, E., & Ingwersen, P. (2000). Dimensions of relevance. *Information Processing & Management*, *36*(4), 533–550. https://doi.org/10.1016/S0306-4573(99)00072-2

Costello, K. L. (2017). Social relevance assessments for virtual worlds: Interpersonal source selection in the context of chronic illness. *Journal of Documentation*, *73*(6), 1209–1227. https://doi.org/10.1108/JD-07-2016-0096

Creswell, J. W. (2012). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research* (Fourth edition). Boston : Pearson.

Creswell, J. W., & Poth, C. N. (2018). *Qualitative inquiry & research design: Choosing among five approaches* (Fourth edition). SAGE Publications.

Crystal, A., & Greenberg, J. (2006). Relevance criteria identified by health information users during Web searches. *Journal of the American Society for Information Science and Technology*, *57*(10), 1368–1382. https://doi.org/10.1002/asi.20436

Cuadra, C. A., & Katter, R. V. (1967). Opening the black box of "relevance." *Journal of Documentation*, *23*(4), 291–303. https://doi.org/10.1108/eb026436

Curty, R. G. (2015). *Beyond "data thrifting": An investigation of factors influencing research data reuse in the social sciences* [Dissertation, Syracuse University]. https://surface.syr.edu/etd/266

Cyganiak, R., Wood, D., & Lanthaler, M. (Eds.). (2014, February 25). *RDF 1.1 Concepts and Abstract Syntax*. W3C. https://www.w3.org/TR/rdf11-concepts/

Darby, R., Lambert, S., Matthews, B., Wilson, M., Gitmans, K., Dallmeier-Tiessen, S., Mele, S., & Suhonen, J. (2012). Enabling scientific data sharing and re-use. *2012 IEEE 8th International Conference on E-Science*, 1–8. https://doi.org/10.1109/eScience.2012.6404476

De Sabbata, S., Mizzaro, S., & Reichenbacher, T. (2015). Geographic dimensions of relevance. *Journal of Documentation*, *71*(4), 650–666. https://doi.org/10.1108/JD-12-2013-0167

Dix, A., Levialdi, S., & Malizia, A. (2006). Semantic halo for collaboration tagging systems. In *Workshop on the social navigation and community based adaptation technologies*. http://www.sis.pitt.edu/%7Epaws/SNC_BAT06/crc/malizia.pdf

Dobreski, B. (2020). Common usage as warrant in bibliographic description. *Journal of Documentation*, *76*(1), 49–66. https://doi.org/10.1108/JD-05-2019-0094

Duff, W. M. (1998). Harnessing the Power of Warrant. *The American Archivist*, *61*(1), 88–105. JSTOR. http://www.jstor.org/stable/40294076

Duffy, J., & Feltovich, N. (2002). Do Actions Speak Louder Than Words? An Experimental Comparison of Observation and Cheap Talk. *Games and Economic Behavior*, *39*(1), 1–27.

https://doi.org/10.1006/game.2001.0892

Duke-Margolis Center for Health Policy. (2018). *Characterizing RWD quality and relevancy for regulatory purposes*. https://healthpolicy.duke.edu/publications/characterizing-rwd-quality-and-relevancy-regulatory-purposes-0

*Fact Sheet UMLS Semantic Network*. (n.d.). Retrieved April 7, 2021, from https://lhncbc.nlm.nih.gov/semanticnetwork/

Fairthorne, R. A. (1969). Content analysis, specification and control. *Annual Review of Information Science and Technology*, *4*, 73–109. https://babel.hathitrust.org/cgi/pt?id=mdp.39015008929476&seq=87

Faniel, I. M., & Jacobsen, T. E. (2010). Reusing scientific data: How earthquake engineering researchers assess the reusability of colleagues' data. *Comput. Supported Coop. Work*, *19*(3–4), 355–375. https://doi.org/10.1007/s10606-010-9117-8

Faniel, I. M., Kriesberg, A., & Yakel, E. (2016). Social scientists' satisfaction with data reuse. *Journal of the Association for Information Science and Technology*, *67*(6), 1404–1416. https://doi.org/10.1002/asi.23480

Flick, U. (2018). *An introduction to qualitative research*. SAGE Publications.

Floridi, L. (2010). *Information: A very short introduction*. Oxford University Press. https://doi.org/10.1093/actrade/9780199551378.001.0001

Foskett, A. C. (1982). The subject approach to information. In *The subject approach to information* (Fourth edition). London: C. Bingley; Hamden, Conn.: Linnet Books.

Foskett, D. J. (1972). A note on the concept of "relevance." *Information Storage and Retrieval*, *8*(2), 77–78.

Fraser, W. (1978). Literary, user and logical warrants as indexing constraints. *The Information Age in Perspective: Proceedings of the ASIS Annual Meeting Knowledge Industry Publications*, 130–132.

Furner, J. (2002). Bibliographic relationships, citation relationships, relevance relationships, and bibliographic classification: An integrative view. *13th ASIS SIG/CR Classification Research Workshop*, 29–37. https://doi.org/10.7152/acro.v13i1.13833

Galeffi, A., Bertolini, M. V., Bothmann, R. L., Rodríguez, E. E., & McGarry, D. (2017). *Statement of International Cataloguing Principles (ICP)*. International Federation of Library Associations and Institutions. https://www.ifla.org/wp-content/uploads/2019/05/assets/cataloguing/icp/icp_2016-en.pdf

Gaukroger, S. (2012). *Objectivity: A Very Short Introduction*. Oxford University Press.

https://doi.org/10.1093/actrade/9780199606696.001.0001

Gitelman, L. (Ed.). (2013). *"Raw Data" Is an Oxymoron*. The MIT Press. https://doi.org/10.7551/mitpress/9302.001.0001

Glaser, B. G., & Strauss, A. L. (1967). The discovery of grounded theory: Strategies for qualitative research / Barney G. Glaser and Anselm L. Strauss. In *The discovery of grounded theory: Strategies for qualitative research*. Aldine Pub. Co.

Glushko, R. J. (2013). The Discipline of Organizing. In *The Discipline of Organizing*. The MIT Press.

Goffman, W. (1964). On relevance as a measure. *Information Storage and Retrieval*, *2*(3), 201–203. https://doi.org/10.1016/0020-0271(64)90004-X

Goldkuhl, G. (2012). Pragmatism vs interpretivism in qualitative information systems research. *European Journal of Information Systems*, *21*(2), 135–146. https://doi.org/10.1057/ejis.2011.54

Green, R. (1995). Topical relevance relationships. I. Why topic matching fails. *Journal of the American Society for Information Science*, *46*(9), 646–653. https://doi.org/10.1002/(SICI)1097-4571(199510)46:9<646::AID-ASI2>3.0.CO;2-1

Green, R. (2001). Relationships in the Organization of Knowledge: An Overview. In C. A. Bean & R. Green (Eds.), *Relationships in the Organization of Knowledge* (pp. 3–18). Springer Netherlands. https://doi.org/10.1007/978-94-015-9696-1_1

Green, R. (2008). Relationships in knowledge organization. *Knowledge Organization*, *35*(2–3), 150–159. https://doi.org/10.5771/0943-7444-2008-2-3-150

Green, R., & Bean, C. A. (1995). Topical relevance relationships. II. An exploratory study and preliminary typology. *Journal of the American Society for Information Science*, *46*(9), 654–662. https://doi.org/10.1002/(SICI)1097-4571(199510)46:9<654::AID-ASI3>3.0.CO;2-3

Gregory, K. M., Cousijn, H., Groth, P., Scharnhorst, A., & Wyatt, S. (2020). Understanding data search as a socio-technical practice. *Journal of Information Science*, *46*(4), 459–475. https://doi.org/10.1177/0165551519837182

Gupta, M., Li, R., Yin, Z., & Han, J. (2011). An Overview of Social Tagging and Applications. In C. C. Aggarwal (Ed.), *Social Network Data Analytics* (pp. 447–497). Springer US. https://doi.org/10.1007/978-1-4419-8462-3_16

Halliday, M. A. K. (1978). Language as social semiotic: The social interpretation of language and meaning. In *Language as social semiotic: The social interpretation of language and meaning*. Edward Arnold.

