EXPLORATION OF RDA-BASED MARC21 SUBJECT METADATA IN WORLDCAT DATABASE AND ITS READINESS TO SUPPORT LINKED DATA FUNCTIONALITY

Vyacheslav Igorevich Zavalin

Dissertation Prepared for the Degree of

DOCTOR OF PHILOSOPHY

UNIVERSITY OF NORTH TEXAS

August 2020

APPROVED:

Shawne D. Miksa, Committee Co-Chair
Brian C. O'Connor, Committee Co-Chair
Barbara Schultz-Jones, Committee Member
Muzhgan I. Nazarova, Committee Member
Jiangping Chen, Chair of the Department of Information Science
Kinshuk, Dean of the College of Information
Victor Prybutok, Dean of the Toulouse Graduate School

Zavalin, Vyacheslav Igorevich. *Exploration of RDA-Based MARC21 Subject Metadata in Worldcat Database and Its Readiness to Support Linked Data Functionality*. Doctor of Philosophy (Information Science), August 2020, 238 pp., 31 tables, 20 figures, 2 appendices, references, 204 titles.

Subject of information entity is one of the fundamental concepts in the field of information science. Subject of any document represents its intellectual potential -- 'aboutness' of the document. Traditionally, subject (along with title and author) is the one of three major ways to access information, so subject metadata plays a central role in this process and the role is constantly growing. Previous research concluded that the larger bibliographic database is, the richer subject vocabularies and classification schemes are needed to support information discovery. Further, a high proportion of information objects are unretrievable without subject headings in metadata records. This exploratory study provides the analysis of the subject metadata in MARC 21 bibliographic records created in 2020; and develops understanding of the level and patterns of 'aboutness' representation in the MARC 21 bibliographic records. Study also examines how these records apply the recent RDA and MARC21 guidelines and features intended to support functionality in a Linked Data environment. Methods of Social Network Analysis were applied along with content analysis, to answer research questions of this study. Suggestions for future research, implications for education, and practical recommendations for library metadata creation and management are discussed.

ACKNOWLEDGEMENTS

I would like to thank my loving family and my committee for support and patience. I dedicate this work to the benefit of all sentient beings.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION

1.1 Problem Statement

Helping users to satisfy their information needs and obtain needed information resources is a top priority and the main principle in library and information practice and research (e.g., Dervin & Nilan, 1986). Metadata, especially structured metadata, is crucial for providing access to recorded knowledge collected and organized in various databases, including library databases. The most common kinds of metadata that have traditionally been included as entry points or "main access points" in metadata records are names of creators, titles of works, and subjects of works. In the distant past, it was possible in principle to find all or most of the information the user needed -- assuming it was available through a library, museum, archive, or other collection -- based on knowing the title and/or the name of the author of the work. The exponential growth of scientific information (Price, 1963) and information in general in the 19th, 20th and 21st centuries, especially intensified since the emergence of the Internet and the Web, has changed the situation. The current information age can be characterized by rapid increase of the amount of generated data, as well as published information, often referred to as information explosion or Big Data environment. This leads to problems with understanding and making decisions under the pressure of a large amount of information, resulting in information overload or information anxiety (e.g., Yang, Chen, & Hong, 2003).

In the information explosion environment, discovery by the known item (title or author) is seriously limited by the information overload. Therefore, information discovery by subject becomes more and more important (e.g., Bates, 2002), and this places an increasing emphasis

on functionality of subject metadata, the parts of metadata records that represent the intellectual content or "aboutness" (e.g., Fairthorne, 1969; Wilson, 1968) of information objects. The creation of subject metadata is a very time-consuming process that involves analysis of subject matter, relationships among topics, form, and genre in the context of the intended audience and possible uses of information objects (Joudrey, Taylor and Miller, 2015).

The process of metadata creation, including subject metadata creation, is guided by several types of standards. The first type is the data content standards, and in the library community it is currently represented primarily by the Resource Description and Access (RDA), (RDA Steering Committee, 2010; RDA Co-publishers, 2010). The library data content standard that provides guidelines specifically on the subject metadata creation is the Library of Congress Subject Headings Manual (Library of Congress, 2020b). The second type is data value standards, and these are represented by controlled vocabularies (e.g., thesauri, lists of subject terms and codes, etc.) and by classification schemes. In the United States of America library community, the widely used subject data value standards include Library of Congress Subject Headings (LCSH), Faceted Application of Subject Terminology (FAST), MARC Geographic Area Codes, Library of Congress Classification, Dewey Decimal Classification, etc. The third type is data encoding and transmission standards, which makes library metadata shareable and interoperable. This type of standards is currently represented in the library cataloging community by the Machine Readable Cataloging (MARC) standard, and to a lesser extent with MARCXML, Metadata Object Description schema (MODS), and the emerging Bibliographic Framework Initiative (BIBFRAME) standard.

According to Buckland (1999), when the user is attempting discovery by subject, the

processing of a search query involves a number of different vocabularies. These vocabularies can include authors, documents, searchers, indexers, syndetic structures, and queries and the complexity of each vocabulary greatly increases the chances of mismatch. Thus, supporting the main function of providing adequate answers to user search queries through functional subject metadata is a very complex task. Subject access plays a central role in information retrieval systems. For example, according to Aluri, Kemp, and Boll (1991), various existing information retrieval systems fall into four major categories, three of which retrieve documents based on their subject. The first, natural language type, indexes documents based using the words contained in the documents themselves. The second of these three groups relies on controlled vocabulary of terms (words or phrases), and the third on controlled vocabulary of notations in classification systems (pp.28-29). The authors note that subject access is the most complex type of information access that continues its evolution and has yet to reach its full potential (p.298). Subject access studies over the years have revealed mixed success in exploration by subject (cf.., Krikelas, 1972; Larson, 1991, Markey, 1984, etc.). Negative user experiences (e.g., Markey, 1984) in subject searches have been identified as the major reason for the under-utilization of controlled vocabularies in subject searching. Markey and Demeyer (1986) recommended expanding search strategies in online catalogs by adding searches based on subject representation through classification schemes. Drabenstott and Weller (1996) and Drabenstott (1996) reported results aimed at developing a new approach to the design of library online catalogs that would improve subject access by utilizing search trees based on user queries for subjects. Based on their experiments, researchers concluded that catalogs enhanced by the implementation of search trees worked well in selecting more effective subject-searching

approaches for the users and substantially increased the functionality of online catalogs. The Semantic Web ideas that emerged from computer science in the late 1990s (Berners-Lee, Handler, & Lasilla, 2001) and related technologies that have been developed since then hold promise for greatly improving information access in general, including subject access, when applied to metadata at the global scale.

Creators of the Semantic Web seek to connect pieces of information in a logical way that is more understandable and processable by machines in order to improve information retrieval (Berners-Lee, 2007). This ability to connect data is called Linked Data, and one of the main steps to achieving this is the inclusion of unique Uniform Resource Identifiers (URIs) that lead the user to the openly available information on the entity identified via a URI using the HTTP protocol. The library metadata community is developing and applying ways to support the Linked Data functionality of metadata. For example, the MARC21 standard has been expanded to facilitate inclusion of URIs into bibliographic and authority records by adding subfields $0 to MARC metadata fields containing subject terms and names from the controlled vocabularies such as LCSH, Library of Congress Name Authority File, Virtual International Authority File (VIAF), etc. The stakeholders in the United States library metadata community, including the Library of Congress, OCLC, National Library of Medicine, various academic libraries, etc., are working to enrich the vast body of existing MARC metadata records with URIs and to provide library metadata as a Linked Data (e.g., Boehr & Bushman, 2018; Godby & Denenberg, 2015; Shieh & Reese, 2015).

The BIBFRAME initiative that began in the early 2010s and is led by the United States' Library of Congress, is building upon Semantic Web principles and applying them to metadata

generated by the library community (BIBFRAME, 2011). The BIBFRAME metadata element set is projected to replace the MARC standard, which was originally developed in the 1960s and which has been applied by the library community in its metadata ever since. While BIBFRAME metadata record creation tools such as BIBFRAME Editor, developed by the BIBFRAME Initiative, are being explored by the early adopters in the library community, the integrated library systems software developers are starting to build BIBFRAME into their tools. Once these tools gain wide adoption and become mainstream, a majority of the newly created library metadata records will originate as BIBFRAME records.

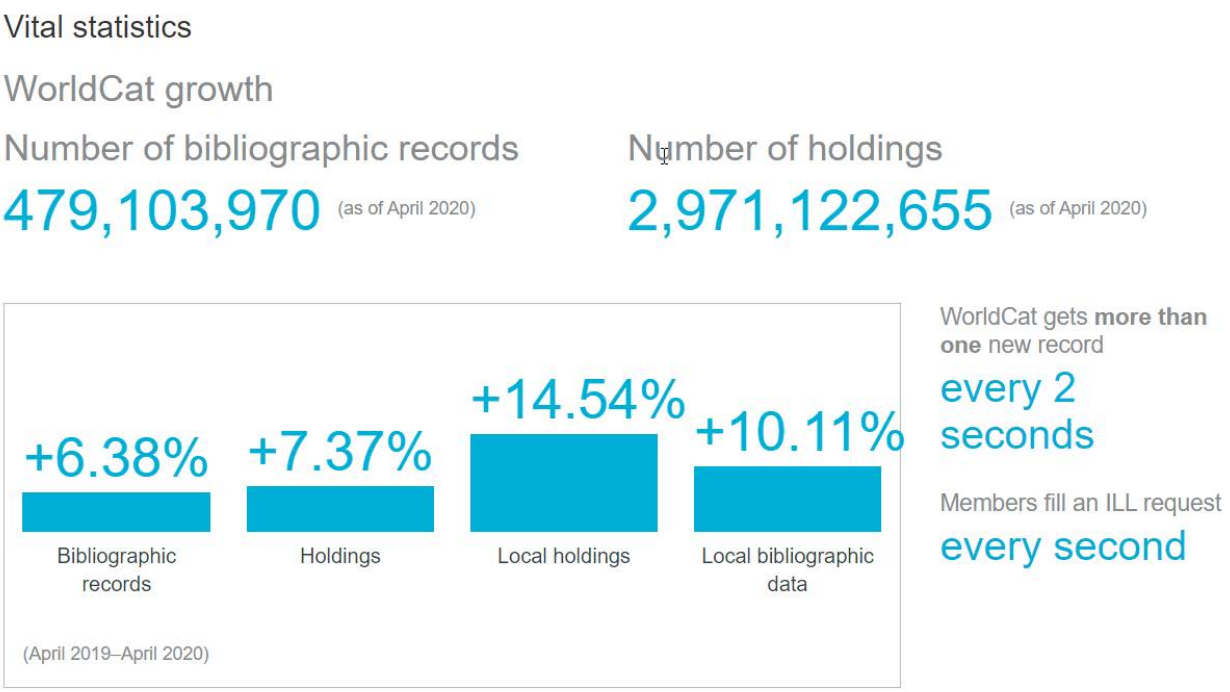


Figure 1.1: OCLC WorldCat vital statistics (OCLC, 2020, April).

At the same, hundreds of millions of existing MARC records that collectively represent and provide access to the vast body of recorded knowledge will need to be reformatted or converted from MARC to BIBFRAME. As of now, the OCLC WorldCat database contains over 479 million metadata records which have been created and edited collaboratively by the

international library community for nearly 50 years (Figure 1.1). A substantial proportion of these records are already being converted to BIBFRAME.

Due to the sheer volume of that conversion task, the reformatting of the millions of metadata records from MARC21 to BIBFRAME will be automatic. As the output quality in automatic conversion processes relies greatly on the input quality, to ensure the conversion is producing meaningful and functional results, the input metadata (data values in MARC 21 bibliographic records) needs to support that functionality. However, it is unclear as to what extent the Semantic Web functionalities can be realized when the records are converted automatically from MARC 21 to BIBFRAME. This study sought answers to this question, with a focus on the subject metadata fields in the records.

1.2 Background of the Problem

Information technologies are intended to help overcome different kinds of difficulties in processing enormous amounts of data and allow users to store, retrieve, manipulate, and transmit information (IT, 2016). Information and communication technologies as a concept does not have a universal definition but is usually perceived as a combination of different types of computing hardware, software, operating systems, systems for audio-video processing, storage, and telecommunications. Information technologies constantly and quite rapidly evolve under the pressure of demands for effectiveness. All these developments and innovations have facilitated changes and transformations in the global information society, which is critically reliant on convenient access to and distribution of information. So, the influence of information technology on society can be observed ubiquitously in all aspects of human life from grocery shopping to business administration, from scientific research and data analysis to visual

representations. These technologies require users to be flexible and have the ability to quickly learn new technological concepts and tools and adopt new information seeking patterns.

Developments in information technologies have substantially contributed to the evolution of information science, which evolved as a discipline throughout the mid- to late 20th century. There is no unified definition of information science (IS). However, as discussed by Buckland (2012), information science can be considered from different perspectives as an academic discipline that is concerned with either information and communication technologies or information physics and information entropy. Alternatively, it can be considered as a science and professional practice that evolved from such existing disciplines as documentation and librarianship and focused on meaning, knowledge, and information recorded in documents. Buckland's viewpoint correlates with the early definition of information science provided by Borko (1968) and used by Tefko Saracevic (2009) in the *Encyclopedia of Library and Information Science*. This definition expresses the idea of the nature of IS as a science and practice "dealing with effective communication of information and information objects, particularly knowledge records, among humans in the context of social, organizational, and individual need for and use of information" (Saracevic, 2009, p. 1). Bates (1999) suggests looking at information science as a multidisciplinary field that involves all kinds of knowledge with a focus on recorded human information as a main concept. Methods and technologies and theoretical concepts for recording, describing, organizing, retrieving, and using information have been very important human activities for centuries. Despite the fact that information retrieval and information organization academic disciplines remain divided in the way they are taught, within design problems and activities these disciplines are interdependable and re-converging (Glushko,

2013). Explanation of this is based on the fact that organization enables retrieval; and the better information is organized, the more efficient retrieval can be (Glushko, 2013).

Helping users to satisfy their information needs and obtain needed information resources is the top priority and principle used in the field of Library and Information Science. The user-centered approach is the most preferred approach in research and practice since the 1970s (e.g. Bates,1972; Dervin & Nilan,1986). As recorded knowledge continues to grow with geometric progression (Price, 1975), finding accurate and relevant information becomes very difficult without functioning metadata in such a dynamic information environment, often referred to as Big Data. With advances in information and communication technologies', significant issues come to light about the quality of metadata and integrated library systems (ILSs) or asset management systems that help institutions to manage, organize, preserve and provide access to information resources stored in their collections. For such activities, these systems use different types of metadata schema to markup data. This metadata is created and aggregated in large volumes. As a part of the data management lifecycle there might arise logical questions about the assessment of the level of representation of metadata and the measurement of its quality.

As an important part of metadata, subject metadata deals with intellectual content of information objects through the use of words and phrases from controlled vocabularies or natural language that represents "aboutness" (e.g., Fairthorne, 1969; Wilson, 1968). Creation of subject metadata or subject representations is a very time-consuming process that involves analysis of subject matter, relationships among topics, form, and genre in the context of the intended audience and possible uses of information objects (Joudrey, Taylor & Miller, 2015).

This process is called subject analysis and encompasses the gathering of data through familiarization with the information object's intellectual content, creating a list of concepts and/or summarization of the content, conversion of these representations into appropriate metadata for the information system and the user, and finally a reexamination of the terms' accuracy and consistency — in other words, evaluation of quality of assigned subject access points (Joudrey, Taylor and Miller, 2015). Traditionally, this metadata is available in library systems coded in the MARC21 standard.

MARC21 is currently the dominant family of machine-readable cataloging encoding formats and international metadata standards (ISO 2709/ANSI Z39.2 standard) for description of information objects and exchange of metadata among and between libraries and other entities. This encoding standard was developed as Machine Readable Cataloging (MARC) in the 1960s at the Library of Congress by computer scientist Henriette Avram (Avram, 1976). Over the years this standard has been adopted and adapted by multiple countries. The current version (MARC21) resulted from an integration of American, British, Canadian MARC formats and UNIMARC, widely used in Europe and Asia. MARC 21 bibliographic standard currently includes over 30 fields for subject representation.

Resource Description and Access (RDA) was developed to replace Anglo-American Cataloguing Rules, 2nd Edition Revised (AACR2r) cataloging rules. The development of RDA started in 2005 was initially released in 2010, but only officially implemented by the Library of Congress in March of 2013. RDA is widely used as the standard for descriptive cataloging by libraries and other institutions. Since its implementation, RDA continues to evolve and grow to meet the end user needs and is currently in the process of major revision called 3R (RDA Toolkit

Restructure and Redesign) that accommodates the recently adopted newly aggregated conceptual model entitled the Library Reference Model (LRM). LRM (IFLA, 2017) replaces a family of functional requirements models: Functional Requirements for Bibliographic Records (FRBR) (IFLA, 1998; 2009), Functional Requirements for Authority Data (FRAD) (IFLA, 2013), and the Functional Requirements for Subject Authority Data (FRSAD) (IFLA, 2010) developed between 1997 and 2013 and serving as a part of the foundation of RDA. These developments brought to attention some of the limitations of MARC 21 as an encoding standard. Understanding the need to create a more flexible framework for bibliographic description that can be useful not only within, but also outside the library community, the Library of Congress is developing BIBFRAME, a Linked Data model for bibliographic description to eventually replace MARC 21 (El-Sherbini, 2018). This development of Linked Data functionality potentially improves discoverability of information through metadata records, including subject access through subject metadata.

Both RDA and BIBFRAME are developed to support Linked Data and Semantic Web development with the ultimate goal of improving discoverability of information objects through increased functionality of and interconnectedness between the metadata records. To that end, RDA and BIBFRAME place emphasis on expressing relations, and using unique identifiers such as Uniform Resource Identifiers (URIs) to support expression of these relationships between various works, their instances, and important entities related to work such as subjects and agents.

While BIBFRAME currently remains an initiative, MARC 21 continues to be a useful encoding standard (El-Sherbini, 2018). MARC 21 also constantly evolves to reflect the changes

in library cataloging practice: new data elements (fields and subfields) are added regularly to support the functionality of RDA and BIBFRAME. For example, according to the content designator history published by the Library of Congress MARC Standards Office for each group of MARC 21 bibliographic fields (e.g., https://www.loc.gov/marc/bibliographic/bd01x09x.html), in the 15 years between the beginning of the RDA development and the date of the most recent revision to MARC 21 bibliographic standard (May 2020), several new fields and subfields have been added to MARC21 bibliographic standard for subject representation and to expand functionality and support Linked Data.

1.3 Research Questions

This exploratory study sought to answer the following research questions:

1. What extent and variety of subject representation do the library metadata records (i.e., MARC21 bibliographic records) currently provide? How are the most recent RDA and MARC21 guidelines and features intended to support functionality in Linked Data environment and BIBFRAME conversion applied in subject metadata elements in the records?
2. How does the application of existing subject metadata in the most recently created MARC21 library metadata records affect relations between these records as measured by social network analysis?
3. How does the subject representation in the newly created MARC21 bibliographic records carry over into BIBFRAME records resulting from automated conversion from MARC21? What implications does such a conversion have for interconnectedness of records based on subject metadata?

This study relied on the combination of research methods: quantitative and qualitative content analysis with application of graph methods as a part of social network analysis (SNA).

Chapter 3 provides details on each of the research questions and the methods used to answer them.

1.4 Conceptual Framework

The conceptual frameworks that provided context for this investigation focus on the functionality of library metadata. They include the Library Reference Model (LRM) adopted by the International Federation of Library Associations and Institutions (IFLA), the three models that preceded LRM—Functional Requirements for Bibliographic Records (FRBR), Functional Requirements for Authority Data (FRAD), and Functional Requirements for Subject Authority Data (FRSAD)—and the Bibliographic Framework (BIBFRAME) model.

1.5 Assumptions, Limitations, Delimitations

This study was intended to explore readiness of the RDA-based MARC21 bibliographic data created according to the most recent official version of RDA rules (last updated in 2018) and the latest version of MARC 21 Bibliographic Format standard (last updated in 2020) to support BIBFRAME and Linked Data functionalities. For that reason, this study did not attempt to examine non-RDA records, records that were partially converted into RDA from existing AACR2 records, or RDA records created in the early stages of RDA adoption. Because the focus of the study was on subject representation of information objects, the project examined only bibliographic records and there was no attempt to examine authority records. Also, the application of MARC 21 bibliographic fields that do not carry any subject metadata were not analyzed (i.e., no descriptive metadata was analyzed).

This study explored the level of application of various components in subject metadata that are expected to support meaningful BIBFRAME conversion of existing MARC 21 bibliographic records. The project maintained the focus on the actual metadata records. The guidelines for metadata creation that were considered as a context for metadata evaluation

were the MARC 21 Bibliographic Format guidelines (https://www.loc.gov/marc/bibliographic/). The study did not include examination of policies and procedural manuals developed and used locally by individual institutions that create RDA-based MARC 21 bibliographic records and convert them to BIBFRAME records for institution-specific guidelines on subject representation.

1.6 Significance of the Study

This study is expected to make a contribution to the understanding of subject metadata application practices in the bibliographic metadata records collaboratively created and shared by libraries and other heritage institutions. It is the first study to examine in-depth (combining quantitative and qualitative data analysis approaches) the MARC21 metadata created according to the most recent version of the library data content standard Resource Description and Access (RDA) and the MARC21 standard. Such a focus allows for estimating the overall readiness of this metadata for supporting Linked Data functionality when converted to the BIBFRAME standard that is expected to replace MARC21 standard in the future.

Recently, there has been a growing research interest in the topic of library metadata as Linked Data. This study is expected to make a contribution to the emerging research of relationships between metadata records in the manner of Linked Data. As part of the study, common social network analysis measurements and methods were used to help develop a common understanding of connections between metadata records based on the data values in the subject metadata fields.

CHAPTER 2

LITERATURE REVIEW

2.1 Subject Metadata in Subject Access: Introduction

In part due to the complexity of the phenomenon, the library and information science field did not develop universally accepted definitions of subject matter and subject access. However, these topics have long been explored, especially in relation to information seeking and information retrieval (Hjørland, 1997) and have become fundamental concepts in the field (e.g., Golub, 2014). Some definitions, such as Fairthorne's (1969) definition of subject matter as "aboutness" of an information object are widely cited in the literature. Cochrane (1979) coined the operational definition of subject access as both the user exploration of the database by subject and the subject cataloging using systematic (e.g., classification system), topical (e.g., subject headings), and natural (e.g., title, abstract words) approaches to the subject matter in a collection.

Miksa (1983), in his fundamental work on historical development of subject representation in catalogs, discusses the subject cataloging as a "scope-matching" process, where the extent of the topical content of an information object has to be matched by either a single subject heading or a set of subject headings used to represent this information object in a metadata record (p.7).

According to Bates (1999), the "information explosion (with us since the invention of printing) has driven most of the major innovations in information organization and access", when the average collection of recorded knowledge grew to a next level, the need evolved for developing new access methods (p. 1048). Charles Cutter's efforts at developing the Library of

Congress Subject Headings and the guidelines on using them in the late 19th -early 20th century are used by Bates as a vivid example of such an innovation that was necessitated and brought to life by information explosion.

As early as the last quarter of the 19th century and the early 20th century, the *Rules for a Dictionary Catalog* formulated by Cutter (1904) highlighted providing subject access as an important function of a library catalog. Subject access is an integral component of all three Cutter's objectives of a library catalog:

1. To enable a person to find a book of which [...] the subject is known
2. To show what the library has [...] on a given subject
3. To assist in choice of a book as to [...] its character (literary or topical). (Cutter, 1876, p.10)

Cutter emphasized the finding principle, in which each information object is represented in a uniquely identifiable way, and collocation, in which similar information objects are brought together, for example those works that are on the same subject or similar/related subjects.

Despite the influence of Cutter's ideas on the library and information science research and on librarianship as a profession, his ideas on the importance of subject access have not been fully realized in the library practice. In the era of card catalogs, the amount of bibliographic information, including subject headings was restricted by the size of 3X5 inch cataloging cards. Moreover, the lack of resources has always been an issue impeding subject access. For example, the crisis in cataloging of late 1930s-1940s exemplified by huge cataloging backlogs – almost 29% of the total Library of Congress collection by 1944 -- called for giving up contents notes, series entries and added name entries, and further limiting subject access for the sake of providing at least some access through author and title fields, lowering costs of

cataloging and decreasing backlogs (MacLeish, as cited in Ercegovac, 1998). Because the crisis in cataloging has remained a reality since 1940s, this approach was later incorporated in AACR2 as minimal-level cataloging and was transferred from card to online public access catalogs with a naive expectation that the power of online catalogs would compensate for simplified cataloging in terms of retrieval (Ercegovac, 1998). For many years cataloging has been guided more by practical considerations of librarians than by the needs of users. Cutter's nineteen century limitations in the view of library catalogs were critically reviewed by Wilson (1983) and Miksa (1983).

In Svenonius' (2000) definition, the "subject language" depicts what a document is about. Similarly, Soergel (2009) defines subject metadata as information concerning what the information object is about and why it is relevant.

Subject metadata creation is based on subject analysis. There are several different models of subject analysis (e.g., Beghtol, 1986; Hjørland, 1998; Langridge, 1989; Šauperl, 2002; Wilson, 1968). These models suggest examining, in addition to an information object's content, intentions of its creators, their viewpoints and biases (if any), and to account for the intellectual and educational level of the intended group of users and possible uses of information in the subject headings applied. Wilson (1968) stated that most works are multifaceted and cover a number of subjects and that it is often impossible to determine "the subject" of a work as a single choice from multiple possible subject descriptions. Similarly, Hjørland (1992; 1997) suggested that subjects of a document are the informative or epistemological intellectual potentials of that document that can change over time and differ between domains. Hjørland's idea points to the need for periodical changes to subject terms in metadata records.

2.2 Subject Access Tools in Library Community: Controlled Vocabularies and MARC Metadata Fields

Subject metadata is an important part of library metadata that contributes greatly to the findability of information objects and that powers the subject search. Library of Congress Subject Headings (LCSH) controlled vocabulary of subject terms has traditionally been used for describing aboutness of information objects in the library community. While a number of other subject controlled vocabularies exist and are used by memory institutions (e.g, AAT, BISAC, MESH etc.), LCSH is by far the largest and most used. According to Frank and Hoshy (2007), it had over 300000 subject authority records as of February 2007 and 6000-8000 new records were added annually. The latest (41st) edition of LCSH published in April of 2020 includes 348246 subject authority records and, according to the introduction to this edition, approximately 4000 new records are currently added to LCSH on an annual basis (https://www.loc.gov/aba/publications/FreeLCSH/LCSH42%20Main%20intro.pdf). This prevailing controlled vocabulary has been translated into various languages and its processes adapted as a model for developing subject headings systems by many countries.

However, despite all its advantages, including richness of subject representation, LCSH has inherent problems with its structure, explained in part by its origin as a controlled vocabulary developed gradually in response to "literary warrant" (Barite, 2018). In the detailed guidelines for assigning subject headings included in the *Rules for a Dictionary Catalog,* Cutter (1904) discussed many of the issues with the structure of LCSH that were found to complicate subject access in the century following: for example, inversion of a phrase in compound subject headings, specificity of subject terms, treatment of synonyms by LCSH, and formulation of geographic headings (p. 66-80). Over the years, studies have often demonstrated that

imperfections of LCSH that result in user confusion and dissatisfaction with subject searching in library catalogs and other bibliographic databases. The mass automation of library catalogs in 1980s brought in new possibilities (e.g., remote access; new access points such as keyword, ISBN etc.; ability to combine multiple search terms in one search query; proximity searching, truncation, etc.) that improved discoverability of information objects in online catalogs compared to traditional card catalogs. However, early online catalogs lacked some of functionality found in card catalogs (e.g., non-Latin scripts support), and transferred some of those limitations of the card catalog into the machine-readable environment. Larson (1991, p.185) summarized major LCSH problems that negatively affected subject search performance and resulted in subject search failures as of the late 1980s-early 1990s:

- *Specificity*. LCSH subject headings were found to be too broad overall, but in some cases, they were too specific for the user's needs.
- *Exhaustivity*. Most works were treated as single-topic work which did not accurately reflect the reality of subject coverage in the published documents and an average of only 1.4 subject headings were included in a bibliographic record which was considered inadequate for subject representation of a monograph (Larson, 1991, p.185). This level of subject representation did not follow the Library of Congress Subject Heading Manual (SHM) guidelines which encouraged providing up to five subject headings for a work, one heading for each 20% of the work's content (Library of Congress, 2008 as cited in Hjørland, 2018).
- *Inconsistency* in the structure of headings (e.g., the use of both inverted and direct phrase forms) and in the practice of adding subdivisions to the main heading (e.g., position of geographic subdivision and topical subdivision in relation to the main heading)

- *Problems with syndetic structure*: incomplete and inconsistent cross-references showing hierarchical structure of LCSH; the use of synonymous terms in different headings, etc.
- *Bias and lack of currency* (use of outdated, racist, sexist, generally disagreeable terms). Although LCSH subject headings are often changed to remove bias and to reflect the contemporary use of the terms in publications, the no-longer-valid former subject headings remain in many records that were created before the change. (Larson, 1991, p.185)

In part due to the tremendous size of this controlled vocabulary, LCSH was updated too slowly to meet evolving requirements so library cataloging activists like "radical librarian" Sanford Berman did not want to wait for LC to make necessary changes; they started to use their own improved models of LCSH, as well as improvements to DDC. Major changes in cataloging practices advocated—and implemented at Hennepin County Public Library—by Sanford Berman for over three decades included creation and use of new subject headings based on new terminology appearing in the media and on users' subject search requests. These changes were incorporated into LCSH with a significant delay; some were added to SAF only decades later (e.g., Berman & Gross, 2017).

S. R. Ranganathan's student Pauline Cochrane was among the active proponents of updating LCSH to facilitate catalog searches. In 1982, she led a project to establish a procedure for submitting suggestions for new LCSH cross-references by other libraries to improve its retrieval functions (Graham, 2004). This project gradually evolved into the Subject Authority Cooperative Project (SACO) for cooperative maintenance of the LCSH subject authority file coordinated by the Library of Congress. Cochrane (1986) identified three major features of LCSH that needed improvement: scope notes, structure of relationships between headings in

the list, and links from the LCSH subject headings to the Library of Congress classification (LCC) class numbers and notations, as well as to terms in other controlled vocabularies. Almost 15 years later, soon after the launch of the Library of Congress Subject Authority File (SAF) database in 1999, Cochrane (2000) examined the progress made since the mid-1980s in improving these three major features and found that 30% of LCSH headings in SAF had links to either LCC or Dewey Decimal Classification (DDC) class numbers. She concluded that as a result, LCSH became much more useful in online catalogs that made links to LCC and other controlled vocabularies (e.g., ERIC, MeSH) more visible and useful. However, Cochrane's study revealed that little progress had been made with adding scope notes to subject authority records in the 22nd edition of LCSH (1999). For example, she found that only 2% of subject headings had scope notes.

The so-called pre-coordinated structure of subject headings (i.e., LCSH subject strings consisting of the main subject term and subdivisions appended to it in a certain order) has often been named as a factor that complicates users' experiences in subject searching of library catalogs and other databases (e.g., Taube, 1953; Farradine, 1970; Weinberg, 1995).

The problems with LCSH and the prevalence of a simple keyword search resulting from user confusion with subject searches observed by many studies led some experts in the field, including those associated with the Library of Congress, to suggest not assigning subject headings from controlled vocabularies in order to save time and money (e.g., Calhoun, 2006; Schniderman, 2006). However, research demonstrates that despite the imperfections, LCSH remains highly valuable and the user experience would be significantly degraded without it. For example, a transaction log study by Gross and Taylor (2005) revealed that 36% of user

keywords searches in a library online catalog would not retrieve records representing English-language documents, and 80%-100% relevant foreign-language publications would be impossible to retrieve if LCSH subject terms were not present in the bibliographic records. Ten years later, in a replication of that study, Gross, Taylor and Joudrey (2015) discovered that although addition of summary notes and tables of contents to the library catalog records resulted in reduced proportion of user search queries that did not retrieve results, the absence of LCSH subject headings in those records led to an average of 27% of user searches not being matched by metadata records. The importance of LCSH was demonstrated not only for information retrieval in library catalogs but also in the full-text environment. For example, in article databases and digital libraries Garret's (2007) study of user search experience in full-text databases of historical materials, Zavalina's (2007) transaction log analysis that observed similar to Gross and Taylor's (2005) results for LCSH success matching user keyword search terms in the digital aggregation of cultural heritage content.

Another reason why an extensive controlled vocabulary like LCSH retains its value is the subject access demands of large-scale collections and databases. Bates (2002) warned against ignoring size-sensitivity of information retrieval databases and claimed that with the rapid expansion of databases, small-scale subject vocabularies and classification schemes fail, and that the larger the collection is (or is projected to be in future) a more sophisticated subject scheme is required to facilitate subject access to it. From this point of view, as the most extensive controlled vocabulary, LCSH will continue to hold promise for describing large (and especially online) collections.

By the year 2000, Chan and Hodges (2000, p. 232) found that the need for providing a post-coordinate faceted approach to subject metadata became particularly important for several reasons:

1. Relative ease of display and use of an online thesaurus based on faceted principles
2. Compatibility in structure and syntax of a post-coordinate subject vocabulary with most other controlled vocabularies
3. Easier mapping of single terms (as opposed to strings) to terms from other controlled vocabularies (both thesauri or lists of subject terms and classification systems), and to equivalent terms in other languages
4. Interoperability between MARC 21 and other metadata standards

To address this need, the Library of Congress partnered with OCLC to develop and apply a method that allowed for more efficient use of the rich data in the subject headings of the millions of records in OCLC's WorldCat and the Library of Congress catalog. This initiative involved the creation of the Faceted Application of Subject Terminology (FAST) headings by parsing the existing LCSH subject strings into separate facets and adding these FAST terms to the existing MARC 21 bibliographic records alongside the LCSH strings. As part of the project, FAST tools to be used by catalogers have been developed: assignFAST, FAST Converter, FAST LinkedData, importFAST, mapFAST, and searchFAST (http://fast.oclc.org/searchfast/). After a smaller-scale pilot in 2013 OCLC started mass-scale application of FAST headings by automatically augmenting large numbers of MARC 21 records in WorldCat with English language of cataloging (https://www.oclc.org/bibformats/en/0xx/040.html). OCLC Research team's user study into the effect of such metadata record augmentations by adding terms from the FAST-controlled vocabulary demonstrates improvement of subject access (Mixter & Childress, 2013). FAST controlled vocabulary is currently used in 1.8 million records (OCLC

Research, 2020). It is important to note here that FAST is a derivative controlled vocabulary that relies on continued development and maintenance of LCSH controlled vocabulary. It cannot and does not aim to provide an alternative to LCSH; rather it provides an added level of functionality, an augmentation to LCSH. FAST makes LCSH easier and more flexible in application, and the requirement to include the MARC 21 subfield $0 with a unique identifier of the FAST authority record for each FAST heading included in the bibliographic record provides important steps towards making subject metadata more usable in Linked Data environment.

In addition to the controlled vocabularies of general applicability such as LCSH and FAST, several other subject controlled vocabularies provide the lists of subject headings that are used for subject representation in more specific contexts and for certain kinds of information objects. For example, the Medical Subject Headings (https://www.nlm.nih.gov/mesh/meshhome.html) (MeSH) vocabulary, developed and maintained by the National Library of Medicine is used to represent the works originating in the biomedical knowledge domain in library catalogs and article databases, including the PubMed portal. Another example of such specialized controlled vocabulary of subject headings that is widely used is the Children's Subject Headings (https://www.loc.gov/aba/cyac/childsubjhead.html) for representing aboutness of works for children and young adults. The Agricultural Thesaurus developed and maintained by the National Agricultural Library (https://agclass.nal.usda.gov/) is another major subject heading list used for representing subject matter of the works in the agricultural knowledge domain. Also, BISAC subject headings, developed by Book Industry Study Group (https://bisg.org/page/BISACFaQ) are often used in library catalog records. Last, but not least,

there are subject headings lists developed and maintained outside of the United States (e.g., a set of four bilingual thesauri Répertoire de vedettes-matière (RVM) (https://rvmweb.bibl.ulaval.ca/) maintained by the University of Laval in Quebec (Canada), and Canadian subject headings maintained by Library and Archives Canada (http://www.bac-lac.gc.ca/eng/services/canadian-subject-headings/Pages/canadian-subject-headings.aspx)).

In addition to subject controlled vocabularies that provide means for verbal subject representation in library metadata (e.g., FAST, LCSH, MeSH, RVM, etc.), a number of subject controlled vocabularies exist that provide non-verbal representation through codes and classification numbers. The largest and the most influential classification systems include the Library of Congress Classification (LCC), an alphanumeric classification scheme developed based on the "literary warrant" and maintained by the United States Library of Congress since the late 19th century; the Dewey Decimal Classification (DDC) numeric classification system that originated in the United States at the turn of the 20th century and received worldwide adoption; the Universal Decimal Classification (UDC), a synthetic faceted classification scheme that was developed by the International Institute of Documentation in Europe based on DDC in the early 20th century and is also widely used worldwide. Library and Archives Canada develops and maintains its own classification systems. Beyond these universal classification systems representing the entirety of human knowledge, many other classification systems exist with the focus on specific knowledge domains. For example, the US National Library of Medicine and the US National Agricultural Library have their classification systems, as do the US Department of Defense, US Government Printing Office and Government of Canada.

In his meta-analysis reviewing the reasons of subject search failures in online catalogs,

Larson (1991) summarized proposals made by numerous researchers detailing necessary improvements to online catalogs. In relation to subject headings, suggestions included assigning more LCSH subject headings per record, supplementing LCSH terms in the records with terms from specialized thesauri (e.g., MeSH), exploiting a machine-readable version of LCSH to provide expanded lead-in vocabulary for the records. In relation to classification, recommendations included providing fuller (more specific) class notations in records, assigning additional class numbers to represent multiple facets of a work, adding terms derived from classification schedules and indexes to record based on its assigned class, and using special indexes such as classification clusters in the online catalogs.

The MARC bibliographic standard, a data encoding standard that was developed in the United States, and, since 1973 serves as the international standard for dissemination of bibliographic data, provides creators of library metadata records with the tools to include terms and codes from these various subject controlled vocabularies in specifically designated MARC 21 fields and subfields or with the help of assigned field indicators. MARC was developed using the old techniques of data management of the 1960s and as such is not aligned with modern programming approaches (Library of Congress, 2008). MARC has been criticized for being designed as a display standard and not storage and retrieval standard, and for not fully supporting machine-readability (e.g., Tennant, 2002; Thomale, 2010). However, the entire MARC 21 family of standards, which includes bibliographic standard, is continually updated in response to the needs brought to life by what Thornburg and Oskins (2007) call “environmental changes” that metadata records need to keep up with, such as changes in data content standards (e.g., transition from AACR to RDA), in controlled vocabularies, etc. New MARC21

bibliographic fields and subfields have been added, while existing fields and subfields are being redefined as needed (e.g., https://www.loc.gov/marc/bibliographic/bd6xx.html). Since the RDA testing and early adoption stage in 2009-2011, new updates to the MARC21 standard were released at least twice a year (the most recent is update No. 30 released in May of 2020). For example, MARC21 bibliographic standard currently includes 16 standard fields and a group of local variable fields in the 6XX block intended for verbal subject representation using controlled vocabularies (LCSH, FAST, and others), as well as free-text keywords:

- 600 - Subject Added Entry - Personal Name
- 610 - Subject Added Entry - Corporate Name
- 611 - Subject Added Entry - Meeting Name
- 630 - Subject Added Entry - Uniform Title
- 647 - Subject Added Entry - Named Event
- 648 - Subject Added Entry - Chronological Term
- 650 - Subject Added Entry - Topical Term
- 651 - Subject Added Entry - Geographic Name
- 653 - Index Term - Uncontrolled
- 654 - Subject Added Entry - Faceted Topical Terms
- 655 - Index Term - Genre/Form
- 656 - Index Term - Occupation
- 657 - Index Term - Function
- 658 - Index Term - Curriculum Objective
- 662 - Subject Added Entry - Hierarchical Place Name
- 688 - Subject Added Entry - Type of Entity Unspecified

- 69X - Local Subject Access Fields (https://www.loc.gov/marc/bibliographic/bd6xx.html)

Three of these bibliographic fields were added to MARC 21 in response to the development of RDA:

- 647 - Subject Added Entry-Named Event (added in 2016)
- 662 - Subject Added Entry - Hierarchical Place Name (added in 2005)
- 688 - Subject Added Entry - Type of Entity Unspecified (added in 2019).

The MARC 21 standard in its latest version also includes a number of additional fields -- 14 standard fields and a group of local variable fields in the 01X-09X range -- that are intended for subject representation using classification codes from various classification schemes, call numbers based on them, and codes from other subject controlled vocabularies (e.g., the MARC geographic area codes):

- 043 - Geographic Area Code
- 045 - Time Period of Content
- 050 - Library of Congress Call Number
- 052 - Geographic Classification
- 055 - Classification Numbers Assigned in Canada
- 060 - National Library of Medicine Call Number
- 070 - National Agricultural Library Call Number
- 072 - Subject Category Code
- 080 - Universal Decimal Classification Number
- 082 - Dewey Decimal Classification Number
- 083 - Additional Dewey Decimal Classification Number
- 084 - Other Classification Number

- 085 - Synthesized Classification Number Components
- 086 - Government Document Classification Number
- 09X - Local Call Numbers (https://www.loc.gov/marc/bibliographic/bd01x09x.html)

Two of these bibliographic fields were added to MARC 21 in response to the development of RDA:

- 083 - Additional Dewey Decimal Classification number (added in 2008)
- 085 - Synthesized classification number components (added in 2008).

The 09X group of Local Call Numbers fields includes the followings specific fields defined by OCLC:

- 090 - Locally Assigned LC-type Call Number
- 092 - Locally Assigned Dewey Call Number
- 096 - Locally Assigned LM-type Call Number
- 098 – Other Classification Schemes
- 099 Local free-text call number (https://www.oclc.org/bibformats/en/0xx.html).

In addition, MARC 21 Bibliographic Format standard includes the 522 Geographic Coverage Note field for representing geographical aboutness of an information object using free-text description. Thus, the total number of subject information bearing MARC 21 fields is 35.

Also, a number of new subfields were added to existing MARC21 fields or redefined to support Linked Data functionality in both subject metadata fields and other metadata fields:

- $0 - Authority record control number or standard number
- $1 - Real World Object URI
- $2 - Source of heading or term
- $4 – Relationship (e.g.,, https://www.loc.gov/marc/bibliographic/bd650.html)

"To promote the creation of unique original cataloging according to a mutually agreed upon standard" (Thomas, 1996, p. 499), the Library of Congress led the creation of the Program for Cooperative Cataloging (PCC) consortium initiative in 1995. The PCC collaborative project developed standards for levels of description in MARC 21 bibliographic records, including BIBCO (the Bibliographic Component of the PCC). BIBCO full-level record guidelines require provision of subject access points. PCC developed first the BIBCO Standard Record (BSR) Metadata Application Profiles (BSR MAPs) for AACR-based MARC21 bibliographic records, and then later for RDA-based MARC21 bibliographic records, which essentially makes the full-level record the minimum standard of description (Library of Congress, 2020a).

2.3 User Subject Knowledge in Subject Access

Beyond data content standards, data values standards, and the data encoding and transmission standards that determine the content and functionality of bibliographic records, including subject metadata fields, another important component of subject access is the user interaction with library catalogs and other databases through searching or browsing (e.g., Cochrane, 1979). A large body of research in information science deals with the subject knowledge (often referred to as domain knowledge) and its role in the effectiveness of access to information. In addition, domain knowledge and background knowledge, two aspects of subject-related knowledge that affect user's success in searching, have been studied and distinguished by Zhang, Liu, and Cole in 2013. This section presents some of the important relevant findings from these studies.

A large-scale catalog use study conducted in the 1960s (Tagliacozzo & Kochen, 1970) revealed that graduate students and faculty using the domain-specific Medical Library

conducted subject search substantially more often than the students using the general Undergraduate Library. Because these results were different from the findings of the earlier studies which had observed graduate students and faculty preference for known-item search (cf., Jackson, 1958), Tagliacozzo and Kochen (1970) suggested that search type selection correlates with the level of knowledge of subject headings in one's field. Bates (1972) tested the effects of familiarity with the subject area on the success of subject searching and concluded that the library catalog was not designed to take advantage of subject expert's knowledge).

Borgman (1986; 1996) formulated a knowledge model which represents the information search as a complex task. This task requires expression of information needs that are often ambiguous with precise terms and relationships which should also match the structure of the information system being searched. "Conceptual knowledge" (including relations between different topics within domain) is the first of the Borgman's three layers of knowledge needed to perform online library catalog searching. Across different types of information retrieval systems, the majority of user search problems occurs at the conceptual layer of knowledge.

Research demonstrates important differences in the domain knowledge and information seeking behavior (including subject searching) of novices and domain-experts. As summarized by LaFrance (1989), expert searchers have greater episodic memory than novices, are schema-driven rather than data-driven, and focus on goals rather than effects. Expert searchers' knowledge is more functional, more complex, and is arranged differently from that of novices, and "sometimes behave like robots" (pp.7-9). Studies of undergraduate students (Allen, 1991) and elementary school children (Hirsh, 1996) observed that information seeking

behavior (i.e., selection of search strategy and tactics) and the outcomes of the subject searching depend to a large extent on a searcher's level of knowledge of both a specific search topic and the broader subject domain. Users with higher domain knowledge were found to use a wider variety of search options and search expressions and to be more successful in finding the records. Researchers also noticed the influence of domain knowledge on how long the searchers prepared for searching and monitored the searches, as well as on the frequency of combining search terms (Hsieh-Yee, 1993).