Harmon, R. B. (1989). *Elements of bibliography: A simplified approach* (Revised edition.).

Metuchen, N.J. : Scarecrow Press.

Harpring, Patricia. (2010). *Introduction to controlled vocabularies: Terminology for art, architecture, and other cultural works*. Getty Research Institute.

Harter, S. P. (1971). The Cranfield II Relevance Assessments: A Critical Evaluation. *The Library Quarterly: Information, Community, Policy*, *41*(3), 229–243. JSTOR. http://www.jstor.org/stable/4306095

Harter, S. P. (1992). Psychological relevance and information science. *Journal of the American Society for Information Science*, *43*(9), 602–615. https://doi.org/10.1002/(SICI)1097-4571(199210)43:9<602::AID-ASI3>3.0.CO;2-Q

Heymann, P., & Garcia-Molina, H. (2006). *Collaborative creation of communal hierarchical taxonomies in social tagging systems* (Technical Report Nos. 2006–10). Stanford / Stanford InfoLab. http://ilpubs.stanford.edu:8090/775/

Hillman, D. J. (1964). The notion of relevance (I). *American Documentation*, *15*(1), 26–34. https://doi.org/10.1002/asi.5090150107

Hjørland, B. (1992). The Concept of 'Subject' in Information Science. *Journal of Documentation*, *48*(2), 172–200. https://doi.org/10.1108/eb026895

Hjørland, B. (1997). *Information seeking and subject representation: An activity-theoretical approach to information science*. Greenwood Press.

Hjørland, B. (2000). Relevance Research: The Missing Perspective(s): "Non-Relevance" and "Epistemological Relevance." *Journal of the American Society for Information Science*, *51*(2), 209. https://doi.org/10.1002/(SICI)1097-4571(2000)51:2<209::AID-ASI14>3.0.CO;2-B

Hjørland, B. (2001). Towards a theory of aboutness, subject, topicality, theme, domain, field, content. . . And relevance. *Journal of the American Society for Information Science and Technology*, *52*(9), 774–778. https://doi.org/10.1002/asi.1131

Hjørland, B. (2002). Epistemology and the socio-cognitive perspective in information science. *Journal of the American Society for Information Science and Technology*, *53*(4), 257–270. https://doi.org/10.1002/asi.10042

Hjørland, B. (2010). The foundation of the concept of relevance. *Journal of the American Society for Information Science and Technology*, *61*(2), 217–237. https://doi.org/10.1002/asi.21261

Hjørland, B. (2013). Theories of Knowledge Organization—Theories of Knowledge. *Knowledge Organization*, *40*(3), 169–181. https://doi.org/10.5771/0943-7444-2013-3-169

Hjørland, B. (2015). Are relations in thesauri "context-free, definitional, and true in all possible worlds"? *Journal of the Association for Information Science and Technology*, *66*(7), 1367–1373. https://doi.org/10.1002/asi.23253

Hjørland, B. (2016). Knowledge Organization (KO). *Knowledge Organization*, *43*, 475–484. https://doi.org/10.5771/0943-7444-2016-6-475

Hjørland, B., & Christensen, F. S. (2002). Work tasks and socio-cognitive relevance: A specific example. *Journal of the American Society for Information Science and Technology*, *53*(11), 960–965. https://doi.org/10.1002/asi.10132

Hodge, G. (2000). *Systems of knowledge organization for digital libraries: Beyond traditional authority files* (Electronic ed.). Digital Library Federation, Council on Library and Information Resources.

Hoffman, G. L. (2009). Meeting Users' Needs in Cataloging: What is the Right Thing to Do? *Cataloging & Classification Quarterly*, *47*(7), 631–641. https://doi.org/10.1080/01639370903111999

Holley, R. M., & Joudrey, D. N. (2021). Aboutness and conceptual analysis: A review. *Cataloging & Classification Quarterly*, *59*(2–3), 159–185. https://doi.org/10.1080/01639374.2020.1856992

Houser, N., & Kloesel, C. (Eds.). (1992). How to Make Our Ideas Clear (1878). In *The Essential Peirce, Volume 1* (pp. 124–141). Indiana University Press; JSTOR. https://doi.org/10.2307/j.ctvpwhg1z.13

Hu, X., Ng, J. T., Tang, E. M., Zhu, J., & Xia, S. (2018). User-centered evaluation of metadata schema for murals: Preliminary results from a survey study. *iConference 2018 Proceedings*. https://hdl.handle.net/2142/100276

Huang, X. (2009). *Relevance, rhetoric, and argumentation: A cross-disciplinary inquiry into patterns of thinking and information structuring*. ProQuest Dissertations Publishing.

Huang, X., & Soergel, D. (2006). An evidence perspective on topical relevance types and its implications for exploratory and task-based retrieval. *Information Research*, *12*(1). http://InformationR.net/ir/12-1/paper281.html

Huang, X., & Soergel, D. (2013). Relevance: An improved framework for explicating the notion. *Journal of the American Society for Information Science and Technology*, *64*(1), 18–35. https://doi.org/10.1002/asi.22811

Huo, Y., Zhou, S., Hu, R., Ren, Y., & Xia, J. (2020). Guarantee Mechanism of Data Continuity for Electronic Record Based on Linked Data. *Journal of Cyber Security*, *2*(3). https://doi.org/10.32604/jcs.2020.010963

Hussein, M., Hirst, S., Salyers, V., & Osuji, J. (2014). Using Grounded Theory as a Method of Inquiry: Advantages and Disadvantages. *The Qualitative Report*, *19*(27), 1–15. https://doi.org/10.46743/2160-3715/2014.1209

Hutchins, W. J. (1977). On the problem of 'aboutness' in document analysis. *Journal of Informatics*, *1*(1), 17–35.

Hutchins, W. J. (1978). The concept of 'aboutness' in subject indexing. *Aslib Proceedings*, *30*(5), 172–181. https://doi.org/10.1108/eb050629

IFLA Study Group on the Functional Requirements for Bibliographic Records. (2009). Functional Requirements for Bibliographic Records: Final report. In *IFLA*. https://www.ifla.org/files/assets/cataloguing/frbr/frbr_2008.pdf

IFLA Working Group on Functional Requirements and Numbering of Authority Records (FRANAR). (2013). *Functional Requirements for Authority Data—A Conceptual Model*. Berlin/Munich: De Gruyter Saur. https://repository.ifla.org/handle/20.500.14598/757

IFLA Working Group on Guidelines for Multilingual Thesauri. (2009). *Guidelines for Multilingual Thesauri* (No. 115; IFLA Professional Reports,). International Federation of Library Associations and Institutions. https://repository.ifla.org/server/api/core/bitstreams/7e784620-80cb-4c71-9a8e-3618ce8d9623/content

IFLA Working Group on the Functional Requirements for Subject Authority Records (FRSAR). (2011). *Functional requirements for Subject Authority Data (FRSAD)—A Conceptual Model* (M. L. Zeng, M. Žumer, & A. Salaba, Eds.). De Gruyter Saur. https://repository.ifla.org/handle/20.500.14598/835

Im, D.-H., & Park, G.-D. (2015). Linked tag: Image annotation using semantic relationships between image tags. *Multimedia Tools and Applications*, *74*(7), 2273–2287. https://doi.org/10.1007/s11042-014-1855-z

Ingwersen, P. (1992). *Information retrieval interaction*. Taylor Graham Publishing.

Isaac, A., & Summers, E. (2009, August 18). *SKOS Simple Knowledge Organization System Primer*. W3C. https://www.w3.org/TR/skos-primer/#secconcept

Isaac, A., Waites, W., Young, J., & Zeng, M. (2011). *Library Linked Data Incubator Group: Datasets, Value Vocabularies, and Metadata Element Sets*. https://www.w3.org/2005/Incubator/lld/XGR-lld-vocabdataset-20111025/#Library_Linked_Data_at_CKAN

Iyer, H. (1995). *Classificatory Structures: Concepts, Relations and Representation*. Indeks Verlag. https://archive.org/details/classificatoryst0002iyer/page/172/mode/2up

Janes, J. W. (1991). Relevance judgments and the incremental presentation of document representations. *Information Processing & Management*, *27*(6), 629–646. https://doi.org/10.1016/0306-4573(91)90004-6

Joudrey, D. N., Miller, D. P., & Taylor, A. G. (2015). *Introduction to cataloging and classification* (Eleventh edition). Libraries Unlimited.

Kagolovsky, Y., & Möhr, J. R. (2001). A new approach to the concept of "relevance" in information retrieval (IR). *Studies in Health Technology and Informatics*, *84*, 348–352. https://doi.org/10.3233/978-1-60750-928-8-348

Kaushik, V., & Walsh, C. A. (2019). Pragmatism as a Research Paradigm and Its Implications for Social Work Research. *Social Sciences (Basel)*, *8*(9), 255. https://doi.org/10.3390/socsci8090255

Kehoe, A., & Gee, M. (2011). Social tagging: A new perspective on textual 'aboutness.' In *Methodological and Historical Dimensions of Corpus Linguistics* (Vol. 6). VARIENG. https://varieng.helsinki.fi/series/volumes/06/kehoe_gee/

Kemp, D. A. (1974). Relevance, pertinence and information system development. *Information Storage and Retrieval*, *10*(2), 37–47. https://doi.org/10.1016/0020-0271(74)90002-3

Khoo, C. S. G., & Na, J.-C. (2006). Semantic relations in information science. *Annual Review of Information Science and Technology*, *40*(1), 157–228. https://doi.org/10.1002/aris.1440400112

Kim, H.-N., Rawashdeh, M., Alghamdi, A., & El Saddik, A. (2012). Folksonomy-based personalized search and ranking in social media services. *Information Systems*, *37*(1), 61–76. https://doi.org/10.1016/j.is.2011.07.002

Kipp, M. E. I. (2005). Complementary or discrete contexts in online indexing: A comparison of user, creator, and intermediary keywords. *Canadian Journal of Information and Library Science*, *29*(4), 419–436. https://doi.org/10.29173/cais174

Kipp, M. E. I. (2011). Controlled vocabularies and tags: An analysis of research methods. *Proceedings from North American Symposium on Knowledge Organization*, *3*, 23–32. https://doi.org/10.7152/nasko.v3i1.12787

Kipp, M. E. I., Beak, J., & Choi, I. (2017). Motivations and intentions of flickr users in enriching flick records for library of congress photos. *Journal of the Association for Information Science and Technology*, *68*(10), 2364–2379. https://doi.org/10.1002/asi.23869

Kipp, M. E. I., & Campbell, D. G. (2006). Patterns and Inconsistencies in Collaborative Tagging Systems: An Examination of Tagging Practices. *Proceedings of the American Society for Information Science and Technology*, *43*(1), 1–18. https://doi.org/10.1002/meet.14504301178

Kipp, M. E. I., & Campbell, D. G. (2010). Searching with Tags: Do Tags Help Users Find Things? *Knowledge Organization*, *37*(4), 239–255. Library, Information Science & Technology Abstracts with Full Text. https://doi.org/10.5771/0943-7444-2010-4-239

Kitiyadisai, K. (1991). *Concepts of relevance in a semiotic framework applied to ISAD (information systems analysis and design)* [Dissertation, London School of Economics and Political Science]. https://etheses.lse.ac.uk/1127/

Kiu, C.-C., & Tsui, E. (2011). TaxoFolk: A hybrid taxonomy–folksonomy structure for knowledge classification and navigation. *Expert Systems with Applications*, *38*(5), 6049–6058. https://doi.org/10.1016/j.eswa.2010.11.014

Kivunja, C., & Kuyini, A. B. (2017). Understanding and Applying Research Paradigms in Educational Contexts. *International Journal of Higher Education*, *6*(5), 26. https://doi.org/10.5430/ijhe.v6n5p26

Klimisch, H.-J., Andreae, M., & Tillmann, U. (1997). A Systematic Approach for Evaluating the Quality of Experimental Toxicological and Ecotoxicological Data. *Regulatory Toxicology and Pharmacology*, *25*(1), 1–5. https://doi.org/10.1006/rtph.1996.1076

Kome, S. (2005). *Hierarchical Subject Relationships in Folksonomies* [Master's, University of North Carolina at Chapel Hill]. https://doi.org/10.17615/br4v-gc16

Königová, M. (1971). Mathematical and statistical methods of noise evaluation in a retrieval system. *Inf. Storage Retr.*, *6*(6), 437–444. https://doi.org/10.1016/0020-0271(71)90009-X

Kopak, A. taxonomy of link types for use in hypertext. (2000). *A taxonomy of link types for use in hypertext* [Doctoral thesis, University of Toronto]. http://hdl.handle.net/1807/14406

Kotkov, D., Maslov, A., & Neovius, M. (2021). Revisiting the tag relevance prediction problem. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1768–1772. https://doi.org/10.1145/3404835.3463019

Kwasnik, B. H. (2010). Semantic warrant: A pivotal concept for our field. *Knowledge Organization*, *37*(2), 106–110. https://doi.org/10.5771/0943-7444-2010-2-106

Langevin, H. M., & Wayne, P. M. (2018). What is the point? The problem with acupuncture research that no one wants to talk about. *The Journal of Alternative and Complementary Medicine*, *24*(3), 200–207. https://doi.org/10.1089/acm.2017.0366

Langridge, D. Wilton. (1992). *Classification–its kinds, elements, systems, and applications*. Bowker in association with the Centre for Information Studies, Charles Short University, Wagga Wagga, N.S.W.

Lee, A., & Brewer, T. (2004). *ITL Bulletin: Information Security Within the System Development Life*

*Cycle (SDLC)*. National Institute of Standards and Technology. https://csrc.nist.gov/files/pubs/shared/itlb/itlbul2004-09.pdf

Lee, J. H., Clarke, R. I., & Perti, A. (2015). Empirical evaluation of metadata for video games and interactive media. *Journal of the Association for Information Science and Technology*, *66*(12), 2609–2625. https://doi.org/10.1002/asi.23357

Lee, K., Kim, H., Shin, H., & Kim, H.-J. (2009). FolksoViz: A Semantic Relation-Based Folksonomy Visualization Using the Wikipedia Corpus. *2009 10th ACIS International Conference on Software Engineering, Artificial Intelligences, Networking and Parallel/Distributed Computing*, 24–29. https://doi.org/10.1109/SNPD.2009.80

Legg, C., & Hookway, C. (2021). Pragmatism. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2021). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2021/entries/pragmatism/

Lesk, M. E., & Salton, G. (1968). Relevance assessments and retrieval system evaluation. *Information Storage and Retrieval*, *4*(4), 343–359. https://doi.org/10.1016/0020-0271(68)90029-6

Li, Q., Wang, P., Liu, C., Li, X., & Hou, J. (2024). Integration patterns in the use of metadata for data sense-making during relevance evaluation: An interpretable deep learning-based prediction. *Journal of the Association for Information Science and Technology*, *n/a*(n/a). https://doi.org/10.1002/asi.24961

Li, X., Uricchio, T., Ballan, L., Bertini, M., Snoek, C. G. M., & Bimbo, A. D. (2016). Socializing the semantic gap: A comparative survey on image tag assignment, refinement, and retrieval. *Acm Computing Surveys*, *49*(1). https://doi.org/10.1145/2906152