Palmer's (1996) examination of interdisciplinary researchers gathering and disseminating information outside of their primary knowledge domain and learning of new subject areas revealed that they often have to rely on intermediaries to help collect and interpret documents from unfamiliar subject areas. Connaway, Johnson and Searing (1997) found that university faculty and graduate students who participated in the focus group study reported that Library of Congress subject headings were often too broad to pinpoint their specialized research interests and that they used controlled-vocabulary-based subject search only when working outside of their knowledge domains.

Comparative analysis of subject searches conducted by two kinds of experts -- domain experts and search experts (e.g., information professionals) -- revealed that domain experts focused on the answers to search questions, and had clear expectations for both answer and context it would appear in, and search experts tended to focus on the problem statement and query formulation because their goal was to find information for the end user (Marchionini et al., 1993). Similarly, library science students were found to use more of the self-constructed

terms in the subject search within their native knowledge domain and to use thesaurus words and synonyms for search in other subject domains (Hsieh-Yee, 1993).

Research suggests that the choice of subject search terms and search tactics change over time, with increased domain knowledge. For example, Pennanen, Serola, and Vakkari's (2003) longitudinal observations of psychology students searching PSYCHINFO database in the process of developing a research proposal revealed that as students acquired more domain knowledge on their research topic, they started to use wider and more specific vocabulary in their subject search, although their use of search operators remained relatively constant. Similarly, Wildemuth (2003) observed that medical students changed their search tactics over time. When the domain knowledge was low, a high number of searches per session was observed due to the inability to initially choose appropriate terms; more domain-knowledgeable students added more concepts in their subject searches but made fewer changes to their searches. Engineering and science students were also found to conduct more searches and to formulate longer search queries with increase in the level of domain knowledge, although their search effectiveness did not necessarily increase with a higher level of domain knowledge (Zhang, Anghelescu, and Yuan, 2005). Hembrooke et al. (2005) observed that domain experts conducted more complex searches, used more unique subject terms, and employed more effective strategies of elaboration. Studies that compared domain-expert understanding of the subject matter of the document and subject descriptions made by information specialists (e.g., Boserup & Krarup 1982) observed that because domain-experts tend to evaluate documents in relation to their scientific value they make "the most precise and useful" judgments about document's subjects.

2.4 Metadata, Big Data, and Linked Data

Electronic data today plays an important role in human society where information and communication technologies have been rapidly evolving for the last 40 years. Related evolution resulted in information systems adopting a variety of Web technologies of information organization that critically rely on data and information quality and level of representation, because this aspect seriously impacts the efficiency and effectiveness of information systems retrieval. In terms of data classification, there are many different types of data that have been involved in the process of generating enormous amounts of information and that have been used in modern information systems. Data can be different. If one looks at data from the perspective of its structure and use, data can be structured, unstructured, semi-structured, spatiotemporal, time-stamped, open, linked, social, operational or big data, and so on. Based on its localization, data can be seen as ‘clear’ or ‘dark’ (Varma, 2019).

There is a sizable layer of technical data and information that describes other data. This type of data is commonly known as metadata. Metadata consists of two words: “meta-”, which, according to Merriam-Webster dictionary (2019), means "transcending" or going beyond the limits of the concept of another word “data”. There are many different types of metadata, including bibliographic metadata that is widely used in bibliographic information systems in the form of database records that represent millions of information resources. To remain efficient, metadata standards and schemas evolve and change in order to improve metadata level of completeness and quality; and thus, improve information systems’ retrieval abilities to satisfy users’ information needs.

Historically, the field of library and information science emerged from two fields: field of

library science, which evolved from schools of librarianship and documentation science, and from interdisciplinary field of information science (Borko, 1968). Inheriting practices from all emerged disciplines, the field is primarily concerned with storage, retrieval, collection, organization, management, preservation, description and use of recordable information in the context of interaction between people and information or information retrieval systems (Borko, 1968; Saracevic, 2009).

One of the key concepts of information retrieval and information organization is the concept of information representation which refers to the description of both the content and the carriers of information: information objects or recorded information. The predominant practice of creating representations of information objects in a library collection is called cataloging. There are several activities related to cataloging including descriptive cataloging, subject cataloging, and classification. These activities result in a bibliographic or catalog record that represents a particular information object. Traditionally, these representations are manually created according to a set of community standards and agreements (e.g., descriptive cataloging rules as found in the Resource Description and Access (RDA) and its predecessor, the Anglo-American Cataloguing Rules), and the resulting records are considered trustworthy by the users. The library catalog holds these individual records. The process of manual creation of all kinds of bibliographic metadata (information representations) is labor-intensive; and relying on this method alone is becoming notably insufficient.

In contrast to the retrieval systems that are based on partial representations of information objects (descriptive metadata), technologies of full-text indexing and keyword search became to some extent a solution to increase the effectiveness of retrieval systems. In

their databases, systems of full-text retrieval store collections of all words (except stop words), predominantly stemmed, from the original texts. In a full-text search, a retrieval system examines all words in each stored document and matches it with a search query specified by the user. Full-text indexing and search techniques became common in online bibliographic databases in the beginning of 1990s. For example, in library databases, the natural language, or keyword, search option, which is an alternative to traditional library's subject search based on controlled vocabulary terms assigned to represent the content of an information object by metadata creators, was added and gained a high level of popularity among users. Research shows that keyword searching as an implementation of full-text indexing has some deficiencies for information retrieval, for example in representing foreign language materials (Gross & Taylor, 2005; Gross, Taylor, & Joudrey, 2015), and non-textual information objects, etc.

The deficiencies of searching based on full-text indexing have been addressed either by providing users with tools that enable them to express their search questions more specifically, or by improving querying tools and search algorithms that can help to increase retrieval precision (e.g. Boolean logic, regular expressions, proximity search, concordance search, etc.). For non-textual materials such as images and films some alternatives to full-text indexing were developed, for example, Anderson and O'Connor's (2009) use of Bellour's structural and functional analysis.

While in the past published information was traditionally available mostly in textual forms (e.g., printed periodicals or monographs), today it is presented in a myriad of less tangible, mostly electronic formats and forms. This demands comprehensive technological approaches in information organization and information representation for effective

information retrieval. Information representation as a technique of information organization ideally should cover as much semantic information as needed to access and retrieve the information stored in a system by specific communities of users. The demand for better information representations was especially obvious during the mid-20th century information explosion (e.g. Vannevar Bush's "memex" memory machine) and once again becomes crucial in the new information explosion related to emergence of ubiquitous Big Data of the 21st century.

According to Park and Brenza (2015), rapid increase of digital repositories and explosion of Big Data leads to development of a variety of tools and technologies (e.g. automated indexing, meta-tag harvesting, content extraction, text and data mining technologies, social tagging, extrinsic data auto-generation) allowing for automatic and semi-automatic generation of information representations. Adoption of these technologies is crucial for libraries, because the ability to provide access to information resources to the libraries' communities remains a main concern. However, researchers have come to the conclusion that semi-automatic indexing tools existing as of 2015 can only solve experimental problems; they are not yet developed sufficiently for full-scale implementation by the library community in a meaningful way (e.g., Park & Brenza, 2015).

Current technological advances developed under the influence of ideas that are commonly known as principles of Web 2.0, allow users not only to create tons of intellectual content, but also to generate massive volumes of descriptive metadata. This calls for expansion of the focus of the field to include the vast amounts of information released not only by official publishers, but also by the public (Sugimoto, Ding & Thelwall, 2012). Rapid growth of published

information leads to an information explosion, and which then requires new technological advances to process such Big Data.

As a phenomenon, Big Data has generated strong interest among a wide range of academic, business, and government organizations; and has spurred extensive discussions between enthusiasts of Big Data and its skeptics. These debates mostly unfolded within specific academic fields, such as information science (Ekbia et al., 2014). One of the most notable characteristics of Information Science is the user-oriented approach or perspective from which Information Science researchers look at the information itself (Fidel, 2000). The focus on interaction between human and information brings into the field interdisciplinarity and allows for the absorption of knowledge from multiple domains (Sugimoto, Ding & Thelwall, 2012). As stated by Borgman (2007; 2015) because Big Data makes new questions possible and thinkable due to its scale it has a potential of serving as "the glue of collaborations" (p. 3), which facilitates interdisciplinarity.

Even though the wide interest and discussion of Big Data appeared relatively recently, the term Big Data is not new. The term and concept of Big Data has been in use in the field of computer science for more than 25 years and was initially introduced by John Mashey in the 1990s (Fan & Bifet, 2012). The data defined as big was generated after WWII by hard sciences, government organizations, and the military-industrial complex as a result of the information explosion. This large-scale data was mainly produced by "big science" (Weinberg, 1961; Price, 1963). The online Oxford English Dictionary (2020) defines Big Data as "data of a very large size, typically to the extent that its manipulation and management present significant logistical challenges; (also) the branch of computing involving such data"

(https://www.oed.com/view/Entry/18833#eid301162178). The term Big Data is also used by some researchers to identify the research that was made possible by using data at unprecedented scale (Borgman, 2015).

Several important attributes of Big Data have been proposed. In 2001, Douglas Laney, Gartner analyst and VP, defined Big Data in a three-dimensional model, consisting of three attributes – three Vs: Volume, amounts of data generated and transferred; Variety, range of data types and sources being generated; and Velocity, speed of data transferring. After almost eleven years, De Mauro and colleagues refined this model by specifying these three Vs as high volume, high variety, and high velocity and emphasizing that specific technology and analytical methods are required for Big Data's transformation into Value -- data usefulness as a business asset (De Mauro, Greco, & Grimaldi, 2016). Recently, in addition to volume, variety, velocity and value, another important attribute -- veracity -- has emerged in discussions of Big Data. Veracity refers to data quality, accuracy and trustworthiness (Ekbia et al., 2014). Big Data attributes Veracity and Value were added after the emergence of social media and popularization of Web 2.0, when a substantial share of generated content originated not only by science, little or big, but also by consumers of information and products. Thus, it is legitimate to assume that the Five-Vs-definition developed under the influence of implementations of Web 2.0 principles and the explosion of e-commerce.

There are several dimensions in which Big Data can be viewed. The Big Data Working Group as part of The Cloud Security Alliance (CSA) describes extensive infrastructure that includes the following six categories (2014): Data, Compute infrastructure, Storage infrastructure, Analytics, Visualization, and Security and privacy. From the perspective of

information science, Ekbia et al. (2014) conceptualize Big Data and provide a comprehensive critical review. The authors use four viewpoints or perspectives. One perspective emphasizes such data physical characteristics as “size, speed and structure” and is called product oriented. Another perspective focuses on the uniqueness of processes that are involved in Big Data and is called process oriented. The third perspective is cognition-oriented and is concerned with cognitive challenges associated with limitations of human beings to mentally process Big Data. The final perspective, described by the authors, is social movement-oriented and is about various possible motivational changes made by Big Data (p.1527), such as information cascades in social media networks. Information cascades are the situations where a series of individuals make their decisions based on the observations of others’ decisions while ignoring their personal knowledge and information (Anderson & Holt, 1997).

Access to recorded information has traditionally been provided through information representations: those found in library catalogs, archival guides, digital library metadata, search indexes generated by Web crawling software. However, due to the scope of Big Data these approaches alone are not sufficient to provide access. IBM (2014) presented Big Data not only as a group of new technological solutions, but also as a shift of paradigms in traditional data mining and analytics. In contrast to traditional approach, the approach of Big Data analytics allows for mining meaningful information simultaneously from all available large amounts of messy data in motion (Manby, 2014). A key component of such an implementation expressed by imminent computer scientist Grace Hoper -- “In pioneer days they used oxen for heavy pulling, and when one ox couldn’t budge a log, they didn’t try to grow a larger ox. We shouldn’t be trying for bigger computers, but for more systems of computers'' (Schieber, 1987, p. 9). This

idea is a foundation for a new approach of using distributed systems of multiple computers – clusters – to manipulate with and process Big Data. The Hadoop family of technologies which makes processing large amounts of data more accessible.

According to Dempsey (2012), Hadoop and related cloud-based computing technologies offered by Amazon, Microsoft, and others that were initially built in response to Big Data requirements of web scale companies, are now becoming more used in the broader environment. Information science and practice are changing to meet the challenges of Big Data. For example, computational approaches are gaining more prominence (Dempsey, 2012). Varian (2008) states that as data is becoming ubiquitous, analysis has to be emphasized in Library and information science education through courses teaching future information scientists how to manipulate and analyze data – machine learning, data visualization, data modeling, statistics (Varian, 2008).

Information Science which has traditionally been an interdisciplinary field, operates in three different ways: as an engineering or technical discipline, as a human-cognition discipline, and as a social science discipline (Cibangu, 2010, para. 1). Linked Data applications that make use of rich controlled vocabularies such as name authority files, lists of subject headings, and thesauri developed in information science and practice has a strong potential for making Big Data more structured and therefore increasing its Veracity and Value. This is one contribution that information science can make to a collaboration between different disciplines in tackling Big Data. Natural Language Processing (NLP) and information visualization as important sub-disciplines of information science offer other valuable contributions to facing the challenges of Big Data. Last, but not least, information science's long-term research into information behavior

can offer insights into research on human interaction with Big Data.

Tim Berners-Lee, who invented the World Wide Web in 1989 and later (in 1994) became the founder and director of the Web standards organization World Wide Web Consortium (W3C), is known as the father of the Semantic Web, a project started by W3C for realizing the idea of having a web in which the machines can fully process the data, to make connections and inferences, and to deliver intelligent answers to user's questions (W3C, 2013).

Semantic Web is often viewed as Web 3.0, a third iteration in the evolution of the World Wide Web. According to Calaresu and Shiri (2015), Web 1.0 connects documents, Web 2.0 connects people with the same shared intended human audience in mind, and Web 3.0 meaningfully connects nodes of information for the base audience of computer applications.

Ontologies, thesauri and possibly taxonomies compose the systemic core of Semantic Web (Calaresu & Shiri, 2015). There are many variations of different applications of ontology as a term across the range of academic disciplines, which leads to the lack of universal definition. Guizzardi (2007) provides an explicit analysis of relationships between Ontology (with the capital "O"), as a philosophical discipline studying reality, categories of being and relationships between them, and ontologies as a structure in the domain of computer science. He formulates ontology as an explicitly documented contextual mapping of a series of interrelated elements. These elements include classes (of entities), properties (of these entities) and their relationships (Breitman, Casanova, & Truszkowski, 2007).

The term taxonomy comes from Greek words taxis "arrangement", -nomia "method" or -nomos "managing" (Taxonomy, 2017) and literally means practice of classification. Taxonomic concepts underlie ontological classification structures. Being a hierarchical system of

classification of things and concepts, taxonomy represents parent-child relations between elements and is often used in science, for example in biology (Breitman, Casanova, & Truszkowski, 2007). In information science, some of the bibliographic classifications (e.g., Dewey Decimal Classification) are also built on hierarchical or taxonomic principles. One of the apparent characteristics of this classification structure is propagation -- the notion that when a parent class is assigned with a certain attribute, all its child elements must also have the same attribute. Taxonomical classification is a substantial basis for ontological structures; however, describing mostly parent-child relations is insufficient for description of more complex relationships.

Thesaurus is another system traditionally used in information science, among other academic disciplines. A thesaurus is a structured controlled vocabulary that in comparison with taxonomic system establishes more complex relationships between elements or terms within it. It provides information about each term in the system and specifies wide spectrum of relations, such as synonymic, broader, or narrower relationships between terms. According to Calaresu and Shiri (2015), there is vague differentiation between thesaural and ontological classification structures; and thesauri "can be viewed as forms of ontologies" (p. 90).

Principal technological components of Semantic Web fit into the layered technical model, which is called Semantic Web Stack or Cake. This layered cake illustrates the hierarchy of coding technologies (languages), where each layer takes advantage of the features of the layers below to make Semantic Web possible. There are several variations of Semantic Stack evolved from the originally created technical model by Berners-Lee in 2000. For example, there were editions of the stack known as Bratt's model (2007), Crowther's model (2008), and

Nowack's model (2009). Each of these models or stacks is nothing more than simple illustration that reflects evolutionary developments in technologies. The stack is still evolving and represents realized and unrealized technologies of the Semantic Web.

Semantic Web is built on principles of Linked Data; thus, often these terms are used interchangeably. According to Berners-Lee (2009), the main principles of Linked Data are as follows:

- Use URIs as names for things
- Use HTTP URIs so that people can look up those names.
- When someone looks up a URI, provide useful information, using the standards (RDF*, SPARQL)
- Include links to other URIs. So, that they can discover more things (para. 3)

Uniform Resource Identifier (URI) and a Unicode character set are basic "addressing and identifying" (Calaresu & Shiri, 2015) technologies for Semantic Web implementation. A URI is used for uniquely identifying "things" on the Web. A Unicode character set is used to address issues with representation and manipulation with textual information written in different languages. Extensible Markup Language (XML), XML Namespaces and XML Schema consequently enable creation of structured data, provide references to other information sources inside a document, and provide predefined document structure and semantic markup of the document. All these technologies could be combined in a group of Hypertext Web technologies.

Another group includes standardized semantic technologies such as Resource Description Framework (RDF), RDF Schema (RDFS), Web Ontology Language (OWL), Semantic Protocol and RDF Query Language (SPARQL), and Semantic Web Rule Language (SWRL) / Rule

Interchange Format (RIF). RDF is an XML-built framework (e.g., Bikakis, 2016) for creating semantic statements in the form of triples. RDF triples is the foundation of the Semantic Web framework. They follow a subject–predicate–object structure, where the subject always means the resource and the predicate represents relationships between the subject and the object. In case of information organization, such triples can be represented as Information Entity – Property – Value. The underlying vocabulary for RDF is RDF Schema. RDF Schema usually stores hierarchies of RDF Classes and RDF Properties. This new data model, provided by this RDF standard, represents structured and semi-structured data formats. Web Ontology Language (OWL), which is based on description logic, accompanied by RDF, allows for providing constraints and creating more advanced structural constructs to convey meaning. SPARQL is used for querying RDF databases. The remaining group of Semantic Web components includes technologies such as Cryptography and User Interfaces, as well as an abstract layer of Trust, Logic and Proof.

There are not many conceptual models that represent the Semantic Web. In 2015, Calaresu and Shiri proposed an experimental conceptual model that is built on the findings of human information interaction research. The term “human information interaction” has been commonly used in the information science community since 1995. The human information interaction area of research focuses more on complexity of users’ relationships with information through a variety of computing devices and interfaces, rather than on technology as in “human computer interaction”, which is sometimes used interchangeably (Calaresu & Shiri, 2015). According to Marchionini (2004), human information interaction should be viewed

as a process where users are more engaged in the process of interaction with information retrieval systems.

Calaresu and Shiri's (2015) model of Semantic Web helps in understanding different levels of interactions between humans and information. The model summarizes three layers derived from analysis of previous research: a) the layer of groups of human users; b) the layer of groups of software applications; and c) the layer of groups of digital documents. The fourth layer proposed by Calaresu and Shiri introduces the concept of "archetype documents" -- the way through which information and data in a Semantic Web setting can be better understood by software applications. According to the authors, archetype documents can be imagined as idealized nodal elements (p.94) and from the technological standpoint can be perceived as RDF structures (p.96).

RDF which evolved from XML language inherits not only its simplicity, but also its power and flexibility. In contrast to traditional management of multiple datasets organized by tables, records and columns (relational databases), the standardized Semantic Web technologies provide great infrastructure for effective Big Data management. Due to these advantages, not only open source communities, but also commercial vendors, such as IBM, Oracle and others have found RDF/SPARQL technologies beneficial for Big Data implementations (DuCharme, 2013).

Since 2007 it has become possible to track the evolution of Linked Data nodes (datasets) created and made available by different institutions including government agencies, corporations, non-profit organizations, libraries, museums, and archives. There is statistical data available that represents datasets, similar to "open source" and "open access" that are

distributed on principles of “open data”, without restrictions of copyright, which means that data is freely available. According to Linked Open Data statistics (Abele et al., 2019), the number of published by contributors of Linked Open Data community datasets in RDF format is quickly increasing: only 12 datasets in May 2007, then 540 in August 2014, 1146 in 2017, 1239 datasets in 2019, and 1255 datasets in May 2020. Each Linked Open Data dataset includes hundreds of thousands and millions of RDF triples. For example, a dataset of Library of Congress Subject Headings (authority data) contains 7332816 triples (https://datahub.io/dataset/lcsh); British National Bibliography (BNB) dataset contains 4.25 million descriptions of books (bibliographic records), which are represented in 148596955 triples (https://datahub.io/dataset/bluk-bnb); the dataset of Open Archives Initiative Harvest contains 24206591 triples representing descriptive metadata records (https://datahub.io/dataset/rkb-explorer-oai).

As a phenomenon, Big Data is drawing the attention of science, industry and the public. As discussed, Big Data is commonly defined through at least three characteristics: High Volume, High Variety, and High Velocity. With that in mind, most of the Linked Open Data datasets are large enough, non-static, and represent an extensive variety of information considered to be Big Data. However, the velocity attribute may need clarification to avoid misleading future discussions.

As mentioned, Big Data exists in a large variety of forms. In the domain of library and information science, there are growing numbers of data generated by collaboration of libraries, archives and museums (LAM). LAMs exist in different organizational settings, including universities and other large institutions. In 2010, Linked Open Data in Libraries, Archives, and

Museums (LODLAM), was created as an informal network of information science enthusiasts to act as a central hub for sharing resources and collaborating and connecting with other interested professionals (http://lodlam.net). Examples of data produced by cultural heritage institutions include, but are not limited to, national bibliographies, catalogs, registries, collections of metadata for datasets, special collection portals, digitized materials, data from Web crawling, tagged resources and so forth. So, data provided by LAMs could have very high value for the researchers with different areas of interest.

Schöch (2013) and Zeng (2016) discuss this generated data in the context of digital humanities and information science. They propose to view data in two dimensions: clean, explicit and structured versus unclean, varied, and large. As opposed to "Smart Data", which refers to clean, structured or semi-structured data (Schöch, 2013), Big Data is viewed as messy, implicit, and unstructured. However, advanced technologies such as the Semantic Web and Big Data allow access to and use of relatively fast growing and large amounts of messy data to discover previously hidden access points, connections, and patterns that can reveal more valuable information and knowledge (Zeng, 2016). In particular, approximately 75% to 90% of generated information, associated with Big Data, is unstructured text – Big Text -- where traditional analytical approaches do not work (MarkLogic Webinar as cited by Zeng, 2016). This requires new tools and approaches for data mining. Kent State University, a member of LODLAM, is involved in The Semantic Analysis Method (SAM) Project (http://lod-lam.slis.kent.edu/SemanticAnalysis.html) and provides aids and resources for identifying and analyzing unstructured textual data from special collections and archives and for generating access points for Linked Data applications. There are also several tools, including COGITO

(http://www.intelligenceapi.com/) and Open Calais (http://www.opencalais.com/), that work under the umbrella of the Semantic Web and Big Data and help in Big Text mining.

The Semantic Web is commonly believed to be the future of the information field. As revealed in the literature reviewed thus far, technologies and principles that are offered by Linked Data have the potential to enable adding components of structure to Big Data, especially to its textual segment. According to Hitzler and Janowicz (2013), Big Data changes the landscape of science and introduces the new fourth paradigm of science – exploration. Link Data reduces Big Data variability and is considered an ideal testbed for researching the challenges of Big Data and experiencing the fourth paradigm of science. Eventually, Big Data will become part of the Semantic Web. The Big Data environment also poses cyberinfrastructure-related challenges for libraries and the need to develop a solid understanding of the ways to support curation, sharing and reuse of data (e.g., Salo, 2017; Xie & Fox, 2017, etc.). The "big and smart" metadata and leveraging the "metadata capital" can offer solutions to the Big Data environment challenges (e.g., Greenberg, 2017) through interdisciplinary research that involves using data science approaches to work on information science and information practice problems. Greenberg (2017) defines data science and metadata. In the context of data science, she presents the “concepts of big metadata, smart metadata, and metadata capital as part of a metadata lingua franca” (p.20).

2.5 Relevant Conceptual Models and Frameworks and Their Discussion in Literature

The conceptual models and frameworks most relevant to this study emerged and were widely adopted over the last 22 years. They include the conceptual models of functional requirements on which the current version of the library cataloging code, Resource Description

and Access (RDA), is partially based (i.e., Library Reference Model (LRM) (IFLA, 2017) and its predecessors FRBR (IFLA, 1998; 2009), FRAD (IFLA, 2008; 2013), and FRSAD (IFLA, 2010)) and the Bibliographic Framework model BIBFRAME. This section briefly presents these models, with special attention paid to how they represent subject access, and then discusses the recent publications on the implementation of BIBFRAME.

The functional requirement models that have been recently integrated into the object-oriented IFLA-LRM model, belong to the category of entity-relationship models. That modeling approach has its origins in computer science and was proposed by Chen (1976) as a generalization or extension of existing data models: network models, relational models, and entity set models. The entity-relationship modeling approach adopts a more natural logical view of data that reflects the real world which consists of entities and their relationships (Chen, 1976). The functional requirements entity models developed within library and information science and applied in the library community are based on the principles and, to a large extent, informed by the Cutter's objectives of a library catalog (Cutter, 1904) in that the main function of the library metadata is to support the tasks of the end-user of information: finding, identifying, selecting, and obtaining information, as well as navigating or exploring information.

The first model in this family of models, Functional Requirements for Bibliographic Records model or FRBR (IFLA, 1998; 2009), defined the user tasks "find", "identify", "select", and "obtain". It identified the set of ten entities normally represented in bibliographic metadata ("work", "expression", "manifestation", "item", "person", "corporate body", "concept", "object", "event", and "place"), the various attributes of these entities and relationships between them (see Figure 1.2.).

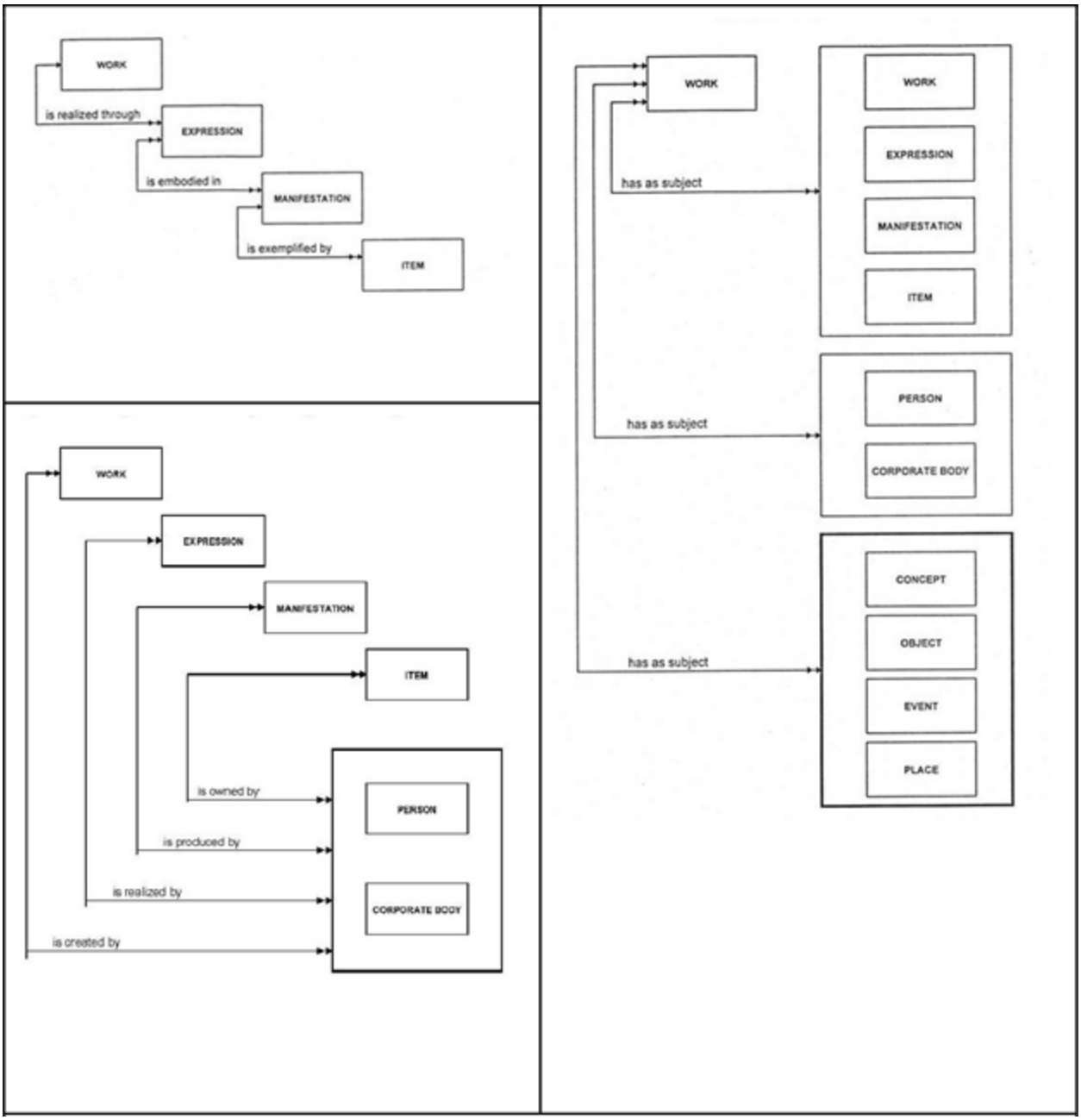


Figure 2.1: FRBR model: groups of entities [adapted from Tillett (2004)]

FRBR mapped the entities, attributes, and relationships to the user tasks they support. In the model, a one-way subject relationship “hasSubject” exists between a “work” and any of the 10 entities (including another “work”), and in particular to Group 3 entities: “concept”, “object”, “event”, and “place”. However, the FRBR model has been criticized for not providing sufficient

modeling of subject access. In particular, researchers pointed out that FRBR omitted important possible subject entities or sub entities such as time, process, situation, genre/form, concrete and abstract concepts, community/society, family, ethnic group, and class of persons (Delsey, 2005; Maxwell, 2008; Zavalina, 2012; Zeng & Salaba, 2005), and provided a very limited pool of attributes for Group 3 entities compared to the attributes defined for Group 1 and Group 2 entities (e.g., Zavalina, 2012).

The limitations of FRBR model are explained by the fact that the model only focused on bibliographic metadata. The International Federation of Library Associations and Institutions FRBR working group worked closely with the two working groups that developed the two related models of functional requirements, Functional Requirements for Authority Data FRAD (IFLA, 2008; 2013) and Functional Requirements for Subject Authority Data FRSAD (IFLA, 2010), with the focus on authority metadata: data about agents represented in name authority records and data about subjects represented in the subject authority records. The FRAD model adapted two FRBR user tasks ("find" and "identify") for authority data context and defined two additional user tasks, "contextualize" and "justify", that the catalogers creating name authority records and those applying the data in these records in the process of creation of bibliographic records needed to support. The FRAD model also combined the 10 FRBR entities, together with a "family" entity, into a group of bibliographic entities and added attributes for Group 3 and some Group 1 and Group 2 FRBR entities. Additionally, FRAD defined new entities common in the context of authority records: "name", "identifier", "controlled access point", "rules", and "agency". Finally, FRAD defined the attributes and relationships for the new entities and mapped them to user tasks. However, FRAD entirely omitted subject relationships, leaving this

task to the FRSAD model. The FRSAD model similarly included two FRBR user tasks ("find" and "identify") and defined additional user tasks relevant to subject authority data: "select", and "explore relations". It combined all of the FRBR/FRAD bibliographic entities, and any other possible entities into the overarching "thema" major entity and defined a new entity "nomen". The FRSAD model focused on defining the attributes of a "thema" and "nomen" and relationships between the three major entities— "work", "thema" and "nomen", between various "themas", and between various "nomens". Similar to FRBR and FRAD, FRSAD model also mapped entities, attributes, and relations to user tasks and identified the level of support for the user tasks as low, moderate, and or high for each one.

Collectively, the three functional requirements models FRBR, FRAD, and FRSAD provided a more or less complete functional modeling of metadata (bibliographic and authority), with some important omissions such as representing collections or aggregates. However, the complexity of the structure of the three interrelated models impeded comprehension and application in the library community, so the need for one consolidated model was quickly realized. Hence, the Library Reference Model (LRM) was created. Most importantly, by the time the development of LRM started, there was a need to create a Linked-Data-ready model, so LRM was designed to meet this need (e.g., Riva & Zumer, 2017). The development of an integrated FRBR-Library Reference Model (FRBR-LRM) started in 2010, led by the International Federation of Library Associations and Institutions (IFLA). A world-wide reviewed version of the model was published at the end of 2017.

The complex process of consolidation and alignment of integrated model design with existing relevant object-oriented models (FRBRoo, CIDOC-CRM) took about six years, and after

the worldwide review and discussion by the global cataloging community, the new model was officially approved in 2017 to replace FRBR, FRAD, and FRSAD. The LRM model defined five user tasks of “find”, “identify”, “select”, “obtain”, and “explore”. The integrated model introduced hierarchical structure, with subclasses and superclasses. For example, the new “res” entity (defined as “any entity in the universe of discourse”) was introduced as a superclass with “work”, “expression”, “manifestation”, “item”, “nomen”, “agent” (with its subclasses “collective agent” and “person”), “place”, and the newly-added entity “timespan” as subclasses. All relations in IFLA-LRM are reciprocal (i.e., two-way), so “isSubject” relationship has an inverse relationship “hasSubject”. The new model established that relations applying to a superclass also apply to all of its subclasses (i.e., so called relation chains consisting of two entities connected by a relation). For example, a “person” “isA” “agent” and “agent” “created” “work” therefore a “person” “created” “work”. The LRM adopted only a small proportion of the attributes out of a pool of hundreds collectively defined in FRBR, FRAD, or FRSAD. These 367 major attributes in LRM are considered the most significant ones. The model also defined a new attribute “manifestation statement” (LRM, 2017, p. 10).

The Bibliographic Framework model BIBFRAME, the development of which began in 2012 (Miller, Ogbuji, Mueller, & MacDougall, 2012), is an entity-relationship model, unlike IFLA-LRM, and like FRBR, FRAD, and FRSAD. It is an implementation type of a model that is also intended to function as a data encoding and transmission standard and is intended to eventually replace MARC21. From the very beginning, the support of Linked-Data functionality is the main principle of BIBFRAME development, which is being designed on the basis of the major building blocks of the Semantic Web -- Resource Description Framework (RDF) and

ontologies (Kroeger, 2013). RDF data is usually semantically annotated using RDF Schema (RDFS) and Web Ontology Languages (OWL) syntaxes. Both RDFS and OWL are World Wide Web Consortium (W3C) specifications for Linked Data.

Version 1 of BIBFRAME model was formally expressed using an RDFS ontology syntax. It included the following major components:

- Two entities Work and Instance (which are named "classes" in BIBFRAME 1.0 model)
- The relation "hasInstance" between Work and Instance
- Relations "subject" and "creator" for a Work class (called properties)
- Relations "publisher", "publishedAt", and "format" for an Instance class (called properties)
- Various smaller classes that Work and Instance are related to through the major relations (known as properties) listed above and other properties. (cf. Schreur, 2018)

The development of BIBFRAME Version 2 which replaced Version 1 in 2017 was necessitated in part by the need for better alignment with cataloging norms (Library of Congress, n.d.). The model was also redesigned to be formally expressed in a more powerful and flexible OWL ontology syntax. In BIBFRAME 2.0, compared to BIBFRAME 1.0, the new major class (Item) was added to accompany the Work and Instance, with a relation "hasItem" from Instance to Item. Two major properties were defined for Item: "heldBy" and "barcode". In BIBFRAME 2.0, Instance kept two of the three BIBFRAME 1.0 major properties -- "format" and "publisher" -- while "publishedAt" was removed. For BIBFRAME 2.0 Work entity, the major property "creator" was redefined as "agent", and the new major property "event" was added. Some smaller classes (e.g., Authority, Annotation) were eliminated in BIBFRAME 2.0, while other new classes

were introduced (e.g., Contribution). Four groups of smaller classes -- titles, identifiers, notes, and roles -- were remodeled.

The tool for creating new BIBFRAME metadata records is the BIBFRAME Editor, which is currently available for download on GitHub (https://github.com/lcnetdev/bfe/) and also usable through a demo version (http://bibframe.org/bfe/index.html). It was created and tested, then revised to align with version 2 of BIBFRAME model. In addition, the algorithms and tools for automatic conversion from MARC21 to BIBFRAME have also been developed and tested. The BIBFRAME.org website and Library of Congress website provided demo versions of these tools for conversion to BIBFRAME version 1.0. For example, the MARC-to-BIBFRAME Comparison tool http://id.loc.gov/tools/bibframe/compare-id/full-ttl allows for the comparison of the same metadata record in two encodings side-by-side. Two serializations of BIBFRAME are available through this tool: Turtle and RDF XML. Another online demo tool converts a MARC21 record by copying and pasting the entire record as MARCXML, or links to an externally stored MARCXML document, into a transformation window and then into a BIBFRAME record in various serializations such as RDFXML, N3, and JSON. However, with the transition from BIBFRAME 1.0 to BIBFRAME 2.0, some of these transformation tools are currently being redesigned and are no longer available.

The Library of Congress has been piloting application of BIBFRAME in its metadata creation for several years. Pilot 1 for BIBFRAME 1.0 ran from 2015-2016 and Pilot 2 for BIBFRAME 2.0 running since 2017. The Library of Congress has developed and published on its website the conversion specifications for converting MARC 21 bibliographic records and MARC21 title authority records to BIBFRAME 2.0 (Library of Congress, 2019a). The

programmatic tools for such conversions, for example, the Extensible Stylesheet Language Transformation (XSLT), are based on the Library of Congress conversion specifications and have been made available on GitHub https://github.com/lcnetdev/marc2bibframe2. In 2018-2019, over 56 million MARC21 bibliographic records from the Library of Congress online catalog were converted to BIBFRAME 2.0 using the Library of Congress conversion specifications. The resulting records have been made available as a BIBFRAME Database in two parts—BIBFRAME Works and BIBFRAME Instances—through the Library of Congress Linked Data Services search interface at http://id.loc.gov/resources/works.html and http://id.loc.gov/resources/instances.html respectively. As part of Pilot 1 and Pilot 2, Library of Congress developed a number of training materials and guidelines documents, including the *BIBFRAME Editor and BIBFRAME Database Manual* (Library of Congress, 2019b).

Several BIBFRAME-related studies have been published. Taniguchi (2017) reviewed and compared two models—BIBFRAME and RDA—in the context of metadata transferability. He highlighted issues with mapping of RDA elements and BIBFRAME properties and with the conversion of MARC21 to BIBFRAME bibliographic records. Taniguchi reveals that the absence of corresponding BIBFRAME properties in RDA-based bibliographic records leads to potential data loss during the process of mapping. He explains this by occurrences of many-to-one and many-to-many RDA-BIBFRAME mappings. He expresses the hope that implementation of LRM model in new RDA development would to some extent solve the issues related to interoperability.

Balster, Rendall and Shrader (2018) described the results of mapping metadata elements of MARC 21 records representing continuing resources—CONSER Standard Record

(CSR) and BIBFRAME Version 1—and further conversion of these mappings into version 2.0 within the frame of the CONSER BIBFRAME project (2015-2018). During this project, investigators were primarily focused on exploring the readiness of CSR metadata records to be converted into BIBFRAME with the added ability to explore relationships among and between other related entities. It was determined that mapping of metadata elements between these two models is possible. However, as stated by Balster, Rendall and Shrader (2018), such mapping alone is not sufficient for exploring relationships between linked records because BIBFRAME as linked-data model offers much better potential than existing representations.

The only published report to date that examined the results of the scrupulous processes of preparation of bibliographic and authority data for linked-data environment is the recent work of Boehr and Bushman (2018). The authors described creation, transformation, and addition of RDF URIs to MeSH-based records and the following assessment of the readiness of MARC21 metadata for supporting BIBFRAME's Linked-Data functionality upon such transformations. They studied the use of RDF URIs in the United States National Library of Medicine MARC 21 bibliographic and authority records and realized that simply adding URIs does not make these records real Linked Data. The authors expressed their hope that this study would become a starting point for future practice and research into how the vast valuable existing resource of available MARC 21 metadata could be efficiently used in Linked Data environment. This includes exploration of potential transformations of MARC21 metadata to comply with the requirements of the Linked Data and to enable exploring between linked information entities.

2.6 Library Subject Metadata Studies

There is an overall lack of studies focusing on the application of subject metadata in library catalog records in recent years. However, multiple studies of library metadata that examined entire metadata records in general have produced findings relevant to the topic of this study. The level of subject representation in card catalog records was examined by Hitchcock (1940). Hitchcock found that subject headings were omitted for many different types of material. The results of her study indicated that a significant number of academic libraries did not provide subject cards in card catalogs for the four broad categories of information objects. According to Hitchcock, catalogers at the time believed that the subject approach to search for these types of materials was less effective compared to search by title or author name and placed low priority at assigning subject headings for these documents:

1. Material not useful to the majority of users
2. Material decentralized from the main collection by departmental libraries and by special physical format
3. Self-cataloging material
4. Material which is represented in catalog under subject by proxy (pp.75-76)

The first category of information objects not receiving adequate subject representation in library card catalogs at the time Hitchcock conducted her study included materials in non-Roman alphabet languages, children's nonfiction and juvenile collections, periodicals, and publications of societies. The second category included newspapers, maps, films, pamphlets, manuscripts, rare books, and censored material. The third category included legislative proceedings, annual reports of institutions (including administrative reports of colleges and universities) and government agencies, conference proceedings, and academic dissertations.

The fourth category included autobiographies, literary works, works with titles worded the same as subject headings that would be used to represent these works, and other (mostly subsequent) editions of the works that were already cataloged.

The background to place these findings in context is the "crisis in cataloging" (the lack of catalogers and resources to spend on cataloging) in the 1930s-1940s that resulted in rapid growth of cataloging backlogs (Osborn, 1941). For example, the Library of Congress backlogs grew to almost 29% of the total collection by 1944 (e.g., Ercegovac, 1998). Such a situation brought up the practice of triage in library cataloging, with a focus on providing at least some access through author and title fields while giving up contents and series notes, added name entries and further limiting subject access in the effort of lowering costs of cataloging and decreasing backlogs. Pierson (1934) claimed that such minimal-level cataloging without subject headings and authorized forms of the names is "the most expensive […] for it leads to endless confusion." He concluded that "time spent on making simple, unverified entries […] is time and money thrown away." (Pierson, 1934, p.313).

This cost-cutting approach had long-term consequences as it was later incorporated in AACR2 as minimal-level cataloging in the 1970s and was carried over from card catalogs to online catalogs with an ungrounded expectation that the information retrieval power of online catalogs would compensate for these deficiencies in bibliographic records (Ercegovac, 1998). This, of course, negatively affected the findability of information objects, including discovery based on collocation, which as defined by Miksa (1983) is gathering related terms together in close proximity between each other. For example, Mann (1991) argued that minimal-level cataloging records significantly complicate the work of reference librarians, who rely on the

connections that subject access points (i.e., subject headings) establish between items in a collection. Even the databases that presumably focused on quality of their bibliographic records—such as Research Libraries Information Network (RLIN)—were found to be underrepresenting subject matter of information objects. For example, Intner (1989), in the comparative content analysis of 430 records from OCLC and RLIN, discovered the lack of subject headings or classification numbers in some of the records in both databases (p. 39). Taylor and Simpson (1986) compared Cataloging-In-Publication (CIP) bibliographic records created by the Cataloging-In-Publication Office of the Library of Congress, generally believed to be of higher quality, with non-CIP records (a total of almost 2000 records) and determined the level of errors and omissions. Taylor and Simpson's definition of consequential or "significant errors" include errors in subject headings, Dewey Decimal Classification (DDC), and Library of Congress Classification (LCC) codes (p. 385). They found that between 11.7% and 12.3% of records had mistakes or omissions in the subject headings, between 6.4% and 10.6% in DDC, and between 4.3% and 5.5% in LCC. Taylor and Simpson also found that between 2.7% and 4.1% of records were missing the geographic area code (MARC field 043).

Similarly, Snow (2012) and Schultz-Jones, Snow, Miksa, and Hasenyager (2012) reported on the results of a survey ranking MARC21 fields important for evaluation of quality of library metadata records. According to this study, 91% of all respondents put the 650 Subject Added Entry--Topical Term field on the 3rd place in the list of top ten "very important" MARC21 fields. Similarly, 85% of that survey respondents put the 651 Subject Added Entry--Geographic Name field on the 5th place; 84% put the 600 Subject Added Entry--Personal Name field on 6th place;

and 80% put 610 Subject Added Entry--Corporate Name field on 8th place in the ranking of "very important" MARC21 fields.

In the late 1990s some studies examined the application of subject headings in online catalogs. Hoffman (1998; 2001) examined the practice of facilitating subject access through creation of individual bibliographic records with more specific subject headings for each work in a multi-work item instead of assigning more general subject headings in a single record describing the whole item. Ercegovac (1998) analyzed effectiveness of minimal-level cataloging records for retrieval of cartographic materials.