Liu, J., Wang, J., Zhou, G., Wang, M., & Shi, L. (2020). How Do People Make Relevance Judgment of Scientific Data? *Data Science Journal*. https://doi.org/10.5334/dsj-2020-009

Lóscio, B. F., Burle, C., & Calegari, N. (2017). *Data on the Web Best Practices*. https://www.w3.org/TR/dwbp/

Lu, K., & Kipp, M. E. I. (2014). Understanding the retrieval effectiveness of collaborative tags and author keywords in different retrieval environments: An experimental study on medical collections. *Journal of the Association for Information Science and Technology*, *65*(3), 483–500. https://doi.org/10.1002/asi.22985

Luo, S., Wang, T., & Kuang, H. (2018). Hierarchical tagging system and folksonomy construction from hierarchical tags. *Systems Engineering - Theory & Practice*, *38*(7), 1862–1869. https://doi.org/10.12011/1000-6788(2018)07-1862-08

Ma, Z., Dou, Z., Xu, W., Zhang, X., Jiang, H., Cao, Z., & Wen, J.-R. (2021). Pre-training for ad-hoc retrieval: Hyperlink is also you need. *Proceedings of the 30th ACM International*

*Conference on Information & Knowledge Management*, 1212–1221. https://doi.org/10.1145/3459637.3482286

Maglaughlin, K. L., & Sonnenwald, D. H. (2002). User perspectives on relevance criteria: A comparison among relevant, partially relevant, and not-relevant judgments. *Journal of the American Society for Information Science and Technology*, *53*(5), 327–342. https://doi.org/10.1002/asi.10049

Mai, J.-E. (2001). Semiotics and indexing: An analysis of the subject indexing process. *Journal of Documentation*, *57*(5), 591–622. https://doi.org/10.1108/EUM0000000007095

Mai, J.-E. (2005). Analysis in indexing: Document and domain centered approaches. *Cross-Language Information Retrieval*, *41*(3), 599–611. https://doi.org/10.1016/j.ipm.2003.12.004

Marcus, R. S., Kugel, P., & Benenfeld, A. R. (1978). Catalog information and text as indicators of relevance. *Journal of the American Society for Information Science*, *29*(1), 15–30.

Marlow, C., Naaman, M., Boyd, D., & Davis, M. (2006). HT06, tagging paper, taxonomy, Flickr, academic article, to read. *Proceedings of the Seventeenth Conference on Hypertext and Hypermedia*, 31–40. https://doi.org/10.1145/1149941.1149949

Maron, M. E. (1977). On indexing, retrieval and the meaning of about. *Journal of the American Society for Information Science*, *28*(1), 38–43. https://doi.org/10.1002/asi.4630280107

Maron, M. E., & Kuhns, J. L. (1960). On Relevance, Probabilistic Indexing and Information Retrieval. *J. ACM*, *7*(3), 216–244. https://doi.org/10.1145/321033.321035

Martínez-Ávila, D., & Budd, J. M. (2017). Epistemic warrant for categorizational activities and the development of controlled vocabularies. *Journal of Documentation*, *73*(4), 700–715. https://doi.org/10.1108/JD-10-2016-0129

Martínez-Ávila, D., & Kipp, M. E. I. (2014). Implications of the Adoption of BISAC for Classifying Library Collections. *Knowledge Organization*, *41*(5), 377–392. https://doi.org/10.5771/0943-7444-2014-5-377

Mathes, A. (2004). *Folksonomies – Cooperative Classification and Communication Through Shared Metadata [Online Report]*. Computer Mediated Communication, LIS590CMC, University of Illinois Urbana-Champaign. https://adammathes.com/academic/computer-mediated-communication/folksonomies.html

Mayring, P. (2014). *Qualitative content analysis: Theoretical foundation, basic procedures and software solution*. http://nbn-resolving.de/urn:nbn:de:0168-ssoar-395173

Miles, A., & Bechhofer, S. (Eds.). (2009, August 18). *SKOS Simple Knowledge Organization System Reference*. https://www.w3.org/TR/skos-reference

Mizzaro, S. (1997). Relevance: The whole history. *Journal of the American Society for Information Science*, *48*(9), 810–832. https://doi.org/10.1002/(SICI)1097-4571(199709)48:9<810::AID-ASI6>3.0.CO;2-U

Mizzaro, S. (1998). How many relevances in information retrieval? *HCI and Information Retrieval*, *10*(3), 303–320. https://doi.org/10.1016/S0953-5438(98)00012-5

Morrow, G., Swire-Thompson, B., Polny, J. M., Kopec, M., & Wihbey, J. P. (2022). The emerging science of content labeling: Contextualizing social media content moderation. *Journal of the Association for Information Science and Technology*, *73*(10), 1365–1386. https://doi.org/10.1002/asi.24637

Motic, B., Patel-Schneider, P. F., & Parsia, B. (Eds.). (2012, December 11). *OWL 2 Web Ontology Language Structural Specification and Functional-Style Syntax (Second Edition)*. W3C. https://www.w3.org/TR/owl2-syntax/

Noh, Y. (2011). A study on metadata elements for web-based reference resources system developed through usability testing. *Library Hi Tech*, *29*(2), 242–265. https://doi.org/10.1108/07378831111138161

*OBO Relation Ontology*. (n.d.). Retrieved April 7, 2021, from http://www.ontobee.org/ontology/ro

Olson, H. A., & Wolfram, D. (2008). Syntagmatic relationships and indexing consistency on a larger scale. *Journal of Documentation*, *64*(4), 602–615. https://doi.org/10.1108/00220410810884093

O'Neill, E. T., Kammerer, K. A., & Bennett, R. (2017). The aboutness of words. *Journal of the Association for Information Science and Technology*, *68*(10), 2471–2483. https://doi.org/10.1002/asi.23856

*Overview of the BIBFRAME 2.0 Model*. (2016). The Library of Congress. https://www.loc.gov/bibframe/docs/bibframe2-model.html

Palinkas, L. A., Horwitz, S. M., Green, C. A., Wisdom, J. P., Duan, N., & Hoagwood, K. (2015). Purposeful Sampling for Qualitative Data Collection and Analysis in Mixed Method Implementation Research. *Administration and Policy in Mental Health and Mental Health Services Research*, *42*(5), 533–544. https://doi.org/10.1007/s10488-013-0528-y

Park, H. (1997). Relevance of science information: Origins and dimensions of relevance and their implications to information retrieval. *Information Processing & Management*, *33*(3), 339–352. https://doi.org/10.1016/S0306-4573(96)00072-6

Park, M. S. (2016). *Exploring social semantic relationships for knowledge representation in health through mining social data* [Florida State University]. http://purl.flvc.org/fsu/fd/FSU_2016SU_Park_fsu_0071E_13307

Park, T. K. (1993). The Nature of Relevance in Information Retrieval: An Empirical Study. *The Library Quarterly*, *63*(3), 318–351. https://doi.org/10.1086/602592

Park, T. K. (1994). Toward a theory of user-based relevance: A call for a new paradigm of inquiry. *Journal of the American Society for Information Science*, *45*(3), 135–141. https://doi.org/10.1002/(SICI)1097-4571(199404)45:3%3C135::AID-ASI3%3E3.0.CO;2-1

Park, T. K., & Morrison, A. M. (2017). The Nature and Characteristics of Bibliographic Relationships in RDA Cataloging Records in OCLC at the Beginning of RDA Implementation. *Cataloging & Classification Quarterly*, *55*(6), 361–386. https://doi.org/10.1080/01639374.2017.1319451

Pauman Budanović, M., Švab, K., & Žumer, M. (2021). The gap between users and cataloguers: Perception of importance of bibliographic data for required reading and leisure reading. *Libellarium Journal for the Research of Writing Books and Cultural Heritage Institutions*, *11*(1–2), 71–87. https://doi.org/10.15291/libellarium.v1i1.322