The large-scale MARC Content Designation Utilization (MCDU) Project examined the extent of application not only for MARC 21 fields but also for subfields in its analysis of 56 million bibliographic records in OCLC WorldCat (the entire population of the records in this database at the time). In one of the publications resulting from the project (Eklund, Miksa, Moen, Snyder, & Polyakov, 2009), the authors compared various groups of fields and subfields observed in their dataset with requirements in MARC21 Bibliographic Format record standards: national, core, and minimal level. The researchers reported the level of application for only one of the subject fields—655—which was observed in 5.1 % of records in the sound recordings set. In another MCDU publication (Moen, Miksa, Eklund, Polyakov, & Synder, 2006), the researchers reported percentages of records that included at least one instance of various fields, including six subject fields: 043, 050, 082, 600, 650, and 651. Interestingly, field 600 (Subject – Personal Name), of all the subject-related fields, was found most often in the records. Earlier, Moen and Benardino (2003) reported on the results of preliminary analysis of approximately 400000 MARC 21 records and demonstrated that out of 184 combinations of MARC 21 subject fields

plus subfields (e.g., $650 $a, 651 $y, etc.) 122 combinations were observed in the dataset, for a total of 1.9194 million of instances. Among subject metadata fields, Moen and Benardino reported on specific subfield-level results only for the subfields of field 650: they observed over 602000 instances of use of the required subfield $a, almost 327000 instances of subfield $x, and almost 231500 instances of subfield $z. For the remaining nine subfields of the 650 field the level of application varied between one instance and 83600 instances in a dataset of 400000 records. No data on the level of application of the subfields of subject metadata fields beyond field 650 was presented in published results of the MCDU project team, however some data does exist in the unpublished technical reports.

Similar to the MCDU project, Mayernik's (2009) metadata study found that subject access fields were included in a high proportion of records. Mayernik examined the distribution of almost 30000 instances of 144 MARC21 fields in 1500 randomly selected bibliographic records in the Library of Congress online catalog with the wide range of dates of record creation. His findings reflected the Zipf distribution of MARC21 field occurrences, with a small proportion of fields (23 out of 144) appearing in nearly all the records. The MCDU project arrived at a similar conclusion—only about 5% of fields appear in 80% of the records. The MARC 21 field 650 appeared in 66% of records and exhibited the largest total number of occurrences (1817) in Mayernik's sample. Fields 050 and 082 containing LCC classification codes and DDC classification codes respectively also belonged to the top most frequently occurring fields and were found by Mayernik in 99.13% and 28.87% of the records. Another subject access field -- geographic area code 043 -- appeared in 33.4% of records in Mayernik's study sample. Overall, subject metadata fields occupied 4 out of 23 positions in the list of most frequently used fields.

Mayernik also reported that records containing the field 650 included on average 1.84 instances of it; field 651 occurred 1.38 times per record containing it, on average; for fields 600 and 655 that number constituted 1.22 and 1.55 respectively.

Smith-Yoshimura et al. (2010) examined patterns of MARC21 field usage in the entire population of 146 million records in WorldCat and implications on metadata practices for an OCLC Research group project. This study revealed that only a small subset of available MARC21 fields (22) occurred in 10% or more of WorldCat records—a finding similar to the MCDU project team's findings regarding WorldCat records and Mayernik's study findings regarding the Library of Congress records. These 22 most frequently occurring fields in OCLC research study included four subject metadata fields: 650 (46% of records), 050 (20%), 043 (19%), and 082 (14%). Some additional subject fields were observed infrequently overall but much more often in records representing certain formats of materials (e.g., field 600 was found to be used in 40% of records representing mixed materials, field 655—in 27.93% of records representing visual materials). Smith-Yoshimura et al. also separately examined the application of fields added to the MARC21 standard in the 2000s, including subject metadata fields 648 and 662. Field 648 added in 2002 was found to be used in 0.07% of records, and field 652 added in 2005 was found to be used in less than 0.1% of records. In the list of six factors that researchers proposed for practitioners to consider in their decision-making regarding creation of MARC21 records, the importance of subject access was emphasized:

> The number of full-text documents available on the Web will substantially increase over the next few years, and the need for surrogate 'descriptive metadata' will decrease. Focus instead on the authorized names, classifications, and controlled vocabularies that key word searching of full-text will not provide. (p. 13).

Foreign-language cataloging is prone to errors in various fields of bibliographic records due to limitations in the level of foreign language skills possessed by the catalogers. The inaccurate and misleading subject representation resulting from this limitation seriously impedes subject access. For example, Soglasnova (2018) discussed the problems with the accuracy of subject headings in the English-language-of-cataloging MARC21 records representing materials in different groups of the Slavic family of languages and reported on the Slavic cataloging community initiatives to help tackle these problems.

Subject metadata in library cataloging records has been evaluated in several recent studies, published after 2010, and most of these studies were conducted by researchers at the University of North Texas. Two of these studies however have analyzed MARC21 bibliographic records but only one focused on subject metadata. Zavalina, Shakeri and Kizhakkethil (2016) examined the change-over-time in the application of subject fields of a sample of 369 RDA-based MARC 21 bibliographic records that represent video recordings in DVD and BluRay format and were created by English-language-of-cataloging institutions. The same records were collected in 2013 and 2015 and comparatively analyzed using quantitative and qualitative content analysis. The findings of that study revealed a high level of change in 6XX MARC fields over a period of 2 years: mostly additions of new fields/subfields and instances of fields/subfields, but also modifications (amendments and replacements) of data values and some deletions. Overall, the observed trend was towards an increase of the average number of subject headings per record. In a related longitudinal study, Zavalina, Zavalin and Miksa (2016) quantitatively examined change in the same 369 MARC 21 bibliographic records between four points in time: 2013, 2014, 2015, and 2016. That study did not specifically focus on subject

metadata fields but has revealed changes (in some cases a year-to-year increase, but in others a decrease or fluctuation) in the number of instances of these fields in the records included in the sample. For example, the level of application of MARC 21 field 650 (topical subject access point) was found to gradually increase from 91.3% of records in the sample to 91.9%. On the other hand, the level of application of MARC 21 field 651 (geographical subject access point) slightly increased from 2013 to 2014, but then substantially decreased from 51.1% of records in 2014 to 42.7% in 2015, but then slightly increased to 45% in 2016. This study produced inconclusive results with regards to the trends in subject metadata application.

Any network can be defined as a group of connected objects that usually referred as nodes or vertices, and connections between the nodes referred as edges. The New York Public Library (NYPL) Lab research team experimented with applying social network analysis tools and techniques to visualize how the MARC 21 metadata records in the New York Public Library Catalog were related based on the subject terms in 6XX fields (Miller, 2014). The resulting network included 430000 nodes based on subject terms, and over 11 million edges that represented relationships between records based on the shared subject terms.

While not looking at MARC 21 bibliographic records, Phillips, Tarver, and Zavalina (2019) have recently conducted a study that is relevant to this literature review. They tested the application of network analysis techniques in analysis of the interconnectedness of metadata in the Dublin-Core-based metadata records in several different collections that are part of a large-scale aggregation of digital collections hosted by the University of North Texas academic library. They compared various network analysis indicators for data values in 15 different metadata fields and found that subject metadata (i.e., data values in Subject and Coverage fields) is the

most promising for improvement of this interconnectedness between metadata records, which in turn improves discoverability of information objects. The same team also examined the data values in the Subject fields in over 8 million metadata records in the Digital Public Library of America (DPLA), using a Big Data quantitative content analysis approach (Tarver et al., 2015). The authors found a wide variation in practices of assigning subject headings among institutions (i.e., hubs) contributing their metadata records to DPLA. For example, the average number of subject headings per metadata record was found to vary from 0.3 to 11, and only 12 out of the 23 DPLA hubs had no any subject headings in less than 10% of metadata records submitted by them.

2.7 Conclusion

The review of the literature indicates that discussions of subject access and subject metadata in library science literature are developed in the larger context of the user needs and behaviors, evolution of library cataloging profession, organizations and cooperative efforts, conceptual frameworks, and emergence and development of professional tools and other related technologies. This review also identifies several important gaps in the literature. There are no recent studies that examine the application of MARC content designation in bibliographic records at a more granular level beyond fields, in other words the use of subfields. Also, none of the available studies of subject metadata application looked at the data values in those fields and subfields qualitatively (beyond identifying change in the data value over time). Importantly, so far, there are no known studies that seek to investigate how a major environmental change such as the gradual (perhaps?) transition of cataloging practices and library databases from MARC21 to BIBFRAME—a standard that is intended to provide higher

functionality in helping meet user needs through metadata—is affecting subject access in library catalogs.

CHAPTER 3

METHODOLOGY

3.1 Introduction

This study was planned and executed to address the gap in the literature on subject access by examining the current state of subject metadata in existing library bibliographic records and its readiness to support the functions associated with Linked Data and Semantic Web. In this chapter, the research methodology used in this study is explained. This includes an examination of the research approach and design, the tools used to collect data for the study, and justification for approaches. A description of the study population is also provided, as well as a description of data collection, data processing, and data analysis techniques.

3.2 Research Questions and Research Approach

To help address the need and research gap identified in the review of the relevant research in Chapter 2, this study sought answers to the following research questions:

1. What extent and variety of subject representation do the library metadata records (i.e., MARC21 bibliographic records) currently provide? How are the most recent RDA and MARC21 guidelines and features intended to support functionality in Linked Data environment and BIBFRAME conversion applied in subject metadata elements in the records?

2. How does the application of existing subject metadata in the most recently created MARC21 library metadata records affect relations between these records as measured by social network analysis?

3. How does the subject representation in the newly created MARC21 bibliographic records carry over into BIBFRAME records resulting from automated conversion from MARC21? What implications does such a conversion have for interconnectedness of records based on subject metadata?

This mixed method research project relied on a combination of qualitative and

quantitative content analysis with Social Network Analysis (SNA), in examination of subject metadata in RDA-based MARC 21 metadata records. The project used a combination of Big Data analytics approach and traditional statistical sampling approach in the unobtrusive data collection. The rest of this section reviews the literature on the selection of research approaches as well as data collection and data analysis methods.

There are various data processing tools, techniques and conceptual approaches, and these constantly evolve, because human's understanding of the world is changing. Traditionally, there are three research approaches: a) quantitative, b) qualitative, and c) mixed methods. Essentially, each of the research approaches is based on a certain set of philosophical assumptions, thoughts, beliefs or ideas that remain at the back end of research (Kuhn, 1962; Guba, 1990; Slife & Williams, 1995). These philosophical beliefs and ideas that guide not just research, but all daily actions (Guba, 1990) are often called paradigms (Kuhn, 1962; Guba, 1990), epistemologies/ontologies (Crotty, 1998) or worldviews (Creswell, 2014).

A quantitative research approach is a number-based type of research that is used to test predefined constructs, concepts, and hypotheses that compose an objective theory by examining the relationships between different variables that can be measured and analyzed statistically. Quantitative analysis is primarily a deductive type of reasoning where conclusions are certain and logically derived from one or more premises (top-down logic). The data used in quantitative research is numeric or binary, which is also numeric. The data is collected through surveys, structured interviews, structured observations, and analysis of documents for numeric information. Validity and reliability of quantitative research greatly depends on instruments and measuring devices used. For example, according to Borgman (2015), studies of Big Data

using statistical methods and computational modeling achieve high reliability due to the scale of the dataset while surveys that rely on samples achieve adequate reliability only if the sample is large.

A qualitative research approach is a text-based type of research that includes focus groups, in-depth unstructured individual interviews, or document analysis for non-numerical information. To formulate theories and hypotheses, this type of research utilizes inductive type of reasoning in which all statements and premises only support the evidence of the truth of a conclusion at some extent of probability (bottom-up logic). This type of research does not involve statistical analysis. The data used in qualitative research is either completely unstructured or semi-structured. Validity and reliability of qualitative research depends mainly on skills and accuracy of the researchers. According to Borgman (2015), qualitative studies that depend on close analysis (e.g. ethnographies) provide rich descriptions of phenomena, but their results are harder to anonymize and share; these studies are concerned more with validity than with reliability.

Structure and highly systematic rules and procedures that already exist for quantitative studies are often named as advantages of quantitative research. The lack of flexibility and room for innovation are disadvantages of quantitative design. Qualitative research approach on the other hand is more flexible and allows researchers to work with self-designed frameworks, but is more time-consuming and, as mentioned above, sometimes lacks generalization or reliability. Therefore, a mixed-methods approach is more preferable (Creswell, 2014).

Since a mixed-methods approach combines quantitative and qualitative approaches, there is an assumption that it should provide a more comprehensive understanding of the

research problem. According to Newman and Benz (1998), if quantitative and qualitative approaches should be seen as representations of different sides of the continuum, but not as something firmly and directly opposite, mixed methods belong to the middle part of this continuum and combine features of both quantitative and qualitative approaches. The data collected and analyzed in mixed-methods research is also both quantitative and qualitative. Mixed-methods research requires more time and resources to collect two types of data and assumes that the researcher is familiar with both quantitative and qualitative approaches (Creswell, 2014).

According to Borgman (2015), social sciences articulate their research methods more explicitly than most other fields. Social sciences research practices strive to balance the description of human behavior at the most detailed level by respecting the rights of human subjects and communities or institutions that are studied. In addition to the distinction between quantitative and qualitative research methods, there are two other important distinctions to keep in mind when designing a study in social sciences, including information science. One that is especially important for research in social sciences is a distinction between obtrusive or unobtrusive data collection methods. In obtrusive studies the human subject is aware of being studied and needs to provide informed consent for participation. Unobtrusive studies collect data without interfering with human subject's behavior. Instead, researchers analyze the recorded, documented human activities instead. Another important distinction is between idiographic studies – those focusing on particular event or place or context – and nomothetic studies – those looking to identify some cause-effect relationships in a broader class of events/places/contexts (Babbie, 2013).

3.3 Design

Methodology literature reveals numerous frameworks for research and discusses its stages from different perspectives. Compared to “scaffolded learning”, a concept used in education inspired by Lev Vygotsky, Crotty (1998) provides his framework in conjunction with four elements: epistemology (paradigm), theoretical perspective, methodology, and methods (p.4). Grounded in epistemology, each of these elements provides the foundation for the next one. Krathwohl (2009) explains the process of a research by using the following three stages: description, explanation, prediction or generalization (p.26). Creswell (2014) provides a model of research framework that involves three major components: philosophical assumptions or paradigms, research design, and specific methods (p. 5). There are four widely discussed paradigms in the literature: positivism, constructivism, pragmatism and transformative.

Despite the fact that there is no universal solution or way to decide which method is more applicable for particular research, researchers may ask themselves how much is already known about the phenomenon. This sort of “maturity of knowledge surrounding the phenomenon” (Krathwohl, 2009) has practical value at the initial stage of study on the path “from discovery to accepted as applicable general knowledge” (pp. 25-26). In practice, with a lack of information and uncertainty about the point of interest, researchers usually begin to collect all the available information and create a vocabulary or language that is needed to describe a phenomenon or point of interest. Such description helps others to understand more deeply what the research is about. Once the point of interest is described, researchers move to the phase of investigation for explanations of the phenomenon; and come up with either

accurate predictions or determined generalizations based on discovered explanations (Krathwohl, 2009).

This study used an exploratory mixed-method design with unobtrusive data collection. It consisted of the two stages. Stage 1 involved the larger dataset of records and utilized high-level quantitative content analysis using data mining and Big Data approaches, in addition to Social Network Analysis. Stage 2 relied on the in-depth manual content analysis of a small purposive sample of metadata records from the larger dataset. Each of the two stages was intended to address—in different ways—all three research questions. However, because of the specifics of the unit of analysis (i.e., on the subject metadata in the entire dataset of records that meet the study criteria in Stage 1 and on subject representation in the subset of individual metadata records in Stage 2) and the size limitation of each of the sets of data, Stage 1 data analyses focus on addressing Research Questions 1 and 2, while Stage 2 analysis provides insight into Research Questions 1 and 3.

The content analysis research method and its two types (qualitative and quantitative) are widely used in the social sciences and information science (cf., Allen & Reser, 1990; Weare & Lin, 2000). Regardless of the type, the first step in content analysis is usually the preliminary exploration of documents or other textual objects to identify categories (e.g., language patterns that represent investigated phenomena). Identification of categories is generally followed by selecting the unit of analysis that would appropriately represent the investigated categories (this could range from a single word to the whole document). Unit of analysis is often defined prior to conducting the study, however there is often the need for so called thematic units or meaning units, the size of which may vary to more accurately represent the studied

phenomena under exploration (e.g., Henri, 1992). Next, the coding stage takes place. Coding is the process of finding and labeling categories. In qualitative content analysis, after the coding procedure is completed, researchers deduce trends or specific phenomena from the coded text. In quantitative content analysis, investigators count the number of instances of each category and apply various statistical tests to determine the weight of individual categories and relationships between categories. In comparative content analysis, these results are then compared across datasets.

Social Network Analysis (SNA) is a research method that is frequently used in information science. This method emerged from sociology, psychology and social anthropology in the first half of the 20th century. According to Case (2012), SNA was popularized among information behavior researchers by Caroline Haythornthwaite in 1996. SNA focuses on social networks – the array of people with whom a person interacts and shares resources. Social networks are believed to influence opinions, attitudes and behaviors, including information behavior. SNA measures relationships and relationship changes between actors -- knowledge-possessing entities, such as humans, groups, and organizations. In SNA, these actors are mapped or visualized with nodes and relationships with connectors among the nodes. The SNA structure is made up of node entities, such as humans, and ties, such as relationships (edges).

Social Network Analysis often relies on a combination of quantitative and qualitative research approaches. According to Prell (2012), sometimes a single approach could be used in social network analysis; for example, qualitative approach alone is used for understanding network evolution and friendship dynamics. It is also possible to proceed to collect quantitative social network data without a pilot study or without including any qualitative component if one

has a clear theoretical perspective to guide the research. However, Prell claims that high quality quantitative social network studies either include a qualitative component or are preceded by a preliminary qualitative research. For example, Uzzi's (1996) well-known ethnographic field work study characterized social ties of respondents and developed quantitative measures of network ties informed by the field work. Also, in quantitative SNA studies researchers often rely on qualitative methods, such as unstructured interviews and focus groups to collect ideas that inform surveys. Just as in other kinds of social science research, SNA qualitative data is used as a means of methodological triangulation.

3.4 Methods of Data Collection

This section describes the data collection methods used in this study. It also describes the techniques of data processing applied in preparation for data analysis.

3.4.1 Population and Sampling

WorldCat was created by the OCLC consortium of the libraries in the United States in 1971 and has evolved to become the largest global centralized database of MARC 21 bibliographic records. It currently aggregates over 479 million records representing information objects in almost 500 languages; the database of bibliographic records currently grows at a rate of 6.38% per year (https://www.oclc.org/en/worldcat/inside-worldcat.html). The WorldCat bibliographic records are submitted by thousands of institutions worldwide (including the almost 16000 libraries that are members of OCLC). Data is available for analysis and freely accessible through Z39.50 client-server communication protocol and other protocols.

The primary population for this study is defined as all of the MARC 21 bibliographic records in OCLC's WorldCat that are assumed to have the necessary components and features for obtaining functional BIBFRAME records when MARC21 records are automatically converted to BIBFRAME using the latest conversion tools based on BIBFRAME version 2 that was adopted in 2017 http://id.loc.gov/resources/works.html). To meet this criterion, MARC21 records should conform to the latest official version of RDA cataloging code, in other words, records that have been created in 2019 and/or 2020 and that self-identify as RDA records. The first stage of this study adopted the Big Data approach in analyzing the entire population specified above as opposed to random sampling of MARC21 records in WorldCat.

For Stage 2 analysis, a small sample of these records was selected for in-depth manual content analysis. The initial plan for Stage 2 was to conduct manual in-depth comparative content analysis of a subsample of 370 most recently created RDA-based MARC 21 bibliographic records and their BIBFRAME equivalents in the Library of Congress database of BIBFRAME 2.0 work records (http://id.loc.gov/resources/works.html). However, due to results of the pilot study (discussed below in section 3.4.2.1), the MARC 21 records collected and analyzed in Stage 1 of this research were all created in 2020, as opposed to 2019. As a result, there were no equivalents to these records in the Library of Congress database of BIBFRAME records which was last updated in June of 2019. For this reason, Stage 2 was redesigned to focus on in-depth manual comparative analysis of a purposefully selected subsample from the sample of 10004 records collected and analyzed in Stage 1.

The purposive subsample of records analyzed in Stage 2 included the 100 records with highest numbers of holdings as of the time of data collection, and the full level of encoding as

indicated by code I or blank in the ELvl subfield of the fixed field (MARC Leader, byte 17). These records were selected because they have the most impact on the discoverability of information objects in the library catalogs and are representative of the overall collection of bibliographic records included in online catalogs of the libraries worldwide. Last, but not least, these records were assumed to be representative of records because they were fully encoded records. Specifically, they make use of all or most of the fields and subfields that were recently added to MARC 21 Bibliographic Standard in order to meet evolving needs and increase the functionality of MARC 21 records in supporting Linked Data and meaningful automatic conversion to BIBFRAME. A record that is Fully-encoded implies high quality cataloging, especially in regards to completeness of MARC21 metadata records (https://www.oclc.org/bibformats/en/fixedfield/elvl.html). The refocusing of Stage 2 also allowed me to obtain insight into the subject representation in the records created by a variety of different institutions, as opposed to only those created by the United States Library of Congress, as would be the case in MARC-to-BIBFRAME record comparison since the database of BIBFRAME 20 records currently only includes those created by LC.

3.4.2 Data Collection

Data was collected with the help of Z39.50 client-server protocol. This protocol is developed for searching and retrieving information from remote databases through Transmission Control Protocol/Internet Protocol (TCP/IP) supporting networks. The justification for using Z39.50 is that this protocol was developed by the Library of Congress Maintenance Agency (https://www.loc.gov/z3950/agency/agency.html), a trusted source, and one that is

widely used in many library settings and it is often incorporated into integrated library systems or asset management systems.

For the benefit of using a user-friendly GUI-based application, the most current version of MARCEdit, a metadata manipulation and editing software suite, was used for the data collection management and for some of the clean-up procedures. This software was developed by Terry Reese in 1999 and after its deployment as a part of a database cleanup project at Oregon State University in 1999, MARCEdit was released for other applications among librarians for wider use in the field of library and information science (MARCEdit Development, 2013).

Z39.50 interactions with databases support different services, such as INIT, a short word for initialization, which may have slightly different meanings depending on the environment. For example, in Unix based operating systems INIT is the initial process started during booting, and it should precede the SEARCH command or service that enables a start to query databases. The querying database is supported by implementation of Reverse Polish Notations (RPN), Prefix Query Format (PQF), and BIB-1 attribute set (https://www.loc.gov/z3950/agency/defns/bib1.html). An example of simple PQF query for finding all documents by a specific attribute -- for example, documents that have the term "information" in the title field -- can be composed as follows:

@attr 1=4 information

Data collection may be limited by such parameters as client-server traffic that refers to the data transfer that takes place between OCLC—database vendor—and a machine outside the local network of OCLC collects the data. Interrupted connections and resulting runtime

errors may occur during data collection that can lead to potential data loss. Under conditions of stable connectivity, all up-to-date available records were retrieved according to search query parameters. Thus, the intention of this study was to collect and analyze the entire collection of most recently created and/or most recently modified (in 2019 and/or 2020) RDA-based MARC21 bibliographic records available at the moment of downloading. The pilot study was conducted in early 2020 and resulted in revision to the initial plan based on the limitations of Z39.50 protocol and other technical issues. The pilot study, its findings, and effect on the study design, are described in the next section

3.4.3 Pilot Study

In January 2020, metadata records were collected based on the advanced database query that combined two search criteria—a year of 2019 and a code "rda" in required if applicable subfield $e Description Convention of MARC21 field 040. The following Z39.50 query was used in OCLC WorldCat using MarcEdit SRU/Z39.50 client:

@and @attr 1=5067 rda @attr 1=1002 2019,

where

- @and is a Boolean search operator AND
- @attr 1 is an operator that defines search attribute of Z39.50
- 5067 is a search attribute that searches MARC21 subfield 040$e Description Conventions
- 1002 is a search attribute that searches MARC21 variable fields
- rda is the keyword value for searching with attribute 5067
- 2019 is a keyword value for searching with attribute 1002.

Based on preliminary testing on small samples, it was assumed that this approach would result in high recall and relatively high precision as it would retrieve all of the RDA-based MARC 21 records that were created or last updated in 2019, albeit a small proportion of the retrieved records might be created and last updated in 2018 (e.g., records created by publishers several months prior to publication date). However, close examination of the 47879 records retrieved this way as part of the pilot study revealed that neither recall nor precision were high: the resulting dataset was seriously incomplete and over 36% of the records in it were not created or last updated in 2019. Moreover, analysis of collected records demonstrated that even among records last updated in 2019 many included only a minor update (e.g., automatically updated holdings information) and did not reflect the latest RDA and MARC 21 metadata creation practices.

It was also observed in the manual evaluation of the records obtained through the pilot study data collection, that even the records last updated in 2019 often did not provide information to answer research questions of the study—many of them were created decades ago and because OCLC WorldCat does not collect or make available metadata versioning data, it was impossible to evaluate the extent of edits that were made to the records in 2019, how many edits were made to the records prior to 2019, and when they were made. Many of these revisions made in 2019 might have been very minor (e.g., correction of a typographical error or adding a table of contents or a note) and would not reflect the latest RDA and MARC 21 standard practices.

The negative outcome of the preliminary analysis of the data collected in the pilot served as a basis for refining the search query for recollecting the data. Upon the trial-and-error

process of formulating the query that would ensure both high precision and high recall in querying the OCLC WorldCat database and that would result in collecting the dataset that accurately reflects the study goals and allows answering its research questions, a solution was found. This solution resulted in refining the raw query the following way:

@and @attr 1=5067 rda @attr 1=5991 XXXX????

where

- @and is a Boolean search operator AND
- @attr 1 is an operator that defines search attribute of Z39.50
- 5067 is a search attribute that searches MARC21 subfield 040$e Description Conventions
- 5991 is a search attribute that searches MARC21 fixed field 008 bytes 00-05 (date record entered in database in the form of yyyymmdd)
- rda is a keyword value for searching with attribute 5067
- XXXX???? is a keyword value for searching with attribute 5991 where XXXX is the year and question marks are used as truncation signs for month and date information (see https://help.oclc.org/Metadata_Services/Z3950_Cataloging/Use_Z39.50_Cataloging/Search_tips_for_OCLC_Z39.50_Cataloging and https://www.loc.gov/z3950/lcserver.html)

After refining the query approach this way, the size of the dataset of RDA-based MARC 21 bibliographic records created in 2019 was estimated to be over 1.7 million of records and the size of the dataset of records created in 2020 was estimated to be 308000 records based on information provided by MARC Edit Z39.50 SRU client when starting the metadata harvesting. According to the findings of recent related studies (e.g., Phillips, 2020), sets of over 2 million of Dublin Core records—which are much more concise than MARC 21 metadata records—present significant and often insurmountable computational challenges in analysis of connections

between the records by the shared data values in metadata fields. Based on my experience working with MARC and Dublin Core metadata, on average, RDA-based MARC 21 bibliographic records are at least five times the size of Dublin Core records majority of which include only eight fields (e.g., Shreeves et al., 2005). Therefore, it was estimated that technical barriers would prevent efficient retrieving, processing, and analyzing a dataset that consists of more than 400000 RDA-based MARC 21 bibliographic records. To obtain a manageable dataset that the most closely reflects the criteria of this study—in other words, records that are most-recently created according to the latest versions of RDA and MARC 21—the decision was made to narrow the focus of data collection by two measures:

- Choosing the year of 2020 (January through April) as opposed to entire set of records from 2019 and/or
- Including only the records that were created in 2020 (but not those that might have been created much earlier and last updated in 2020 with often a minor update)

The manual evaluation of several metadata records retrieved from OCLC WorldCat using MARC Edit Z39.50 SRU client tool as part of the pilot study also allowed me to determine that the only order in which the records are retrieved and downloaded is starting with the records that have the highest number of holdings attached.

3.4.4 Data Processing

The downloaded dataset was not human-readable and required transformations and clean up procedures. An example of pure downloaded MARC21 machine-readable data (a file with the. mrc filename extension) can be seen in Figure 3.1, as record retrieved as a pilot study.

An example of a result of simple lossless transformation of a pure machine-readable MARC21 record (a file with the. mrc filename extension) that is human-readable and is the

structured form of this same data (a file with the .mrk filename extension) is given in the Figure 3.2.

Figure 3.1: Pure representation of MARC21 bibliographic record in .mrc file format

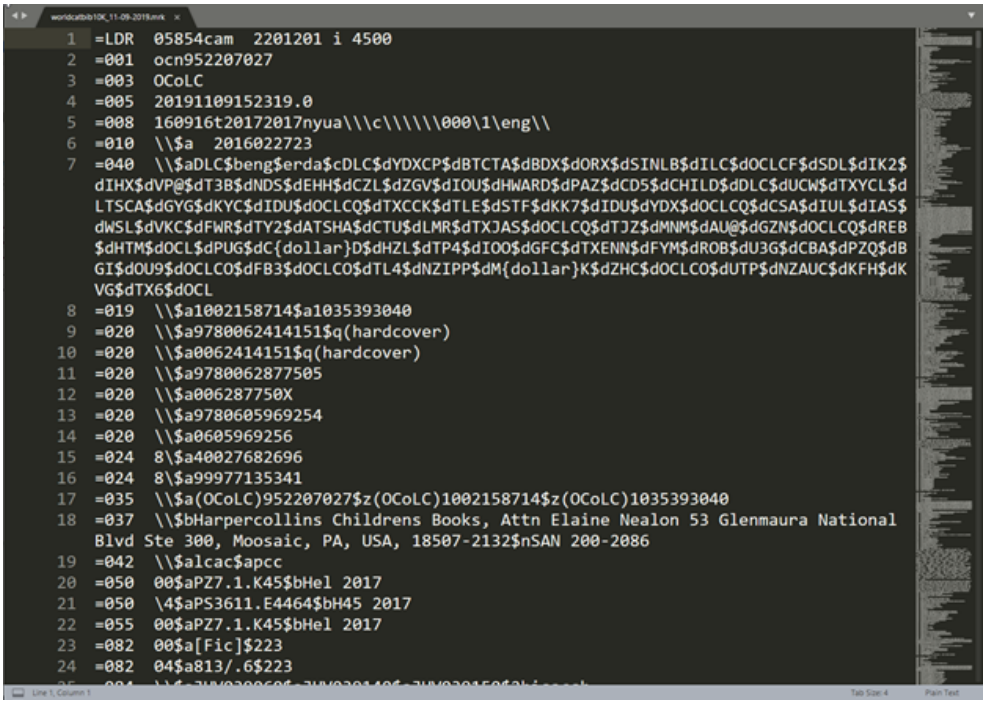
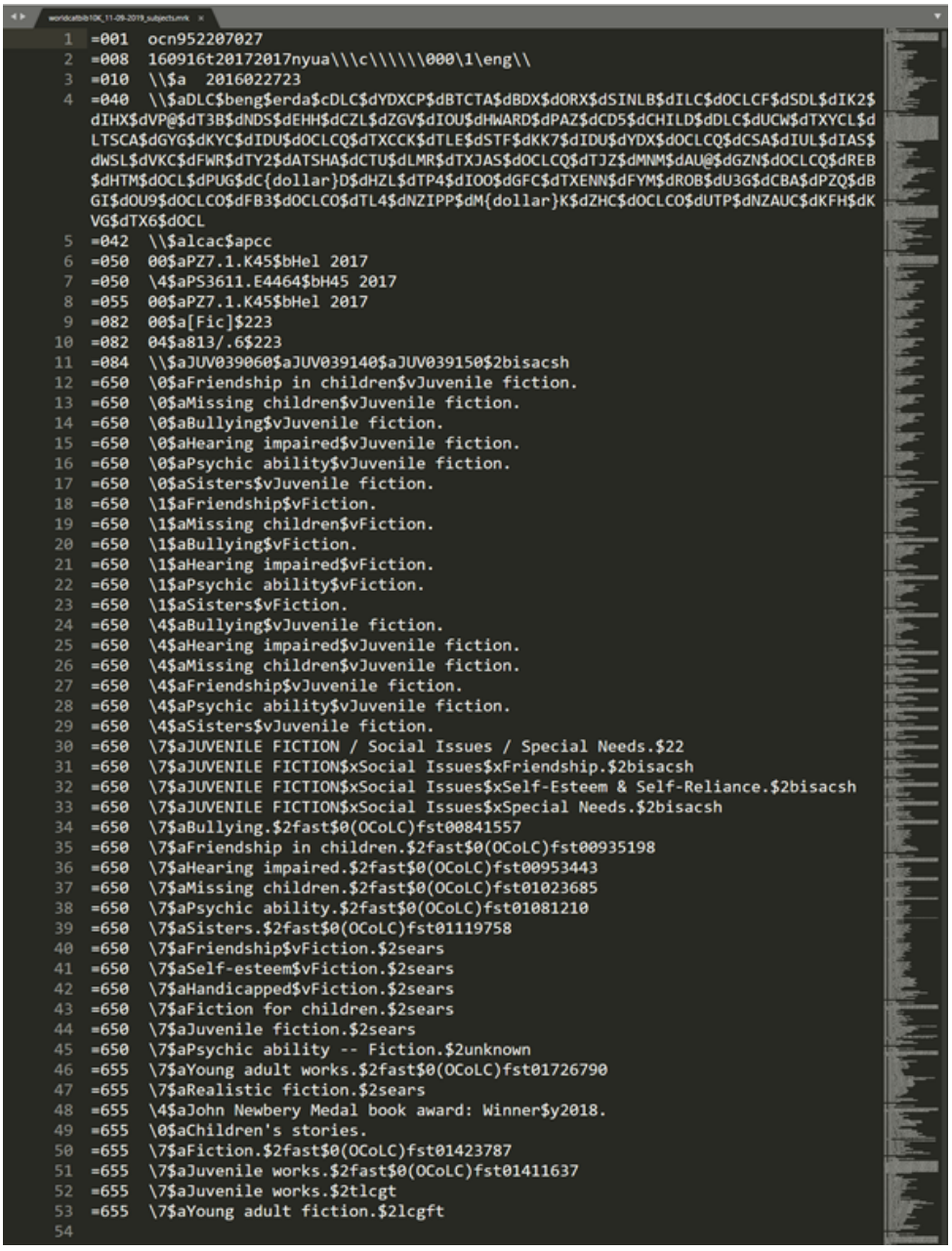


Figure 3.2: Human-readable form of pure MARC21 bibliographic record in .mrk file format

Data clean up procedures were performed with the help of such instruments as MARCEdit, Regex and GREP. The transformed dataset contained complete bibliographic records in human-readable format. To increase performance of calculations and further analysis, clean up procedures were applied. For cleaning data, MARCEdit in conjunction with Regex and GREP utilities was used. An example of clean record in human-readable format is given in Figure 3.3.



```
=001  ocn952207027
=008  160916t20172017nyua\\\c\\\\\\000\1\eng\\
=010  \\$a   2016022723
=040  \\$aDLC$beng$erda$cDLC$dYDXCP$dBTCTA$dBDX$dORX$dSINLB$dILC$dOCLCF$dSDL$dIK2$
dIHX$dVP@$dT3B$dNDS$dEHH$dCZL$dZGV$dIOU$dHWARD$dPAZ$dCD5$dCHILD$dDLC$dUCW$dTXYCL$d
LTSCA$dGYG$dKYC$dIDU$dOCLCQ$dTXCCK$dTLE$dSTF$dKK7$dIDU$dYDX$dOCLCQ$dCSA$dIUL$dIAS$
dWSL$dVKC$dFWR$dTY2$dATSHA$dCTU$dLMR$dTXJAS$dOCLCQ$dTJZ$dMNM$dAU@$dGZN$dOCLCQ$dREB
$dHTM$dOCL$dPUG$dC{dollar}D$dHZL$dTP4$dIOO$dGFC$dTXENN$dFYM$dROB$dU3G$dCBA$dPZQ$dB
GI$dOU9$dOCLCO$dFB3$dOCLCO$dTL4$dNZIPP$dM{dollar}K$dZHC$dOCLCO$dUTP$dNZAUC$dKFH$dK
VG$dTX6$dOCL
=042  \\$alcac$apcc
=050  00$aPZ7.1.K45$bHel 2017
=050  \4$aPS3611.E4464$bH45 2017
=055  00$aPZ7.1.K45$bHel 2017
=082  00$a[Fic]$223
=082  04$a813/.6$223
=084  \\$aJUV039060$aJUV039140$aJUV039150$2bisacsh
=650  \0$aFriendship in children$vJuvenile fiction.
=650  \0$aMissing children$vJuvenile fiction.
=650  \0$aBullying$vJuvenile fiction.
=650  \0$aHearing impaired$vJuvenile fiction.
=650  \0$aPsychic ability$vJuvenile fiction.
=650  \0$aSisters$vJuvenile fiction.
=650  \1$aFriendship$vFiction.
=650  \1$aMissing children$vFiction.
=650  \1$aBullying$vFiction.
=650  \1$aHearing impaired$vFiction.
=650  \1$aPsychic ability$vFiction.
=650  \1$aSisters$vFiction.
=650  \4$aBullying$vJuvenile fiction.
=650  \4$aHearing impaired$vJuvenile fiction.
=650  \4$aMissing children$vJuvenile fiction.
=650  \4$aFriendship$vJuvenile fiction.
=650  \4$aPsychic ability$vJuvenile fiction.
=650  \4$aSisters$vJuvenile fiction.
=650  \7$aJUVENILE FICTION / Social Issues / Special Needs.$22
=650  \7$aJUVENILE FICTION$xSocial Issues$xFriendship.$2bisacsh
=650  \7$aJUVENILE FICTION$xSocial Issues$xSelf-Esteem & Self-Reliance.$2bisacsh
=650  \7$aJUVENILE FICTION$xSocial Issues$xSpecial Needs.$2bisacsh
=650  \7$aBullying.$2fast$0(OCoLC)fst00841557
=650  \7$aFriendship in children.$2fast$0(OCoLC)fst00935198
=650  \7$aHearing impaired.$2fast$0(OCoLC)fst00953443
=650  \7$aMissing children.$2fast$0(OCoLC)fst01023685
=650  \7$aPsychic ability.$2fast$0(OCoLC)fst01081210
=650  \7$aSisters.$2fast$0(OCoLC)fst01119758
=650  \7$aFriendship$vFiction.$2sears
=650  \7$aSelf-esteem$vFiction.$2sears
=650  \7$aHandicapped$vFiction.$2sears
=650  \7$aFiction for children.$2sears
=650  \7$aJuvenile fiction.$2sears
=650  \7$aPsychic ability -- Fiction.$2unknown
=655  \7$aYoung adult works.$2fast$0(OCoLC)fst01726790
=655  \7$aRealistic fiction.$2sears
=655  \4$aJohn Newbery Medal book award: Winner$y2018.
=655  \0$aChildren's stories.
=655  \7$aFiction.$2fast$0(OCoLC)fst01423787
=655  \7$aJuvenile works.$2fast$0(OCoLC)fst01411637
=655  \7$aJuvenile works.$2tlcgt
=655  \7$aYoung adult fiction.$2lcgft
```

Figure 3.3: Clean MARC 21 bibliographic record in .mrk file format after clean-up procedures

As a result of clean-up procedures, a set of bibliographic records contained the following four groups of structured alphanumeric data:

- Group #1: MARC21 subject added entry and index terms fields:
    - 600 - Subject Added Entry - Personal Name
    - 610 - Subject Added Entry - Corporate Name
    - 611 - Subject Added Entry - Meeting Name
    - 630 - Subject Added Entry - Uniform Title
    - 647 - Subject Added Entry - Named Event
    - 648 - Subject Added Entry - Chronological Term
    - 650 - Subject Added Entry - Topical Term
    - 651 - Subject Added Entry - Geographic Name
    - 653 - Index Term - Uncontrolled
    - 654 - Subject Added Entry - Faceted Topical Terms
    - 655 - Index Term - Genre/Form
    - 656 - Index Term - Occupation
    - 657 - Index Term - Function
    - 658 - Index Term - Curriculum Objective
    - 662 - Subject Added Entry - Hierarchical Place Name
    - 688 - Subject Added Entry - Type of Entity Unspecified
    - 69X - Local Subject Access Fields
- Group #2: MARC21 classification and call numbers fields:
    - 050 - Library of Congress Call Number
    - 052 - Geographic Classification
    - 055 - Classification Numbers Assigned in Canada

- 060 - National Library of Medicine Call Number
- 070 - National Agricultural Library Call Number
- 072 - Subject Category Code
- 080 - Universal Decimal Classification Number
- 082 - Dewey Decimal Classification Number
- 083 - Additional Dewey Decimal Classification Number
- 084 - Other Classification Number
- 085 - Synthesized Classification Number Components
- 086 - Government Document Classification Number
- 09X - Local Call Numbers (090, 092, 096, 098, and 099)

- Group #3: Additional MARC21 subject metadata fields:
  - 043 - Geographic Area Code
  - 045 - Time Period of Content
  - 522 - Geographic Coverage Note
- Group #4: MARC21 fields with contextual information. To be used for evaluation of general characteristics of the collected metadata records (detailed in the section 3.5.3 Measures below) and for Social Network Analysis purposes. Group #4 includes:
  - 001 - Control Number
  - 008 - Fixed-Length Data Elements
  - 010 - Library of Congress Control Number
  - 040 - Cataloging Source
  - 042 - Authentication Code

3.4.4.1 Language of Materials Data Extraction

To retrieve language codes from the 008 MARC21 fixed field, the following regular

expression was used to generate a report: (=008.{2}.{35})(.{3}). This regular expression consists of two groups: group 1 (=008.{2}.{35}) and group 2 (.{3}), for the purpose of sorting collected language codes alphabetically. The first group in the regular expression reads information located in 008 MARC21 field, skips the first two bytes, which are blank spaces, and then reads forward 35 bytes. Based on understanding the structure of MARCEdit mnemonic format. mrk, this part can be simplified to the expression: (=008.{37}); however, a longer expression is easier for reading and understanding. The second part of expression, group 2, reads the next three bytes, which are the language code itself. Thus, grouping results by group 2 provided a clean sorted list of language codes. Output is presented in tab delimited format that includes a header with regular expression search criteria and the specified count values.

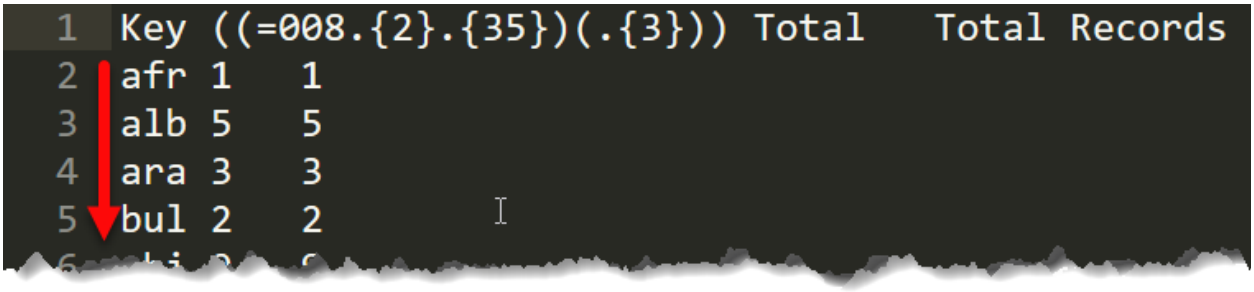


Figure 3.4: Example of language report

3.4.4.2 Data Processing for Building Networks of MARC 21 Bibliographic Records based on Shared Subject Terms

The process of building a network of bibliographic records that share similar attribute values and the application of network graph methods to bibliographic data analysis is not as straightforward as it might seem. One of the main challenges in applying SNA methods is the representation of bibliographic data. Originally, bibliographic data taken directly from machine-readable MARC databases and converted to human-readable mnemonic form is presented in

the form of a list of string values that needs to be structured, cleaned and normalized, in other words to be prepared for the future analysis. There are a number of tools available for data preparation and analysis. Retrieving of subsets from bibliographic dataset is done specifically with the help of PyMARC, a Python programming language library that is designed for working with bibliographic data encoded in MARC21 standard. Refining extractions, data normalization and subjects clustering are done with help of Google OpenRefine, General Refine Expressions Language (GREL) and RegEx expressions. OpenRefine is a powerful web browser application that runs a server on a personal computer and is designed for exploring, cleaning, transforming, reconciling and matching messy data (Topham, 2018). In addition to GREL and RegEx, the application works with codes written on Jython, a version of the Python programming language that runs on Java platform, and the Clojure programming language.

Normally, the process of network graph creation requires a list of vertices (e.g. V= {V1, V2, V3,...Vn}) and a list of edges (e.g. E = {{V1, V2},{V2,V3}, …{Vn-1, Vn}) with corresponding attributes that describe relationships and directions between these vertices. In the case of graph method application to bibliographic data, there is no need to create directed network graphs, because relationships between records are rather undirected; therefore, analysis of such relationships can be built based on adjacency matrices that represent weighted values of relationships and do not include any information about directions between records.

One of the ways to create such adjacency matrix for subject headings is to create a pivot table that represents aggregated values of shared subject headings per each record and then process it in order to create an adjacency correlation matrix of weighted values of shared attributes, in other words, subject headings. There are several options for doing this. One

option is to use Pandas, one of the libraries in Python created for data analysis that is frequently used for work with dataframes. One of the advantages of using Python Pandas is resource efficiency; the computing environment easily processes large volumes of data in relatively short periods of time and requires minimal resources. Another option to work on adjacency matrix is to involve Rapidminer, a data science software platform developed for data preparation, predictive analytics, text mining, machine learning and big data analytics. The system also integrates Radoop, one of the deployments available for harnessing Hadoop clusters (Mierswa & Klinkenberg, 2020). Rapidminer is a fast developing and evolving product with SAS Enterprise Miner-like graphical user interface that allows for multitasking and performing several analytical calculations at the same time. It is decided to use Rapidminer at this stage.

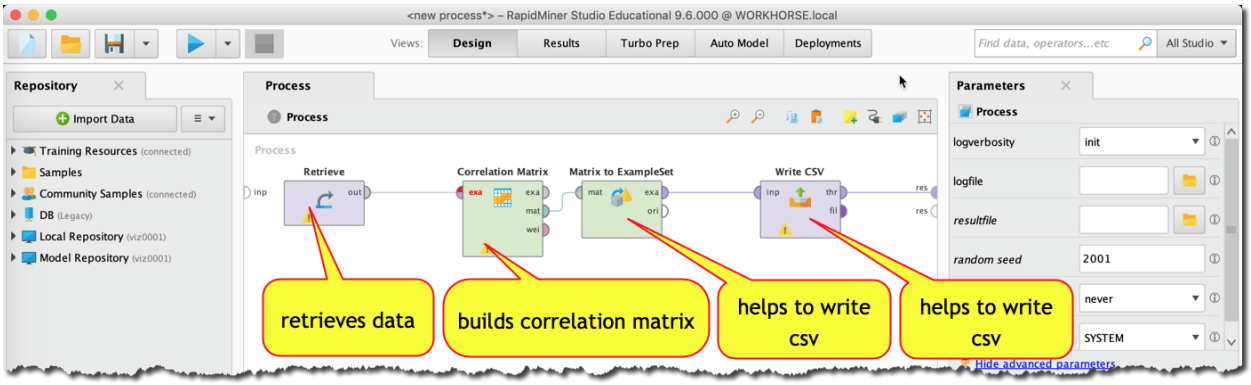


Figure 3.5: Process of correlation matrix creation in Rapidminer

The process of adjacency matrix creation in Rapidminer involves the following modules or steps (Mierswa & Klinkenberg, 2020):

1. Database retrieval—at this step the csv file is loaded to the system.
2. Correlation matrix creation—this step determines correlation between all attributes or variables and produces a vector of weights (table of correlation coefficients) based on these correlations. Correlation is a statistical method that can show whether, and how, strongly pairs of attributes or variables are related.

3. Conversion—this step is important in order to help in writing data output to a csv file; at this step, a matrix object is transformed into an object that is an exact representation of the original matrix and can be written into csv format.

4. Writing CSV file—this step is required for recording data output in comma-separated values (CSV) file that stores tabular data in plain-text form.

Resulting data output contains a table of weights—correlation coefficients—between sets of variables, between bibliographic records. Usually, correlation coefficients or weights can help to evaluate and measure relationships between variables: perfect relationships (0.9<1 for positive and -1<-0.9 for negative correlations), strong relationships (0.5<0.9 for positive and -0.9<-0.5 for negative correlations), weak relationships (0.1<0.5 for positive and -0.5<-0.1 for negative correlations) and uncorrelated relationships between variables (0<0.1 for positive and -0.1<0 for negative correlations). However, in this study correlations adjacency matrix is used to create a network graph and to visualize relationships between variables (records). In order to complete this, one more procedure has to be done that concerns the treatment of missing values. For the matrix used in SNA application, missing values should be imputed with zeros. The most efficient way to do data imputation in terms of time and resource management is to use Pandas, one of the popular libraries of programming language Python for data analysis. Method .fillna() from Pandas with parameter (0) applied to dataframe structure helps to replace null values with zeros. This tremendously aids further network analysis that can be done with help of NodeXL, a SNA software, that allows for the creation of a network graph based on adjacency matrix or with the help of other tools, such as NetworkX, a Python library for SNA or Gephi, a Java Script application for graph analysis. In this study, NodeXL Pro version was chosen due to better functionality—multiple processes and analytics can be performed and managed at the same time. In contrast to Python applications, such as NetworkX library that

allows for consequent processes, NodeXL Pro provides complex analytics and visualizations simultaneously, which is more preferable for the research (Smith et al, 2010); the relatively small size of data sets used for the analysis allow for GUI applications.

3.5 Data Analysis

This section details the steps of data analysis, and measures and metrics assessed to answer the research questions.

3.5.1 Stages of Analysis

The study consisted of two stages of data analysis, both of which utilized content analysis as the primary research method. The first stage used quantitative and qualitative content analysis (semi-automatic, with the help of various computational tools such as MARCEdit and PowerGREP, Python, Rapidminer, and NodeXL) of recently created RDA-based MARC 21 bibliographic records. The second stage involved in-depth manual content analysis of a sample of those MARC21 bibliographic records.

In addition to content analysis, the first stage of the study made use of the social network analysis and graph theory to visualize and analyze collocation of metadata records by creating networks of records that share the same properties — subject metadata values available in MARC21 metadata records. Implementation and use of SNA techniques and graph theory is not new for information science studies; these methods have been used for quite a while in the field. Case (2012) refers to the Social Network Analysis (SNA) as one of the research methods that are frequently used in information science. Scientific Citation Index (SCI) is one of the examples of such implementation (Price, 1965). Analysis of information behavior through

social networks is another example of such implementation (Schultz-Jones, 2009). Phillips, Tarver and Zavalina (2019) described the use of Social Network Analysis (SNA) in "metadata network analysis research at the University of North Texas (UNT) Libraries" (p. 1).

3.5.2 Measures

This study assessed general characteristics of the recently created RDA-based MARC 21 bibliographic records. The distribution of records by the following:

- Characteristics of cataloging: level of cataloging (based on data values in the ELvl subfield of the fixed field), language of cataloging (based on data values in the subfield $b of field 040), location and types of institutions that created the records (based on data values in the subfield $a of field 040), and number of holdings attached to bibliographic records (based on data values in the field 948 Local Data included in all records harvested from OCLC WorldCat).

- Characteristics of materials represented by records: materials types represented (based on data values in field 006 and 007), language of materials (based on data values in the Lang subfields of the fixed field).

Specific characteristics related to subject representation in metadata records were assessed:

- Level of application of the fields used for subject representations

- Level of application of the subfields used for subject representation (including Linked-Data-enabling subfields)

- Level of application of the controlled vocabularies used for subject representation based on the field indicators for and/or data values in certain subfields (e.g., 655 with the second indicator 7 and a specific data value from this list http://www.loc.gov/standards/sourcelist/genre-form.html in subfield $2; 650/651/655 with the second indicator 3 for Medical Subject Headings, 5 for Canadian Subject Headings, etc.).

- Co-occurrence levels for various subject metadata fields and subfields, including, but not limited to, those that are intended for representation of similar types of information:

    - 651 Subject Added Entry Geographic Name and/or 650 $z Subject Added Entry Topical Term Geographic subdivision with 043 Geographic Area Code and/or a field 522 Geographic Coverage Note.

- 650 $y Subject Added Entry Topical Term Temporal subdivision and/or 651 $y Subject Added Entry Geographic Name Temporal Subdivision with 045 Time Period of Content.
- The level of application of different controlled vocabularies in 6XX subject fields

- Co-occurrence levels for various subject controlled vocabularies used in the records
- Distribution of subject terms used in the 6XX subject terms
- Network characteristics of the networks formed by shared data values in 6XXs subject added entry and indexing fields 650 and 655 and the most consistently applied classification fields 050 and 082:
  - Network density, vertex degree, average geodesic distance between vertices

The following descriptive statistics measures for subject metadata were assessed in the study:

- Central tendency measures (mean, median, mode, minimum, maximum, range) and variability measures (variance and standard deviation) of numbers of different subject fields and instances of such fields (e.g., 650, 651, 082) per record
- Central tendency measures (mean, median, mode, minimum, maximum, range) and variability measures (variance and standard deviation) of the number of instances for each repeatable subject metadata field (e.g., 090) per record
- Central tendency measures (mean, median, mode, minimum, maximum, range) and variability measures (variance and standard deviation) of the number of instances for each subject metadata field/subfield combination (for example, 650$z, 082 $b) per record
- In particular, the Linked-data-enabling subfields of subject metadata fields (Table 3.1):

Table 3.1: Linked-Data-enabling subfields in subject metadata fields of MARC21 bibliographic records

MARC 21 Subject Metadata Fields	Subfields: codes, names, and repeatability information
600, 610, 611, 630, 650, 651, 654, 662, 688	$0 - Authority record control number or standard number (R) $1 - Real World Object URI (R) $2 - Source [of heading, name, title, term] (NR) $4 - Relationship (R)
043, 052, 055, 080, 084, 086, 647, 648, 655, 656, 657, 658	$0 - Authority record control number or standard number (R) $1 - Real World Object URI (R)

MARC 21 Subject Metadata Fields	Subfields: codes, names, and repeatability information
	$2 - Source [of heading, name, title, term] (NR)
050, 060, 070, 085	$0 - Authority record control number or standard number (R) $1 - Real World Object URI (R)
072, 082, 083, 092	$2 - Source (NR) or Edition number (NR) for DDC numbers
090, 096, 098, 099, 522, 653	None

To analyze relations between the records based on the subject metadata, networks of records were created with the help of software for data analysis and visualization. Metadata records represented by record IDs played the role of vertices. Shared properties, such as data values in MARC21 subject metadata fields played the role of edges and represented network connections between records. The following graph metrics were used to evaluate resulting networks:

- *Cardinality*—number of vertices (bibliographic records)
- *Degree*—number of edges connected to a vertex (bibliographic record)
- *Network or graph density*—a portion of potential connections (that could potentially exist between two vertices) in the actual network.
- *Clustering coefficient*—how vertices are embedded in their neighborhood
- *Betweenness (centrality)*—reflects the degree to which vertex stand between each other
- Closeness (centrality)—a measure of how fast information spreads from a given vertex to other reachable vertices in the network
- *Eigenvector (centrality) or prestige score*—the influence of a vertex in a network
- *(Average) geodesic distance*—the (average) distance between two vertices in a graph—the number of edges in a shortest possible distance from one vertex to another
- *Network diameter*—the maximal geodesic distance within the network
- *Connected components*—the number of connected components in the network

- *Modularity*—a metric for measuring the network structure—the strength of division of a network into groups, clusters or communities—modules. The higher modularity value, the higher density of vertices within groups or communities.

The unit of analysis of Stage 1 was mostly at the level of the dataset, as opposed to the level of individual records or the subset of MARC 21 data elements—subject fields and subfields—in these records. In-depth manual content analysis of a subsample of 100 records analyzed in Stage 1 at the dataset level was conducted in Stage 2 to supplement Stage 1 analyses and provide triangulation.

A number of measures that provide valuable insight into the subject representation in metadata records can only be assessed at the record level first, and then in some cases can be aggregated to the sample level. These measures include, for example, analysis of co-occurrences of certain subject fields within the same records. Also, analysis of central tendency measures (minimum, maximum, median, mode), and variability measures (variance and standard deviation) for numbers of different subject fields (e.g., 650, 651, 082) and subfields, as well as their instances per record can only be meaningfully done if individual metadata records are closely examined. Therefore, these measures were assessed in Stage 2 of this study.

3.5.3 Reliability and Validity

Predictions and generalizations that could be drawn from research (Kratwohl, 2009) relate to such important measures of research quality as reliability and validity. Reliability refers to the likelihood of obtaining the same results in repeated studies of the same phenomenon in the same setting. Validity is a degree with which research design captures and measures what is studied (Babbie, 2013).

The design of this study included a methodological triangulation to ensure reliability and validity. Four types of triangulation are described by Denzin (2006): 1) data triangulation which involves time, space and persons; 2) investigator triangulation, which involves multiple researchers; 3) theory triangulation, which involves using two or more theories in interpretation of phenomenon; and 4) methodological triangulation, which involves using multiple methods of data collection or analysis. In this study, different approaches stemming from two different research perspectives were used for data analysis: descriptive and social network analysis.

The study was designed to make use of the Big Data approach in collecting and analyzing the entire dataset to meet the criteria that supports answering the research questions. According to Borgman (2015), studies of Big Data using statistical methods and computational modeling achieve high reliability due to the scale of the dataset, while surveys that rely on samples achieve adequate reliability only if the sample is large. The analysis of the entire dataset that matches the set of criteria which are based on research questions eliminates the need for statistical tests such as t-test or calculating Pearson r that are needed for ensuring that the findings are generalizable in studies that rely on samples. Manual content analysis of a small sample of MARC 21 records in Stage 2 was designed to assess only objective characteristics, with quantitative or binary measures. This allowed to eliminate the typical for manual content analysis limitation -- researcher bias -- that causes problems with reliability of findings (Neuendorf, 2002).

CHAPTER 4

FINDINGS

4.1 Findings Obtained in Stage 1

This chapter details the findings obtained in the Stage 1 of this study: analysis of the most- recently-created RDA-based MARC 21 records that are available through OCLC WorldCat database.

4.1.1 Introduction

OCLC WorldCat database of MARC 21 bibliographic records was queried via MarcEdit SRU/Z39.50 client using the following raw query that combined two search criteria: @and @attr 1=5067 rda @attr 1=5991 2020????. According to MARCEdit Z39.50/SRU client, approximately 308000 metadata records that met the search criteria existed in the OCLC WorldCat database at the time of data collection. There was no indication as to what proportion of these records was unique and how many duplicates were included among 308000 records. After at least five separate attempts to collect the records that match the search criteria defined by this study, only 141310 records (45.88% of the total dataset at the time of data collection) were downloadable for analysis. Collecting this large dataset involved many computational challenges. The automated process of retrieving records took more than 80 hours and was possible only in increments of no more than 100000 records and at one point it was terminated by OCLC WorldCat server due to technical error which occurred when retrieving the 3rd batch of records. The downloaded dataset was processed using the MARCEdit record deduplication tool to remove identical records. The deduplication resulted in significant change in the collected dataset as it brought its size down to 10014 unique records. This

indicates that only approximately 7% of all records in OCLC WorldCat that match the search criteria set up by this study design are unique. This large sample allows me to draw conclusions that are highly-generalizable—with a confidence level of 99% and confidence interval of 0.25—to the entire population of RDA-based MARC 21 bibliographic records in OCLC WorldCat database that were created in 2020.

The remainder of section 4.1. reports the findings of Stage 1: descriptive statistical measures of various quantitative and qualitative characteristics of the collected RDA-based MARC 21 bibliographic records overall in the dataset. The presentation starts with the results regarding the general characteristics of records: distribution of records by material type, by cataloging source, encoding level, holdings, language of cataloging, and by languages of materials represented in records. Results regarding specific characteristics of subject metadata fields and subfields follow: overall level of application of subject metadata fields and subfields across the dataset, including Linked-Data enabling data elements, correlation between the overall number of occurrences of subject metadata elements representing the same content in different ways, level of use of MARC 21 tools to indicate primary and secondary subject headings, and various controlled vocabularies used in the data values of 6XX fields, as well as in fields 072 and 084. Next are presented the findings regarding the most widely used subject terms across the dataset for each of the eleven 6XX fields. Section 4.1 is concluded with the presentation of the findings regarding the connectivity between MARC 21 bibliographic records in the sample (n=10014) based on the shared data values in the two 6XX subject metadata fields—650 and 655—which are included in at least 50% of all records in the dataset and network analysis characteristics or the resulting metadata networks.

4.1.2 General Characteristics of MARC 21 Records

4.1.2.1 Material Types

Records in the collected dataset represented 7 out of 8 broad types of materials as defined by MARC standard. Two types of materials were represented by 10% or more of collected records: books of various kinds (59.41%), and sound recordings (20.48%). Three more types of materials were represented by over 5% of records each: computer files (7.54%), continuing resources (6.21%), and visual materials (5.90%). The proportion of records representing two additional broad types of materials, maps and scores, was low, with under 0.5% each. One broad type of records as defined by MARC standard -- records representing mixed materials -- was not observed in the collected dataset. Table 4.1 provides details on the distribution of these general types of materials in the sample, based on information in the MARC Leader field. Table 4.1 also provides data on the distribution of some of the subtypes of these broad material types based on MarcEdit Material Type Report which mines information encoded in MARC21 bibliographic fixed and control fields such as 006, 007, and 008.

Table 4.1: Distribution of records by material type with subtypes (n=10014)

Material Type	no. of records	% of all records
Books:	**5949**	**59.4068%**
Online	3932	39.2650%
Large Print	80	0.7989%
Microform	64	0.6391%
Electronic	3	0.0300%
Microfilm	3	0.0300%
Direct Electronic	1	0.0100%
Others	1866	18.6339%
Sound Recordings:	**2051**	**20.4813%**

Material Type	no. of records	% of all records
Online	1693	16.9063%
Direct Electronic	21	0.2097%
Others	337	3.3653%
Computer Files:	**755**	**7.5394%**
Online	743	7.4196%
Direct Electronic	12	0.1198%
Others	0	0.0000%
Continuing Resources:	**621**	**6.2013%**
Online	607	6.0615%
Others	14	0.1398%
Visual Materials:	**591**	**5.9017%**
Online	259	2.5864%
Direct Electronic	6	0.0599%
Videorecording	1	0.0100%
Others	325	3.2455%
Maps:	**33**	**0.3295%**
Online	10	0.0999%
Others	23	0.2297%
Scores:	**14**	**0.1398%**
Online	1	0.0100%
Others	13	0.1298%
TOTAL	**10014**	**100%**

4.1.2.2 Cataloging Sources, Encoding Levels, Holdings, and Languages of Cataloging

Data codes contained in MARC 21 field 040 Cataloging Source were analyzed to obtain information about institutions that created the unique records in the collected sample and official language of cataloging indicated in the records. MARC 21 Bibliographic standard defines non-repeatable field 040 as

> The MARC code for or the name of the organization(s) that created the original bibliographic record, assigned MARC content designation and transcribed the record into machine-readable form, or modified (except for the addition of holdings symbols) an existing MARC record. These data and the code in 008/39 (Cataloging source) specify the parties responsible for the bibliographic record. (https://www.loc.gov/marc/bibliographic/bd040.html).

Both the mandatory non-repeatable subfield 040 $a Original Cataloging Agency and the system-supplied non-repeatable subfield 040 $c Transcribing Agency contain codes from the Directory of OCLC Members (https://www.oclc.org/en/contacts/libraries.html), or MARC Code List for Organizations (https://www.loc.gov/marc/organizations/), which are maintained by US Library of Congress or national code assigning institutions in Canada, United Kingdom, Germany, and Estonia. The subfields serve for capturing information on which institution created the bibliographic record and which institution transcribed the record into machine-readable form respectively. Based on OCLC guidelines for contributing records to WorldCat, a transcribing institution that contributes to the record it did not create (e.g., as part of the retrospective conversion process), is expected to transcribe the record exactly as found on the source original cataloging record, without making any substantial changes (https://www.oclc.org/bibformats/en/0xx/040.html). It was found that records in the sample were created by a total of 398 institutions worldwide and transcribed by a total of 333 institutions. Most (n=320) institutions transcribed their own original cataloging records, while some were indicated in field 040 only as creators (n=78) or only as transcribers (n=13). Data values in mandatory 040 subfield $a Original Cataloging Agency were used as search terms in searching the lists of institution codes maintained by OCLC and Library of Congress. Figures 4.1 and 4.2 show distribution of original cataloging agencies by the type of institution and by the country where the institution is located respectively.

Five institution codes were not found in the Directory of OCLC Members or MARC Code List of Organizations. These institutions are presented in Figure 4.1 as "not identified". The remaining 393 institutions that served as original cataloging agencies belonged to 12 categories: academic, association/foundation, corporate/business, federal/national government, public, publisher/producer/supplier, school, state/municipal government, state/national library, vendor, votech/community colleges, and other. The three largest groups of institutions by type were public libraries (n=177), academic libraries (n=119), and vendors (n=45).

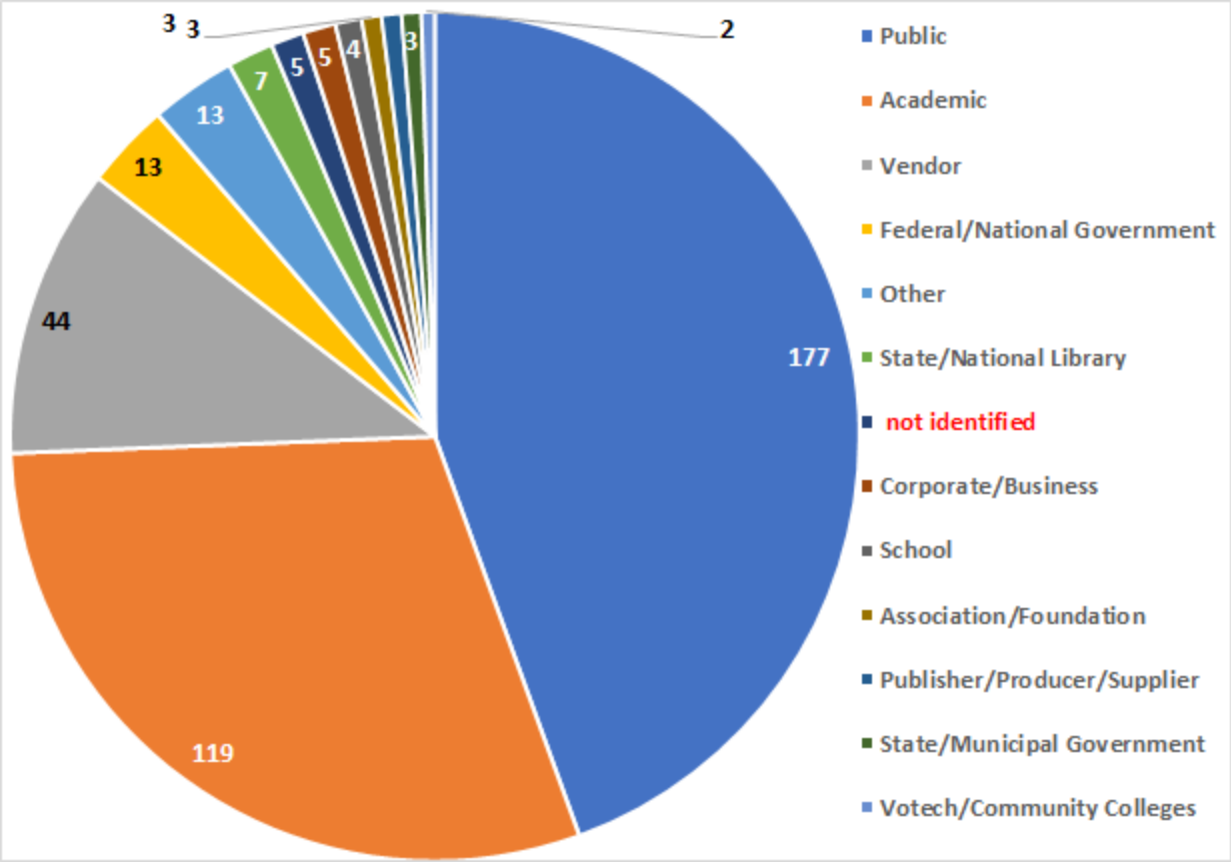


Figure 4.1: Original cataloging agencies by institution type (n=398)

Since five institution codes were not found in the Directory of OCLC Members or MARC Code List of Organizations, the countries of location for these institutions are presented in

Figure 4.2 as “not identified”. The remaining 393 institutions that served as original cataloging agencies for records in the sample are located in 12 countries: Australia, Canada, Egypt, France, Germany, Hong Kong, Italy, Lebanon, Netherlands, New Zealand, Singapore, South Africa, Switzerland, Thailand, United Kingdom, and the United States. Eighty-two percent of record-contributing institutions were located in the United States.

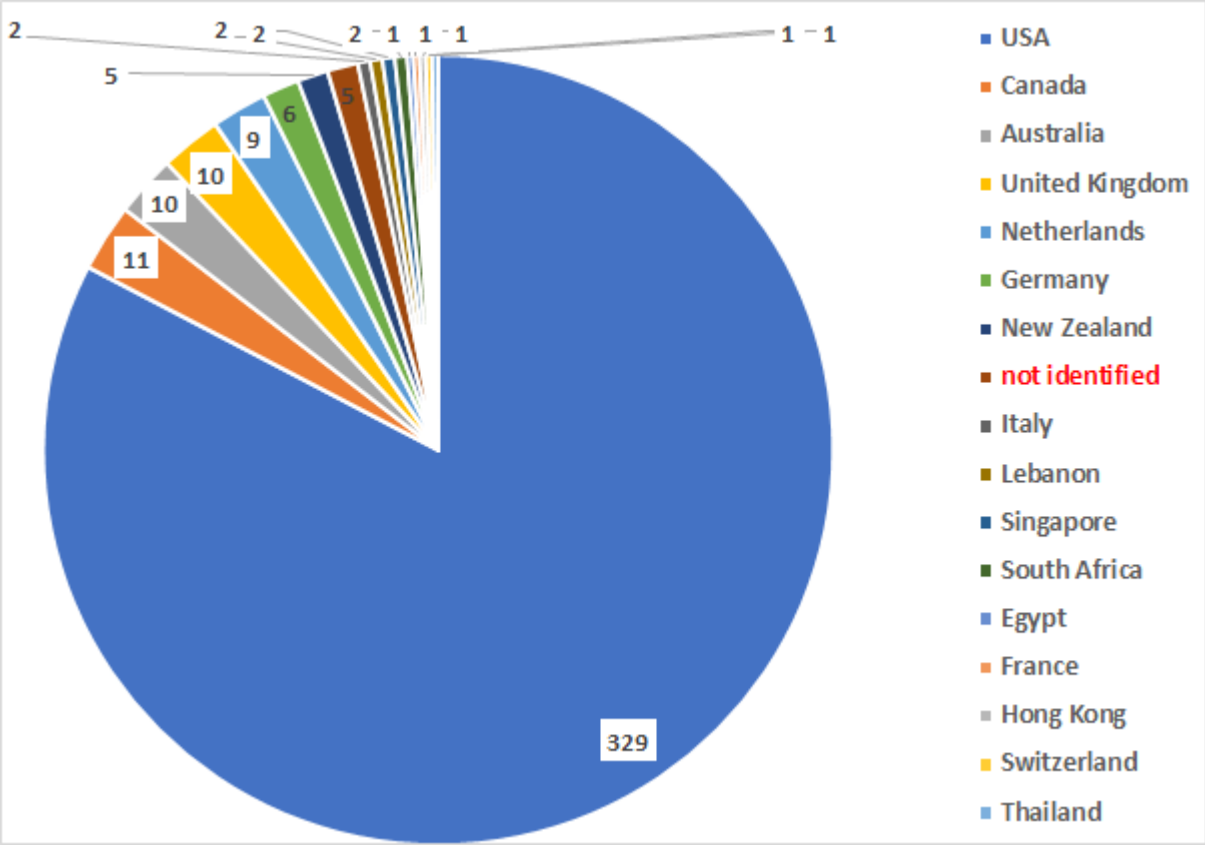


Figure 4.2: Original cataloging agencies by country of location (n=398)

A total of 7622 records (76.11% of all records in the collected dataset) were created by institutions located in the USA. Another 11.63% of records (1165 records in the collected dataset) were created in Germany. A total of 19 institutions created more than 100 records each; collectively these institutions created 80% of records in the sample. This group included two US-based institutions with over 1500 of records contributed by each: a vendor Naxos of

America Incorporated (n=1597), and a federal government agency US Government Publishing Office (n=1535). The Bibliotheks Verbund Bayern (Bavarian Library Association located in Germany)—self-identified in the Directory of OCLC members as an "other" type of institution—contributed the third largest number of records (n=795), closely followed by the United States Library of Congress (n=767). Two other libraries that contributed more than 100 records each in a sample were the Libraries Australia (national library agency of Australia), and Ohio University library. Eleven out of 19 institutions that contributed 100 or more records each were those identified as vendors or "corporate/business" in the Directory of OCLC members. In addition to Naxos of America Incorporated, these vendors included Australian-based Ebook Library, Germany-based Walter De Gruyter GMBH &Co KG, US-based Alexander Street, Baker and Taylor, Gobi Library Services, JSTOR, Midwest Tape, Netlibrary, Overdrive, and Taylor & Francis Group. Each of the remaining 379 institutions contributed between one and 87 records.

Results of the analysis of data values consisting of 1-character-long codes in the ELVl (Encoding Level) subfield of the fixed field of all records in the sample are presented in Table 4.2. Over half of the records (51.38%) follow the highest standards of cataloging: full encoding level. This includes records created by the Program for Cooperative Cataloging authorized participants (code empty, n=2333), full encoding level by the members of OCLC consortium (code I, n=2776, and full level with materials not examined (code 1, n=36). The core level of cataloging (i.e., higher than minimal but lower than full (code 4)) was observed in four records in the sample. Almost 27% of records followed a minimal level standard of encoding. This included the minimal level input by OCLC participants (code K, n=2683), and minimal level (code 7, n=5). An abbreviated level of cataloging which does not meet the minimum level standard as

defined by the MARC 21 Format for Bibliographic Data, National Level Full and Minimal Requirements (https://www.loc.gov/marc/bibliographic/nlr/) was observed in nine additional records (code 3). Finally, an additional 2170 records had code 8 (prepublication level, n=338) or M (added from a batch process, n=1832). Unlike the other six codes observed in the sample and discussed above, codes 8 and M do not refer to the quality, or completeness, of bibliographic records but simply to the context in which these records are created, therefore records with these codes in the ELvl subfield of the fixed field can vary in quality.

Table 4.2: Distribution of records by the encoding level

Code		Number of Records
Full Level	I	2776
	blank	2333
	1	36
	subtotal	5145
Core Level	4	2
	subtotal	2
Minimal Level	K	2683
	7	5
	subtotal	2688
Abbreviated Level	3	9
	subtotal	9
Other	8	338
	M	1832
	subtotal	2170
Total		10014

Holdings data in OCLC WorldCat database contributes to facilitating access to information objects. It represents how many institutions own the item in their collections. An institution is automatically coded as the one holding the item in the OCLC WorldCat database when it chooses the "Update holdings" option in the Action menu of the OCLC Connexion cataloging tool (or performs an equivalent action in alternative cataloging tools that work with OCLC WorldCat database) for the MARC 21 bibliographic record that an institution contributed to the database as original cataloging agency. This also happens when updating holdings if an institution's catalogers do not create the record themselves but use an existing record in the copy cataloging process and export the record to the institution's local online catalog. The number of institutions that updated holdings to add the record to their catalogs is automatically calculated and reported in MARC 21 field 948 when the record is exported to the local online catalog or otherwise harvested (e.g., through Z39.50 protocol for harvesting MARC metadata).

As shown in Figure 4.3, bibliographic records in the sample ranged widely in the number of holdings: from 14 to 1574 (1517 was the maximum number of holdings) at the time of data collection. The largest group of records (n=120) had between 134 and 254 holdings, followed by those having between 14 and 134 holdings (n=113), and those having between 254 and 374 holdings (n=91). Overall, the higher the number of holdings, the lower the number of records had these holdings.

A small proportion of records (1.6%) was found to have a very wide reach and significant effect on the bibliographic metadata contained in online catalogs of various institutions: 166 records have 500 or more holdings each and thus are included in a high proportion of all library

catalogs worldwide that use OCLC WorldCat in original and copy cataloging workflows. This analysis also revealed that the vast majority of records with the highest number of holdings (98%) were the records that conform to the full level of cataloging as defined by the MARC 21 Format for Bibliographic Data, National Level Full and Minimal Requirements (https://www.loc.gov/marc/bibliographic/nlr/): those with codes blank or I in the ELvl subfield of the fixed field.

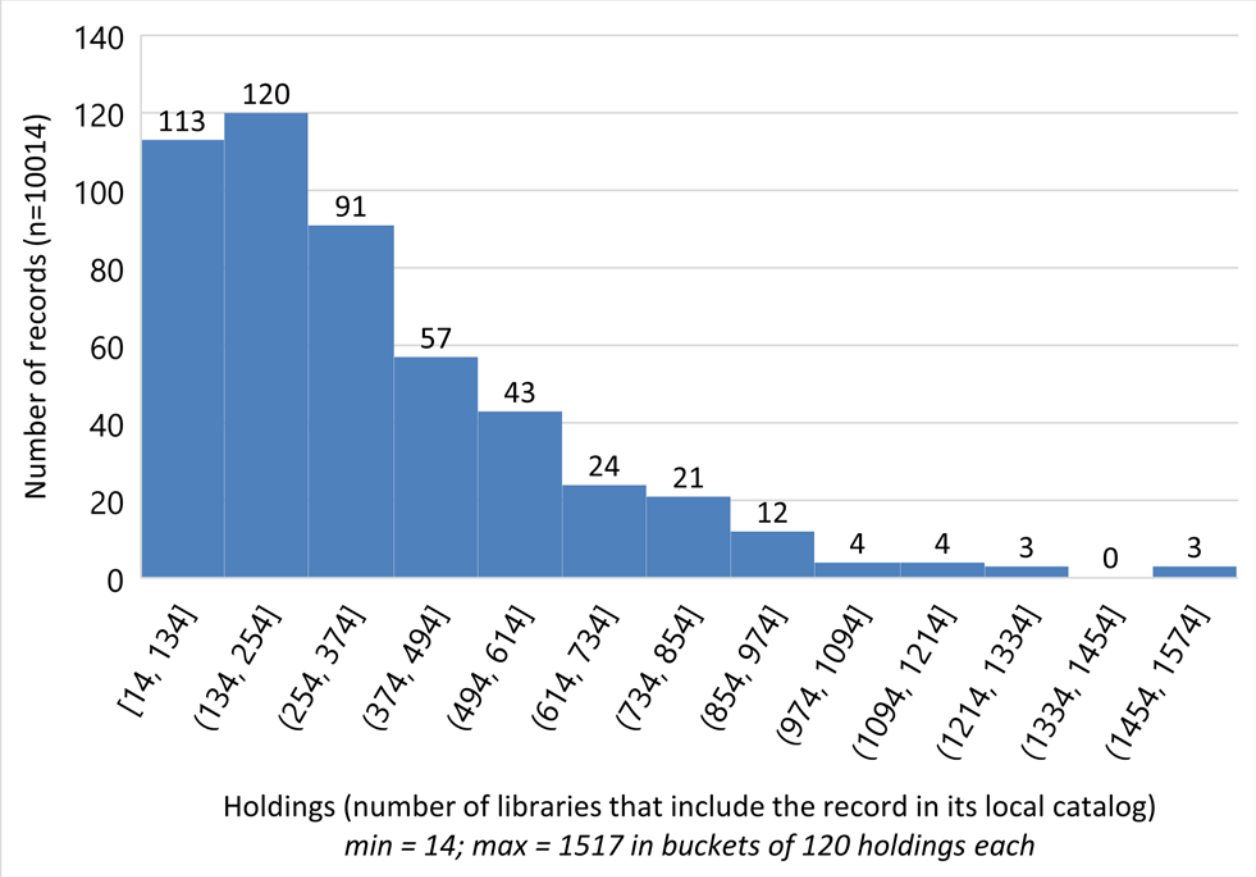


Figure 4.3: Distribution of the number of holdings in the records (n=10014)

Analysis of data values in field 040 subfield $b Language of Cataloging using codes from US Library of Congress MARC Code list for Languages http://www.loc.gov/marc/languages/language_code.html shows that 88.76% of all records in the sample were created by cataloging agencies that use English as a language of cataloging,

followed by German language of cataloging (9.43% of records). A much smaller proportion of records were created by cataloging agencies using Dutch as the official language of cataloging (1.79% of records). Only one record in the sample was created by the agency that uses French as the official language of cataloging. No languages of cataloging beyond English, German, Dutch and French were represented in the collected dataset.

4.1.2.3 Languages of Materials in the Collected Data Sample

Data codes in the Language subfield (bytes 35-37) of the MARC 21 Bibliographic Format fixed field 008 indicate the language of the information object represented by a metadata record. The codes included in this subfield are drawn from the Library of Congress MARC Code list for Languages.

Based on the distribution of these codes in the collected dataset (Table 4.3), 1124 or 11.22% of MARC 21 records in the sample represent materials without linguistic content (MARC Language code zxx in the fixed field 008). Common examples of materials without any linguistics content include instrumental music, silent movies, paintings, etc. An additional 21 records in the sample (0.21%) had a MARC Language code und (i.e., undetermined) in the fixed field 008. Majority of the records (over 88%) represented materials that include any content in any language (e.g., text of various kinds, recorded songs, etc.). As shown in Table 4.3, a total of 62 records (or 0.62%) represented multilingual materials (MARC Language code mul in the fixed field 008). Materials in 47 different languages were observed in the sample of metadata records. A large proportion of records in the sample represented materials in English (7401 or 73.9%), followed by four other Western languages: German (637 records or 6.36%), Dutch (164

records or 1.64%), Spanish (153 records or 1.53%) and French (1.47%). Records representing materials in 42 other languages occurred in less than 1% of the sample each.

Table 4.3: Distribution of records by the language of material

MARC Language Code	Language Name	Number of Records	% of All Records
eng	English	7401	73.9065%
zxx	*no linguistic content*	1124	11.2243%
ger	German	637	6.3611%
dut	Dutch	164	1.6377%
spa	Spanish	153	1.5279%
fre	French	147	1.4679%
mul	*multilingual*	62	0.6191%
ita	Italian	61	0.6091%
lat	Latin	58	0.5792%
por	Portuguese	36	0.3595%
und	*undetermined*	21	0.2097%
rus	Russian	17	0.1698%
fin	Finnish	14	0.1398%
cze	Czech	11	0.1098%
ind	Indonesian	11	0.1098%
nor	Norwegian	9	0.0899%
swe	Swedish	8	0.0799%
pol	Polish	7	0.0699%
rum	Romanian	7	0.0699%
bul	Bulgarian	6	0.0599%
hun	Hungarian	6	0.0599%
chi	Chinese	5	0.0499%
dan	Danish	5	0.0499%
jpn	Japanese	5	0.0499%
tur	Turkish	4	0.0399%
grc	Greek, Ancient (to 1453)	3	0.0300%

MARC Language Code	Language Name	Number of Records	% of All Records
gre	Greek, Modern (1453-)	3	0.0300%
aze	Azerbaijani	2	0.0200%
bos	Bosnian	2	0.0200%
heb	Hebrew	2	0.0200%
hin	Hindi	2	0.0200%
kor	Korean	2	0.0200%
ukr	Ukrainian	2	0.0200%
arn	Mapuche	1	0.0100%
bur	Burmese	1	0.0100%
fro	French, Old (ca. 842-1300)	1	0.0100%
gla	Scottish Gaelic	1	0.0100%
gle	Irish	1	0.0100%
haw	Hawaiian	1	0.0100%
ice	Icelandic	1	0.0100%
kaz	Kazakh	1	0.0100%
lit	Lithuanian	1	0.0100%
luo	Luo (Kenya and Tanzania)	1	0.0100%
map	Austronesian (Other)	1	0.0100%
mlg	Malagasy	1	0.0100%
per	Persian	1	0.0100%
slo	Slovak	1	0.0100%
srp	Serbian	1	0.0100%
tib	Tibetan	1	0.0100%
vie	Vietnamese	1	0.0100%

4.1.3 Subject Representation in MARC 21 Records

4.1.3.1 Subject Metadata Fields

A total of 26 subject metadata fields were identified in the collected dataset (Table 4.4). Eight individual fields intended for subject representation—083, 085, 522, 656, 657, 658, 662,

and 688—did not appear in any of the records in the sample. Also, no fields from the field group 69X Local Subject Access Fields occurred in any of the records. The records in the sample did not include four out of five of the new subject metadata fields added to MARC 21 Bibliographic Standard between 2005 and 2019 to meet emerging requirements: 083 Additional Dewey Decimal Classification Number, 085 Synthesized Classification Number Components, 662 Subject Added Entry - Hierarchical Place Name, and 688 Subject added Entry - Type of Entity Unspecified). However, another newly added subject field 647 Subject Added Entry - Named Event was observed in 125 records in the dataset (1.25%).

Field 650 Subject Added Entry - Topical Term appeared in the largest proportion of records (92.04%). Field 655 Index Term--Genre/Form was the second most commonly occurring subject field overall (79.12%), closely followed by fields containing two types of classification numbers: 050 Library of Congress Call Number (65.56%) and 082 Dewey Decimal Classification Number (53.03%). Five more subject fields appeared in more than 10% of records in the sample: 651 Subject Added Entry - Geographic Name (36.83%), 043 Geographic Area Code (33.96%), 086 Government Document Classification Number (16.47%), 072 Subject Category Code (10.49%) and 045 Time Period of Content (10.13%). All other fields appeared in less than 10% of records. Two of these subject fields - 096 Locally Assigned NLM-type Call Number and 654 Subject Added Entry - Faceted Topical Terms -- appeared in only one record.

Most of the 26 MARC 21 subject fields appeared in records in multiple instances (Table 4.4). Field 653 Index Term--Uncontrolled exhibited the largest number of instances in the records that it was observed in: 9.72 per record. The fields that occurred in the second and third largest number of instances per record were 650 Subject Added Entry--Topical Term (5.78

instances) and 654 Subject Added Entry -- Faceted Topical Terms (3.00 instances). A total of 318 records had 15 or more instances of field 650. The one record in the sample representing a graphic novel included a stunning 46 instances of field 650.

Three additional fields in the 6XX block appeared in the records between 2.3 and 2.6 times on average: 655 Index Term - Genre/Form, 600 Subject Added Entry - Personal Name, and 610 Subject Added Entry - Corporate Name. Only one of the classification fields—field 072 Subject Category Code—appeared more than twice per record on average (2.5876). Only five fields appeared exactly once on average per record that included them: 043 Geographic Area Code, 045 Time Period of Content, 070 National Agricultural Library Call Number, 092 Locally Assigned Dewey Call Number, and 096 Locally Assigned NLM-type Call Number.

Table 4.4: Distribution of subject fields in the records (n=10014)

MARC 21 Subject Field Tag	MARC 21 Subject Field Name	Total Records	Percentage of All Records in the Sample	Total Instances	Average Instances Per Record with Field
043	Geographic Area Code	3401	33.96%	3401	1.0000
045	Time Period of Content	1014	10.13%	1014	1.0000
050	Library of Congress Call Number	6565	65.56%	6651	1.0131
052	Geographic Classification	32	0.32%	39	1.2188
055	Classification Numbers Assigned in Canada	84	0.84%	86	1.0238
060	National Library of Medicine Call Number	137	1.37%	150	1.0949
070	National Agricultural Library Call Number	13	0.13%	13	1.0000
072	Subject Category Code	1050	10.49%	2716	2.5867
080	Universal Decimal Classification Number	12	0.12%	16	1.3333

MARC 21 Subject Field Tag	MARC 21 Subject Field Name	Total Records	Percentage of All Records in the Sample	Total Instances	Average Instances Per Record with Field
082	Dewey Decimal Classification Number	5310	53.03%	5398	1.0166
084	Other Classification Number	753	7.52%	1110	1.4741
086	Government Document Classification Number	1649	16.47%	1658	1.0055
090	Locally Assigned LC-type Call Number	15	0.15%	21	1.4000
092	Locally Assigned Dewey Call Number	22	0.22%	22	1.0000
096	Locally Assigned NLM-type Call Number	1	0.01%	1	1.0000
600	Subject Added Entry - Personal Name	876	8.75%	2075	2.3687
610	Subject Added Entry - Corporate Name	703	7.02%	1632	2.3215
611	Subject Added Entry - Meeting Name	97	0.97%	117	1.2062
630	Subject Added Entry - Uniform Title	172	1.72%	292	1.6977
647	Subject Added Entry - Named Event	125	1.25%	144	1.1520
648	Subject Added Entry - Chronological Term	977	9.76%	979	1.0020
650	Subject Added Entry - Topical Term	9217	92.04%	53243	5.7766
651	Subject Added Entry - Geographic Name	3688	36.83%	6642	1.8010
653	Index Term - Uncontrolled	181	1.81%	1760	9.7238
654	Subject Added Entry - Faceted Topical Terms	1	0.01%	3	3.0000
655	Index Term--Genre/Form	7923	79.12%	20550	2.5937

4.1.3.2 Application of Subject Metadata Subfields, Including Linked-Data Supporting Ones

A total of 115 subfields of the 26 MARC21 subject metadata fields were observed in the

records in the sample. This is approximately 41% of the total number of 286 subfields that are defined for these 26 subject metadata fields in MARC 21 Bibliographic Format standard. Table A in the Appendix A shows the list of all subfields defined by MARC 21 Bibliographic Format standard for subject metadata fields and the levels of application for each of them. Table 4.5 shows the top 20 most frequently applied subject metadata subfields that are not defined by this study as Linked-Data-enabling subfields. Figure 4.4 shows the comparative level of use of Linked-Data-enabling subject metadata subfields which were observed in at least one instance in the dataset.