Perreault, J. M. (1966). Documentary relevance and structural hierarchy. *American Documentation*, *17*(3), 136–139. https://doi.org/10.1002/asi.5090170307

Polit, D. F. (2014). Getting serious about test–retest reliability: A critique of retest research and some recommendations. *Quality of Life Research*, *23*(6), 1713–1720. https://doi.org/10.1007/s11136-014-0632-9

Quintarelli, E. (2005). *Folksonomies: Power to the people*. ISKO Italy-UniMIB meeting, Milan, Italy. https://www.iskoi.org/doc/folksonomies.htm

Rafferty, P. (2018). Tagging. *Knowledge Organization*, *45*(6). https://doi.org/10.5771/0943-7444-2018-6-500

Rafferty, P., & Hidderley, R. (2007). Flickr and Democratic Indexing: Dialogic approaches to indexing. *Aslib Proceedings*, *59*(4), 397–410. https://doi.org/10.1108/00012530710817591

Raper, J. (2007). Geographic relevance. *Journal of Documentation*, *63*(6), 836–852. https://doi.org/10.1108/00220410710836385

Rapley, T. (2014). Sampling Strategies in Qualitative Research. In U. Flick, *The SAGE Handbook of Qualitative Data Analysis*. SAGE Publications. https://doi.org/10.4135/9781446282243

Rauen, F. J. (2009). Relevance and Genre: Theoretical and Conceptual Interfaces. In *Genre in a Changing World* (pp. 57–77). The WAC Clearinghouse; Parlor Press. 10.37514/PER-B.2009.2324

Redigolo, F. M., Lopes Fujita, M. S., & Gil-Leiva, I. (2022). Guidelines for subject analysis in subject cataloging. *Cataloging & Classification Quarterly*, *60*(5), 424–443.

https://doi.org/10.1080/01639374.2022.2093300

Riva, P., Le Bœuf, P., & Žumer, M. (2017). *IFLA Library Reference Model: A conceptual model for bibliographic information*. Ifla. https://www.ifla.org/publications/node/11412 https://www.ifla.org/files/assets/cataloguing/frbr-lrm/ifla-lrm-august-2017.pdf

Rousi, A. M., Savolainen, R., Harviainen, M., & Vakkari, P. (2018). Situational relevance of music information modes: An empirical investigation among Doctor of Music students. *Journal of Documentation*, *74*(5), 1008–1024. https://doi.org/10.1108/JD-10-2017-0149

Rowley, J. (2007). The wisdom hierarchy: Representations of the DIKW hierarchy. *Journal of Information Science*, *33*(2), 163–180. https://doi.org/10.1177/0165551506070706

Ruthven, I. (2021). Resonance and the experience of relevance. In *J. Assoc. Inf. Sci. Technol.* (Vol. 72, Issue 5, pp. 554–569). John Wiley & Sons, Inc. https://doi.org/10.1002/asi.24424

Saarti, J. (2002). Consistency of subject indexing of novels by public library professionals and patrons. *Journal of Documentation*, *58*(1), 49–65. https://doi.org/10.1108/00220410210425403

Saracevic, T. (1975). RELEVANCE: A review of and a framework for the thinking on the notion in information science. *Journal of the American Society for Information Science*, *26*(6), 321–343. https://doi.org/10.1002/asi.4630260604

Saracevic, T. (1996). Relevance reconsidered. *Proceedings CoLIS 2 : Second International Conference on Conceptions of Library and Information Science: Integration in Perspective*, 201–218.

Saracevic, T. (2007). Relevance: A review of the literature and a framework for thinking on the notion in information science. Part II: nature and manifestations of relevance. *Journal of the American Society for Information Science and Technology*, *58*(13), 1915–1933. https://doi.org/10.1002/asi.20682

Saracevic, T. (2016). *The notion of relevance in information science: Everybody knows what relevance is. But, what is it really?* Synthesis Lectures on Information Concepts, Retrieval, and Services. https://www.morganclaypool.com/doi/abs/10.2200/S00723ED1V01Y201607ICR050

Šauperl, Alenka. (2002). Subject determination during the cataloging process. In A. Taylor (Ed.), *Subject determination during the cataloging process*. Scarecrow Press.

Savolainen, R., & Kari, J. (2006). User-defined relevance criteria in web searching. *Journal of Documentation*, *62*(6), 685–707. https://doi.org/10.1108/00220410610714921

Schamber, L. (1991). Users' criteria for evaluation in multimedia information seeking and use situations [Ph.D., Syracuse University]. In *ProQuest Dissertations and Theses* (303938286).

ProQuest Dissertations & Theses Global. https://surface.syr.edu/it_etd/43

Schamber, L. (1994). Relevance and information behavior. *Annual Review of Information Science and Technology*, *29*, 3–48. https://babel.hathitrust.org/cgi/pt?id=mdp.39015052604249&seq=23

Schamber, L., & Eisenberg, M. (1988). Relevance: The search for a definition. *Proceedings of the 51st Annual Meeting of the American Society for Information Science*. 51st Annual Meeting of the American Society for Information Science, Atlanta, Georgia. Library, Information Science & Technology Abstracts with Full Text.

Schamber, L., Eisenberg, M. B., & Nilan, M. S. (1990). A re-examination of relevance: Toward a dynamic, situational definition. *Information Processing and Management*, *26*(6), 755–776. https://doi.org/10.1016/0306-4573(90)90050-C

Schreier, M. (2012). *Qualitative content analysis in practice*. SAGE Publications.

Schutz, A. (1970). Reflections on the problem of relevance. In *Reflections on the problem of relevance.* Yale University Press.

Sedghi, S., Sanderson, M., & Clough, P. (2008). A study on the relevance criteria for medical images. *Image CLEF 2007*, *29*(15), 2046–2057. https://doi.org/10.1016/j.patrec.2008.07.003

Shiri, A. (2014). Linked data meets big data: A knowledge organization systems perspective. *Advances in Classification Research Online*, *24*, Article 1. https://doi.org/10.7152/acro.v24i1.14672

Smiraglia, R. P. (2016). Empirical methods for knowledge evolution across knowledge organization systems. *Knowledge Organization*, *43*(5), 351–357. https://doi.org/10.5771/0943-7444-2016-5-351

Soergel, D. (2009). *Knowledge Organization System Development*. https://www.researchgate.net/publication/308874089_Knowledge_Organization_System_Development

Sperber, D., & Wilson, D. (1995). Relevance: Communication and cognition / Dan Sperber and Deirdre Wilson. In *Relevance: Communication and cognition* (Second edition.). Blackwell Publishers.

Stankovic, M., Breitfuss, W., & Laublet, P. (2011). Linked-data based suggestion of relevant topics. *Proceedings of the 7th International Conference on Semantic Systems*, 49–55. https://doi.org/10.1145/2063518.2063525

Subirats, I., & Zeng, M. L. (2020). *Linked Open Data Enabled Bibliographical Data (LODE-BD) 3.0 – A practical guide on how to select appropriate encoding strategies for producing Linked Open Data Enabled Bibliographical Data*. FAO.

http://www.fao.org/documents/card/en/c/cb2209en

Sundin, O., & Johannisson, J. (2005). Pragmatism, neo-pragmatism and sociocultural theory. *Journal of Documentation*, *61*(1), 23–43. https://doi.org/10.1108/00220410510577998

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

Swanson, D. R. (1977). Information Retrieval as a Trial-And-Error Process. *The Library Quarterly: Information, Community, Policy*, *47*(2), 128–148. JSTOR. http://www.jstor.org.ezproxy.lib.uwm.edu/stable/4306788

Swanson, D. R. (1986). Subjective versus Objective Relevance in Bibliographic Retrieval Systems. *The Library Quarterly (Chicago)*, *56*(4), 389–398. https://doi.org/10.1086/601800

Taylor, A. (2013). Examination of work task and criteria choices for the relevance judgment process. *Journal of Documentation*, *69*(4), 523–544. https://doi.org/10.1108/JD-12-2011-0054

Taylor, A., Zhang, X., & Amadio, W. J. (2009). Examination of relevance criteria choices and the information search process. *Journal of Documentation*, *65*(5), 719–744. https://doi.org/10.1108/00220410910983083

Tennis, J. T. (2005). Experientialist Epistemology and Classification Theory: Embodied and Dimensional Classification. *Knowledge Organization*, *32*(2), 79–92. https://ssrn.com/abstract=2878207

*The Getty Vocabularies*. (n.d.). Getty Vocabularies: LOD. Retrieved April 13, 2025, from https://vocab.getty.edu/

Tillett, B. (1987). *Bibliographic relatioships: Toward a conceptual structure of bibliographic information used in cataloging*. University of California, Los Angeles.

Tillett, B. (2001). Bibliographic Relationships. In C. A. Bean & R. Green (Eds.), *Relationships in the Organization of Knowledge* (pp. 19–35). Springer Netherlands. https://doi.org/10.1007/978-94-015-9696-1_2

Trabelsi, C., Ben Jrad, A. ̈ıcha, & Ben Jrad, S. (2010). Bridging Folksonomies and Domain Ontologies: Getting Out Non-taxonomic Relations. *2010 IEEE International Conference on Data Mining Workshops*, 369–379. https://doi.org/10.1109/ICDMW.2010.72

Trant, J. (2009). Studying Social Tagging and Folksonomy: A Review and Framework. *Journal of Digital Information*, *10*(1).

Trigg, R. H. (1983). A Network-Based Approach To Text Handling For The On-Line Scientific Community [Ph.D., University of Maryland, College Park]. In *ProQuest Dissertations and*

*Theses* (303173380). ProQuest Dissertations & Theses Global.

Vakkari, P., & Hakala, N. (2000). Changes in relevance criteria and problem stages in task performance. *Journal of Documentation*, *56*(5), 540–562. https://doi.org/10.1108/EUM0000000007127

Van der Veer Martens, B., & Van Fleet, C. (2012). Opening the black box of "relevance work": A domain analysis. *Journal of the American Society for Information Science and Technology*, *63*(5), 936–947. https://doi.org/10.1002/asi.21699

van Dijk, T. A. (1979). Relevance assignment in discourse comprehension. *Discourse Processes*, *2*(2), 113–126. https://doi.org/10.1080/01638537909544458

Van Setten, M., Brussee, R., van Vliet, H., Gazendam, L., van Houten, Y., & Veenstra, M. (2006). On the importance of "who tagged what." *The Proceedings of Workshop on the Social Navigation and Community-Based Adaptation Technologies*, *6*. https://sites.pitt.edu/~paws/SNC_BAT06/crc/vansetten.pdf

Vander Wal, T. (2005). *Explaining and Showing Broad and Narrow Folksonomies*. https://www.vanderwal.net/random/entrysel.php?blog=1635

Vander Wal, T. (2007). *Folksonomy*. https://vanderwal.net/folksonomy.html

Vellucci, S. L. (1997). *Bibliographic relationships in music catalogs*. Lanham, Md. : Scarecrow Press.

Vickery, B. C. (1959a). Subject Analysis for Information Retrieval. *Proceedings of the International Conference on Scientific Information: Two Volumes*, 855–866. https://doi.org/10.17226/10866

Vickery, B. C. (1959b). The Structure of Information Retrieval Systems. *Proceedings of the International Conference on Scientific Information: Two Volumes*, 1275–1290. https://doi.org/10.17226/10866

Vig, J., Sen, S., & Riedl, J. (2009). Tagsplanations: Explaining recommendations using tags. *Proceedings of the 14th International Conference on Intelligent User Interfaces*, 47–56. https://doi.org/10.1145/1502650.1502661

Wang, P., & Soergel, D. (1998). A cognitive model of document use during a research project. Study I. Document selection. *Journal of the American Society for Information Science*, *49*(2), 115–133. https://doi.org/10.1002/(SICI)1097-4571(199802)49:2%3C115::AID-ASI3%3E3.0.CO;2-T

Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems*, *12*(4), 5–33.

Wang, Y., Tai, Y., & Yang, Y. (2018). Determination of semantic types of tags in social tagging system.

*Knowledge Organization*, *45*(8), 653–666. https://doi.org/10.5771/0943-7444-2018-8-653

Wetzker, R., Zimmermann, C., Bauckhage, C., & Albayrak, S. (2010). I tag, you tag: Translating tags for advanced user models. *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, 71–80. https://doi.org/10.1145/1718487.1718497

White, H. D. (2011). Relevance theory and citations. *Journal of Pragmatics*, *43*(14), 3345–3361. https://doi.org/10.1016/j.pragma.2011.07.005

White, H. D. (2017). Relevance theory and distributions of judgments in document retrieval. *Information Processing & Management*, *53*(5), 1080–1102. https://doi.org/10.1016/j.ipm.2017.02.010

Wildemuth, B. M., Marchionini, G., Fu, X., Oh, J. S., & Yang, M. (2019). The usefulness of multimedia surrogates for making relevance judgments about digital video objects. *Information Processing & Management*, *56*(6), 102091. https://doi.org/10.1016/j.ipm.2019.102091

Wilson, P. (1968). *Two kinds of power: An essay on bibliographical control.* University of California Press.

Wilson, P. (1973). Situational relevance. *Information Storage and Retrieval*, *9*(8), 457–471. https://doi.org/10.1016/0020-0271(73)90096-X

Wu, C., & Zhou, B. (2009). Semantic relatedness in folksonomy. *2009 International Conference on New Trends in Information and Service Science*, 760–765. https://doi.org/10.1109/NISS.2009.150

Xu, Z., Fu, Y., Mao, J., & Su, D. (2006). *Towards the semantic web: Collaborative tag suggestions*. Collaborative web tagging workshop at WWW2006, Edinburgh, Scotland. https://ra.ethz.ch/CDstore/www2006/www.rawsugar.com/www2006/13.pdf

Yu, L. (2015). Back to the fundamentals again: A redefinition of information and associated LIS concepts following a deductive approach. *Journal of Documentation*, *71*(4), 795–816. https://doi.org/10.1108/JD-12-2014-0171

Zappavigna, M. (2015). Searchable talk: The linguistic functions of hashtags. *Social Semiotics*, *25*(3), 274–291. https://doi.org/10.1080/10350330.2014.996948

Zavalina, O. L. (2011). Contextual Metadata in Digital Aggregations: Application of Collection-Level Subject Metadata and Its Role in User Interactions and Information Retrieval. *Journal of Library Metadata*, *11*(3–4), 104–128. https://doi.org/10.1080/19386389.2011.629957

Zeng, M. L. (2008). Knowledge organization systems (KOS). *Knowledge Organization*, *35*(2–3), 160–182. https://doi.org/10.5771/0943-7444-2008-2-3-160

Zeng, M. L., & Mayr, P. (2019). Knowledge Organization Systems (KOS) in the Semantic Web: A multi-dimensional review. *International Journal on Digital Libraries*, *20*(3), 209–230. https://doi.org/10.1007/s00799-018-0241-2

Zhang, G., Wang, J., Liu, J., & Pan, Y. (2021). Relationship between the Metadata and Relevance Criteria of Scientific Data. *Data Science Journal*. https://doi.org/10.5334/dsj-2021-005

Zhiwei Chen & Yunjie Xu. (2005). User-Oriented Relevance Judgment: A Conceptual Model. *Proceedings of the 38th Annual Hawaii International Conference on System Sciences*, 101b–101b. https://doi.org/10.1109/HICSS.2005.654

Zins, C. (2007). Conceptual approaches for defining data, information, and knowledge. *Journal of the American Society for Information Science and Technology*, *58*(4), 479–493. https://doi.org/10.1002/asi.20508

Appendix A: IRB Approval

UNIVERSITY of WISCONSIN
UWMILWAUKEE

Institutional Review Board

uwm.edu/irb
irbinfo@uwm.edu
414-662-3544

**Date:** August 25, 2022

**To:** Margaret Kipp
**Dept:** School of Information Studies
**CC:** Li Yang - Co-Investigator, Margaret Kipp - Faculty Advisor

**IRB #:** 23.027
**Title:** Data relevance: conceptualization and theorization in the perspective of knowledge organization

The University of Wisconsin-Milwaukee Institutional Review Board has granted your protocol Exempt Status under Category 2, 3 as governed by 45 CFR 46.104(d).