As shown in Table 4.5 and Appendix A, the most frequently occurring subfields logically included the mandatory subfields $a in the widely applied subject added entry and index term fields: 650, 655, and 651. As these subfields are nonrepeatable, the number of instances of these subfields in the dataset equaled the number of appearances of their respective fields. Repeatable and required if applicable subfields $x General Subdivision, $v Form Subdivision, and $z Geographic Subdivision of the field 650 -- also belonged to the top 20 most frequently occurring. These subfields, along with subfield $y Chronological Subdivision, are also defined by MARC 21 bibliographic standard for nine fields beyond 650—600, 610, 611, 630, 647, 648, 651, 654 (except $x), and 655—but appeared much less frequently. The top 20 most frequently appearing subfields also included mandatory subfields $a in three classification fields: 050 Library of Congress Call Number (repeatable $a), 082 Dewey Decimal Classification Number (repeatable $a), and 072 Subject Category Code (non-repeatable $a). Required if applicable non-repeatable subfield $b Item Number of a 050 field also was observed in a high number of instances. In addition, subfields of two non-classification fields 0XX, the 043 and 045, were

found to occur often (almost 4000 instances each): mandatory repeatable 043 $a Geographic Area Code and required if applicable repeatable 045 $b Formatted 9999 B.C. through C.E. Time Period.

Table 4.5: Top 20 most frequently occurring subject metadata subfields (except Linked-Date-enabling ones)

Field	Subfield Code	Subfield Name	Total Subfield Instances	Average No. of Subfield Instances Per Record with Field
650	$a	Topical term or geographic name as entry element (NR)	53243	5.7766
655	$a	Genre/form data or focus term (NR)	20548	2.5935
650	$x	General subdivision (R)	13873	1.5052
650	$v	Form subdivision (R)	9487	1.0293
650	$z	Geographic subdivision (R)	8855	0.9607
050	$a	Classification number (R)	6655	1.0137
651	$a	Geographic name (NR)	6642	1.8010
082	$a	Classification number (R)	5584	1.0516
050	$b	Item number (NR)	5476	0.8341
045	$b	Formatted 9999 B.C. through C.E. time period (R)	3987	3.9320
043	$a	Geographic area code (R)	3918	1.1520
072	$a	Subject category code (NR)	2716	2.5867
600	$a	Personal name (NR)	2075	2.3687
651	$x	General subdivision (R)	1925	0.5220
653	$a	Uncontrolled term (R)	1825	10.0829
072	$x	Subject category code subdivision (R)	1756	1.6724
610	$a	Corporate name or jurisdiction name as entry element (NR)	1632	2.3215
084	$a	Classification number (R)	1349	1.7915
600	$d	Dates associated with a name (NR)	1236	1.4110
651	$z	Geographic subdivision (R)	1056	0.2863

As shown in Appendix A, Linked-Data enabling subfields that facilitate expression of relations between entities in the metadata records took eight spots on the list of top 20 most frequently occurring subfields. As shown in Figure 4.4, non-repeatable subfield $2 was included in 5 fields: required if applicable $2 Source of Heading or Term in fields 650 and 651, mandatory $2 Source of Term in 655, required if applicable $2 Source in 072, and required if applicable $2 Edition Number in 082. Overall, in the records analyzed in this study, Linked-Data-enabling subfield $2 was observed in 13 subject metadata fields: 072, 084, 082, 092, 600, 610, 611, 630, 647, 648, 650, 651, and 655. It was not observed in other 4 out of 17 subject metadata fields for which it is defined by the MARC 21 Bibliographic standard: 043 ($2 Real World Object URI), 052 ($2 Code Source), 055 ($2 Source of Call/Class Number), and 080 ($2 Edition Identifier).

Optional repeatable subfield $0 Authority Record Control Number or Standard Number in fields 650, 651, and 655 was the other Linked-Data enabling subfield included in the top 20 most frequently observed subfields of subject fields in this study (Table 4.5). Overall, in the analyzed records, Linked-Data-enabling subfield $0 was observed in nine subject metadata fields: 084, 600, 610, 611, 630, 647, 650, 651, and 655. It was not observed in other nine out of 18 subject metadata fields for which it is defined by the MARC 21 Bibliographic standard: 043, 050, 052, 055, 060, 070, 080, 648, and 654.

The optional repeatable Linked-Data-enabling subfield $1 Real World Object URI is defined by MARC 21 bibliographic standard for 16 subject metadata fields out of those included in the records analyzed in this study: 050, 052, 055, 060, 070, 080, 600, 610, 611, 630, 647, 648, 650, 651, 654, and 655. However, this subfield was not found in any of the 10014 RDA-based

MARC 21 bibliographic metadata records in the sample.

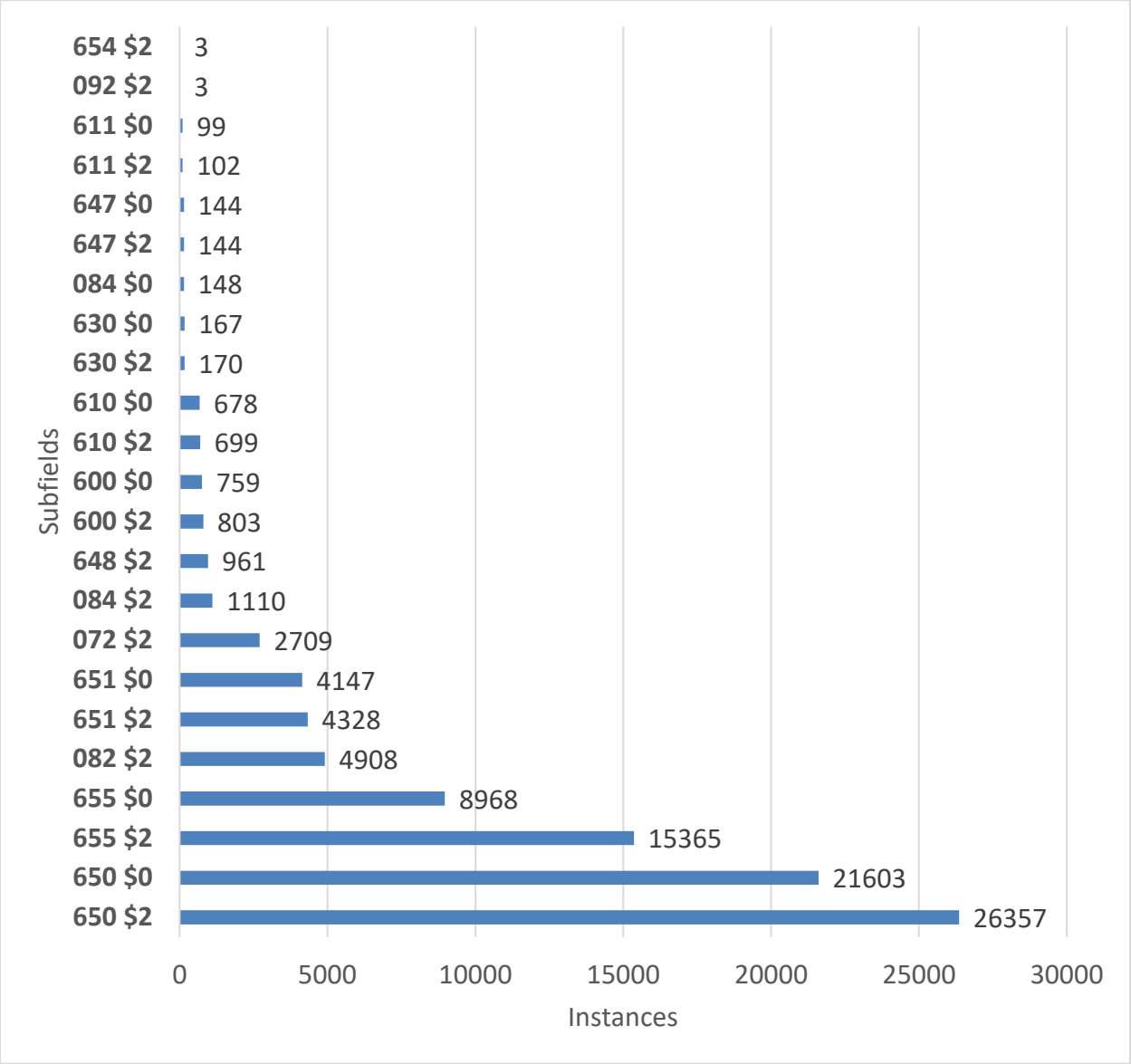


Figure 4.4: Number of observed instances of Linked-Data-enabling subject metadata subfields

The optional repeatable Linked-Data enabling subfield $4 Relationship is defined by MARC 21 bibliographic standard for seven subject metadata fields—all subject added entry fields—out of those included in the records analyzed in this study: 600, 610, 611, 630, 650, 651, and 654. It also was not found in any of the records in this study.

Another subfield for encoding similar information to that contained in subfield $4, optional repeatable subfield $e Relator Term, is defined by MARC 21 standard for five subject metadata fields: subject added entry fields 600, 610, 650, 651, and 654. Likewise, this Linked-Data-enabling subfield was not observed in any of the 61520 total instances of fields 610, 650, 651, or 654. Subfield $e Relator Term was observed in a single instance of field 600 (out of 2075 field instances in 876 records).

The level of co-occurrences of subject fields intended for representing the same type of information in different ways was analyzed in Stage 1 at the level of the entire dataset, and in Stage 2 at a much more refined level of each individual metadata records in a small subsample. This includes the following combinations:

- 651 Subject Added Entry Geographic Name (or 650 $z Subject Added Entry Topical Term Geographic Subdivision) with 043 Geographic Area Code (or 522).
- 650 $y Subject Added Entry Topical Term Temporal Subdivision and/or 651 $y Subject Added Entry Geographic Name Temporal Subdivision with 045 Time Period of Content.

As shown by its absence in Table 4.4, field 522 Geographic Coverage Note was not observed in any of the records in the dataset of 10014 records collected and analyzed in this study, therefore it was excluded from co-occurrence analysis.

MARC 21 Bibliographic Format standard defines field 043 as "Geographic area codes associated with an item" (https://www.loc.gov/marc/bibliographic/bd043.html). The OCLC Bibliographic Formats and Standards guide narrows down that definition to focus it on subject representation: "Contains the geographic area code (GAC), which is an aid to a subject approach to the item. It provides a hierarchical breakdown of geographical and/or political entities" and makes this field widely applicable to representing any type of material

(https://www.oclc.org/bibformats/en/0xx/043.html). MARC Bibliographic Standard Entry for field 043 states "choice of geographic area code is usually based on the geographic names and/or subdivisions in 6XX subject added entry and index term fields", which includes field 651 and subfield $z in nine other MARC 21 subject added entry fields. Thus, the optional non-repeatable field 043 ideally should be included in the record every time one or more instances of the repeatable field 651 are included. However, the level of application of 043 was not similar to the level of application of 651. It was found (see Table 4.4) that overall, field 651 occurred at least once in a substantially higher number of records (3688 or 36.83%) than field 043 (3401 or 33.96%). A total of 287 records, or almost 3% of all analyzed 10014 records, included field 651 but did not include field 043.

Also, based on the MARC 21 Bibliographic standard recommendations cited above, field 043 is expected to be included in the record whenever subfield $z Geographic Subdivision is included in any instances of the following 10 subject added entry MARC 21 fields: 600, 610, 611, 630, 647, 648, 650, 651, 654, and 655. As shown in the previous Table 4.5, the sample analyzed in this study did not include any records containing subfield $z in five of these fields: 611, 630, 647, 648, or 654. The total number of instances of repeatable subfield $a of field 043 (n=3918) was substantially higher than the total number of instances of repeatable subfield $z in fields 600 (n=31), 610 (n=25), 655 (n=11), and 651 (n=1056). However, it was significantly lower than the total number of instances of subfield $z in the subject added entry fields 650 (n=8815).

Similarly to field 043, the MARC 21 Bibliographic Standard definition of field 045 Time Period of Content is broad "A time period code and/or a formatted time period associated with the item" and covers not only subject representation (e.g., "the period depicted by the content

of the item" for books, continuing resources, motion pictures etc.) but also other uses (e.g., "time period of composition" for sound recordings and printed music scores) (https://www.loc.gov/marc/bibliographic/bd045.html). Field 045 is expected to be included in the records that include a field 650 Subject Added Entry - Topical Term with the data value that indicates a time period (e.g., "$a Shenandoah Valley Campaign, 1864 (May-August)") and/or include any subject added entry fields with subfield $y Chronological Subdivision (e.g., "$a Egypt $x Economic conditions $y To 332 B.C."). The OCLC Bibliographic Formats and Standards guide instructs catalogers not to use field 045 in records that represent "any item for which a chronological approach would not be a logical or common approach to the subject matter, [including:]

- Biography, unless a time period is specified on the piece or in a subject heading
- Collections or anthologies of literature, unless they indicate a clearly delineated time period
- Comprehensive histories of a subject or a country that cover more than 500 years
- Dictionaries, encyclopedias, glossaries, catalogs, and gazetteers intended to be nonhistorical in approach
- Genealogies and family histories
- Handbooks, manuals, and "how-to" books." (https://www.oclc.org/bibformats/en/0xx/045.html)

Non-repeatable optional field 045 appeared in 1014 RDA-based MARC 21 bibliographic records analyzed in this study (10.13%). Field 045 has two non-repeatable and five repeatable subfields, two of which were observed in the records collected and analyzed in this study: $a Time Period Code (22 instances total) and $b Formatted 9999 B.C. through C.E. Time Period (3987 instances total). As evident from data in Appendix A, subfield $y was included a total of

1772 times: in 1051 instances of field 650, 672 instances of field 651, 29 instances of field 655, 19 instances of field 610, and 1 instance of field 600. This indicates the level of application of field 045 comparable to the cumulative level of application of subfield $y in subject added entry fields.

4.1.3.3 Prioritization of Subject Headings

MARC 21 bibliographic records usually include multiple subject fields in the 6XX block of fields and usually more than one instance of specific fields from this block. As only one of the subject terms in multiple subject added entry fields or index fields in the record -- the one that most accurately represented the aboutness of an information object -- should be used for determining classification codes in 05X-09X classification number fields and call number fields, MARC 21 bibliographic standard has provisions for indicating which of the subject headings is the primary one, and which ones are the secondary ones. This is done with the help of the 1st MARC field indicator in fields 650 Subject Added Entry - Topical Term, 653 Index Term - Uncontrolled, and 654 Subject Added Entry - Faceted Topical Terms. OCLC Bibliographic Formats and Standards expands these guidelines to field 690 Local Subject Added Entry - Topical Term (https://www.oclc.org/bibformats/en/6xx/690.html). For these fields, 1st indicator Level of Subject can have one of the 4 following values:

- \ [empty] - No information provided (default 1st indicator value)
- 0 - No Level Specified
- 1 - Primary
- 2 - Secondary

Indicators in MARC21 bibliographic 6XX subject added entry fields and index term fields

in all metadata records in the dataset collected for this study were analyzed to determine the level of use of each of these values for the 1st indicator. Since no records in the dataset contained field 690 this analysis, results of which are shown in Table 4.6, focused on fields 650, 653, and 654. As shown in Table 4.6, the vast majority of instances of fields 650 (98.45%), 653 (99.32%), and 654 (100%) did not use 1st indicators 0, 1, or 2, and kept the default blank indicator value. The 1st indicator 0 No Level Specified was used in 321 (0.6%) of all instances of field 650, in nine (0.51%) of all instances of field 653 and was not used in field 654 at all. Only three instances of field 653 (0.17%) used 1st indicator 1 and none used 1st indicator 2. Similarly, a very small proportion of instances of the widely used field 650 made use of 1st indicator 1 (0.57%) or second indicator 2 (0.38%). Because field 650 was included in one or more instances in majority of records in the collected dataset -- 9217 (92.04%) -- this finding indicates that the level of use among record-contributing institutions in their cataloging of an option to prioritize one of the subject headings in the record is minimal.

Table 4.6: 1st indicators in fields 650, 653, and 654

Fields	Instances of 1st Indicator Blank: No Information Provided	% of All Instances of this Field	Instances of 1st Indicator 0: No Level Specified	% of All Instances of this Field	Instances of 1st Indicator 1: Primary	% of All Instances of this Field	Instances of 1st Indicator 2: Secondary	% of All Instances of this Field
650	52417	98.45%	321	0.60%	305	0.57%	200	0.38%
653	1748	99.32%	9	0.51%	3	0.17%	0	0.00%
654	3	100.00%	0	0.00%	0	0.00%	0	0.00%
TOTAL	54168		330		308		200	

4.1.3.4 Controlled Vocabularies Used for Subject Representation

Indicators in MARC21 bibliographic 6XX subject added entry fields and index term fields were also analyzed to determine the level of use of different controlled vocabularies for subject representation. MARC 21 standard defines indicators for metadata fields: 1st and 2nd. The second indicator is defined in the standard for nine subject added entry and index fields observed in the cords analyzed in this study: 600, 610, 611, 630, 647, 648, 650, 651, and 655. The second indicator between 0 and 3, and between 5 and 6 in these fields is a code that represents a specific controlled vocabulary from the list included in MARC 21 Bibliographic standard:

- 0 - United States Library of Congress Subject Headings (LCSH)
- 1 - United States Library of Congress subject headings for children's literature
- 2 - United States National Library of Medicine Medical Subject Headings (MESH)
- 3 - United States National Agricultural Library subject authority file
- 5 - Canadian Subject Headings
- 6 - Répertoire de vedettes-matière

A second indicator of 7 denotes other controlled vocabulary. If a 2nd indicator of 7 is used, the controlled vocabulary is specified in a data value or code in the subfield $2 (e.g., “viaf” for Virtual International Authority File in field 600, “fast” for Faceted Application of Subject Terminology in field 650, etc.). Finally, a second indicator of 4 indicates the controlled vocabulary that the data value in the field is taken from is not specified. The definition of the first indicator in the subject added entry MARC 21 fields varies. For example, a 1st indicator of 1 in field 600 Subject Added Entry - Personal Name indicates that the order of name components in the data value of this field starts with the surname, and a 1st indicator of 1 in field 650

Subject Added Entry - Topical Term means that the data value included in this instance of repeatable field 650 is the primary topical subject heading that presents "Main focus or subject content of the material" (https://www.loc.gov/marc/bibliographic/bd650.html). For three of these nine subject added entry and index fields—647, 648, and 651—the 1st indicator is not defined at all.

As shown in Table 4.7, the most often used 2nd indicator is 7 which means that the controlled vocabulary used is specified in the subfield $2. This second indicator was observed in each of the nine subject added entry fields, for a total of 48929 instances. Most often, a 2nd indicator of 7 is used in fields 650 (26358 times or 49.52% of all field 650 instances), 655 (15365 times or 74.77% of all field 655 instances), and 651 (4328 times or 65.16% of all field 651 instances). In fields 600, 610, 611, 630, and, 647, and 648, it was used in less than 1000 records each; however, the percentage of all instances of a field was high for all fields, with 38.70% the lowest (field 600). Table 4.7 shows distributions of data values found in subfield $2 of subject added entry fields with a 2nd indicator of 7.

The Library of Congress controlled vocabulary LCSH is widely used (see Table 4.7). A 2nd indicator of 0 is observed in seven out of nine subject added entry and index MARC 21 fields: 600 Subject Added Entry - Personal Name, 610 Subject Added Entry - Corporate Name, 611 Subject Added Entry - Meeting Name, 630 Subject Added Entry - Uniform Title, 650 Subject Added Entry - Topical Term, 651 Subject Added Entry - Geographic Name, and 655 Index Term - Genre/Form. Second indicator 0 appears in a significant number of field instances overall: 30875. Most often, a 2nd indicator of 0 is used in fields 650 (23870 times or 44.83% of all field 650 instances). In three fields, it is used in more than 1000 field instances: 655 (2555 times or

12.43% of all field 655 instances), 651 (2195 times or 33.05% of all field 651 instances), and 600 (1197 times or 57.69% of all field 600 instances). A second indicator of 0 was observed at a lower level in fields 611 (15 times or 12.8% or all field 611 instances) and 630 (122 times or 41.78% of all field 630 instances). It was not found in any instances of fields 647 or 648.

The third most often used subject added entry field second indicator is 4, which means that controlled vocabulary is not specified. It was observed in a total of 3665 field instances (see Table 4.7). This second indicator was found in five out of nine subject added entry fields: 600, 610, 648, 650, and 655. However, a substantial level of use was observed only for two fields: 655 (2568 times or 12.5% of all field 655 instances) and 650 (1070 times or in 2.01% of all field 650 instances). In the other three fields it was used under 20 times and in under 2% of all respective field instances.

The fourth and final often used second indicator in subject added entry and index fields was 1 which represents the Library of Congress Subject Headings of Children's Literature (see Table 4.7). It was observed in a total of 1473 instances of five fields: 600, 610, 650, 651, and 655. However, the only field in which this 2nd indicator was widely used is 650: 1345 times or 2.53% of all field 650 instances. It was observed in 61 instances of field 651 (0.92%), and in 51 instances of field 600 (2.46%). In two other fields—651 and 655—this controlled vocabulary was used 10 or less times and in under 0.5% of all respective field instances.

A second indicator of 2, which represents Medical Subject Headings (MeSH), was observed in the total of 588 field instances in 5 fields: 600, 610, 650, 651, and 655 (see Table 4.7). Approximately 85% of its use was observed in field 650. A second indicator of 6, which represents the Répertoire de vedettes-matièr controlled vocabulary, was observed in the total

of 131 field instances in 4 fields: 600, 610, 650, and 651. Almost 75% of its use was observed in field 650. A second indicator of 3, which represents the National Agricultural Library subject authority file, was observed in only four field instances—all in field 650. A second indicator of 5, which stands for the Canadian Subject Headings, was observed in a single instance of field 650.

Table 4.7: Application of 2nd field indicators in MARC 6XX subject fields for which 2nd indicator is defined

Fields	600	610	611	630	647	648	650	651	655	TOTAL
instances of 2nd indicator 0	1197	921	15	122	0	0	23870	2195	2555	30875
instances of 2nd indicator 1	51	6	0	0	0	0	1345	61	10	1473
instances of 2nd indicator 2	2	1	0	0	0	0	497	36	52	588
instances of 2nd indicator 3	0	0	0	0	0	0	4	0	0	4
instances of 2nd indicator 4	6	3	0	0	0	18	1070	0	2568	3665
instances of 2nd indicator 5	0	0	0	0	0	0	1	0	0	1
instances of 2nd indicator 6	16	2	0	0	0	0	98	15	0	131
instances of 2nd indicator 7	803	699	102	170	144	960	26358	4328	15365	48929
TOTAL	2075	1632	117	292	144	978	53243	6635	20550	

As shown in Table 4.8, a total of 64 different data values were found in the non-repeatable subfields $2 of the 349412 instances of subject added entry and index fields that had a second indicator 7. This indicator stands for “other” controlled vocabulary (i.e., not one of those 6 for which second indicators 0, 1, 2, 3, 5, and 6 are reserved in MARC 21 Bibliographic standard). These data values represent a total of 62 controlled vocabularies as one data value --

"unknown" (appeared in 1 instance of subject field 6XX) does not indicate any controlled vocabulary and one additional -- "lcgft\" (appeared in one instance of a subject field) -- is clearly a mistyped "lcgft." Local controlled vocabulary (as indicated by subfield $2 with a data value of "local") was used in 21 field instances in 21 different records: in fields 600 Subject Added Entry - Personal Name and 655 Index Term - Genre/Form. All other instances of 6XX subject fields with second indicator 7 used data values from the 61 different standard controlled vocabularies designed by various entities worldwide and for various knowledge domains: from Art and Architecture Thesaurus (code "aat" in 3 instances of field 655 in a single record) to OLAC Video Game Genre Terms (code "olacvggt" in 31 instances of field 655), to Book Industry Communication (BIC) UK Standard Library Categories (code "ukslc" in fields 650 in 8 records).

The most often used "other" controlled vocabulary was the Faceted Application of Subject Terminology: code "fast" was found in subfields $2 of 35908 subject field instances (Table 4.8). Three more controlled vocabularies served as the source of data values in more that 1000 of 6XX subject field instances collectively: Gemeinsame Normdatei by the German National Library (code "gnd" found in 1278 field instances), Library of Congress Genre/Form Terms for Library and Archival Materials (code "lcgft" found in 5848 subject field instances), and Book Industry Standards and Communications (BISAC) Subject Headings by Book Industry Study Group (code "bisacsh" in 2774 subject field instances). Interestingly, 254 instances of topical subject fields with a second indicator of 7 included the "overdrive" data value in subfield $2. These records are likely created by the OverDrive company—one of the major suppliers of materials (including ebooks, audiobooks, magazines and more) to libraries, and one of the institutions that contributed a substantial proportion of records in the sample as seen from the

previous Figure 4.1 and Figure 4.2. According to OverDrive press release (Sockel, 2016), the company adds BISAC subject headings to MARC 21 records, however there is no information about OverDrive's in-house controlled vocabulary. If these records created by OverDrive use BISAC headings, it is unclear why the code "bisacsh" was not used instead in subfield $2.

Seven additional "other" controlled vocabularies served as the source of data values in between 103 and 772 instances of the 6XX subject fields (Table 4.8). They included:

- OverDrive, as discussed above
- Sears List of Subject headings used since 1923 for subject cataloging in small and medium-sized libraries (code "sears" observed in 762 instances of subject fields: mostly 650 and 651, but also occasionally in 600 and 655)
- Guidelines on Subject Access to Individual Works of Fiction, Drama, Etc. controlled vocabulary http://experimental.worldcat.org/gsafd/browseGSAFD.html (code "gsafd" observed in 623 instances of field 655)
- NBD Biblion Trefwoordenthesaurus, the Dutch Keyword Thesaurus by NBD Biblion Foundation https://www.nbdbiblion.nl/product/abonnement-trefwoordenthesaurus (code "nbdbt" found in 252 instances of field 650)
- United States National Aeronautics and Space Administration (NASA) Thesaurus of subject terms on the topics of aerospace engineering and supporting areas of engineering and physics, astronomy, astrophysics, planetary science, Earth sciences, and biological sciences (code "nasat" observed in 153 instances of field 650)
- Gemeinsame Normdatei: Beschreibung des Inhalts https://wiki.dnb.de/download/attachments/106042227/AH-007.pdf controlled vocabulary by the German National Library (code "gnd-content" found in 152 instances of field 655)
- Centraal Bestand Kinderboeken (Central File Children's Books Theme Keyword) http://support.oclc.org/ggc/richtlijnen/?id=12&ln=nl&sec=k-556X controlled vocabulary developed and used in the Netherlands (code "cbk" observed in 103 instances of fields 655)

The remaining 50 controlled vocabulary codes were used in a low number of 6XX subject fields with a second indicator of 7: between one and 96 field instances total.

This analysis also revealed some instances of the incorrect use of second indicator 7 (Table 4.8). For example, in situations where the controlled vocabulary is unknown, MARC 21 Bibliographic standards uses the configuration of 6XX fields with a second indicator of 4. However, the data value "unknown" was used instead in the subfield $2 in one instance of field 651. Another example is the use of the code "lcsh" in subfield $2 in 14 instances of field 655 with a second indicator of 7 in 13 records in the sample, even though a second indicator of 0 is indicated by MARC 21 Bibliographic standard for representing the fact that the genre/form term in field 655 is derived from the Library of Congress Subject Headings controlled vocabulary (https://www.loc.gov/marc/bibliographic/bd655.html).

Table 4.8: Level of application of "other" controlled vocabularies based on data values in 6XX \7 $2

Controlled Vocabulary (Data Value in 6XX \7 $2)	Number of 6XX Field Instances
aat	3
agrovoc	3
ascl	3
bcl	12
bic	5
bicssc	78
bidex	32
bisac	7
bisach	2
bisacsh	2774
blmlsh	3
btr	88
cbk	103
cct	24
clams	1
eclas	20

Controlled Vocabulary (Data Value in 6XX \7 $2)	Number of 6XX Field Instances
eflch	6
eurovoc	5
fast	35908
fmesh	9
gmgpc	3
gnd	1278
gnd-content	152
gsafd	623
gtlm	4
gtt	96
hilcc	3
idszbz	14
idszbzes	1
iptcnc	50
larpcal	14
lcgft	5848
lcgft/	1
lcsh	13
local	21
mim	2
mup	1
naf	2
nasat	153
nbc	2
nbdbt	252
netc	4
olacvggt	31
overdrive	254
qlsp	13
ram	67

Controlled Vocabulary (Data Value in 6XX \7 $2)	Number of 6XX Field Instances
rasuqam	3
rbgenr	17
renib	2
reo	1
rero	4
rvmgf	3
sao	19
sears	762
sfit	2
shsples	3
stw	46
swd	40
tekord	1
thema	18
tlcgt	2
ukslc	8
unknown	1
TOTAL	**349412**

Similar to 6XX fields, the non-repeatable subfield $2 can be used to specify a controlled vocabulary (i.e., a list of subject category codes or a classification scheme respectively) used in the data values in two other subject metadata fields: 072 Subject Category Code, and 084 Other Classification Number. The overall level of use of this subfield in fields 072 and 084 was reported in Table 4.5 and discussed above. Analysis of data values in all 3819 instances of subfield $2 in fields 072 and 084 reveals that a total of 29 different controlled vocabularies were used in subject fields 072 and 084 (Table 4.9). The most widely used was the code list associated with BISAC subject headings (code “bisacsh”, n=2080, and misspelled code “bisach”,

n=2), and it was the only one that occurred in both fields. Seven codes out of 30 were also observed in $2 of 6XX fields with a second indicator of 7 (Table 4.8): “bicssc”, “bisacsh”, “bisach”, “eflch”, “mup”, “rero”, “thema”, and “ukslc”.

Table 4.9: Level of application of controlled vocabularies in 072 and 084 subject metadata fields

Subfield $2 Data Value	Number of Instances
bisacsh	2080
bicssc	527
sdnb	421
rvk	223
nur	79
bcl	67
lacc	64
brclbps	60
siso	38
thema	38
pim	37
cbkcd	36
fid	32
stub	25
lcco	24
clc	15
msc	11
blsrissc	9
ssgn	9
moys	4
ukslc	4
bkl	3
kssb/8	3
bisach	2
eflch	2

Subfield $2 Data Value	Number of Instances
rero	2
bcmc	1
clasbcud	1
lu-luope	1
mup	1
TOTAL	**3819**

4.1.3.5 Subject Terms Used in 6XX Fields

Subject terms present in the subfield $a of MARC 21 bibliographic fields 600, 610, 611, 630, 647, 648, 650, 651, 653, 654, and 655 were analyzed. Analysis focused on subfields $a as mandatory subfields for all of these fields. For nine of the 11 6XX subject metadata fields observed in this dataset, the subfield $a is non-repeatable, and for two fields—653 and 654—it is repeatable. The distribution of subject terms is reported below.

4.1.3.5.1 600$a Personal Name (NR)

As reported above, out of the 10014 records in the sample included a total of 2075 instances of field 600 subfield $a in 876 records (Table 4.5). A total of 862 unique terms contained in 610$a were observed in the dataset. Fifty-four of the personal name terms appeared in at least 0.05% of all records in the dataset. The most widely used name was the name of a United States President, which appeared in 0.93% of records. Another US president name was the third most widely used at 0.18% and shared this position with the name “Jesus Christ”. William Shakespeare’s name was the second most widely used name, appearing in 0.25% of all records in the dataset. The entire distribution of 600$a subject terms that appeared in at least 0.05% of all records is presented in Table 4.10.

Table 4.10: Distribution of subject terms used in 600$a: terms found in at least 0.05% of all records (n=10014)

600 $a Data Values	Count(oclcn)	Percentage
Trump, Donald	93.0	0.93%
Shakespeare, William	25.0	0.25%
Washington, George	18.0	0.18%
Jesus Christ	18.0	0.18%
Hua, Mulan	16.0	0.16%
Franklin, Benjamin	15.0	0.15%
Quixote	12.0	0.12%
Napoleon	11.0	0.11%
Simpson, Jessica	10.0	0.10%
Potter, Harry	10.0	0.10%
Granger, Hermione	10.0	0.10%
Weasley, Ron	10.0	0.10%
Walter, Bruno	10.0	0.10%
Einstein, Albert	9.0	0.09%
Dallas, Eve	9.0	0.09%
Rogers, Fred	8.0	0.08%
Banks, Alan	8.0	0.08%
Hegel, Georg Wilhelm Friedrich	8.0	0.08%
Beckett, Samuel	8.0	0.08%
Rimbaud, Arthur	7.0	0.07%
Batman	7.0	0.07%
Heidegger, Martin	7.0	0.07%
Peter Pan	7.0	0.07%
Bugs Bunny	7.0	0.07%
Proust, Marcel	7.0	0.07%
Ailes, Roger	6.0	0.06%
Désirée	6.0	0.06%
James	6.0	0.06%
Rawls, John	6.0	0.06%

600 $a Data Values	Count(oclcn)	Percentage
Heracles	6.0	0.06%
Hercules	6.0	0.06%
Paul	6.0	0.06%
Neumann, Hanus Stanislav	6.0	0.06%
Newman family	6.0	0.06%
Lindbergh, Charles A	6.0	0.06%
Aristotle	6.0	0.06%
Daffy Duck	6.0	0.06%
Spider-Man	6.0	0.06%
Mujibur Rahman	6.0	0.06%
Roosevelt, Theodore	5.0	0.05%
Mary	5.0	0.05%
Wagner, Richard	5.0	0.05%
Chomsky, Noam	5.0	0.05%
Tubman, Harriet	5.0	0.05%
Rowling, J. K	5.0	0.05%
Lacan, Jacques	5.0	0.05%
Kant, Immanuel	5.0	0.05%
Siegel, Siena Cherson	5.0	0.05%
Monet, Claude	5.0	0.05%
Augustine	5.0	0.05%
Wittgenstein, Ludwig	5.0	0.05%
Dante Alighieri	5.0	0.05%
Dora	5.0	0.05%
Morrison, Toni	5.0	0.05%

4.1.3.5.2 610$a Corporate Name or Jurisdiction Name as Entry Element (NR)

As reported above, the 10014 records in the sample included a total of 1632 instances of field 610 subfield $a in 703 records (Table 4.5). A total of 275 unique terms contained in

610$a were observed in the dataset. The most widely used corporate name was the United States (8.9%), followed by the name of European Union (0.32%). The Catholic church occurred in 0.24% of total number of records. The distribution of subject terms used in 610$a and found in at least 0.05% of all records is available in Table 4.11.

Table 4.11: Distribution of subject terms used in 610$a: terms found in at least 0.05% of all records (n=10014)

600 $a Data Values	Count(oclcn)	Percentage
United States	891.0	8.90%
European Union	32.0	0.32%
Catholic Church	24.0	0.24%
United States Postal Service	18.0	0.18%
North Atlantic Treaty Organization	15.0	0.15%
Great Britain	14.0	0.14%
United Nations	13.0	0.13%
Geological Survey (U.S.)	12.0	0.12%
Library of Congress	12.0	0.12%
Smithsonian Institution	10.0	0.10%
Housing Choice Voucher Program (U.S.)	10.0	0.10%
U.S. Census Bureau	8.0	0.08%
Deutschland	8.0	0.08%
United States Military Academy	7.0	0.07%
IS (Organization)	7.0	0.07%
Comprehensive Opioid Abuse Program (U.S.)	7.0	0.07%
Fox News	6.0	0.06%
E.I. du Pont de Nemours & Company	6.0	0.06%
Auschwitz (Concentration camp)	6.0	0.06%
Board of Governors of the Federal Reserve System (U.S.)	6.0	0.06%
Supplemental Nutrition Assistance Program (U.S.)	6.0	0.06%
Women Airforce Service Pilots (U.S.)	6.0	0.06%
Great Lakes Inventory and Monitoring Network (U.S.)	5.0	0.05%

600 $a Data Values	Count(oclcn)	Percentage
Franciscans	5.0	0.05%

4.1.3.5.3 611$a Meeting Name or Jurisdiction Name as Entry Element (NR)

The 611 MARC21 field represents the names of meetings or conferences used as subject access points. As reported above, the 10014 records in the sample included a total of 117 instances of field 611 subfield $a in 97 records (Table 4.5). A total of 48 unique terms contained in 611$a were observed in the dataset. The most frequently used name of the meeting was "World War (1939-1945)"; it occurred in 0.20% of all records. The second most frequently used name of the meeting was "Holocaust, Jewish (1939-1945)". This name occurred in 0.08% of all records and was followed in frequency by the name of the "American Revolution (1775-1783)". This term occurred in 0.07% of all records. The distribution of subject terms used in 611$a and found in at least 0.02% of all records is available in Table 4.12.

Table 4.12: Distribution of subject terms used in 611$a: terms found in at least 0.02% of all records (n=10014)

611 $a Data Values	Count(oclcn)	Percentage
World War (1939-1945)	20.0	0.20%
Holocaust, Jewish (1939-1945)	8.0	0.08%
American Revolution (1775-1783)	7.0	0.07%
World War (1914-1918)	6.0	0.06%
Olympic Games	6.0	0.06%
American Civil War (1861-1865)	4.0	0.04%
Melbourne Cup (Horse race)	4.0	0.04%
Revolution (France : 1789-1799)	3.0	0.03%
Global Financial Crisis (2008-2009)	3.0	0.03%
Civil War (Spain : 1936-1939)	3.0	0.03%
Cold War (1945-1989)	3.0	0.03%

611 $a Data Values	Count(oclcn)	Percentage
Thirty Years' War (1618-1648)	3.0	0.03%
Teheran Conference	2.0	0.02%
Meeting of the ASEAN Heads of Government	2.0	0.02%
Summit of the Americas	2.0	0.02%
Vatican Council	2.0	0.02%
Afghan War (2001-)	2.0	0.02%
French and Indian War (United States : 1754-1763)	2.0	0.02%
Vietnam War (1961-1975)	2.0	0.02%
Burning Man (Festival)	2.0	0.02%
Old Fiddlers' Convention	2.0	0.02%
Masters Golf Tournament	2.0	0.02%
Lewis and Clark Expedition	2.0	0.02%

4.1.3.5.4 630$a Uniform Title (NR)

As reported above, the 10014 records in the sample included a total of 292 instances of field 630 subfield $a in 172 records (Table 4.5). A total of 121 unique terms contained in 630$a were observed in the dataset. The most widely used in the dataset subject added entry uniform title was “Bible”. This title appeared in 0.78% of all records and was followed by “SAP HANA (Electronic resource)”, which appeared in 0.18% of all records. Titles such as “Constitution (United States)”, “Twitter”, “Linux” and “SAP ERP” appeared in 0.06% of all records. Table 4.13 represents subject terms that appeared in 630 $a in at least 0.03% of all records.

Table 4.13: Distribution of subject terms used in 630$a: terms found in at least 0.03% of all records (n=10014)

630 $a Data Values	Count(oclcn)	Percentage
Bible	78.0	0.78%
SAP HANA (Electronic resource)	18.0	0.18%
Constitution (United States)	6.0	0.06%

630 $a Data Values	Count(oclcn)	Percentage
Twitter	6.0	0.06%
Linux	6.0	0.06%
SAP ERP	6.0	0.06%
National Emergencies Act (United States)	4.0	0.04%
Motor Vehicle Information and Cost Savings Act (United States)	4.0	0.04%
Dragon Ball Z (Television program)	4.0	0.04%
Eulenspiegel (Satire)	4.0	0.04%
Patient Protection and Affordable Care Act (United States)	3.0	0.03%
LinkedIn (Electronic resource)	3.0	0.03%
Bachelor (Television program)	3.0	0.03%
Golden girls (Television program)	3.0	0.03%

4.1.3.5.5 647 $a Named Event (NR)

As reported above, the 10014 records in the sample included a total of 144 instances of field 647 subfield $a in 125 records (see previous Table 4.5). A total of 58 unique terms contained in 647$a were observed in the dataset. Similar to the distribution of 611$a names of the meeting, first place was taken by “World War”, appearing in 0.35% of all records. The term “Revolution” appeared in 0.08% of all records. Third place was shared between two terms: “American Revolution” and “Jewish Holocaust”. Both terms appeared in 0.07% of all records in the dataset. The entire distribution of all 647$a subject terms occurred in at least 0.02% of all records can be seen in Table 4.14.

Table 4.14: Distribution of subject terms used in 647$a: terms found in at least 0.02% of all records (n=10014)

647 $a Data Values	Count(oclcn)	Percentage
World War	35.0	0.35%
Revolution	8.0	0.08%

647 $a Data Values	Count(oclcn)	Percentage
American Revolution	7.0	0.07%
Jewish Holocaust	7.0	0.07%
American Civil War	4.0	0.04%
Battle of Iwo Jima	4.0	0.04%
Cold War	4.0	0.04%
German Occupation of Italy	3.0	0.03%
Vietnam War	3.0	0.03%
War on Terrorism	3.0	0.03%
Iraq War	3.0	0.03%
Afghan War	3.0	0.03%
Syrian Civil War	3.0	0.03%
Arab Spring	3.0	0.03%
Battle of Gettysburg	2.0	0.02%
Boston Massacre	2.0	0.02%
German Occupation of France	2.0	0.02%
Great Fire	2.0	0.02%
Hurricane Katrina	2.0	0.02%
Global Financial Crisis	2.0	0.02%
Napoleonic Wars	2.0	0.02%
Cuban Missile Crisis	2.0	0.02%
Oklahoma City Federal Building Bombing	2.0	0.02%
Eruption of Vesuvius	2.0	0.02%

4.1.3.5.6 648 $a Chronological Term (NR)

As reported above, the 10014 records in the sample included a total of 979 instances of field 648 subfield $a in 977 records (Table 4.5). A total of 172 unique terms contained in 648$a were observed in the dataset. The top three places in the distribution of chronological terms used in the dataset were taken by the following periods: 2011-2020 (2.48%), 1900-1999

(1.11%), and 2000-2099 (0.73%). The entire distribution of subject terms used in 648$a and found in at least 0.02% of all records is present in the table 4.15.

Table 4.15: Distribution of subject terms used in 648$a: terms found in at least 0.02% of all records (n=10014)

648 $a Data Values	Count(oclcn)	Percentage
2011-2020	248	2.48%
1900-1999	111	1.11%
2000-2099	73	0.73%
1800-1899	52	0.52%
1939-1945	43	0.43%
2001-2010	33	0.33%
1991-2000	24	0.24%
2020	21	0.21%
Since 2017	18	0.18%
To 1500	17	0.17%
1900-2099	14	0.14%
Since 2000	13	0.13%
1800-1999	13	0.13%
Since 1945	12	0.12%
1700-1799	10	0.10%
Geschichte 1985	10	0.10%
1789-1899	7	0.07%
Geschichte	6	0.06%
2019	6	0.06%
1500-1599	6	0.06%
1898-1951	5	0.05%
Since 2009	5	0.05%
Since 1991	5	0.05%
Since 1989	5	0.05%
1775-1865	5	0.05%
1775-1815	4	0.04%

648 $a Data Values	Count(oclcn)	Percentage
Since 2016	4	0.04%
1914-1918	4	0.04%
1600-1699	4	0.04%
1775-1783	4	0.04%
1981-1990	4	0.04%
1971-1980	4	0.04%
1951-1960	4	0.04%
1861-1865	3	0.03%
1789-1799	3	0.03%
711-1516	3	0.03%
1600-1799	3	0.03%
2008-2009	3	0.03%
2005	3	0.03%
Since 2011	3	0.03%
1500-1700	3	0.03%
1760-1820	3	0.03%
30-600	3	0.03%
1945	3	0.03%
Since 1990	2	0.02%
1996	2	0.02%
1837-1901	2	0.02%
1775-1789	2	0.02%
1288-1918	2	0.02%
1961-1975	2	0.02%
2001-2009	2	0.02%
2016	2	0.02%
2008	2	0.02%
2003-2011	2	0.02%
Since 1917	2	0.02%
Geschichte 1600-1800	2	0.02%

648 $a Data Values	Count(oclcn)	Percentage
1754-1763	2	0.02%
1700-1899	2	0.02%
To 1066	2	0.02%
Since 1980	2	0.02%
1991-2020	2	0.02%
1962	2	0.02%
Since 1948	2	0.02%
1783-1789	2	0.02%
1517-1648	2	0.02%
To 500	2	0.02%
1995	2	0.02%
To 332 B.C	2	0.02%

4.1.3.5.7 650$a Topical Term or Geographic Name as Entry Element (NR)

As reported above, the 10014 records in the sample included a total of 53243 instances of field 650 subfield $a in 9217 records (Table 4.5). A total of 10750 unique topical subject terms contained in 650 $a were observed in the dataset. Group of 33 subject topical terms were used between 1-5% of all records in the dataset. Another group consists of 10717 topical terms used in less than 1% of all records. The most widely used term in the distribution is “Automobiles”, which appeared in 5.11% of all records. This term was followed by the term “Jazz”, which appeared in 4.61% of all records. In third place was the term “Man-woman relationships”. This term was not far in terms of distribution from the previous one. It appeared in 4.53% of all records, with a distance of only eight records. Distribution of 650$a topical terms that were found in at least 1% of all records is presented in Table 4.16.

Table 4.16: Distribution of subject terms used in 650$a: terms found in at least 1% of all records (n=10750)

Field 650 Subfield $a Data Values	Count(oclcn)	Percentage
Automobiles	512	5.11%
Jazz	462	4.61%
Man-woman relationships	454	4.53%
Piano music	372	3.71%
Symphonies	338	3.38%
Murder	323	3.23%
Friendship	291	2.91%
Families	221	2.21%
Air	202	2.02%
Orchestral music	180	1.80%
Magic	179	1.79%
Politics and government	174	1.74%
Women	172	1.72%
Operas	165	1.65%
African Americans	160	1.60%
FICTION	144	1.44%
Jazz vocals	143	1.43%
Sonatas (Piano)	142	1.42%
Popular music	141	1.41%
Secrecy	133	1.33%
Environmental monitoring	131	1.31%
Climatic changes	118	1.18%
World War, 1939-1945	113	1.13%
Brothers and sisters	113	1.13%
Motor vehicles	113	1.13%
Schools	112	1.12%
Indians of North America	112	1.12%
Chamber music	112	1.12%

Field 650 Subfield $a Data Values	Count(oclcn)	Percentage
Missing persons	111	1.11%
Interpersonal relations	109	1.09%
Concertos (Piano)	106	1.06%
Presidents	104	1.04%
Children	103	1.03%

4.1.3.5.8 651 $a Geographic Name (NR)

As reported above, the 10014 records in the sample included a total of 6642 instances of field 651 subfield $a in 3688 records (Table 4.5). A total of 729 unique terms contained in 651$a were observed in the dataset. First place in the distribution is taken by geographic name of the United States. This term occurred in 20.94% of all records. Second and third places in the distribution were relatively distant from the first one and represent the same country—the term Great Britain placed second in the distribution (2.15%) and the term England followed in third place (1.6% of all records). Distribution of all geographical subject terms taken from 651$a that occurred in at least 0.2% of all records is available in Table 4.17.

Table 4.17: Distribution of subject terms used in 651$a: terms found in at least 0.2% of all records (n=10014)

651 $a Data Values	Count(oclcn)	Percentage
United States	2097.0	20.94%
Great Britain	215.0	2.15%
England	160.0	1.60%
France	108.0	1.08%
Germany	104.0	1.04%
China	97.0	0.97%
California	80.0	0.80%
Alaska	79.0	0.79%

651 $a Data Values	Count(oclcn)	Percentage
New York (State)	77.0	0.77%
Australia	65.0	0.65%
Japan	64.0	0.64%
India	60.0	0.60%
Mexico	54.0	0.54%
Spain	51.0	0.51%
Washington (D.C.)	51.0	0.51%
Europe	50.0	0.50%
Canada	49.0	0.49%
Italy	48.0	0.48%
Africa	48.0	0.48%
Washington (State)	46.0	0.46%
Russia (Federation)	44.0	0.44%
Scotland	43.0	0.43%
New York (N.Y.)	43.0	0.43%
European Union countries	39.0	0.39%
London (England)	38.0	0.38%
Middle East	37.0	0.37%
Massachusetts	35.0	0.35%
Egypt	35.0	0.35%
Louisiana	34.0	0.34%
Virginia	34.0	0.34%
Latin America	34.0	0.34%
Texas	33.0	0.33%
Ukraine	32.0	0.32%
Poland	32.0	0.32%
Florida	30.0	0.30%
Colorado	28.0	0.28%
Deutschland	28.0	0.28%
West Virginia	27.0	0.27%

651 $a Data Values	Count(oclcn)	Percentage
Brazil	25.0	0.25%
Cuba	24.0	0.24%
North America	24.0	0.24%
Narragansett Indian Tribe	24.0	0.24%
Turkey	23.0	0.23%
Iran	23.0	0.23%
Georgia	22.0	0.22%
Ohio	22.0	0.22%
Arizona	22.0	0.22%
Los Angeles (Calif.)	21.0	0.21%
Ireland	21.0	0.21%
Rhode Island	21.0	0.21%
Russia	20.0	0.20%
Israel	20.0	0.20%
Syria	20.0	0.20%

4.1.3.5.9 653 $a Uncontrolled Term (R)

As reported above, the 10014 records in the sample included a total of 1825 instances of field 653 subfield $a in 181 records (Table 4.5). A total of 1589 unique terms contained in 653$a were observed in the dataset. Distribution of uncontrolled subject terms was headed by the term “Australian”, appearing in 0.13% of all records. This was followed by the group of terms that appeared in 0.10% of all records. These terms were “Environment”, “Politics”, “Science and technology”, and “Fachpublikum/ Wissenschaft”, which translates from the German language as “Specialist audience / science”. The third place in the distribution was taken by “Hardback” and “Education (General)”. These terms occurred in 0.09% of all records in

the dataset. Table 4.18 presents the entire distribution of 653$a subject terms that occurred in at least 0.03% of all records.

Table 4.18: Distribution of subject terms used in 653$a: terms found in at least 0.03% of all records (n=10014)

653 $a Data Values	Count(oclcn)	Percentage
Australian	13.0	0.13%
Environment	10.0	0.10%
Politics	10.0	0.10%
Science and technology	10.0	0.10%
Fachpublikum/ Wissenschaft	10.0	0.10%
Hardback	9.0	0.09%
Education (General)	9.0	0.09%
Economy, business and finance	5.0	0.05%
Literature (General)	5.0	0.05%
Political institutions and public administration (General)	5.0	0.05%
Religion (General)	5.0	0.05%
Society	4.0	0.04%
Paperback / softback	4.0	0.04%
Soziologie	4.0	0.04%
Sociology	4.0	0.04%
n/a	4.0	0.04%
Environmental technology. Sanitary engineering	4.0	0.04%
Social sciences (General)	4.0	0.04%
International relations	3.0	0.03%
Chamber Music	3.0	0.03%
Disaster, accident and emergency incident	3.0	0.03%
Political science (General)	3.0	0.03%
Science (General)	3.0	0.03%
Economy	3.0	0.03%
Neoliberalism	3.0	0.03%

653 $a Data Values	Count(oclcn)	Percentage
Colonies and colonization. Emigration and immigration. International migration	3.0	0.03%
Philosophy (General)	3.0	0.03%
History (General)	3.0	0.03%
climate change	3.0	0.03%
temperature	3.0	0.03%
mechanical properties	3.0	0.03%

4.1.3.5.10 654 $a Focus Term (R)

As reported above, the 10014 records in the sample included a total of three instances of field 654 subfield $a in a single record (Table 4.5). Focus terms were present in only three records from the entire dataset (n=10014). In looking at the following presentation of the subject terms populated with help of BISAC ($2 bisacsh) controlled vocabulary it is evident that these strings of subject terms were populated without following bibliographic formats and standards guidelines.:

- “FICTION / Mystery & Detective / Historical”,
- “FICTION / Mystery & Detective / Traditional British”, and
- “FICTION / Historical”

Examples of the correct use of the bibliographic formats and standards for the field 654 as provided by Library of Congress are as follows:

- 654##cmalimestone.$2aat
- 654##cfbFrench colonialcvaportraitsczbUnited StatesczbNew Jersey.$2aat
- 654##cfbRomanesquecmbstonecrachurchesckarenovation.$2aat (https://www.loc.gov/marc/bibliographic/bd654.html)

4.1.3.5.11 655 $a Genre/Form Data or Focus Term (NR)

As reported above, the 7923 records in the sample included a total of 20550 instances of field 655 subfield $a (Table 4.4). A total of 714 unique terms contained in 655$a were observed in the dataset. A group of three genre subject terms (0.42% of total number of genre subject terms) that are used in the range between 10 and 29% of all records include electronic books, streaming audio and fiction. “Electronic books” are present as the main genre term in 29.12% of all records; “Streaming audio” is present in 15.92% of all records; “Fiction” as a main genre subject term present in 10.65% of all records in the dataset. The next group of genre subject terms is used in the range 1-5% of all records in the data set and includes 33 genre subject terms. The last group of subject terms is used in less than 1% of all records and include 678 genre subject terms. Distribution of the 36 genre subject terms that are found in at least 1% of all records is presented in Table 4.19.

Table 4.19: Distribution of subject terms used in 655$a: terms found in at least 1% of all records (n=10014)

655$a Data Values	Number of Records	Percentage
Electronic books	2916	29.12%
Streaming audio	1594	15.92%
Fiction	1066	10.65%
History	564	5.63%
Legislative hearings	544	5.43%
Audiobooks	541	5.40%
Juvenile works	476	4.75%
Romance fiction	428	4.27%
Feature films	384	3.83%
Legislative materials	377	3.76%
Zeitschrift	364	3.63%

655$a Data Values	Number of Records	Percentage
Thrillers (Fiction)	359	3.58%
Video recordings for the hearing impaired	350	3.50%
Detective and mystery fiction	317	3.17%
Biographies	302	3.02%
Historical fiction	293	2.93%
Documentary television programs	254	2.54%
Fiction films	251	2.51%
Drama	243	2.43%
Fantasy fiction	228	2.28%
Live sound recordings	174	1.74%
Criticism, interpretation, etc.	156	1.56%
Novels	145	1.45%
Conference papers and proceedings	140	1.40%
Picture books	134	1.34%
Graphic novels	133	1.33%
Domestic fiction	121	1.21%
Handbooks and manuals	121	1.21%
Film clips	120	1.20%
Autobiographies	114	1.14%
Statistics	108	1.08%
Mystery fiction	106	1.06%
Children's films	106	1.06%
Love stories	103	1.03%
Video recordings for people with visual disabilities	102	1.02%
Science fiction	102	1.02%

4.1.3.6 Metadata Record Networks Formed by Shared Subject Terms in 6XX Fields

In the dataset collected and analyzed in Stage 1, only four subject metadata fields were found to occur in at least 50% of all records and could therefore be meaningfully analyzed by

using Social Network Analysis measures. These subject fields are as follows:

- 650 -- 92.04%
- 655 -- 79.12%
- 050 -- 65.56%
- 082 -- 53.03%

Application of graph methods to the analysis of 050$a, 082$a, 650$a and 655$a did not reveal any connections between 050$a Library of Congress classification number and 082$a Dewey Decimal Classification number. Records with these fields contain only self-looped types of relationships, which indicates that each record with 050$a and 082$a in the dataset has only one edge parameter, this data value is unique and each of these records connect only to themselves. In contrast to 050$a and 082$a, networks built on 650$a topical subject heading and 655$a genre/form subject heading are interconnected within each network. A network built on 650$a subject terms has 9217 vertices, which is 92.04% of the total number of records in the dataset (n=10014). The fact that this parameter is easily observed by looking at the number of occurrences of this field in the dataset was one of the main criteria for choosing this field for network analysis. A network of records created based on shared data values in the 655$a subject term has a cardinality of 7923, which can be similarly observed by the percentage of its occurrences in the dataset (79.12%). The numbers of self-looped relationships in these networks are correspondingly 9217 and 7923 and equal the numbers of vertices. This means that each bibliographic record contains a single subject term that connects this record only to itself. The total number of relationships or edges in both networks is 12764 for 650$a

and 192971 for 655$a, which is not surprising because these MARC21 fields are repeatable according to bibliographic formats and standards.

The number of separate groups of connected vertices is represented by a parameter of *connected components.* The 650$a network has 8119 such connected groups and 655$a network has 6150; the maximum number of vertices in a connected component for the network of 650$a is 572 and for the network of 655$a it is 1482. This means that records that are connected through 650$a have a larger number of smaller groups than records connected through 655$a.

According to calculations, the average number of edges with the shortest possible distance from one vertex to another for the 650$a is 7.06 and is 2.90 for 655$a. These measures represent *geodesic distance* in the networks. If the shortest distance is the minimum number of edges needed to connect two vertices, the longest distance represents the diameter or maximum geodesic distance; and for both networks these parameters equal 21 for 650$a and 10 for 655$a respectively.

Graph density measure for 650$a is 0.000083514 and 0.00298 for 655$a. These parameters represent a portion of potential connections that could exist between two vertices and show how the network vertices are tightly connected. As such, the low parameters of both network densities indicate that most bibliographic records in the dataset are not tightly connected.

Modularity measure was calculated upon application of the motif network simplification. There are three types of motifs used for the graph simplification: fan motif, D-connector motif and clique motif. More information on the simplification is presented further

on in the *graph simplification* paragraph. Modularity is a measure that indicates a number of connections that come out of the group to connect the other vertices in a different group. These numbers for the network 650$a (0.34) and 655$a (0.26) are relatively low, so these indicate that the groups in both networks are well established.

Table 4.20 represents the network analysis metrics for four networks: 050$a, 082$a, 650$a and 655$a.

Table 4.20: Network analysis measures for subject metadata fields 050, 082, 650, and 655

Graph Metric	050$a	082$a	655$a	650$a
Graph Type	Undirected	Undirected	Undirected	Undirected
Vertices/Cardinality	6565	5310	7923	9217
Unique Edges	6565	5310	9359	12764
Edges With Duplicates	0	0	183612	0
Total Edges	6565	5310	192971	12764
Self-Loops	6565	5310	7923	9217
Connected Components	6565	5310	6150	8119
Single-Vertex Connected Components	6565	5310	6114	7839
Maximum Vertices in a Connected Component	1	1	1482	572
Maximum Edges in a Connected Component	1	1	177876	2678
Maximum Geodesic Distance (Diameter)	0	0	10	21
Average Geodesic Distance	0	0	2.908345	7.06947
Graph Density	0	0	0.002971099	0.000083514
Modularity	n/a	n/a	0.261417	0.400616
NodeXL Pro Version	1.0.1.433	1.0.1.433	1.0.1.433	1.0.1.433

Analysis of other graph metrics, such as vertex degree, betweenness centrality, closeness centrality, eigenvector centrality, page rank and clustering coefficient can be

discussed only in the context of distributions of these metrics among each vertex.

Vertex degree is a graph measure that represents the number of edges coming out of a single vertex. From the 9217 records that contain 650$a only two records (vertices) have 44 (maximum) number of edges and seven records have a degree value ranging from 31 to 40. A total of 81 records have vertex degree values ranging from 21 to 30; 185 records out of 9217 have vertex degree values ranging from 11-20; and 8942 records from the same subset have vertex degree values ranging from 2 to 10.

Betweenness centrality, the distribution of all values for the network of 650$a is in the range from 0 to 70100 and majority of all records (8947 out of 9217) have betweenness centrality values ranging between 0 and 100. Smaller portions (270, around 2.9% of total number of vertices) of records have larger values of this measure and take a central position in the network.

Figure 4.5 represents distribution of vertex degree and betweenness centrality values for the network of records with 650$a.

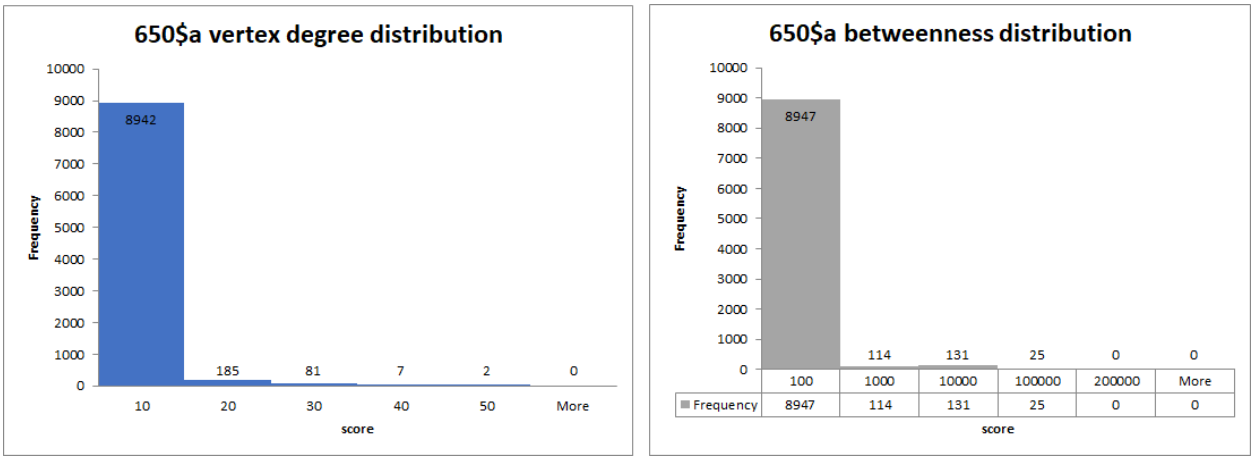


Figure 4.5: Two histograms of degree and betweenness centrality values distribution among records that contain 650$a

Closeness centrality -- average distance to all vertices in the network. Distribution in the network of 650$a shows that the measures allocated in the range between 0 and 1 and majority of all records (7839) have 0 closeness centrality. A total of 968 records involved in the network 650$a have closeness centrality up to 0.6, while the rest of all records (410) have relatively higher values in the range from 0.6 to 1.

Another measure that defines a position of a vertex to all network participants by calculating a weight based on a distance is *eigenvector centrality or "eigencentrality"*. All eigenvector centrality values in the graph of 650$a are close to zero. There are only 30 records that insignificantly deviate from this picture revealing their eigenvector centrality measures in the range up to 0.04, which can be ignored.

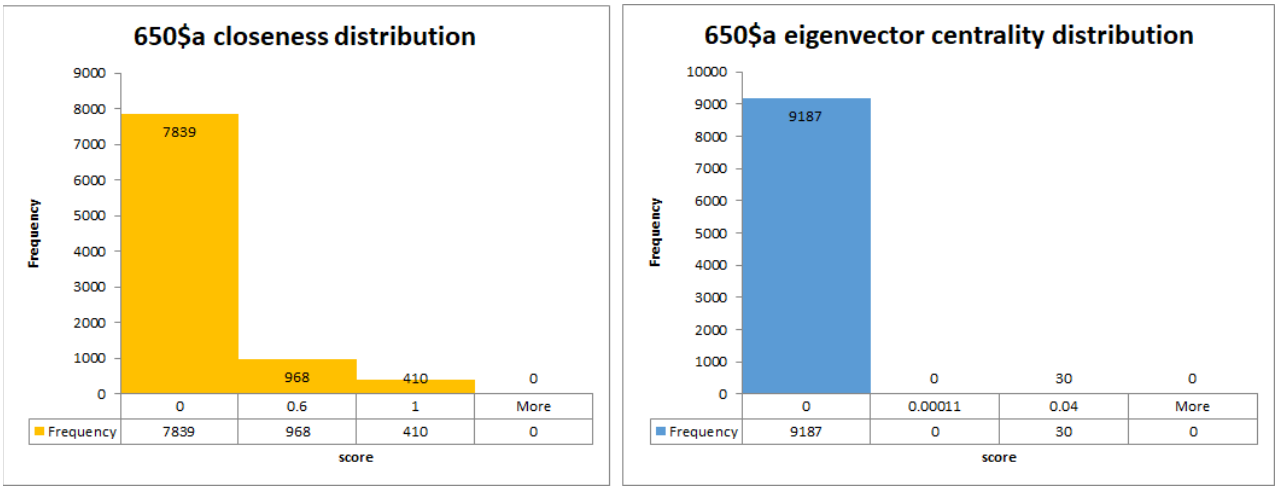


Figure 4.6: Two histograms of closeness centrality and eigencentrality values distribution among records that contain 650$a

In contrast to degree centrality that allows for network evaluation through shortest distances between vertices and eigenvector centrality that measures all distances, *page rank* is an eigenvector-based algorithm that scores the relative importance of all nodes in the network. This measure works better with directed networks; and in an undirected network of records

based on shared data values in the 650$a, the page rank values fall into the 0.3 - 3.15 range. A majority of all records (8781) have their page rank values between 0.5 and 1.

Clustering coefficient differs from measures of centrality and is similar to network density measures. A total of 8422 records showed 0 values of clustering coefficient, while 795 records had their coefficients of clustering in the range between 0.5 to 1.

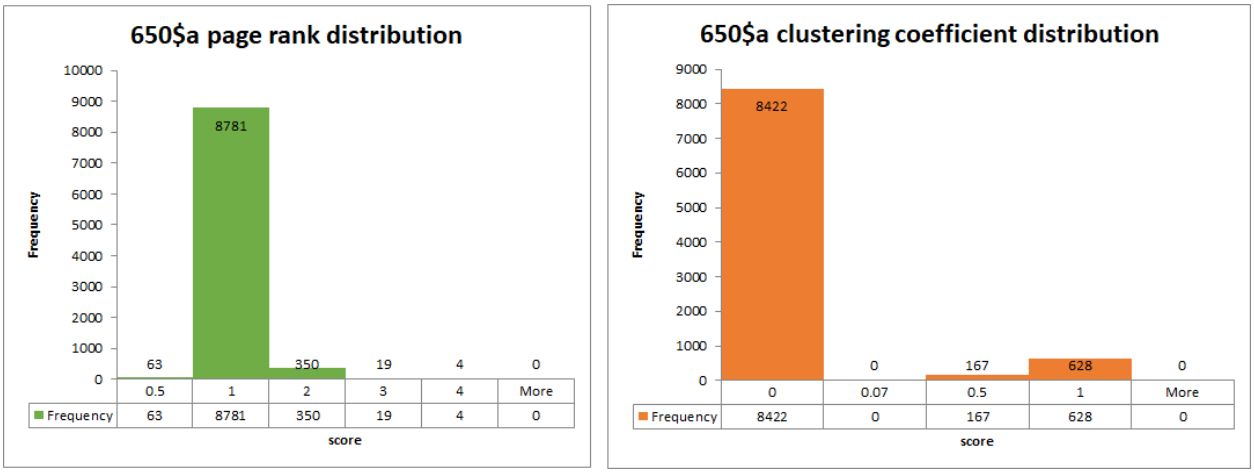


Figure 4.7: Two histograms of page rank and clustering coefficient values distribution among records that contain 650$a

In the 655$a network, the distribution of *vertex degree* values is not as steep and spreads more evenly than the distribution of vertex degree values in 650$a network. Both histograms skewed to the left, representing a large number of records that have vertex degree values ranging from 2 to 10. However, in contrast to the distribution of 650$a vertex degree values where the entire distribution falls into the range between 2 and 44, the distribution of degree values of 655$a is widely spread between 2 and 487.

Betweenness centrality measures for 655$a network members is skewed to the left and a majority of records (7197) have these measures ranging from 0 to 100. A total of 397 records have betweenness centrality ranging from 100 to 1000. For 294 records betweenness centrality

falls in the range of 1000 to 10000. For 35 records, betweenness centrality from 100000 to 153133.34 was observed.

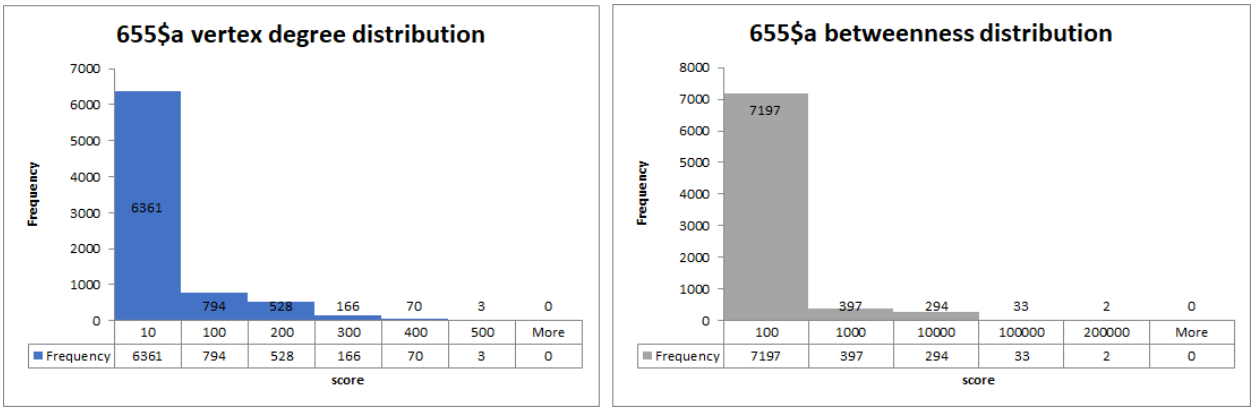
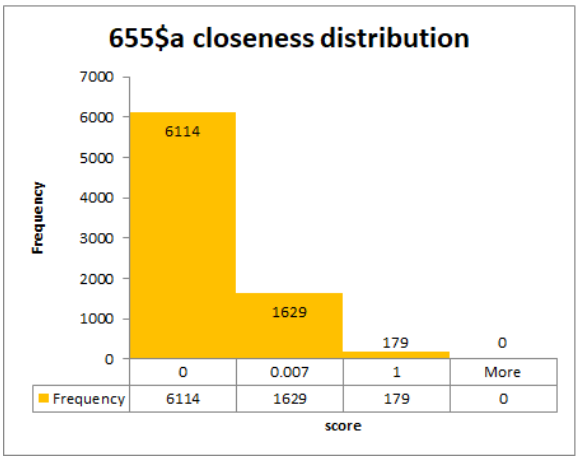


Figure 4.8: Two histograms of degree and betweenness centrality values distribution among records that contain 655$a

Closeness centrality for the network of 655$a spreads more evenly. A total of 6114 records have 0 closeness centrality; 1629 records have their measures ranging from 0 to 0.007, where 0.00676 is the average closeness centrality for the network; and 179 records have maximal closeness centrality ranging from 0.07 to 1.

Eigencentrality measures in the network of 655$a were spread unevenly: 6605 records have 0 values; 556 records have their values around 0.00013, which is average eigenvector centrality and the remaining 762 records have maximal measures ranging from 0.00013 to 0.004.

Page rank distribution shows that the majority of all vertices in the network of 655$a (6863 records) have their page rank values around 1. The second group of records (807 records) in the score of page rank distribution have page rank values ranging from 1 to 2. A total of 228 records have their page rank values in the range of up to 0.5. This is the third group of records in the score of page rank distribution.



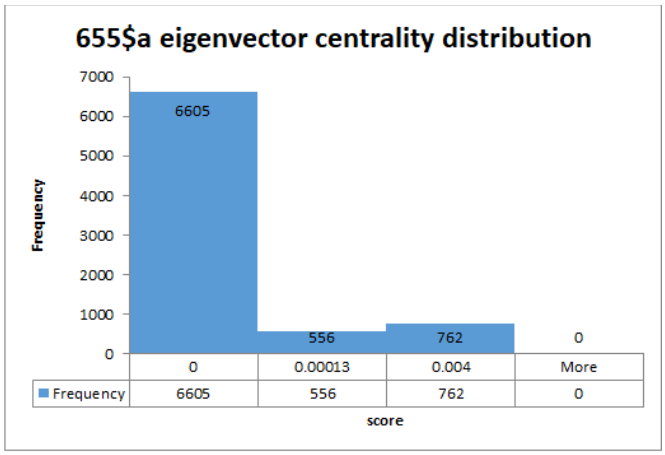
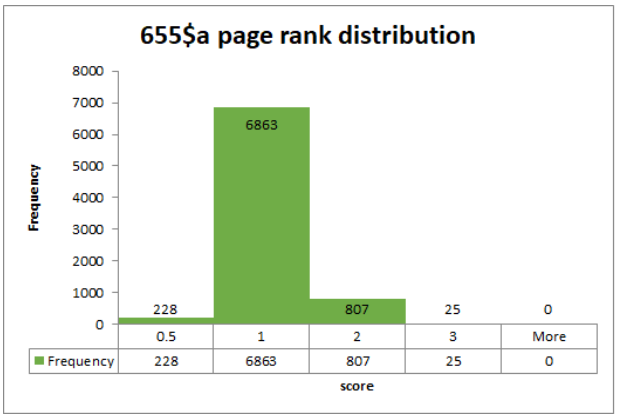


Figure 4.9: Two histograms of closeness centrality and eigencentrality values distribution among records that contain 655$a

Clustering coefficient values in the 655$a network were distributed between two major groups of records: 6168 records have a clustering coefficient of 0; for 1754 records have clustering coefficients ranged from 0.19 to 1. Moreover, 946 records out of 1754 have their clustering coefficients closer to 1 and only one record has a clustering coefficient around the average value of 0.19.



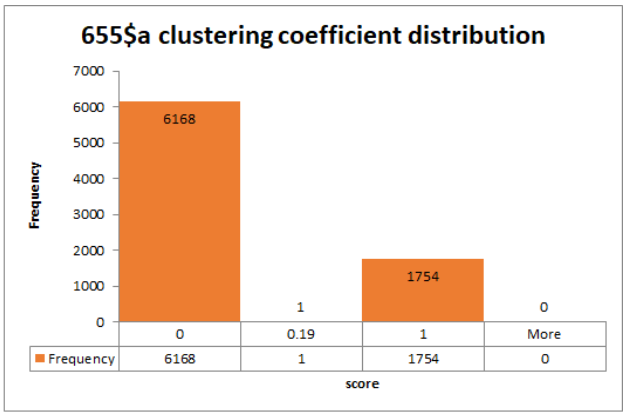


Figure 4.10: Two histograms of page rank and clustering coefficient values distribution among records that contain 655$a

4.1.3.7 Graph Simplifications

A large number of records have a subject heading that creates self-looped relationships

in the containing record, meaning those records do not connect to any other records except themselves. Self-loops are created when network vertices link to themselves and usually appear as circles on a network graph, creating clutter in the visualizations. Such clutter disturbs visual perception and understanding. If there is a need to eliminate visualization of self-loops, they can be filtered through the "visibility" column in the NodeXL worksheet that represents network edges, by using a simple formula: =IF([@[Vertex 1]]=[@[Vertex 2]],0,1). This expression checks if Vertex1 equals Vertex2 and then places a zero in the cell, which acts as if the data has been deleted. If Vertex1 does not equal Vertex2, the formula places a "1" and the edge is visible. Figure 4.11 below represents two 650$a graphs with and without self-loops:

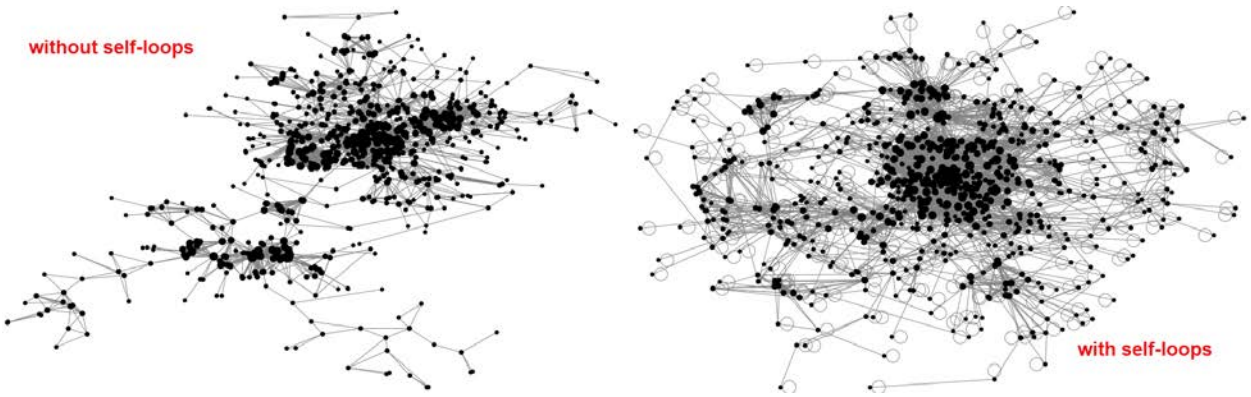


Figure 4.11: 650$a graph with and without self-loops

Another type of simplification that can be applied to the network graph is grouping vertices by types and forms of relations, repeating motifs. This type of simplification increases readability of graphs' visualizations (Dunne & Shneiderman, 2012). Such aggregation by simplifying common repeating network structures is implemented in NodeXL. There are three types of common repeating network structures available for motif simplification: fan, connector, and clique. Figure 4.12 presents motif simplification settings in NodeXL.

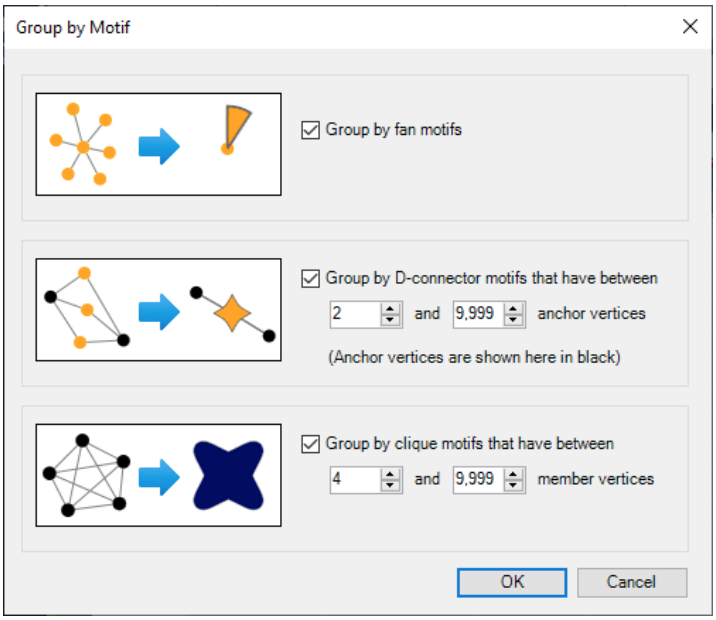
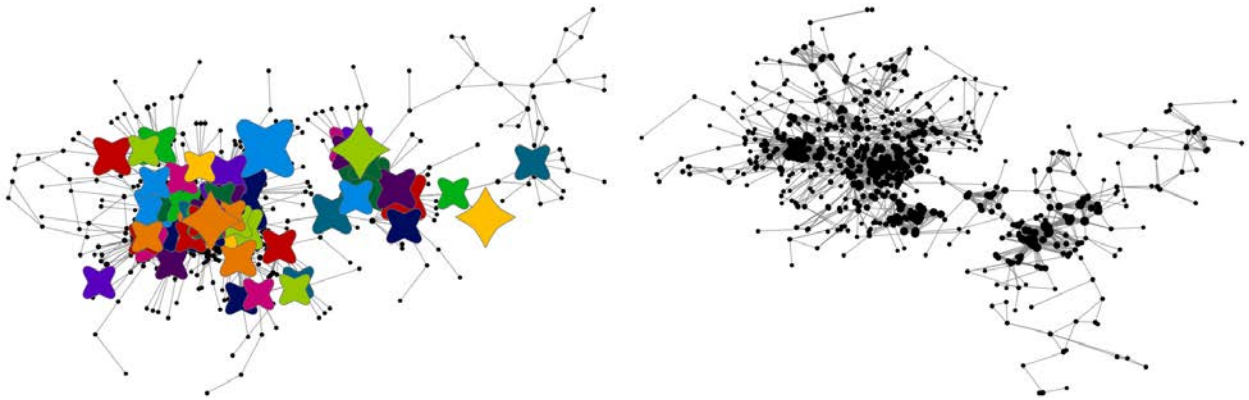


Figure 4.12: Default parameters of motif simplification used in NodeXL

Application of the Harel-Koren fast multi-scale layout and default motif simplification settings provided by NodeXL created visualizations for both 650$a and 655$a networks (Figure 4.13). The reasons for implementation of this Harel-Koren layout for visualization are as follows: the algorithm works extremely fast and provides aesthetic representation of undirected networks with straight-line connections (Harel & Koren, 2001, p.2).

Figure 4.12. Represents 650$a (at the top) and 655$a (at the bottom) networks with and without motif simplification.



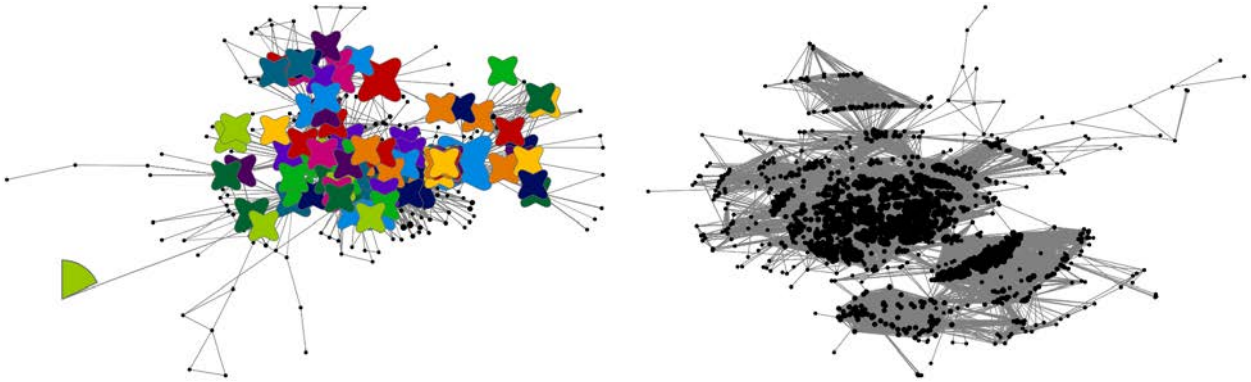


Figure 4.13: 650$a (top) and 655$a (bottom) graphs with motif simplification and without

4.2 Findings Obtained in Stage 2

Stage 2 in this study design was intended to supplement and refine the findings of Stage 1, by shifting the focus of subject metadata analysis in MARC 21 bibliographic records from the database level to the record level. The purposive subsample of records analyzed in Stage 2 included the 100 records with highest numbers of holdings as of the time of data collection, and the full level of encoding as indicated by code I or blank in the ELvl subfield of the fixed field (MARC Leader, byte 17).

The remainder of section 4.2 presents the findings obtained in the Stage 2 of this study. First, general characteristics of the records are reported: distribution of the sample by encoding level codes, by the number of holdings, by the language of cataloging, by institutions that created records, types and languages of materials represented in records, etc. Next, the application of various subject fields and subfields, including Linked-Data supporting data elements, is presented. This is followed by a presentation of the findings regarding co-occurrence between subject fields and subfields intended for the same type of information within a record and the use of an option to indicate primary and secondary subject terms with

the 1st indicator in fields 650, 653, 654, and 655. Finally, the findings regarding the application of various controlled vocabularies for subject representation are reported.

4.2.1 General Characteristics

As indicated by the code "I" in the ELvl subfield of the fixed field, fifty-one of the 100 records analyzed in Stage 2 (51%) were

> [t]he most complete MARC record[s] created from an inspection of the material [that conform] to OCLC full-level input standards, which are based on the MARC 21 Format for Bibliographic Data, National Level Full and Minimal Requirements [and] may also conform to the BIBCO Standard Record (BSR) RDA Metadata Application Profile or the CONSER Standard Record (CSR) RDA Metadata Application Profile. https://www.oclc.org/bibformats/en/fixedfield/elvl.html)

An additional 49 records (49%) were the next most complete full-encoding- level records by OCLC participants as indicated by the code "I" in the ELvl subfield of the fixed field. Due to the high level of use of these records in the copy cataloging, all of them have been edited at least once after their creation, and 97% of them were edited multiple times by multiple institutions as indicated by the field 040 subfield $d Modifying agency.

As indicated by the code "lccopycat" in the MARC field 042 Authentication Code, 40 of the records with the code blank in ELvl were created as part of the Library of Congress Copy Cataloging program. When this code is included in the 042, this means that the record was created by the Library of Congress, "based on another cataloging agency's record, [and h]eadings are verified with the relevant authority file, except those subject headings not from Library of Congress Subject Headings." (https://www.oclc.org/bibformats/en/0xx/042.html). Nine additional records in the sample of 100 had a code "pcc" in field 042 which represents the Program for Cooperative Cataloging and means that the

> "[r]ecord is authenticated under the auspices of the program", "[a]ll name and series headings have been verified through the appropriate national level authority file", "[a]uthority records have been created if they do not already exist", and "[s]ubject headings are checked for authorized forms and combinations supported by the relevant authority." (https://www.oclc.org/bibformats/en/0xx/042.html)

The number of holdings in the sample of 100 records ranged between 571 to 1514. The records in the subsample analyzed in Stage 2 represented 4 material types—books (n=83), continuing resources (n=1), sound recordings (n=3), and visual materials (n=13).

The records were found to represent only the English-language materials. The language of cataloging, as indicated by data value in subfield $b of field 040, was also found to be English for all 100 records in the sample. However, records were created by 31 different institutions in six countries: Australia, Canada, Hong Kong, New Zealand, United Kingdom, and a variety of geographic locations in the United States. The sample also included records created by eight types of institutions: academic libraries (e.g., University of Hong Kong library), school libraries (e.g., Anchorage school district library in Alaska), public libraries (e.g., Winnipeg Public Library in Canada), state/national libraries (e.g., Libraries Australia), federal/national government agencies (e.g., US National Library of Medicine), associations/foundations (e.g., LIBRARIES HOROWHENUA in the New Zealand), vendors (e.g., Baker & Taylor), and other corporate/business organizations (e.g., NetLibrary Incorporated). The number of records in the sample created by each institution ranged from one for 16 institutions to 29 for a single institution (the Baker & Taylor Inc., Electronic Business and Information Services Unit), with the average number of 3.225 records per institution.

4.2.2 Application of Subject Metadata Fields

Table 4.21 presents results of an in-depth manual content analysis of 100 records

analyzed in the Stage 2 with regards to central tendency measures and variability measures in the level of application of repeatable subject fields and non-repeatable field 043. A total of 18 MARC 21 bibliographic fields intended for subject representation were observed in this sample. Five additional subject fields that occurred in the main dataset were not observed in any of the records in the stage 2 purposive sample. One of these non-observed fields includes the 653 Uncontrolled Term which is normally not included in the full-level cataloging records as it is replaced in these records with controlled access points. Four other subject metadata fields not observed in this smaller purposive sample of full-level of cataloging records were observed in relatively low proportion in the larger dataset from which this sample was derived. These include the 070, 086, 096, and 630.

As shown in Table 4.21, only field 650 Subject Added Entry - Topical Term was included in all 100 records. A minimum of two instances and a maximum of 46 instances of that field were included in every record. Three other fields were included in the vast majority (98%) of records in this sample: 050 Library of Congress Call Number, 082 Dewey Decimal Classification Number, and 655 Index Term - Genre/Form. The level of application of the remaining 14 subject metadata fields observed in this dataset ranged widely between 1% of records (fields 080 Universal Decimal Classification Number, 092 Locally Assigned Dewey Call Number, and 654 Subject Added Entry--Faceted Topical Terms) and 59% of records for field 651 Subject Added Entry--Geographic Name.

Data in Table 4.21 demonstrates that fields 650, 655, and 651 had the highest number of instances for the records in which they were included, which is not surprising. The average number of instances of these fields was 13.35, 6.93, and 2.54.

Table 4.21: Statistical indicators for subject metadata fields observed in Stage 2 sample (n=100)

	TOTAL instances in 100 records	no. of records with 1+ instance	average no. of instances per record if present	median no. of instances per record	mode no. of instances per record	max no. of instances per record	min no. of instances per record	variance	standard deviation
043 (NR)	53	53	1	1	0	1	0	0.251616	0.501614
050 (R)	100	98	1.02040816	1	1	2	0	0.040404	0.201008
055 (R)	5	5	1	0	0	1	0	0.04798	0.219043
060 (R)	3	3	1	0	0	1	0	0.029394	0.171447
072 (R)	3	2	1.5	0	0	2	0	0.049596	0.222702
080 (R)	1	1	1	0	0	1	0	0.01	0.1
082 (R)	100	98	1.02040816	1	1	2	0	0.040404	0.201008
084 (R)	13	12	1.08333333	0	0	2	0	0.134444	0.366667
092 (R)	1	1	1	0	0	1	0	0.01	0.1
600 (R)	66	28	2.35714286	0	0	6	0	1.56	1.249
610 (R)	14	7	2	0	0	2	0	0.26303	0.512865
611 (R)	5	4	1.25	0	0	2	0	0.068182	0.261116
647 (R)	7	6	1.16666667	0	0	2	0	0.08596	0.293189
648 (R)	16	16	1	0	0	1	0	0.135758	0.368453
650 (R)	1335	100	13.35	12	12	46	2	56.39141	7.509422
651 (R)	150	59	2.54237288	1	0	8	0	2.858586	1.690735
654 (R)	3	1	3	0	0	3	0	0.09	0.3
655 (R)	679	98	6.92857143	7	8	19	0	14.51101	3.809332

The median was also high for two of them—12 for 650 and seven for 655—while it was moderate for field 651 at one. The highest mode was observed for 650 and 655 (12 and 8 respectively), while for 651 it was 0, as it was for most other fields except fields 050 and 082 (a mode of 1). One more field, the 600 Subject Added Entry--Personal Name field, was on average included in more than two instances in the records where it was observed (with a mean of 2.34).

The highest level of variability was also observed in four fields—650, 655, 651, and 600—as both variance and variability indicators for each were above 1.0: between 1.56 and 56.39 for variance and between 1.249 and 7.51 for standard deviation (see Table 4.21). For the remaining 14 subject metadata fields observed in the records in this purposive sample, the variability was moderate, with both variance and standard deviation indicators below 0.6.

Table 4.22 shows distribution of the number of different subject fields (from the 18 observed in the sample) per record, as well as distribution of the total number of instances of various subject fields per record. On average, a total of six different subject fields were observed in the record (the mean was 5.99, and both median and mode numbers were 6), with the minimum of three and the maximum of 10. The total number of instances of all subject fields combined ranged much more substantially: from 5 to 68 per record. The mean, median, and mode for the number of instances of all subject fields combined per record were similar to each other: 25.7, 26, and 27 respectively. The variability measure analysis demonstrated high variability for the total number of subject field instances per record (variance of 80.29 and standard deviation of 8.96) and relatively moderate variability for the number of subject fields (variance of 2.52 and standard deviation of 1.59)

Section 4.2.4 further reports on the results of the analysis into the way in which the subject fields and some subfields co-occurred in the records.

Table 4.22: Number of subject fields and field instances per record (n=100)

	mean	median	mode	maximum	minimum	variance	standard deviation
number of different subject fields per record	5.99	6	6	10	3	2.51505	1.585891
total number of instances of all subject fields per record	25.7	26	27	68	5	80.2929	8.960632

4.2.3 Application of Subject Metadata Subfields, Including Linked-Data-Enabling

Application of Linked-Data enabling subfields was also evaluated. Subfield $2, which specifies controlled vocabulary from which the term is taken, was found to be used consistently in a variety of 6XX subject fields. It was also found to be used (much less consistently) in classification fields 072 Subject Category Code and 084 Other Classification Number.