This exemption determination is valid for three years and will expire on **August 24, 2025**. Before the expiration date, you will receive an email explaining how to either keep the study open or close it. If the study is completed before the expiration date, you may notify the IRB by sending an email to irbinfo@uwm.edu.

Any proposed changes to the protocol must be reviewed by the IRB before implementation, unless the change is specifically necessary to eliminate apparent immediate hazards to the subjects.

It is your responsibility to:

- promptly report unanticipated problems to the IRB
- maintain proper documentation of study records
- ensure that all study staff receive appropriate training as outlined in the protocol
- adhere to the policies and guidelines set forth by the IRB, UWM, and the UW System, and to all applicable state and federal laws

Contact the IRB office if you have any further questions. Thank you for your cooperation and best wishes for a successful project.

Appendix B: Informed Consent

UNIVERSITY of WISCONSIN
UWMILWAUKEE

**Informed Consent for Research Participation**
**IRB #:** 23.027
**IRB Approval Date:** August 25, 2022

| | |
|---|---|
| **Study title** | Data relevance: conceptualization and theorization in the perspective of knowledge organization |
| **Researchers** | Li Yang, PhD candidate, School of Information studies<br>Margaret E.I. Kipp, Associate Professor, School of Information Studies |

Your participation in this study is voluntary. You may choose to not answer any of the questions or withdraw from this study at any time without penalty. Your decision will not change any present or future relationship with the University of Wisconsin Milwaukee.

### What is the purpose of this study?

The purpose of this study is to inquire about the concept of data relevance from a knowledge organization perspective in the Linked Data context, including discovering semantic relationships in LOD KOS, conceptualizing relevance in terms of LOD KOS, and developing a data relevance theory.

### What will I do?

you will do an offline task first and then an online interview.

Offline:

First, you will be given 2 task datasets and asked to design two metadata schemes to describe the datasets.

Second, you will be asked to compare other metadata schemes for the datasets with the metadata designed by you; then decide whether to revise the metadata schemes you designed or not. You will be asked to record any modifications and the reasonings. We will discuss it in the interview. We recommend you spend no more than 30-40 minutes on each resource in each step.

Online (video or audio):

We will have an interview with you. You will be asked about the reasons for your decision to or not to revise the metadata. You will also be asked to identify the relationships between the modified fields and the original fields you designed, and between the fields you designed and the schemes we provided. The interview will take you around 60 minutes.

Only the online interview is synchronous. The entire project will take you no more than 3 hours.

### Risks

Risks to participants are considered minimal. Conduction of interviews and collection of data using the internet involves the same risks that a person would encounter in everyday use of the internet, such as breach of confidentiality. While the researchers have taken every reasonable step to protect your confidentiality, there is always the possibility of interception or hacking of the data by third parties that is not under the control of the research team. We are minimizing this risk in the following ways by (1) not using any identifying information in data coding and analysis (e.g., your name, email address) and (2) storing all electronic data on a password-protected, encrypted computer.

### Other Study Information

| | |
|---|---|
| **Possible benefits** | • There are no benefits to participants other than further research.<br>• The results will have potential benefits for future policies, techniques, tools, etc. for data reuse, management, and sharing. |
| **Estimated number of participants** | 15 metadata librarians and data engineers |
| **How long will it take?** | 3 hours |
| **Costs** | None |
| **Compensation** | $50 Amazon gift card. Due to UWM policy and IRS regulations, we may have to collect your name, address, social security / tax ID number, and signature to give you this compensation. |
| **Future research** | Your data won't be used or shared for any future research studies. |
| **Recordings** | We will record you. The recordings will be used for data coding and analysis.<br>The recording is necessary for this research. If you do not want to be recorded, you should not participate in this study. |
| **Funding source** | The School of Information Studies, UWM |
| **Removal from the study** | In order for our data to be useful, it is important that you attend both offline and online parts. If you miss any part and can't reschedule, we'll have to take you out of the study. |

## Data Security

| | |
|---|---|
| **What identifying information will be collected and why?** | We have your name and contact information so we can contact you. But the identifying information will be removed in the data. We will replace it with a research ID. |
| **How long will my data be kept?** | Text data will be stored for 5 years. |
| **How is data kept secure?** | • All identifying information is removed and replaced with a research ID.<br>• We'll store all electronic data on a password-protected, encrypted computer.<br>• We'll keep your identifying information separate from your research data, but we'll be able to link it to you by using a research ID. We will destroy this link after we finish collecting and analyzing the data.<br>• As with any data collected online, there is always a risk of data being hacked or intercepted. We're using a secure system to collect this data, but we can't completely eliminate this risk. |

## Who might see my data and why?

| | |
|---|---|
| The researchers | To conduct the study and analyze the data |
| The IRB (Institutional Review Board) at UWM<br><br>The Office for Human Research Protections (OHRP) or other federal agencies | To ensure we're following laws and ethical guidelines |

| | |
|---|---|
| Anyone (public) | We plan to share our findings in publications or presentations. You will not be identified by name. If we quote you, we will use a pseudonym. |

**Contact information:**

| | | |
|---|---|---|
| **For questions about the research, problems, or complaints** | Li Yang | 414-324-2299/ liyang@uwm.edu |
| **For questions about your rights as a research participant, problems, or complaints** | IRB (Institutional Review Board; provides ethics oversight) | 414-662-3544 / irbinfo@uwm.edu |

**Signatures**

If you have had all your questions answered and would like to participate in this study, sign on the lines below. Remember, your participation is completely voluntary, and you're free to withdraw from the study at any time.

_____
Name of Participant (print)

_____ _____
Signature of Participant Date

Appendix C: Pre-questionnaire

# Pre-questions

Please use √ to specify your choices.

1. Do you speak English?

   ☐ Yes

   ☐ No

2. Do/Did you do work related to metadata or knowledge organization systems (vocabularies, ontologies, taxonomies, etc.)?

   ☐ Yes

   ☐ No

3. Are/Were you a metadata/data librarian or data engineer?

   ☐ Yes

   ☐ No

4. Would you like to participate in this study? By choosing Yes, you are indicating that you have read the Consent, you are age 18 or older and you voluntarily agree to participate in this study.

   ☐ Yes

   ☐ No

Appendix D: Structured Information Sample

We have already made the description of Resource 1 structuralized for your reference. For some fields that have many values, we only give examples that are directly excerpted from the original resource. You may consult the original resource if needed.

**Author of the report**: Bruce Mactavish, 37 Waterford Bridge Rd., St. John's NF Canada A1E 1C5.

**Report publication**: Field note, Volume 52 (1998), pp. 430-431.

**Season of observations**: 1998 Summer

**Seasonal weather report**: It was a warmer and drier summer than average across the Region.

**Geographic area of observation**: Atlantic provinces region

**Birds found**: Black-bellied Whistling-Duck, Black-necked Stilt, Bar-tailed Godwit, Black-tailed Gull, Gull-billed Tern, White-winged Dove, Scissor-tailed Flycatcher, and Violet-green Swallow.