The vast majority of MARC records (98%) in the analyzed sample of 100 records included one or more instances of the most important Linked-Data-enabling subfield $0 Authority Record Control Number or Standard Number, according to Shieh and Reese (2015) and others. This was true for a variety of 6XX fields: 600, 610, 611, 650, 651, and 655. Subfield $0 is also defined by the MARC21 Bibliographic Format for field 648. However, none of the instances of the 648 in the 100 full encoded level records, and with the highest numbers of holdings, were found to include it.

Linked-Data enabling subfield $0 was observed only in the instances of 6XX fields that contained FAST headings. It was not observed in any instances of 6XX fields with second

indicators 0, 1, 2, or 6, which are intended for holding terms from the following controlled vocabularies: Library of Congress Subject Headings (LCSH), LC subject heading for children's literature, Medical Subject Headings (MeSH), and Répertoire de Vedettes-Matière. As well, it was not observed in any instances of 6XX fields with a second indicator of 4, which represents "Source not specified". Finally, for other controlled vocabularies commonly used and indicated in subfield $2 of 6XX fields with second indicator 7 -- BISACSH, GSAFD, LCGFT, and SEARS -- as well as for infrequently applied controlled vocabularies GND and GTT, subfield $0 was not included in any of the 6XX field instances in the sample. Section 4.2.6 below includes details on FAST and other controlled vocabularies' application and section 4.2.7 discusses the co-occurrence of controlled vocabularies within the records in the sample.

Records in the analyzed sample did not include any instances of two other Linked-Data-enabling subfields of the MARC 21 subject metadata fields -- $1 Real World Object URI and $4 Relationship.

The application of three additional subfields—repeatable subfield $a in non-repeatable field 043, and 6XX subfield $z Geographic Subdivision and subfield $y Chronological Subdivision—was examined in Stage 2 of this study and compared to the application of other subject metadata elements in MARC 21 bibliographic records that are intended for representing chronological and geographical aboutness of information objects. Table 4.23 presents results of the analysis into the overall level of application of these subfields.

The largest number of instances in a sample was observed for 6XX subfield $z: it occurred 72 times in a total of 33% of records. Subfield $a in the field 043 occurred in a larger proportion of records (53%) but in a smaller overall number of instances at 62. Subfield $y in

various 6XX fields was the least frequently used: 16 instances total were observed in 9% of records. As shown in Table 4.23, the average number of instances of a subfield was the lowest (1.169811) for 043 $a, followed by 6XX $y (1.777778), and 6XX $z (2.181818). The mode number of instances was 0 for all three subfields, and only one subfield (043 $a) had a median number of instances above zero. The widest range in the level of application was observed for 6XX $z: the number of instances per record ranged from 0 to 9. Similarly, the variability measures -- variance and standard variation -- were the highest for 6XX $z.

Section 4.2.7 below discusses results of the analysis and co-occurrence of these and other fields and subfields

Table 4.23: Statistical indicators for three subject metadata subfields (n=100)

	% of records with 1+ instance	TOTAL instances in 100 records	average no. of instances per record if present	median no. of instances per record	mode no. of instances per record	max no. of instances per record	min no. of instances per record	variance	standard deviation
043 $a	53%	62	1.169811	1	0	4	0	0.693112	0.480404
6XX $z	33%	72	2.181818	0	0	9	0	1.484465	2.203636
6XX $y	9%	16	1.777778	0	0	7	0	0.76171	0.580202

4.2.4 Co-Occurrence of Fields and Subfields

Close examination of each of the 100 records in the purposive sample as part of Stage 2 revealed that certain pairs of subject metadata elements (fields and/or subfields) carrying similar or related types of information often co-occurred. As shown in Table 4.24, most records included two classification fields: 050 Library of Congress Call Number and 082 Dewey Decimal Classification Number. There were no records in the sample which excluded both fields, and 94% of records in the sample included both fields, for the total correlation of 94%. The co-

occurrence between these two fields was the highest among all the subject data elements, except a pair consisting of two 6XX fields; 650 and 655. All but one record in the sample (99%) included both 650 and 655 fields. Co-occurrences between other 6XXs subject added entry and index term fields (e.g., 650 and 651, 600 and 610, etc.) and other than 050 and 082 pairs of classification fields were much lower and is not included in Table 4.24.

Analyses conducted in Stage 2 indicate a high level of correlation in the presence (and absence) in the record of the following pairs of subject metadata fields and subfields combinations:

- Fields 050 and 082 occurred together in 94% of records and there were no records in which both fields were missing (overall correlation of 0.94).
- Field 648 and subfield 6XX $y occurred together in 83% of records and were both absent in additional 8% of records, for an overall correlation of 0.91.
- Fields 043 and 651 occurred together in 39% of records and were both absent in additional 51% of records, for an overall correlation of 0.9.
- Field 043 and subfield 6XX $z occurred together in 43% of records and were both absent in additional 29% of records, for an overall correlation of 0.72.
- Fields 648 and 611 occurred together in 4% of records and were both absent in additional 84% of records, for an overall correlation of 0.88.
- Fields 043 and 611 occurred together in 3% of records and were both absent in additional 46% of records, for an overall correlation of 0.49.

Table 4.24: Cooccurrence for selected subject metadata fields/subfields pairs

pairs of fields / subfields	% of records with 0 instances of both fields/subfields	% of records with 1+ instances of both fields/subfields	overall correlation
650 and 655	0%	98%	0.98
043 and 6XX $z	29%	43%	0.72
043 and 651	51%	39%	0.90
043 and 611 $c	46%	3%	0.49

pairs of fields / subfields	% of records with 0 instances of both fields/subfields	% of records with 1+ instances of both fields/subfields	overall correlation
648 and 6XX $y	8%	83%	0.91
648 to 611 $d	4%	84%	0.88
050 to 082	0%	94%	0.94

4.2.5 Use of an Option to Indicate Primary and Secondary Subject Headings

As shown in Table 4.25, only 2% of all records analyzed in Stage 2 made use of the option to indicate primary and secondary subject terms using other-than-default values for the 1st indicator in 6XX field. That option is enabled by MARC 21 Bibliographic Standard for five subject metadata fields: 650, 653, 654, 655, and 690. No instances of fields 690 and 653 were observed in the purposive sample examined in Stage 2. The only record that contained field 654 used the default blank 1st field indicator. Similarly, none of the numerous instances of field 655 included in 98% of all records in the sample, used the non-blank 1st indicator.

Only 2% of records within a total of 9 instances used the non-blank 1st indicator in field 650. One of these two records, created by the US National Library of Medicine with the ELvl code blank (highest level of cataloging), uses non-blank 1st indicator in all six instances of field 650: 1st indicator 1 for the primary subject heading “Macular Degeneration$ -- diet therapy” and 1st indicator 2 for the four secondary subject headings: “Macular Degeneration -- genetics”, “Macular Degeneration -- prevention & control”, “Diet, Mediterranean”, “Cognition”, and “Aged”. This record happened to have the highest number of holdings in the dataset: 1517. The second record of the two, created by the Netlibrary Incorporated with the second highest possible level of cataloging (as indicated by ELvl code I), used 1st indicator 1 for three out of its 6 instances of field 650. All three instances of the field 650 with 1st indicator 1 contained the

terms from the GOO-trefwoorden thesaurus by Koninklijke Bibliotheek in the Netherlands (code “gtt”). This second record had a total of 870 holdings.

Table 4.25: Application of non-empty 1st Feld indicator (n=100)

	TOTAL instances in 100 records	no. of records with 1+ instance	average no. of instances per record if present	median no. of instances per record	mode no. of instances per record	max no. of instances per record	min no. of instances per record	variance	standard deviation
650 with 1st indicator non-blank	9	2	4.5	0	0	6	0	0.446364	0.6681045
Indicator 1 (primary heading)	4	2	2.5	2	1, 3	3	1	2	1.4142
Indicator 2 (secondary heading)	5	1	5	n/a	n/a	5	5	n/a	n/a

4.2.6 Application of Controlled Vocabularies

Table 4.26 shows that, as indicated by second indicator 0 in the 6XX, that within fields intended for holding controlled vocabulary terms, the Library of Congress Subject Headings was observed at the highest level of subject representation overall. All seven records in the sample that contained field 610 included at least one instance of this field with a second indicator of 0. Ninety-nine percent of records containing field 650 included at least one instance of this field with a second indicator of 0. Similarly, 27 out of 28 records (96%) containing field 600 included at least one instance of this field with a second indicator of 0. That was also true, although to the lesser extent, for field 651 (61% or 36 out of 59 records with the field), and field 655 (35%

or 34 out of 98 records with this field). The only 6XX MARC bibliographic metadata field for which the 2nd indicator is defined and that was observed in this sample—field 611 Subject Added Entry--Meeting Name—had a low number of application of the Library of Congress Subject Headings (LCSH) controlled vocabulary as a 2nd indicator of 0 was only included in one record.

The highest level of use of the LCSH controlled vocabulary in the 6XX fields that was observed in the purposive sample in Stage 2 occurred in fields 650, 600, and 651 (Table 4.26). In these three fields, an average of 4.05, 1.44, and 1.13 instances of the field with a second indicator of 0 was observed respectively. Median and mode number of instances of a field with this second indicator equal zero for all but one of the 6XX fields. For the field 650, the median was four and the mean was three.

The highest variability in the level of application of LCSH based on the number of field instances with a second indicator of 0, was observed for field 650 (variance of 5.04 and standard deviation of 2.25). For the remaining five 6XX fields, both variance and standard deviation of the level of application of LCSH were 0.76.

Based on the data values in the 6XX subfield $2, seven additional non-LCSH controlled vocabularies for subject representation were observed in the records in this sample (Table 4.28). Level of application of these controlled vocabularies varied from only one record (1% of the sample) for the GOO-trefwoorden thesaurus by Koninklijke Bibliotheek in the Netherlands (code “gtt”) to 98% of records for Faceted Application of Subject Terminology (code “fast”).

Table 4.26: Level of application of the Library of Congress Subject Headings controlled vocabulary (n=100)

	TOTAL instances in 100 records	no. of records with 1+ instance	average no. of instances per record if present	median no. of instances per record	mode no. of instances per record	max no. of instances per record	min no. of instances per record	variance	standard deviation
600 (R)	66	28	2.35714286	0	0	6	0	1.56	1.249
including LCSH: 600 00 or 600 10	39	27	1.44444444	0	0	3	0	0.563535	0.75069
610 (R)	14	7	2	0	0	2	0	0.26303	0.512865
including LCSH: 610_10 or 610_20	7	7	1	0	0	1	0	0.065758	0.256432
611 (R)	5	4	1.25	0	0	2	0	0.068182	0.261116
including LCSH: 611_20	1	1	1	0	0	1	0	0.01	0.1
650 (R)	1335	100	13.35	12	12	46	2	56.39141	7.509422
including LCSH: 650 _0	401	99	4.05050505	4	3	15	0	5.040303	2.245062
651 (R)	150	59	2.54237288	1	0	8	0	2.858586	1.690735
including LCSH: 651 _0	41	36	1.13888889	0	0	2	0	0.345354	0.587668
655 (R)	679	98	6.92857143	7	8	19	0	14.51101	3.809332
including LCSH: 655_0	35	34	1.02941176	0	0	2	0	0.25	0.5

Table 4.27: Level of application of the non-LCSH controlled vocabularies based on 6XX 2nd indicator values (n=100)

	TOTAL instances in 100 records	no. of records with 1+ instance	average no. of instances per record if present	median no. of instances per record	mode no. of instances per record	max no. of instances per record	min no. of instances per record	variance	standard deviation
6XX with 2nd ind 1	37	90	2.432432	0	0	6	0	2.151515	1.466804
6XX with 2nd ind 2	4	12	3	0	0	8	0	0.692525	0.832181
6XX with 2nd ind 4	12	18	1.5	0	0	4	0	0.351111	0.592546
6XX with 2nd ind 6	2	4	2	0	0	3	0	0.099394	0.315268

Table 4.28: Application of the non-LCSH controlled vocabularies based on subfield $2 data value in 6XX fields with 2nd indicator 7 (n=100)

Non-LCSH controlled vocabulary code	Vocabulary name	no. of records with 1 or more instance of this controlled vocabulary term in 6XX fields
bisacsh	BISAC Subject Headings List	72
fast	Faceted Application of Subject Terminology (FAST)	98
gnd	Gemeinsame Normdatei	3
gsafd	Guidelines on Subject Access to Individual Works of Fiction, Drama, Etc.	56
gtt	GOO-trefwoorden thesaurus	1
lcgft	Library of Congress genre/form terms for library and archival materials	91
sears	Sears List of Subject Headings	64

4.2.7 Co-Occurrence of Controlled Vocabularies

The manual in-depth content analysis of the 100 most widely shared OCLC WorldCat MARC 21 bibliographic records by library catalogs revealed that certain pairs of subject controlled vocabularies were often used in the same records together. Table 4.29 presents these findings for most frequently co-occurring pairs. As demonstrated by data in this table, the pair of controlled vocabularies that co-occur the most often within a record is LCSH and FAST. In 90% of records, both FAST and LCGFT terms are included. Four additional pairs of controlled vocabularies co-occur in more than 50% of records overall as shown by the correlation indicator: SEARS and BISAC subject headings (0.74), FAST and BISAC subject headings (0.73), FAST and SEARS subject headings (0.65), and LCGFT and GSAFD genre headings (0.59).

The lowest levels of overall co-occurrence expressed as correlation (Table 4.29) was observed for the terms from the following pairs of controlled vocabularies: GSAFD and FAST

(0.08), followed by MESH and BISAC subject headings (0.26), and Répertoire de vedettes-matière and BISAC subject headings (0.28).

Although most records in the sample (72%) included one or more instances of 650 field containing BISAC subject headings, only a fraction of these (2% and 10% of records respectively) also included field 084 Other Classification Numbers or 072 Subject Category Code with the corresponding BISAC subject codes for these headings. As a result, the co-occurrence of the two closely related controlled vocabularies—BISAC headings and BISAC subject codes—was low (between 30% and 36% overall).

Table 4.29: Co-occurrence of controlled vocabularies within the same records (n=100)

pairs of controlled vocabularies	% of records with 1+ instances of use of each vocabulary	% of records with 0 instances of use of each vocabulary	correlation
LCSH and FAST	97%	1%	0.98
LCGFT and FAST	90%	1%	0.91
SEARS and BISAC headings	54%	18%	0.74
FAST and BISAC headings	72%	1%	0.73
FAST and SEARS	64%	1%	0.65
LCGFT and GSAFD	56%	3%	0.59
LC subject heading for children's literature and BISAC headings	28%	19%	0.47
LC subject heading for children's literature and SEARS	24%	23%	0.47
Source not specified and SEARS	8%	32%	0.40
Source not specified and BISAC headings	11%	27%	0.38

pairs of controlled vocabularies	% of records with 1+ instances of use of each vocabulary	% of records with 0 instances of use of each vocabulary	correlation
Répertoire de Vedettes-Matière and SEARS	1%	35%	0.36
BISAC headings in 650 and BISAC subject codes in 084	10%	26%	0.36
MESH and SEARS	1%	33%	0.34
BISAC headings in 650 and BISAC subject codes in 072	2%	28%	0.30
Répertoire de Vedettes-Matière and BISAC headings	1%	27%	0.28
MESH and BISAC headings	1%	25%	0.26
GSAFD and FAST	7%	1%	0.08

CHAPTER 5

DISCUSSION AND CONCLUSIONS

5.1 Introduction

This exploratory study intended to answer the following research questions:

1. What extent and variety of subject representation do the library metadata records (i.e., MARC21 bibliographic records) currently provide? How are the most recent RDA and MARC21 guidelines and features intended to support functionality in Linked Data environment and BIBFRAME conversion applied in subject metadata elements in the records?

2. How does the application of existing subject metadata in the most recently created MARC21 library metadata records affect relations between these records as measured by social network analysis?

3. How does the subject representation in the newly created MARC21 bibliographic records carry over into BIBFRAME records resulting from automated conversion from MARC21? What implications does such a conversion have for interconnectedness of records based on subject metadata?

This study was organized into two stages. The first stage examined the dataset consisting of all RDA-based MARC 21 bibliographic records created in 2020 and available for harvesting using Z39.50 protocol from the OCLC WorldCat database. The second stage was intended to refine and supplement the findings from Stage 1 high-level analysis by shifting the focus of analysis from the dataset level to the record level. The two stages together contributed to answering the research questions posed for this study. This chapter discusses findings from the two stages of this research project and provides comparisons to applicable findings of previous studies. It then provides the answers obtained in this study to research questions that guided the investigation. This is followed by the conclusion of the study in which the impact to the field, the limitations and challenges that have been identified, and possible directions for future research are discussed.

5.2 Discussion

Following the most large-scale analysis of MARC 21 bibliographic records to date—146 million of OCLC WorldCat records—Smith-Yoshimura et al. (2010) recommended prioritizing subject access points in MARC21 record creation:

> The number of full-text documents available on the Web will substantially increase over the next few years, and the need for surrogate 'descriptive metadata' will decrease. Focus instead on the authorized names, classifications, and controlled vocabularies that keyword searching of full-text will not provide. (p. 13).

The findings of both Stage 1 and Stage 2 in this study demonstrate that this recommendation has been implemented to some extent. This is evident from the increase in the overall level of several subject fields application (050, 082, 650, 651, 655, 648)—the percentage of records containing fields and the average number of instances of these fields in the records containing them—when compared with applicable findings of the previous studies that examined the level of application of various MARC 21 fields, including some of the subject metadata fields (cf., Eklund et al., 2009; Intner, 1989; Moen et al., 2006; Moen & Benardino, 2003; Mayernik, 2009; Smith-Yoshimura et al., 2010; Taylor & Simpson, 1986).

Table 5.1 compares the findings of the present study -- both Stage 1 and Stage 2 -- to the relevant findings of these previous studies. As shown in Table 5.1, significantly higher levels of application of field 651 were observed in this study. Between 36.81% of the recently (in 2020) created MARC bibliographic records in Stage 1 and 59% in Stage 2 included this field as opposed to only 9.93% in the 2010 study of all WorldCat records (Smith-Yoshimura et al., 2010). The number of instances per record containing this field was also higher in this study than in Mayernik's (2009) study of the Library of Congress catalog records: between 1.8 and 2.54 as opposed to 1.38.

Table 5.1: Comparison of this study findings with applicable findings of previous studies on MARC 21 metadata

Findings for subject metadata fields	Taylor & Simpson, 1986	Intner, 1989	MCDU project (Eklund et al., 2009, Moen et al., 2006, Moen & Benardino, 2003)	Mayernik, 2009	Smith-Yoshimura et al., 2010	The current study
Field 043 (non- repeatable)	Missing in 2.7%-4.8% of records	*n/a*	141409 instances total for 419657 WorldCat records (included in 33.70% of records)	Included in 33.4% of the Library of Congress records	Included in 19% of all 146M WorldCat records	Included in between 33.96% (Stage 1, n=10014) and 53% (Stage 2, n=100) of WorldCat records
Field 050	Errors or omissions in 4.3% to 7.2% of records	*n/a*	300385 instances total for 419657 records (0.71 instances per record)	Included in 99.13% of records	Included in 20% of all 146M WorldCat records	Included in between 65.56% and 100% of records
Filed 082	Errors or omissions in 6.4% to 13.3% of records	*n/a*	274313 instances total for 419657 records (0.65 instances per record)	Included in 28.87% of records	Included in 14% of all 146M WorldCat records	Included in between 53.03% and 100% of records
Field 600	*n/a*	*n/a*	69636 instances total for 4196570 records (0.17 instances per record)	Included in 32% of the Library of Congress records. 1.22 instances per record that included this field	Included in 7.08% of 146M WorldCat records	Included in between 8.75% and 28% of records. 5th highest number of instances per record that includes the field: between 2.36 and 2.37
Field 610	*n/a*	*n/a*	*n/a*	*n/a*	Included in 5.29% of 146M WorldCat records	Included in 7% (Stage 1 and Stage 2) of WorldCat records
Field 611	*n/a*	*n/a*	*n/a*	*n/a*	Included in 0.14% of all 146M WorldCat records	Included in between 0.97% (Stage 1) and4 % (Stage 2) of WorldCat records
Field 630	*n/a*	*n/a*	*n/a*	*n/a*	Included in 1.01% of 146M WorldCat records	Included in 1.72% (Stage 1 only) of WorldCat records
Field 648	*n/a*	*n/a*	*n/a*	*n/a*	Included in 0.07% of all 146M WorldCat records	Included in between 1.25% (Stage 1) and 16 % (Stage 2) of WorldCat records
Field 650	Errors or omissions in 11.7% to 13.9% of records	Errors (including omissions) in subject headings in 1.2% of OCLC and RLIN records overall. Main and added entities errors occur in 23% of records and include LCRI errors in name headings used as subject headings	602362 instances total for 419657 records (1.44 instances per record)	Included in 66% of 1500 records in 1817 instances total (1.84 instances per record with field)	Included in 46% of 146M WorldCat records	Included in between 92.04% and 100% of records: between 5.78 (Stage 1) and 13.35 instances per record in (Stage 2)
Field 651	*n/a*	Errors (including omissions) in subject headings in 1.2% of OCLC and RLIN records overall	113050 instances total for 419657 records (0.27 instances per record)	1.38 instances per Library of Congress record that included this field	Included in 9.93% of all 146M WorldCat records	Included in between 36.83% (Stage 1) and 59% (Stage 2) of WorldCat records. Between 1.8

Findings for subject metadata fields	Taylor & Simpson, 1986	Intner, 1989	MCDU project (Eklund et al., 2009, Moen et al., 2006, Moen & Benardino, 2003)	Mayernik, 2009	Smith-Yoshimura et al., 2010	The current study
						and 2.54 instances per record containing field
Field 653	*n/a*	*n/a*	55311 instances total for 419657 records (0.132 instances per record)	*n/a*	Included in 6.04% of 146M WorldCat records	Included in 1.81% (Stage 1 only) of WorldCat records
Filed 654	*n/a*	*n/a*	*n/a*	*n/a*	Included in 0.04% of 146M WorldCat records	Included in 0.01% (Stage 1 only) of WorldCat records
Field 655	*n/a*	*n/a*	Included in 5.1% of sound recording records in WordCat	1.55 instances per Library of Congress record that included this field	Included in 4.27% of 146M WorldCat records	Included in between 79.12% (Stage 1) and 98% (Stage 2) of WorldCat records. Between 2.59 (Stage 1) and 6.93 (Stage 2 instances per record containing the field
Filed 662	*n/a*	*n/a*	*n/a*		Included in 0.1% of all 146M WorldCat records	Not found in any records in Stage 1 or Stage 2

An even higher increase (Table 5.1) was observed in the level of application of field 655, which was observed in between 79.12% and 98% of records in this study, as opposed to only 4.27% of records in the Smith-Yoshimura et al. study (2010) and 5.1% of records representing sound recordings in MCDU project (Eklund et al, 2009). The number of instances of this field per record containing it was also much higher in this 2020 study than in Mayernik's (2009) study of the Library of Congress catalog records: between 2.59 instances on average observed in Stage 1 and 6.93 instances observed in Stage 2 as opposed to 1.55 instances.

The most drastic increase was observed (Table 5.1) in the level of application of field 650 which is the most widely applicable among all 6XX subject metadata fields, with the exception of 655. This field was observed in 92.04% in Stage 1 and 100% of records in Stage 2 of this study, as opposed to 46% of WorldCat records in Smith-Yoshimura et al. (2010) or 66% of the Library of Congress Catalog records in Mayernik (2009). The number of instances of this field per record containing the field observed in the current study is also much higher than that in Mayernik (2009) and MCDU project studies: between 5.78 in Stage 1 and 13.35 in stage2 as opposed to 1.84 and 1.44 instances observed in these two previous studies respectively. Unlike this study, none of the previous studies examined the co-occurrence of various controlled vocabularies within the same record. However, based on the findings of the manual content analysis in Stage 2 which showed that majority of instances (69.97%) of field 650 included non-LCSH subject headings it is highly likely that this drastic increase in the number of instances of field 650 is at least in part caused by the emerging practice of adding subject terms from controlled vocabularies other than LCSH. This includes not only the addition (often automatic) of terms from FAST, which is basically a faceted subset of LCSH that supports post-coordination,

but also the addition of terms from alternative controlled vocabularies of topical terms such as BISAC and SEARS, the use of which was frequently observed in this study.

This practice of enriching records by adding non-LCSH subject terms from a variety of controlled vocabularies of topical terms significantly expands subject representation in records, and if accompanied with Linked-Data-enabling metadata elements, will greatly increase functionality of bibliographic records in supporting the Explore user task (LRM, 2017) in either MARC 21 or BIBFRAME environment. While not a high proportion of works are about a person, organization, meeting, place, or another work (as represented by MARC 21 6XX fields 600, 610, 611 /648, 630, and 651 respectively), each information object has some kind of topical aboutness (e.g., Wilson, 1968; Hjørland, 1997). Therefore, the MARC 21 field 650 has the highest potential for providing connections between the records and building metadata networks, similarly to the Subject field in non-MARC metadata, as found by Phillips, Zavalina, and Tarver (2019).

As can be seen in Table 5.1, some increase was observed in the application of fields 648, 610, 611, and 630, although not as noticeable when compared to findings of previous studies. The application of field 648 that is relatively new (added to MARC 21 Bibliographic Format standard in 2002) and was only examined by Smith-Yoshimura et al. (2010) study of 146 million of OCLC WorldCat records, at the level of 0.07% of records, has increased to between 1.25% and 16% of records observed in Stage 1 and Stage 2 of the current study respectively. The current study observed that between 8.75% of records analyzed in Stage 1 and 28% of records analyzed in Stage 2 contained field 600, with the average number of instances per record containing the field between 2.36 and 2.37. This represented an increase compared to 7.08% of

records containing this field (Smith-Yoshimura et al., 2010) and 1.22 instances per record containing field 600 in the Library of Congress catalog (Mayernik, 2009). Percentage of records including field 610 in this study (both Stage 1 and Stage 2) was 7%, while the previous study of WorldCat records by Smith-Yoshimura et al. found this field in 5.29% of records. Compared to the only previous study that reported levels of field application results with regards to field 630 (Smith-Yoshimura et al. 2010), this field was used in the somewhat higher proportion of records analyzed by this study of the most-recent created RDA-based MARC 21 bibliographic records: 1.72% of records as opposed to 1.01% of records.

With regards to substantial increase in the level of application of 6XX fields that was observed in this study in comparison with findings of previous large-scale studies such as MCDU project and Smith and Yoshimura's OCLC Research project, it is worth noting that it is possible that a substantial proportion of the observed increase is due to the large-scale efforts by OCLC to automatically generate FAST headings in fields 658, 650, 651, and 655 from the LCSH subject strings includes in the 650 and 651 fields of bibliographic records in OCLC WorldCat database. According to Mixter and Childress (2013), the efforts started in 2013.

The findings regarding non-6XX subject metadata fields revealed a higher level of application than those observed in the Smith-Yoshimura et al. (2010) study but similar levels to those observed in the MARC Content Designation Utilization (MCDU) studies (see Table 5.1). For example, the overall level of application of field 050 has increased substantially in OCLC WorldCat records compared to that found by Smith-Yoshimura et al. in 2010: from 20% to between 65.56% and 100% of records. However, MCDU project that analyzed OCLC WorldCat records collected in 2004, found similar levels of application as this study (approximately 70%).

Similarly, although the findings of this dissertation show a higher level of application of field 043 than in Smith-Yoshimura et al.'s study (33.96%-53% as opposed to 19% of records), Stage 1 findings are very close to those of MCDU project (33.7% of records). Likewise, although the findings of this study show higher level of application of field 082 than in the Smith-Yoshimura et al study (between 53.03% in Stage 1 and 100% records in Stage 2 as opposed to 20% of records), Stage 1 findings are similar to those of MCDU project (approximately 65% of records).

It is important to note that all of these previous studies (see Table 5.1) were conducted prior to transition from AACR2 to RDA, and that several new subject metadata fields have been added to MARC 21 Bibliographic Format standard since the time the latest one was completed. This included fields 083, 085 and 688 which were not observed in the present study, as well as fields 084, 086, and 647, which were observed in the present study.

There is only one recent project that focused on the evolution of RDA-based records conducted by the team of researchers at the University of North Texas (Zavalina, Shakeri, & Kizhakkethil, 2016; Zavalina, Zavalin, & Miksa, 2016; Zavalina, Zavalin, Shakeri & Kizhakkethil, 2016). My study found similar levels of application for some of the subject metadata fields as studies conducted as part of this recent UNT project. For example, Zavalina, Shakeri, and Kizhakkethil (2016) reported on the level of application of field 043 (54.35% of records), which increased from 35.5% of records between 2013 and 2015, which is almost identical to the Stage 2 of this study. Likewise, overall the level of application of MARC 21 field 650 and the average number of instances of this field per record observed in the Stage 1 of my study is similar to that observed by Zavalina, Shakeri, and Kizhakkethil (2016): 92.04% of records and 5.78 instances per record compared to 91.58% or records and 5.038 instances per record.

The recent UNT project was also the only one that looked at the application of Linked-Data-enabling subfields. Most of these subfields were either added to MARC 21 Bibliographic Format standard or redefined for the support of Linked Data functionality only recently. These include subfield $0 Authority Record Control Number or Standard Number (redefined in 2010), subfield $1 Real World Object URI (added in 2017), and subfield $4 Relationship (renamed and redefined in 2017). Unlike the previous study which examined a small sample of English-language-of-cataloging OCLC WorldCat records for English-language video recordings in DVD format, my study provides a much more robust overall understanding of the various RDA-based MARC 21 records in the OCLC WorldCat database. There were some similarities and some differences in the level of application of Linked-Data-enabling subfields of subject metadata fields observed in this study when compared to those observed in the recent UNT project. For example, Zavalina, Shakeri, and Kizhakkethil (2016), similar to this study, did not observe any use of subfields $1 and $4 in the MARC 21 field 600. However, their study found subfield $0 to be used in 32.32% instances of field 600 in the sample of 369 records in 2015, while this study found it to be used much more often—in 86.64% of all instances of field 600 in the sample of 10014 records in Stage 1.

I also observed consistently high levels of application for the Linked-Data-enabling subfield $2 that was used in a high proportion of records in a total of 4 different subject metadata fields: 072, 082, 084, 092, 600, 610, 611, 630, 647, 648, 650, 651, 654, and 655. However, subfields $1 and $4 were never used. Moreover, the study revealed that the most important Linked-Data-enabling subfield in MARC 21 bibliographic metadata -- subfield $0 Authority Record Control Number or Standard Number (e.g., Shieh and Reese, 2015, etc.) --

despite being widely used overall, was not applied to its full capacity. It was observed in Stage 2 of this study that even in the highest quality full-level cataloging records, Linked-Data enabling subfield $0 was consistently used only for FAST subject headings in fields 600, 610, 611, 650, 651, and 655. It was completely omitted for terms from any other subject controlled vocabularies that were used in the records: BISAC, GND, GSAFD, GTT, LCGFT, LCSH, and SEARS. It was also excluded inone of the FAST facets—the chronological facet that is represented in field 648; no instances of field 648 with subfield $2 data value of “fast” had the subfield $0 present. This omission means that when MARC 21 records are converted to BIBFRAME 2.0, URIs for controlled-vocabulary terms would not be included, and for subject representation other than that with FAST (based on LCSH). records would mostly rely on literal data values (strings of characters) that have no Linked Data power.

Stage 1 of this study relied on data mining and Big Data analytics approaches and in order to build an overall understanding of subject representation in the dataset on MARC 21 bibliographic records. Stage 2 involved a manual content analysis of a small purposive sample to supplement and refine Stage 1 findings, and to provide triangulation. The level of subject representation in the 100 records analyzed in Stage 2 was predictably found to be much broader and much more consistent than in the entire data set of 10014 records analyzed in Stage 1. This is explained by the higher standards of cataloging followed by the 100 records in Stage 2 subsample and by the very large number of holdings attached to the subsample that resulted in numerous edits by various libraries that adopted these records into their online catalogs. However even within this small purposive sample of the most complete cataloging records based on the full-level cataloging standard followed, a substantial variability was

observed (as measured by variance and standard deviation), especially for application of fields 650 and 655. These fields represent topical aboutness and format or genre of an information object and, therefore, as discussed above, apply to each information object. For that reason, a higher consistency in the level of their application (e.g., as expressed in the number of instances of field) was expected in the subset of records that follow the strictest cataloging standards, based on codes blank or I in ELvl subfield of the fixed field.

The findings presented and discussed above, allow me to make the following conclusions regarding the answers to research questions addressed by this study.

5.2.1 Research Question 1

What extent and variety of subject representation do the library metadata records (i.e., MARC21 bibliographic records) currently provide? How are the most recent RDA and MARC21 guidelines and features intended to support functionality in Linked Data environment and BIBFRAME conversion applied in subject metadata elements in the records?

Analyses conducted in this study demonstrate that subject representation in MARC 21 bibliographic records created in 2020 based on the RDA data content standard guidelines and the most recent version of the MARC 21 Bibliographic Format standard, has substantially increased in extent and variety compared to earlier-created MARC 21 metadata analyzed in previous studies. Most of the data elements added to MARC 21 Bibliographic Format standard to reflect RDA, BIBFRAME, and overall Linked Data functionality requirements, in the last two decades are applied in one or more records in the studied dataset (with exception of 11 subject metadata fields: 083, 085, 098, 099, 522, 656, 657, 658, 662, 688, and 69X). A total of 26 out of 37 possible subject metadata fields were found in the records collected and analyzed in this study. Both the number of various subject fields included in records, and the number of

instances of these subject fields are substantially higher in the RDA records created in 2020 than what was observed in the past studies of pre-RDA MARC 21 bibliographic records. However, the overall level of application of some of these fields and subfields (e.g., subfields $1, and $4) and/or the consistency of the application of some of these fields and subfields (e.g., subfield $0), is not yet where it should be to fully realize their potential. Moreover, high variability (as measured through variance and standard deviation indicators) was observed in the application of several key subject metadata fields, including 650 and 655.

5.2.2 Research Question 2

How does the application of existing subject metadata in the most recently created MARC21 library metadata records affect relations between these records as measured by social network analysis?

The increased subject representation, especially the addition of topical terms from non-LCSH controlled vocabularies, has substantially improved collocation. It increased connections between the records based on shared subject terms, and allowed for building and examining networks of metadata records using Social Network Analysis measures in Stage 1.

5.2.3 Research Question 3

How does the subject representation in the newly created MARC21 bibliographic records carry over into BIBFRAME records resulting from automated conversion from MARC21? What implications does such a conversion have for interconnectedness of records based on subject metadata?

Due to the adjustments to the study design necessitated by the need to overcome computational challenges and technical issues encountered in the process of data collection (to be discussed in section 5.3.4. Limitations), the current study did not collect the data for an analysis that would provide answers to this question. However, analyses conducted in this

study helped refine the understanding of how the richness of subject metadata in MARC 21 records might affect the resulting BIBFRAME records resulting from conversion of MARC 21 records. This research question is left to be answered by future studies discussed in section 5.3.5. Future Research. However, analysis of MARC 21 metadata records, especially manual analysis conducted as part of Stage 1, allows to partially asmnwer this question – the p[art about readioness for make conclusions

5.3 Conclusion

5.3.1 Contribution

This study was the first to systematically examine RDA-based MARC 21 metadata records using a large dataset. Previous large-scale studies looked at mostly non-RDA studies as these studies were conducted before transition from Anglo-American Cataloging Rules (AACR2) to Resource Description and Access (RDA) in cataloging practice. This study was also the first to examine MARC 21 records created after the most recent (2019) addition of fields to MARC 21 Bibliographic Format standard, a standard that has undergone significant changes and expansions, including in subject metadata elements, in the recent years since the inception of the RDA cataloging code. It is the first study to focus its analyses on subject metadata in MARC 21 bibliographic records, using a large dataset, unlike the handful of previous studies that relied on small samples. Last, but not least, no published studies prior to this one (except Miller, 2014) examined the distribution of subject terms in various 6XX subject added entries and the indexing MARC 21 bibliographic fields and networks formed by these terms and shared by bibliographic records. This is the first study to complete the subject metadata network analysis in a heterogeneous centralized dataset (WorldCat) as opposed to individual library catalogs

(e.g., New York Public library catalog as in Miller's 2014 study) or to non-MARC metadata records (e.g., Digital Public Library of America Dublin-core based records as in Phillips' 2020 study).

Findings of this study provide the much-needed empirical data about the patterns in application in bibliographic records of the 26 various subject metadata fields—and their subfields—as defined in the current version of the MARC 21 bibliographic format standard (1999 edition, Update 30, as of May 2020). It also provides insight into how and to what extent various controlled vocabularies are used for subject representation in the most recently created (i.e., 2020) RDA-based MARC 21 bibliographic records overall, as well as in the subset of these records that follow the highest standard of cataloging (full-level input), all of which implies a high degree of completeness and application of access points.

SNA was applied only to the subject metadata fields that occur in at least 50% of all records: 650 (subject added entry – topical term), 655 (index term – genre/form), 050 (LC classification number) and 082 (Dewey Decimal classification number). The networks were created based on correlation adjacency matrices that require multiple computational iterations. No connections between records based on data values in classification numbers metadata fields were found. Only two subject metadata fields among selected for the analysis revealed connectivity: 650 and 655. The graph based on field 655 data values revealed more density in comparison with the graph build on the field 650 data values. However, overall density of all networks for found relatively low. Thus, effective applicability was found only for subject metadata fields containing terms as opposed to numbers and codes. However, it is worth noting that the results of application of SNA methods were affected by refining and

normalization of data values conducted as part of data processing. Similar recent work applied graph methods to the analyses of non-MARC21 bibliographic metadata (including subject metadata): UNTL and Dublin Core based Digital Public Library of America (DPLA) application profile (e.g. Phillips, 2020). Phillips study found varying levels of network density for different collection of metadata and based on different metadata element and experimented with applying several data normalization algorithms. Future SNA studies of MARC21 bibliographic metadata will need to comparatively explore the effect of different normalization algorithms.

SNA analyses and data preparation completed as part of this study allowed to develop an algorithm for application of SNA to the analysis of the interconnectedness between MARC21 bibliographic records that share similar data values in subject metadata fields. This algorithm needs to be tested and refined in the future research. It will also need to be expanded to application of SNA to MARC21 bibliographic records as a whole beyond subject metadata.

5.3.2 Study Recommendations for Cataloging Practice

Based on empirical data analysis results obtained in this study, the conclusion is made that subject metadata in MARC 21 records at the current stage is not yet ready for meaningful conversion to BIBFRAME and support of BIBFRAME and Linked Data Functionality. Available MARC 21 content designation intended to support this functionality is not used to full capacity. Examination of MARC21 records (especially as part of Stage 2) allows to formulate practical recommendations for catalogers and metadata managers who create and update RDA-based MARC 21 records, as well as the broader library metadata community that includes stakeholders such as Library of Congress Linked Data initiative, OCLC, and developers of controlled vocabularies. Implementation of these recommendations would result in a stronger

support of Linked Data and more meaningful conversion to BIBFRAME 2.0. The recommendations include:

- Including subfield $0 with authority record ID number for all instances of
    - field 043 that contain terms from Geographic Are Code controlled vocabulary (currently available through Library of Congress Linked Data Portal)
    - field 655 that contain LCGFT genre headings (currently available through Library of Congress Linked Data Portal)
    - field 648 chronological term which uses FAST chronological facet terms.
- Adding field 648 with chronological facet terms from FAST controlled vocabulary -- based on data in the field 046 in name authority records -- when a record represents a resource that is about:
    - a person, and a record includes field 600 (regardless of whether subfield $d is included) [FAST headings are not currently generated automatically by running FAST macro from fields 600. The process by which most FAST headings are added to the records is based on subject strings in fields 650 and 651)]
    - an organization, and a record includes field 610 (regardless of whether subfield $d is included) [FAST headings are not currently generated automatically by running FAST macro from fields 610. The process by which most FAST headings are added to the records is based on subject strings in fields 650 and 651)]
    - a conference or other meeting, and a record includes 611 (regardless of whether subfield $d is included) [FAST headings are not currently generated automatically by running FAST macro from fields 611. The process by which most FAST headings are added to the records is based on subject strings in fields 650 and 651)]
    - some phenomenon, family or group of people for which there is a known time period (normally, such an authority record has field 150 topical heading which does not include dates, e.g., “Dionne quintuplets”, etc.)
- Adding field 651 with geographic name facet term from FAST controlled vocabulary, and a corresponding code in 043 field when a record represents a resource that is about:
    - a conference, and a record includes field 611 subfield $c
    - one or more ethnic groups, and a record has subjects heading(s) such as “[The ethnic name] Americans” in field 650

- Consistently including field 084 or field 072 with BISAC subject codes whenever BISAC subject headings are used in field 650
- Consistently using the option available for 4 MARC 21 bibliographic subject fields to indicate primary and secondary subject terms in the record with the 1st indicator values (1 or 2): in most commonly used subject fields 650 and 655, and in much less frequently applied fields 653 and 654.
- Working to add to LC Linked Data Portal the most frequently used non-LCSH-based lists of subject headings -- BISAC and SEARS -- in Linked Data form with unique record IDs. After this is done, add subfields $0 in field 650 instances that contain SEARS and BISAC headings.

5.3.3 Study Recommendations for Cataloging Education

This study revealed insufficient level of Linked-Data-enabling subfields. These data elements are not currently widely included in graduate cataloging courses at the introductory level, as these courses mostly focus on core elements. Because most students completing Masters programs in Library and Information Science only take the introductory cataloging course and advanced cataloging courses are not required for future catalogers and offered less often than introductory courses, it is recommended to consider revising introductory cataloging curricula to emphasize application of Linked-Data -enabling MARC21 data elements.

5.3.4 Study Recommendations for Data Processing and Analysis of MARC 21 Metadata Records

The experience obtained in overcoming various challenges in collecting, processing, and analyzing MARC 21 metadata records as part of this study allows me to suggest recommendations for more efficient workflow for future studies of a similar nature.

The data collection recommendations include careful consideration of the advantages and disadvantages in sampling approach selection when collecting bibliographic data from individual library catalogs or an aggregated database such as OCLC WorldCat using Z39.50

protocol. If the researcher intends to collect a relatively large dataset or a very large sample—over 100000 records—project planning needs to budget sufficient time for data collection as it can be time-consuming and encounter technical errors in the process of downloading records that would require recollection of data. On the other hand, relying on smaller samples – for example, 384 records from a population of at 1 million or more, with the population proportion of 0.50 and a standard error of 0.05 (based on Krejcie and Morgan, 1970 etc.)—assumes that a small sample is representative of a given population if that sample is random. However, as I discovered in this study, the order in which the records are collected from OCLC WorldCat database via Z39.50 protocol is not random: the bibliographic records that are collected first are the ones with the highest number of holdings attached to them. Regardless of the sample size chosen, it is also important to keep in mind that records collected are not necessarily unique and that the sample might include a number of duplicates.

The suggested workflow for the preprocessing of the datasets collected from OCLC WorldCat using the Z39.50 protocol—before any analyses can be started—includes the following steps:

1. Downloading records in the native MARC 21 format in a file with the. mrc filename extension
2. Running the resulting file through the deduplication in the MARC Editor tool in MARC Edit to ensure all duplicates are removed and only unique records remain in the dataset.
3. Additional data extractions and data refining using Python PyMARC, Openrefine, and other tools can be run on the files with the .mrc file name extension to have data structured in .csv file format.
4. Automated analysis using Python Pandas, Rapidminer, and other tools can be run on files with .csv file format.

5. If manual content analysis is planned, conducting lossless transformation from the deduplicated database file with the .mrc filename extension into MARC XML document with the .xml filename extension or the mnemonic human-readable MARC document file with the .mrk file name extension.

After completing these preprocessing steps, data can be analyzed using a variety of analyses and tools. Application of different tools and technologies in data analytics requires definite contribution of time spent for the learning curve and troubleshooting. Using different versions and implementations of Python developments on different platforms may cause some discrepancies in results. For example, running the same script within identical Python environments on Windows-based and Unix-based machines revealed different results of data extractions. During this study, running scripts on Unix platforms generated more accurate and complete results although all computing dependencies had the same version control and were properly updated. In addition, writing the results of data processing and extractions in Python to .csv file format on Windows-based machines required extra Unicode error type treatment in contrast to running the same procedure on Unix-based machines where the process was rather flawless. These types of discrepancies are pretty common and require extra time for troubleshooting.

Another recommendation is to carefully read supporting documentation available for any type of a tool used in the research. Although this might be an obvious suggestion, researchers may encounter some lack of supporting documentation and come through a painful time-consuming trial-and-error process of troubleshooting a problem. Thus, one's contribution into code refining and troubleshooting of errors can be very valuable; and active use of different platforms for collaboration and sharing developers' knowledge, such as GitHub

(https://github.com/), Stack Overflow (https://stackoverflow.com/) or specific Google Groups is highly recommended.

5.3.5 Limitations

This study had several limitations and delimitations that were discussed in the introduction in Chapter 1. This section provides detailed information on the limitations of this study and solutions to address these limitations (if any) in this study.

One of the limitations of content analysis is researcher bias, which is normally alleviated through the use of detailed coding manuals, and coding at least 10% of data by additional coder(s), beyond the principal investigator, and subsequent evaluation of the intercoder agreement (otherwise called intercoder reliability). This study was designed to assess only objective (i.e., mostly quantitative and binary) characteristics and measures and did not include any subjective evaluations. For example, those regarding the accuracy of subject metadata, with the exception of the obvious misspellings in some of the data values that were identified in the controlled-vocabulary codes designating the names of controlled vocabularies found in subfield $2 of 6XX, 072, and 084 fields. For this reason, coding by multiple coders was not needed, as researcher bias was not introduced.