**Numbers of bird:** e.g., fify red-throated Loons were still in migration at GMI, NB.

**Special observation**: e.g.,

The unprecedented spring influx of N. Shovelers into the e. extreme of the Region continued into June, with several breeding records resulting.

A male Ruddy Duck was an unexpected addition to the New Minas, NS, sewage lagoon waterfowl assemblage June 12-24 (JWo).

Early returning migrants were two Semipalmated and White-rumped sandpipers June 27 at Renews, NF (PL, BMt).

A Forster's Tern was most unusual in spring, and even more so for being inland at Youngs Cove, Grand L., N, June 10 (FH).

**Date/time-period of sighting:** e.g.,

One-hundred fifty Manx Shearwaters were seen from the St. Pierre ferry July 13 (RE).

The ad. Black-tailed Gull, appearing for the 2nd consecutive year on Sable I, NS, was present late April to at least the end of July.

A few sub-ad. Thick-billed Murres were seen off Deer I. and Campobello I., NB, in late July and early August (PLe, ST).

This year there was one at Louisbourg, NS, June 1-6 (fide BMy) and Tracadie, NB, July 16 (NB).

**Geographic location of sighting:** e.g., Sable I., NS, St. John's NF,

**Specific location**: e.g.,

A female Black-necked Stilt was photographed in a drained lobster pound in Lower W Pubnico, NS, June 28 for the 5th provincial record (JD et al.)

1

In Newfoundland, a brood was found at St. Paul's Inlet (ST) and a distressed female feigning an injury was at Parson's Pond (JMa).

**Activity (when observed):** e.g. feeding, nesting, breeding,

**Activity evidence:** e.g.,

first breeding evidence was a nest with a newly hatched chick and 2 eggs beside the runway at the Churchill Falls airport July 10 (BMt, KO).

Three juvenile Black-headed Gulls with several adults at a Ring-billed Gull colony in Spaniard's Bay, NF, was the first breeding evidence for the Avalon Pen.

**Bird age:** sub-adults, immatures, newly hatched, juvenile, adult,

**Observer:** e.g.,

....was an extraordinary find by 2 cyclists on Jun 17 (fide SF).

Gert Arkensteyn, Robert Bateman.....

**Observer activity:** e.g., A female Black-necked Stilt was photographed in a drained lobster pound in Lower W Pubnico, NS, June 28 for the 5th provincial record (JD et al.).

**Subregional editor:** e.g.,

Brian Dalzell, Roger Etcheberry.....

**Image information**

**Subject:** Bar-tailed Godwit of the European subspecies lapponica

**Location:** at Cherry Hill Beach, Nova Scotia

**Date:** June 9, 1998

**Photographer:** McLaren

Appendix E: Comparison Table for Stage 2

Please follow the table below to complete the tasks of Stage 2. Please feel free to add more information if you think it is important. If you are not comfortable with the format, you may change it but make sure to include all the fields.

| **Stage-1**<br>**Schema: Element** | **Stage-2(record your changes)**<br>**Schema: Element** | **What I wanted to represent (you may use the text from the resources)** | **Reasons for your changes/no-changes** |
|---|---|---|---|
| | | | |
| | | | |

Examples:

For the first resource

| **Stage-1**<br>**Schema: Element** | **Stage-2(record your changes)**<br>**Schema: Element** | **What I wanted to represent (you may use the text from the resources)** | **Reasons for your changes/no-changes** |
|---|---|---|---|
| dcterms: creator | - | Author of the report | No change: The most frequently used element for author |
| ... | | | |

For the second resource

| **Stage-1**<br>**Schema: Element** | **Stage-2(record your changes)**<br>**Schema: Element** | **What I wanted to represent (you may use the text from the resources)** | **Reasons for your changes/no-changes** |
|---|---|---|---|
| - | dcterms: identifier | The rubbing numbered 'Ms. Pelliot chinois 4508 | I missed the information in Stage 1. |
| dcterms: source | dbpedia-owl:picture<br>bf:count<br>dcterms:source | Image information: three images are shown on the first page of Resource 2. The source of the images is Chen, Tsu-Lung (1958). You may check the reference section for more detail. | The embedded structure makes the description clear and easy to follow.<br>I missed the count information. It is important to describe image information. |

Appendix F: Interview Questions

## Introduction

Hi ***, I am Li Yang. Nice to meet you. Welcome and thank you for participating in this interview. I will be leading the interview today. The interview is about a research project Data Relevance: Conceptualization and Theorization from the perspective of Knowledge Organization. We appreciate your contribution to this project.

I will be assisted by the recording function of TEAMS. The interview will be audio-recorded and will be transcribed for analysis only. We will then destroy the recordings.

I would like to let you know that your participation is totally voluntary. Please let me know if you feel uncomfortable with any questions in the interview or if you want to withdraw from the interview at any time. The questions we will be asking you are not a test. There are no right or wrong answers. We want to know what you think and feel.

Do you have any questions before we start?

TEAMS doesn't have a function to record audio only. Please feel free to close your camera if you would like to.

I am going to record the conversation.

## Common Questions

First, let's review the two-stage tasks briefly. In the first stage, you designed two metadata schemas to describe two resources: an Audubon field note and a Chinese rubbing. In the second stage, you modified your schemas according to the references.

1. Would you please talk about your experience with metadata-related work in general? Do you have metadata schemas that you use frequently in your work?
2. Were you familiar with the two types of resources before you did the tasks?
3. Would you please talk about your general impression of the tasks?
4. Did you use the structuralized information we provided when doing the tasks? Would you please talk more about how you used it?

**The following questions ask about the tasks in the first stage.**

5. You reused terms from different vocabularies. Did you use LOV to look for vocabularies and terms in the first stage? When you designed/modified the metadata schemas, was it easy for you to find suitable terms?
6. Do you have a strategy for creating the schemas for both resources? What do you think were the most difficult parts of the process?
7. When selecting the terms, did you decide the vocabularies first, then choose the appropriate terms from the vocabularies. Or you focused more on terms?
8. Did you intend to design a hierarchy in your schemas in the first stage? Or did you intend to borrow the hierarchy from your reference? Do you prefer a flat schema or a hierarchical one? Why?

**The following questions ask about the tasks in the second stage.**

9. Did you use LOV to look for vocabularies and terms in the second stage? Why or why not?
10. Have you ever worked with locating data on a specific topic or for a project? If not, what do you think is a good way to find data? If yes, would you please talk about it? How did you locate the relevant data? How did you determine that data was relevant to your project? Did you use metadata (especially specific elements) to determine relevance? Do you feel the structures in metadata would affect your impression of relevance?
11. We would like to know more about your thoughts and perspectives on data relevance. According to your experience, how do you define relevant data?
12. Do you have any other questions or comments?

## Individualized Questions (samples)

1. For the first resource, you listed the vocabulary and terms in column 2 "Stage-2 (record your changes). That means you changed to a new term, but you said "no change" in the reasons column. Would you please talk about it?
2. You kept publisher, rights, format, and relation. The information is not provided in the resource description. Would you please talk more about your thoughts and reasons?
3. You made several changes for accuracy. For example, you used Chinese Metadata for Rubbings: currentLocation.geographicLocation instead of dc:location, dwc:lifeStage instead of dc:description. However, you also made changes for generality. For example, you used dc:description instead of dc:date for the first resource . Would you please talk more about the two different decisions?
4. In the Audubon Note, you mentioned twice "Considered use of Dublin Core, but prefer use of Bibframe". Would you please talk more about that?
5. In the line of Finding place, you mentioned "Can use the more granular element in the schema designed for rubbings". There is granular location information in the description. What elements did you consider? Why didn't you use them?
6. You added Description-engraving description and rubbing description. Are they new terms? And which vocabulary are they from?

## Ending

Thank you for your time and effort on this project. The interview is done. I will stop recording. We will send you an E-gift card as an honorarium to your email soon after our conversation. We deeply appreciate it again. Thank you.