This study encountered a number of technical issues in collecting the dataset of MARC 21 bibliographic records from OCLC WorldCat database over Z39.50 protocol: query terminations by the host, for example. Another important challenge was the lack of reliable information on the estimated total number of records that met the search parameters of this study's Z39.50 query, and complete absence of information on the proportion of the records in OCLC WorldCat overall (and specifically for the given search parameters) that are unique, as

opposed to duplicates. For these reasons, it is unclear whether or not the dataset collected and analyzed in Stage 1 of this study is the entire population of unique records that meet the study criteria—RDA-based records created in 2020—or simply a subset of such a population. Therefore, it is not clear whether or not the Stage 1 analyses can be reliably categorized (as originally intended when planning this study) as those following the Big Data analytics approach, under which the whole population is analyzed as opposed to its sample and where sample error is completely avoided.

For the reasons discussed above, as well as due to the particular order in which Z39.50 protocol query collects records from OCLC WorldCat (in the inverted order of the number of holdings attached to the record), it was also not possible to accurately assess how representative the collected dataset is of the whole population of the recently-created RDA-based MARC 21 bibliographic records (e.g., those created in 2019 and 2020 and presumably based on the latest at the time of data collection —May 2019 or November 2019—update of MARC 21 Bibliographic Format standard) in OCLC WorldCat. However, the size of the collected dataset (10014 unique metadata records after deduplication that removed 93% of collected records as duplicates) far exceeded the minimum random sample size of 384 to achieve representativeness and reliability of results in analysis of populations consisting of over 1 million of items (e.g., Krejcie and Morgan, 1970). Thus, if the data collected and analyzed in Stage 1 of this study is a sample, the sheer volume of this sample, although non-random, is expected to ensure that sample error is minimized, and allows for generalizations to be made.

Stage 2 findings demonstrated higher overall completeness of subject metadata. This is explainable by mostly full encoding level (Elvl) and the fact that records in the sample were

created or updated as part of Program for Cooperative Cataloging (PCC) and/or Library of Congress Copy Cataloging Project. These projects and programs have certain requirements for MARC21 bibliographic records. As a result, records analyzed in stage 2 might have a different level of application and variety of subject metadata than an average record in WorldCat.

The purposive sample of 100 most widely held RDA-based MARC 21 bibliographic records created in 2020 with the highest level of cataloging (as indicated by data values blank and I in the ELvl subfield of the fixed field) analyzed in Stage 2 did not include any records for materials in languages other than English or the records with English language of cataloging. Due to this, it was impossible to conduct comparative evaluation of subject metadata for groups of records based on the language of cataloging or language of materials. This study also did not assess measures of central tendency and variability measures for the number of instances for each subject metadata subfield (e.g., 6XX$z), including the Linked-Data enabling subfields (e.g., 6XX $0) comparatively for records representing different types of materials in analog and digital form, and by language of item represented by the record.

Due to the serious computational challenges in collecting, processing, and analyzing large datasets of MARC 21 records (over 400 thousand records), the scope of this study had to be revised down from all RDA-based records created or last updated in 2019 (estimated population size of at least 1.7 million) to all RDA-based records created in 2020. For this reason, the initially planned component of the study -- side-by-side comparative analysis of a sample of these MARC 21 records with their BIBFRAME2 .0 work records counterparts was not possible at this time as the only currently existing database of BIBFRAME 2.0 records made available by the United States Library of Congress (http://id.loc.gov/resources/works.html) was

last updated in June of 2019 and did not include any of the records created in 2020 at the time of data collection and analysis in this study.

The results of application of SNA methods were affected by normalization and cleaning procedures applied to the data values used in the study. For example, only in the subfield $a of all subject fields' values 50193 ending periods and 1148 ending commas were removed. More detailed text clustering and normalization reveals better connectivity between records.

5.3.6 Future Research

This study assessed patterns of subject representation in the entire set of 10014 records collected and in a smaller purposive sample of records from that first set representing different types of materials and created by different institutions worldwide overall. Comparative analysis of the patterns of application of various subject metadata elements (fields and subfields) and controlled vocabularies for records created in different countries (with the same or different languages of cataloging) and records representing different types of materials—all eight broad material types as defined by codes in fields 006 and 007 of MARC 21 bibliographic records and their subtypes (e.g., electronic and print versions of books as indicated by data values "computer" and "unmediated" in RDA-based MARC 21 field 337) would be the next logical step. The high-level semi-automated analysis using large datasets would benefit from supplementing mostly manual and more in-depth analysis of smaller subsamples. This study did not assess measures of central tendency and variability measures for the number of instances for each subject metadata subfield (e.g., 6XX$z), including the Linked-Data enabling subfields (e.g., 6XX $0) comparatively for records representing different types of materials in analog and digital form, and by language of item represented by record. These analyses are not possible through

high-level Big Data analytics and data mining approaches alone, therefore future research would need to include the manual in-depth content analysis to evaluate these indicators.

Future studies are also needed to compare patterns of application of subject metadata in RDA-based MARC 21 bibliographic records and BIBFRAME 2.0 work records that are derived from these records through automated conversion. These studies would evaluate what (if anything) is lost in the process of such conversion and develop suggestions for enhancing the subject metadata in MARC 21 records pre- and post- conversion to ensure the high Linked Data functionality of resulting BIBFRAME 2.0 records. One way to conduct such studies would be to rely on the database of BIBFRAME 2.0 work records (http://id.loc.gov/resources/works.html) and their exact MARC 21 equivalents from which these records were derived in the Library of Congress online catalog. To develop an understanding of the quality of the conversion and the resulting Linked Data functionality for the MARC21 and BIBFRAME 2.0 work records that are not created solely by the Library of Congress, and which are therefore more representative of the entire population of bibliographic records contained in the library catalog worldwide, these future studies would need to rely on centralized databases such as OCLC WorldCat as a source of MARC 21 records, and on conversions to BIBFRAME 2.0 done by researchers themselves using the current conversion specifications and programs (https://www.loc.gov/bibframe/), possibly with the help of tools such as BIBFRAME testbed and Link Identifiers in MARCEdit MARC Next editor.

Studies that examine metadata records in relation to guidelines in policies and procedures manuals developed and used locally by individual institutions that create RDA-based MARC 21 bibliographic records and convert them to BIBFRAME records for institution-

specific guidelines on subject representation, as well as *MARC 21 Format for Bibliographic Data, National Level Full and Minimal Requirements; BIBCO Standard Record (BSR) RDA Metadata Application Profile; CONSER Standard Record (CSR) RDA Metadata Application Profile* would help to generate a more complete picture of the overall BIBFRAME-readiness and Linked Data functionality support in existing library metadata. Such future studies would need to work with much smaller samples of metadata records than the one analyzed in Stage 1 of this study.

The records collected for analysis in this study were created in January-April of 2020. Due to the age of the records, it was initially assumed, based on the findings of previous studies on MARC 21 metadata change (e.g., Zavalina, Zavalin, & Miksa, 2016; Zavalina & Zavalin, 2018), that very little record modification would be found. However, this assumption was not supported by the present study. In-depth manual content analysis conducted in Stage 2 revealed that all 100 records with the highest level of holdings (between 571 and 1514), despite being created no more than 3 months before analysis started in January-February 2020, were edited by institutions other than the original record creator at least once, and 98% of them were edited multiple times by multiple institutions. A promising possible direction for future research would be to analyze metadata change in the new records over time to determine what modifications are made to the subject metadata in them and to examine social networks formed between institutions that create RDA-based metadata records according to the latest versions of MARC standard and those that transcribe and edit them (based on the data in field 040 subfields $a Original Cataloging Agency, subfield $b Transcribing Agency, and subfield $d Modifying Agency).

At present, the majority of records in large databases like OCLC WorldCat fall into the following categories:1) the non-RDA metadata records, 2) those records that were partially converted into RDA from existing AACR2 records using automated algorithms and without human cataloger evaluation and augmentation of results, and 3) RDA-based records created in the early stages of RDA testing and adoption, before a number of new data elements were added to MARC 21 Bibliographic Format standard to support RDA and Linked Data functionality. All of these records will eventually need to be converted to BIBFRAME. To support this conversion with empirical data, future studies will need to extend to these categories of MARC 21 bibliographic records the analysis of readiness for meaningful conversion to BIBFRAME and support of Linked Data functionality; regarding both subject representation and beyond. To obtain a more complete and more accurate picture, these future studies would need to rely on data mining and Big Data analytics approaches and would need to be able to overcome the computational challenges currently experienced with analyses of large datasets of MARC 21 metadata.

Future studies of functional readiness of library metadata to support Linked-Data requirements might address the following research questions:

- What will be the best configuration to provide easier maintenance?
- How can redundant details be eliminated/excluded?
- How can derivation of data can be increased and how this data can be moved into the web environment?

APPENDIX A

OCCURRENCES OF THE SUBFIELDS OF SUBJECT METADATA FIELDS

Field	subfield code	subfield name and repeatability (R=repeatable, NR=non-repeatable)	Total subfield instances	Average no. of subfield instances per record with field
043	$0	Authority record control number or standard number (R)	0	0
043	$2	Real World Object URI (R)	0	0
043	$6	Linkage (NR)	0	0
043	$8	Field link and sequence number (R)	0	0
043	$a	Geographic area code (R)	3918	1.152014113
043	$b	Local GAC code (R)	0	0
043	$c	ISO code (R)	0	0
045	$6	Linkage (NR)	0	0
045	$8	Field link and sequence number (R)	0	0
045	$a	Time period code (R)	22	0.021696252
045	$b	Formatted 9999 B.C. through C.E. time period (R)	3987	3.931952663
045	$c	Formatted pre-9999 B.C. time period (R)	0	0
045	$6	Linkage (NR)	0	0
045	$8	Field link and sequence number (R)	0	0
050	$0	Authority record control number or standard number (R)	0	0
050	$1	Real World Object URI (R)	0	0
050	$3	Materials specified (NR)	0	0
050	$6	Linkage (NR)	0	0
050	$8	Field link and sequence number (R)	0	0
050	$a	Classification number (R)	6655	1.013709063
050	$b	Item number (NR)	5476	0.834120335
052	$0	Authority record control number or standard number (R)	0	0
052	$1	Real World Object URI (R)	0	0
052	$2	Code source (NR)	0	0
052	$6	Linkage (NR)	0	0
052	$8	Field link and sequence number (R)	0	0
052	$a	Geographic classification area code (NR)	39	1.21875

Field	subfield code	subfield name and repeatability (R=repeatable, NR=non-repeatable)	Total subfield instances	Average no. of subfield instances per record with field
052	$b	Geographic classification subarea code (R)	36	1.125
052	$d	Populated place name (R)	0	0
055	$0	Authority record control number or standard number (R)	0	0
055	$1	Real World Object URI (R)	0	0
055	$2	Source of call/ class number (NR)	0	0
055	$6	Linkage (NR)	0	0
055	$8	Field link and sequence number (R)	0	0
055	$a	Classification number (NR)	86	1.023809524
055	$b	Item number (NR)	79	0.94047619
060	$0	Authority record control number or standard number (R)	0	0
060	$1	Real World Object URI (R)	0	0
060	$8	Field link and sequence number (R)	0	0
060	$a	Classification number (R)	150	1.094890511
060	$b	Item number (NR)	28	0.204379562
070	$0	Authority record control number or standard number (R)	0	0
070	$1	Real World Object URI (R)	0	0
070	$8	Field link and sequence number (R)	0	0
070	$a	Classification number (R)	13	1
070	$b	Item number (NR)	13	1
072	$2	Source (NR)	2709	2.58
072	$6	Linkage (NR)	0	0
072	$8	Field link and sequence number (R)	0	0
072	$a	Subject category code (NR)	2716	2.586666667
072	$x	Subject category code subdivision (R)	1756	1.672380952
080	$0	Authority record control number or standard number (R)	0	0
080	$1	Real World Object URI (R)	0	0

Field	subfield code	subfield name and repeatability (R=repeatable, NR=non-repeatable)	Total subfield instances	Average no. of subfield instances per record with field
080	$2	Edition identifier (NR)	0	0
080	$6	Linkage (NR)	0	0
080	$8	Field link and sequence number (R)	0	0
080	$a	Universal Decimal Classification number (NR)	16	1.333333333
080	$b	Item number (NR)	1	0.083333333
080	$x	Common auxiliary subdivision (R)	0	0
082	$2	Edition number (NR)	4908	0.924293785
082	$6	Linkage (NR)	0	0
082	$8	Field link and sequence number (R)	0	0
082	$a	Classification number (R)	5584	1.051600753
082	$b	Item number (NR)	52	0.009792844
082	$m	Standard or optional designation (NR)	2	0.000376648
082	$q	Standard or optional designation (NR)	298	0.056120527
084	$0	Authority record control number or standard number (R)	148	0.196547145
084	$2	Number source (NR)	1110	1.474103586
084	$6	Linkage (NR)	0	0
084	$8	Field link and sequence number (R)	0	0
084	$a	Classification number (R)	1349	1.791500664
084	$b	Item number (NR)	3	0.003984064
084	$q	Assigning agency (NR)	42	0.055776892
090	$a	Classification number (R)	21	1.4
090	$b	Local Cutter number (NR)	21	1.4
090	$e	Feature heading (NR)	0	0
090	$f	Filing suffix (NR)	0	0
092	$2	Edition number (NR)	3	0.136363636
092	$a	Classification number (R)	22	1
092	$b	Item number (NR)	15	0.681818182
092	$e	Feature heading (NR)	0	0

Field	subfield code	subfield name and repeatability (R=repeatable, NR=non-repeatable)	Total subfield instances	Average no. of subfield instances per record with field
092	$f	Filing suffix (NR)	0	0
096	$a	Classification number (R)	1	1
096	$b	Item number (NR)	0	0
096	$e	Feature heading (NR)	0	0
096	$f	Filing suffix (NR)	0	0
600	$0	Authority record control number or standard number (R)	759	0.866438356
600	$1	Real World Object URI (R)	0	0
600	$2	Source of heading or term (NR)	803	0.916666667
600	$3	Materials specified (NR)	0	0
600	$4	Relationship (R)	0	0
600	$6	Linkage (NR)	0	0
600	$8	Field link and sequence number (R)	0	0
600	$a	Personal name (NR)	2075	2.368721461
600	$b	Numeration (NR)	49	0.055936073
600	$c	Titles and other words associated with a name (R)	494	0.563926941
600	$d	Dates associated with a name (NR)	1236	1.410958904
600	$e	Relator term (R)	1	0.001141553
600	$f	Date of a work (NR)	0	0
600	$g	Miscellaneous information (R)	0	0
600	$h	Medium (NR)	0	0
600	$j	Attribution qualifier (R)	0	0
600	$k	Form subheading (R)	0	0
600	$l	Language of a work (NR)	0	0
600	$m	Medium of performance for music (R)	0	0
600	$n	Number of part/section of a work (R)	1	0.001141553
600	$o	Arranged statement for music (NR)	0	0
600	$p	Name of part/section of a work (R)	0	0

Field	subfield code	subfield name and repeatability (R=repeatable, NR=non-repeatable)	Total subfield instances	Average no. of subfield instances per record with field
600	$q	Fuller form of name (NR)	73	0.083333333
600	$r	Key for music (NR)	0	0
600	Ss	Version (R)	0	0
600	$t	Title of a work (NR)	62	0.070776256
600	$u	Affiliation (NR)	0	0
600	$v	Form subdivision (R)	496	0.566210046
600	$x	General subdivision (R)	284	0.324200913
600	$y	Chronological subdivision (R)	1	0.001141553
600	$z	Geographic subdivision (R)	31	0.035388128
610	$0	Authority record control number or standard number (R)	678	0.964438122
610	$1	Real World Object URI (R)	0	0
610	$2	Source of heading or term (NR)	699	0.9943101
610	$3	Materials specified (NR)	0	0
610	$4	Relationship (R)	0	0
610	$6	Linkage (NR)	0	0
610	$8	Field link and sequence number (R)	0	0
610	$a	Corporate name or jurisdiction name as entry element (NR)	1632	2.321479374
610	$b	Subordinate unit (R)	1014	1.442389758
610	$c	Location of meeting (R)	0	0
610	$d	Date of meeting or treaty signing (R)	0	0
610	$e	Relator term (R)	0	0
610	$f	Date of a work (NR)	1	0.001422475
610	$g	Miscellaneous information (R)	0	0
610	$h	Medium (NR)	0	0
610	$k	Form subheading (R)	0	0
610	$l	Language of a work (NR)	0	0
610	$m	Medium of performance for music (R)	0	0

Field	subfield code	subfield name and repeatability (R=repeatable, NR=non-repeatable)	Total subfield instances	Average no. of subfield instances per record with field
610	$n	Number of part/section/meeting (R)	8	0.011379801
610	$o	Arranged statement for music (NR)	0	0
610	$p	Name of part/section of a work (R)	0	0
610	$r	Key for music (NR)	0	0
610	$s	Version (R)	0	0
610	$t	Title of a work (NR)	85	0.120910384
610	$u	Affiliation (NR)	0	0
610	$v	Form subdivision (R)	119	0.169274538
610	$x	General subdivision (R)	572	0.813655761
610	$y	Chronological subdivision (R)	19	0.027027027
610	$z	Geographic subdivision (R)	25	0.035561878
611	$0	Authority record control number or standard number (R)	99	1.020618557
611	$1	Real World Object URI (R)	0	0
611	$2	Source of heading or term (NR)	102	1.051546392
611	$3	Materials specified (NR)	0	0
611	$4	Relationship (R)	0	0
611	$6	Linkage (NR)	0	0
611	$8	Field link and sequence number (R)	0	0
611	$a	Meeting name or jurisdiction name as entry element (NR)	117	1.206185567
611	$c	Location of meeting (R)	5	0.051546392
611	$d	Date of meeting or treaty signing (R)	10	0.103092784
611	$e	Subordinate unit (R)	0	0
611	$f	Date of a work (NR)	0	0
611	$g	Miscellaneous information (R)	0	0
611	$h	Medium (NR)	0	0
611	$j	Relator term (R)	0	0
611	$k	Form subheading (R)	0	0

Field	subfield code	subfield name and repeatability (R=repeatable, NR=non-repeatable)	Total subfield instances	Average no. of subfield instances per record with field
611	$l	Language of a work (NR)	0	0
611	$n	Number of part/section/meeting (R)	4	0.041237113
611	$p	Name of part/section of a work (R)	0	0
611	$q	Name of meeting following jurisdiction name entry element (NR)	0	0
611	Ss	Version (R)	0	0
611	$t	Title of a work (NR)	0	0
611	$u	Affiliation (NR)	0	0
611	$v	Form subdivision (R)	4	0.041237113
611	$x	General subdivision (R)	0	0
611	$y	Chronological subdivision (R)	0	0
611	$z	Geographic subdivision (R)	0	0
630	$0	Authority record control number or standard number (R)	167	0.970930233
630	$1	Real World Object URI (R)	0	0
630	$2	Source of heading or term (NR)	170	0.988372093
630	$3	Materials specified (NR)	0	0
630	$4	Relationship (R)	0	0
630	$6	Linkage (NR)	0	0
630	$8	Field link and sequence number (R)	0	0
630	$a	Uniform title (NR)	292	1.697674419
630	$d	Date of treaty signing (R)	7	0.040697674
630	$e	Subordinate unit (R)	0	0
630	$f	Date of a work (NR)	1	0.005813953
630	$g	Miscellaneous information (R)	1	0.005813953
630	$h	Medium (NR)	0	0
630	$k	Form subheading (R)	1	0.005813953
630	$l	Language of a work (NR)	2	0.011627907
630	$m	Medium of performance for music (R)	0	0

Field	subfield code	subfield name and repeatability (R=repeatable, NR=non-repeatable)	Total subfield instances	Average no. of subfield instances per record with field
630	$n	Number of part/section of a work (R)	0	0
630	$o	Arranged statement for music (NR)	0	0
630	$p	Name of part/section of a work (R)	67	0.389534884
630	$r	Key for music (NR)	0	0
630	$s	Version (R)	1	0.005813953
630	$t	Title of a work (NR)	0	0
630	$v	Form subdivision (R)	23	0.13372093
630	$x	General subdivision (R)	40	0.23255814
630	$y	Chronological subdivision (R)	0	0
630	$z	Geographic subdivision (R)	0	0
647	$0	Authority record control number or standard number (R)	144	1.152
647	$1	Real World Object URI (R)	0	0
647	$2	Source of heading or term (NR)	144	1.152
647	$3	Materials specified (NR)	0	0
647	$6	Linkage (NR)	0	0
647	$8	Field link and sequence number (R)	0	0
647	$a	Named event (NR)	144	1.152
647	$c	Location of named event (R)	60	0.48
647	$d	Date of named event (NR)	144	1.152
647	$g	Miscellaneous information (R)	0	0
647	$v	Form subdivision (R)	0	0
647	$x	General subdivision (R)	0	0
647	$y	Chronological subdivision (R)	0	0
647	$z	Geographic subdivision (R)	0	0
648	$0	Authority record control number or standard number (R)	0	0
648	$1	Real World Object URI (R)	0	0
648	$2	Source of heading or term (NR)	961	0.983623337

Field	subfield code	subfield name and repeatability (R=repeatable, NR=non-repeatable)	Total subfield instances	Average no. of subfield instances per record with field
648	$3	Materials specified (NR)	0	0
648	$6	Linkage (NR)	0	0
648	$8	Field link and sequence number (R)	0	0
648	$a	Chronological term (NR)	979	1.002047083
648	$v	Form subdivision (R)	0	0
648	$x	General subdivision (R)	0	0
648	$y	Chronological subdivision (R)	0	0
648	$z	Geographic subdivision (R)	0	0
650	$0	Authority record control number or standard number (R)	21603	2.3438212
650	$1	Real World Object URI (R)	0	0
650	$2	Source of heading or term (NR)	26357	2.859607247
650	$3	Materials specified (NR)	0	0
650	$4	Relationship (R)	0	0
650	$6	Linkage (NR)	0	0
650	$8	Field link and sequence number (R)	44	0.004773788
650	$a	Topical term or geographic name as entry element (NR)	53243	5.776608441
650	$b	Topical term following geographic name as entry element (NR)	0	0
650	$c	Location of an event (NR)	2	0.00021699
650	$d	Active dates (NR)	0	0
650	$e	Relator term (R)	0	0
650	$g	Miscellaneous information (R)	12	0.001301942
650	$v	Form subdivision (R)	9487	1.029293696
650	$x	General subdivision (R)	13873	1.505153521
650	$y	Chronological subdivision (R)	1051	0.114028426
650	$z	Geographic subdivision (R)	8855	0.960724748
651	$0	Authority record control number or standard number (R)	4147	1.124457701

Field	subfield code	subfield name and repeatability (R=repeatable, NR=non-repeatable)	Total subfield instances	Average no. of subfield instances per record with field
651	$1	Real World Object URI (R)	0	0
651	$2	Source of heading or term (NR)	4328	1.173535792
651	$3	Materials specified (NR)	0	0
651	$4	Relationship (R)	0	0
651	$6	Linkage (NR)	0	0
651	$8	Field link and sequence number (R)	0	0
651	$a	Geographic name (NR)	6642	1.800976139
651	$e	Relator term (R)	0	0
651	$g	Miscellaneous information (R)	14	0.003796095
651	$v	Form subdivision (R)	854	0.231561822
651	$x	General subdivision (R)	1925	0.521963124
651	$y	Chronological subdivision (R)	672	0.182212581
651	$z	Geographic subdivision (R)	1056	0.286334056
653	$6	Linkage (NR)	0	0
653	$8	Field link and sequence number (R)	0	0
653	$a	Uncontrolled term (R)	1825	10.08287293
654	$0	Authority record control number or standard number (R)	0	0
654	$1	Real World Object URI (R)	0	0
654	$2	Source of heading or term (NR)	3	3
654	$3	Materials specified (NR)	0	0
654	$4	Relationship (R)	0	0
654	$6	Linkage (NR)	0	0
654	$8	Field link and sequence number (R)	0	0
654	$a	Focus term (R)	3	3
654	$b	Non-focus term (R)	0	0
654	$c	Facet/hierarchy designation (R)	0	0
654	$e	Relator term (R)	0	0
654	$v	Form subdivision (R)	0	0

Field	subfield code	subfield name and repeatability (R=repeatable, NR=non-repeatable)	Total subfield instances	Average no. of subfield instances per record with field
654	$y	Chronological subdivision (R)	0	0
654	$z	Geographic subdivision (R)	0	0
655	$0	Authority record control number or standard number (R)	8968	1.131894484
655	$1	Real World Object URI (R)	0	0
655	$2	Source of (NR)	15365	1.939290673
655	$3	Materials specified (NR)	0	0
655	$5	Institution to which field applies (NR)	2	0.00025243
655	$6	Linkage (NR)	0	0
655	$8	Field link and sequence number (R)	0	0
655	$a	Genre/form data or focus term (NR)	20548	2.593462072
655	$b	Non-focus term (R)	5	0.000631074
655	$c	Facet/hierarchy designation (R)	0	0
655	$v	Form subdivision (R)	28	0.003534015
655	$x	General subdivision (R)	10	0.001262148
655	$y	Chronological subdivision (R)	29	0.00366023
655	$z	Geographic subdivision (R)	11	0.001388363

APPENDIX B

PYTHON SCRIPTS USED IN THE ANALYSIS

```
 ## --- example of script for extraction 040 abc from .mrc file ---

from pymarc import MARCReader
import csv
#create a CSV file
csv_out = csv.writer(open('200K_40abc_.csv', 'w'), delimiter = ',', quotechar = '"', quoting =
csv.QUOTE_ALL)
#write a header row in your CSV file
csv_out.writerow(['a','b','c'])
# approach for non repeatable fields
with open('200K.mrc', 'rb') as fh:
   reader = MARCReader(fh)
   for record in reader:
      a = b = c = ''
      if record['040'] is not None:
         if record['040']['a'] is not None:
            a = record['040']['a']
         else:
            a = 'None'
         if record['040']['b'] is not None:
            b = record['040']['b']
         else:
            b = 'None'
         if record['040']['c'] is not None:
            c = record['040']['c']
         else:
            c = 'None'

      else:
         a = 'None'

         b = 'None'

         c = 'None'

      #print(a,b,c)

      csv_out.writerow([a,b,c])

# --- end of script ---


# --- example of script for 6XX fields extraction from .mrc file ---
```

```
from pymarc import MARCReader
import csv
#create a CSV file
csv_out = csv.writer(open('200Kdedup_oclsn6xx_may25.csv', 'w'), delimiter = ',', quotechar =
'"', quoting = csv.QUOTE_ALL)
#write a header row in your CSV file
csv_out.writerow(['oclcn','600','610','611','630','647','648','651','653','654'])
#print all
with open('200Kdedup.mrc', 'rb') as fh:
   reader = MARCReader(fh)
   for record in reader:
      #oclc_number = topic = genre = lcn = ddcn = ''
      oclcn = a = b = c = d = e = f = g = h = i = ''
#Check to make sure OCLC number exists in MARC 035 field
      if record['035'] is not None:

 #Check to make sure there's a |a
         if record['035']['a'] is not None:
            oclc_number = record['035']['a']
            #oclc_number = re.sub("[^0-9]", "", oclc_number)
            #print ('RecordID :', oclc_number)
            csv_out.writerow([oclc_number,'','','','','','','','',''])
#get repeatable fields:
      for a in record.get_fields('600'):
         #print('Topic :', topic)
         csv_out.writerow(['',a,'','','','','','','',''])

      for b in record.get_fields('610'):
         #print('Genre :', genre)
         csv_out.writerow(['','',b,'','','','','','',''])

      for c in record.get_fields('611'):
         #print('Place :', place)
         csv_out.writerow(['','','',c,'','','','','',''])

      for d in record.get_fields('630'):
         #print('Place :', place)
         csv_out.writerow(['','','','',d,'','','','',''])

      for e in record.get_fields('647'):
         #print('Place :', place)
         csv_out.writerow(['','','','','',e,'','','',''])
```

```
        for f in record.get_fields('648'):
            #print('Place :', place)
            csv_out.writerow(['','','','','','',f,'','',''])

        for g in record.get_fields('651'):
            #print('Place :', place)
            csv_out.writerow(['','','','','','','','',g,'',''])

        for h in record.get_fields('653'):
            #print('Place :', place)
            csv_out.writerow(['','','','','','','','','',h,''])

        for i in record.get_fields('654'):
            #print('Place :', place)
            csv_out.writerow(['','','','','','','','','',i])

 # --- end of script ---

 # --- example of script for data imputation ---

import pandas as pd
# read csv file as a dataframe
matrix = pd.read_csv('200Kdedup_oclsn655a_may20-matrix-zeros.csv')
# check dataframe shape
matrix.shape
#check if there are missing values a:
matrix.isnull()
#check if there are missing values b:
matrix.isnull().sum()
#replace NULL values with ZEROS, to replace with space: modifiedMatrix=matrix.fillna(" ")
modifiedMatrix=matrix.fillna(0)
# checked modified matrix:
modifiedMatrix.isnull().sum()
#double check the shape of modified matrix
modifiedMatrix.shape
#write modified matrix to csv
modifiedMatrix.to_csv('200Kdedup655a-modifiedMatrix.csv',index=False)

# --- end of script ---

#--- example of script for correlation matrix creation ---
```

```
import pandas as pd
# read data from .csv file
data = pd.read_csv('200Kdedup_oclsn655a_may20-pivot2-perRecord.csv')
   # extra step -- we do not need it because data is in dataframe already
   #df = pd.DataFrame(data)
corr = data.corr()
corr.to_csv('200Kdedup082a-modifiedMatrix-test.csv',index=False)

# --- end of script ---
```

REFERENCES

Abele, A., Buitelaar, P., Cyganiak, R., Jentzsch, A., Andryushechkin, V., Debattista, J., & Nasir. J. (2020). *The Linked Open Data cloud diagram*. Retrieved from: https://lod-cloud.net/#diagram

Allen, B. L., & Williams, M. E. (1991). Cognitive research in information science: implications for design. In *M. E. Williams (Ed.), Annual review of information science and technology. Vol.26.* Place of Publication: Medford, NJ, USA. Country of Publication: USA.: Learned Information.

Allen, B., & Reser, D. (1990). Content analysis in library and information science research. *Library & Information Science Research*, *12*(3), 251-262.

Aluri, R., Kemp, D.A., & Boll, J.J. (1991). *Subject analysis in online catalogs*. Englewood: Libraries Unlimited.

Anderson, L., & Holt, C. (1997). Information cascades in the laboratory. *The American Economic Review*, *87*(5), 847-862. Retrieved from: http://www.jstor.org/stable/2951328

Anderson, R., & O'Connor, B. (2009). Reconstructing Bellour: Automating the Semiotic Analysis of film. *ASIS&T Bulletin*, June-July. Retrieved from: http://www.asis.org/Bulletin/Jun-09/JunJul09_Anderson_OConnor.html

ANSI/NISO Z39.50-2003. (2003). *Information Retrieval (Z39.50): Application Service Definition and Protocol Specification*. Retrieved from: https://www.loc.gov/z3950/agency/Z39-50-2003.pdf

Avram, H. D., & Library of Congress. (1976). *MARC, its history and implications*. Washington, DC: Library of Congress.

Babbie, E. (2013). *The practice of social research* (13th ed.). Belmont, CA: Wadsworth.

Balster, K., Rendall, R., & Schrader T. (2018). Linked serial data: Mapping the CONSER standard record to BIBFRAME. *Cataloging & Classification Quarterly*, *56*(2/3), 251-261.

Bates, M. (1972). *Factors affecting subject catalog search success*. (Unpublished doctoral. dissertation). University of California, Berkeley.

Bates, M. (1999). The invisible substrate of information science. *Journal of the American Society for Information Science*, *50*(12), 1043-1050.

Bates, M. (2002). After the dot-bomb: Getting web information retrieval right this time. *First Monday*, 7(7). Retrieved from: https://firstmonday.org/ojs/index.php/fm/article/view/971/892

Berners-Lee, T., Hendler, J., & Lassila, O. (May 17, 2001). The Semantic Web. In *Scientific American*. Retrieved from: https://www-sop.inria.fr/acacia/cours/essi2006/Scientific%20American_%20Feature%20Article_%20The%20Semantic%20Web_%20May%202001.pdf

Berners-Lee, T. (2007). *Testimony Before the United States House of Representatives Committee on Energy and Commerce Subcommittee on Telecommunications and the Internet Hearing on the "Digital Future of the United States: Part I -- The Future of the World Wide Web"* Retrieved from: http://dig.csail.mit.edu/2007/03/01-ushouse-future-of-the-web.html

Berners-Lee, T. (2009). Linked Data. In *W3C*. Retrieved from: https://www.w3.org/DesignIssues/LinkedData.html

Big Data. (2020). Big Data. In *Oxford English Dictionary*. Retrieved from: https://www.oed.com/view/Entry/18833#eid301162178

Beghtol, C. (1986). Bibliographic classification theory and text linguistics: aboutness analysis, intertextuality and the cognitive act of classifying documents. *Journal of Documentation, 42*(2), 84-113.

Berman, S., & Gross, T. (2017). Expand, humanize, simplify: An interview with Sandy Berman. *Cataloging & Classification Quarterly, 55*(6), 347-360.

Bikakis, N. (2016). *XML and Semantic Web W3C Standards Timeline-History*. Retrieved from: http://www.dblab.ntua.gr/~bikakis/XML%20and%20Semantic%20Web%20W3C%20Standards%20Timeline-History.pdf

Boehr, D., & Bushman, B. (2018). Preparing for the future: National Library of Medicine's project to add MESH RDF URIs to its bibliographic and authority records. *Cataloging & Classification Quarterly, 56*(2/3), 262-272.

Borgman, C. (1986). Why are online catalogs hard to use? Lessons learned from information retrieval studies. *Journal of American Society for Information Science, 37*(6), 387-400.

Borgman, C. (1996). Why are online catalogs still hard to use? *Journal of American Society for Information Science, 47*(7), 493-503.

Borgman, C. (2007). *Scholarship in the Digital Age: Information, Infrastructure, and the Internet*. Cambridge, MA: MIT Press.

Borgman, C. (2015). *Big data, little data, no data: scholarship in the networked world*. Cambridge, MA: MIT Press.

Borko, H. (1968). Information science: What is it? *American Documentation 19*(1), 3-5. Retrieved from: http://cdigital.uv.mx/bitstream/123456789/6699/2/Borko.pdf

Bratt, S. (2007). Semantic Web, and other technologies to watch. In *W3C*. Retrieved from: https://www.w3.org/2007/Talks/0130-sb-W3CTechSemWeb/#(1)

Boserup, I., & Krarup, K. (1982). *Reader-Oriented Indexing: An investigation into the extent to which subject specialists should be used for the indexing of documents by and for professional readers, based on a sample of sociological documents indexed with the help of the PRECIS indexing system*. Copenhagen: The Royal Library.

Breitman, K.K., Casanova, M.A., & Truszkowski, W. (2007). *Semantic Web: Concepts, Technologies and Applications*. Springer, New York, NY.

Buckland, M. (2012). What kind of science can information science be? *Journal of Information Science and Technology 63*(1), 1-7. Retrieved from: http://people.ischool.berkeley.edu/~buckland/whatsci.pdf

Calaresu, M., & Shiri, A. (2015). Understanding Semantic Web: a conceptual model. *Library Review, 64*(1/2), 82-100.

Calhoun, K. (2006). *The Changing Nature of the Catalog and its Integration with Other Discovery Tools: Final Report*. March 17, 2006. Prepared for the Library of Congress. Retrieved from: http://www.loc.gov/catdir/calhoun-report-final.pdf

Castells, M. (2011). A network theory of power. *International Journal of Communication, 5*, 773-787.20.

Chan, L., & Hodges, T. (2000). Entering the millennium: A new century for LCSH. *Cataloging & Classification Quarterly, 29*(1/2), 225-234.

Chen, P.P. (1976). The entity-relationship model: toward a unified view of data. *ACM Transactions on Database Systems, 1*(1), 1-36.

Cibangu, S. K. (2010). Information science as a social science. *Information Research, 15*(3). Retrieved from: http://www.informationr.net/ir/15-3/paper434.html

Cloud Security Alliance. (2014). *Big Data Taxonomy in CSA*. Retrieved from: https://downloads.cloudsecurityalliance.org/initiatives/bdwg/Big_Data_Taxonomy.pdf

Cochrane, P. (1979). Universal Subject Access (USA): can anyone do it? In *Redesign of Catalogs and Indexes for Improved Online Subject Access: selected papers of Pauline A. Cochrane*, Phoenix, Ariz.: Oryx Press, 1985, pp. 223-238.

Cochrane, P. (1986). *Improving LCSH for Use in Online Catalogs*. Colorado Springs, CO: Libraries Unlimited.

Cochrane, P. (2000). Improving LCSH for use in online catalogs revisited: What progress has been made? What issues still remain? *Cataloging & Classification Quarterly, 29*(1/2), 73-89.

Connaway, L., Johnson, D., & Searing, S. (1997). Online catalogs from the user's perspective: the use of focus group interviews. *College and Research Libraries, 58*(September), 403-420.

Creswell J.W. (2014). *Research design: qualitative, quantitative, and mixed-methods approaches*. Thousand Oaks, CA: Sage.

Crotty, M. (1998). *The foundations of social research: meaning and perspective in the research process*. Thousand Oaks, CA: Sage.

Crowther, R. (2008). Planning a Semantic Web Site: prepare your site for structured data. In *IBM*. Retrieved from: https://www.ibm.com/developerworks/library/x-plansemantic/x-plansemantic-pdf.pdf

De Mauro, A., Greco, M., & Grimaldi, M. (2016). A formal definition of big data based on its essential features. *Library Review, 65*(3), 122-135.

Delsey, T. (2005). Modeling subject access: Extending the FRBR and FRANAR conceptual models. *Cataloging & Classification Quarterly, 39*(3/4), 49-61.

Dempsey, L. (2012). Libraries and the information future: some notes. *Information Services & Use, 32,* 203-214.

Denzin, N. (2006). *Sociological Methods: A Sourcebook* (5th ed.). Chicago, IL: Aldine Transaction.

Dervin, B., & Nilan, M. (1986). Information needs and uses. In *M. E. Williams (Ed.), Annual Review of Information Science and Technology*, 21, 3-33.

Drabenstott, K. (1996). Enhancing a new design for subject access to online catalogs. *Library Hi Tech, 14*(1), 87-108.

Drabenstott, K., & Weller, M. (1996). Failure analysis of subject searches in a test of a new design for subject access to online catalogs. *Journal of American Society for Information Science, 47*(7), 519-537.

DuCharme, B. (2013). What do RDF and SPARQL bring to Big Data projects? *Big Data, 1*(1). Retrieved from: http://online.liebertpub.com/doi/pdf/10.1089/big.2012.0004

Dunne, C., & Shneiderman, B. (2013). Motif simplification: Improving network visualization readability with fan, connector, and clique glyphs. In, A*CM SIGCHI Conference on Human Factors in Computing Systems.* Retrieved from: http://www.cs.umd.edu/hcil/trs/2012-29/2012-29.pdf

Ekbia, H., Mattioli, M., Kouper, I., Arave, G., Ghazinejad, A., Bowman, T., Suri, V. R., Tsou, A., Weingart, S., & Sugimoto, C. R. (2014). Big data, bigger dilemmas: A critical review. *Journal of the Association for Information Science and Technology, 66*(8), 1523-1545.

Eklund, A.P., Miksa, S.D., Moen, W.E., Snyder, G., & Polyakov, S. (2009). Comparison of MARC content designation utilization in OCLC WorldCat records with national, core, and minimal level record standards. *Journal of Library Metadata, 9*, 36-64.

El-Sherbini, M. (2018). RDA implementation and the emergence of BIBFRAME. *JLIS.it, 9*(1). Retrieved from: https://www.jlis.it/article/view/66-82

Ercegovac, Z. (1998). Minimal level cataloging: what does it mean for maps in the contexts of card catalogs, online catalogs and digital libraries. *Journal of the Association for Information Science, 49*(8), 706-719.

Fairthorne, R.A. (1969). Content analysis, specification and control. *Annual Review of Information Science and Technology, 4*, 73-109.

Fan, W., & Bifet, A. (2012). Mining big data: current status and forecast to the future. *ACM SIGKDD Explorations Newsletter, 14*(2), 1–5. doi:10.1145/2481244.2481246. Retrieved from: http://www.kdd.org/exploration_files/V14-02-01-Fan.pdf

Farradine, J. (1970). Analysis and organization of knowledge for retrieval. *Aslib Proceedings, 22*(12), 607-616.

Fidel, R. (2000). *The user-centered approach*. Retrieved from: http://faculty.washington.edu/fidelr/RayaPubs/User-CenteredApproach.pdf

Frank, P., & Hoshy, L. (2007). *Library of Congress Report on Subject Cataloging. ALA ALCTS CCS Subject Analysis Committee (SAC) annual meeting*, Washington, DC, June 24, 2007. SAC07-ANN/3.75.

Garrett, J. (2007). Subject headings in full-text environments: the ECCO experiment. *College & Research Libraries, 68*(1), 69-81.

Glushko, R. J. (Ed.). (2013). *The discipline of organizing*. Cambridge, MA: MIT Press.

Godby, C. J., & Denenberg, R. (2015). *Common Ground: Exploring Compatibilities between the Linked Data models of the Library of Congress and OCLC*. Retrieved from: https://www.oclc.org/content/dam/research/publications/2015/oclcresearch-loc-linked-data-2015-a4.pdf

Graham, R.Y. (2004). Subject no-hits searches in an academic library online catalog: an exploration of two potential ameliorations. *College and Research Libraries, 65*(1), 36-54.

Granovetter, M. (1983). The strength of weak ties: a network theory revisited. *Sociological Theory, 1*, 201–233.

Greenberg, J. (2017). Big metadata, smart metadata, and metadata capital. *Journal of Data and Information Science, 2*(3), 19-36.

Gross, T., Taylor, A., & Joudrey, D. (2015). Still a lot to lose: The role of controlled vocabulary in keyword searching. *Cataloging & Classification Quarterly, 53*(1), 1-39.

Gross, T., & Taylor, A. (2005). What have we got to lose? The effect of controlled vocabulary on keyword searching results. *College and Research Libraries, 66*(3), 212-230.

Guba E.G. (1990). *The Paradigm Dialog*. Newbury Park, CA: SAGE Publications, Inc.

Guizzardi, G. (2007). On ontology, ontologies, conceptualizations, modeling languages, and (meta)models, *Proceedings of the 2007 Conference on Databases and Information Systems IV, IOS Press Amsterdam, The Netherlands*, 18-39. Retrieved from: http://www.inf.ufes.br/~gguizzardi/FAIA.pdf

Harel D., & Koren Y. (2001) A fast multi-scale method for drawing large graphs. In: Marks J. (eds) Graph Drawing. GD 2000. *Lecture Notes in Computer Science, vol 1984*. Springer, Berlin, Heidelberg. Retrieved from: http://www.wisdom.weizmann.ac.il/~harel/papers/ms_jgaa.pdf

Henri, F. (1992). *Computer conferencing and content analysis. Collaborative Learning through Computer Conferencing: The Najaden Papers*. A. R. Kaye. New York, Springer, 115-136.

Hembrooke, H., Granka, L., Gay, G., & Liddy, E. (2005). The effects of expertise and feedback on search term selection and subsequent learning. *Journal of the American Society for Information Science and Technology, 56*(8), 861-871.

Hirsh, S. (1996). *The Effect of Domain Knowledge on Elementary School Children's Information Retrieval Behavior on an Automated Library Catalog*. [Unpublished doctoral dissertation]. University of California, Los Angeles.

Hitchcock, J. E. (1940). Subject coverage in university library catalogs. *Library Quarterly, 10*, 69–94.

Hitzler, P., & Janowicz, K. (2013). Linked Data, Big Data, and the 4th paradigm. *Semantic Web, 0*(0). Retrieved from: http://www.semantic-web-journal.net/system/files/swj488.pdf

Hjørland, B. (2018). Subject (of documents). In *Encyclopedia of Knowledge Organization*. Retrieved from: https://www.isko.org/cyclo/subject

Hjørland, B. (1992). The concept of 'subject' in information science. *Journal of Documentation, 48*(2), 172-200.

Hjørland, B. (1997). The concept of subject or subject matter and basic epistemological positions. In *Information Seeking and Subject Representation: An Activity-Theoretical Approach to Information Science*. Westport CT: Greenwood Press, 55-103.

Hjørland, B. (1998). Theory and metatheory of information science: a new interpretation. *Journal of Documentation 54*, 606-621.

Hoffman, H. H. (2001). Subject access to works in online catalogs. *Technicalities, 21* (September/October), 9-11.

Hoffman, H. H. (1998). Evaluation of three record types for component works in analytic online catalogs. *Library Resources and Technical Services, 42*(4), 292-303.

Hsieh-Yee, I. (1993). Effects of search experience and subject knowledge on the search tactics of novice and experienced searchers. *Journal of the American Society for Information Science, 44*(3), 161–174.

IFLA. (2009). *Functional Requirements for Bibliographic Records. Final Report*. Retrieved from: https://www.ifla.org/files/assets/cataloguing/frbr/frbr_2008.pdf

IFLA. (2010). *Functional Requirements for Subject Authority Data (FRSAD): A Conceptual Model*. Retrieved from: https://www.ifla.org/files/assets/classification-and-indexing/functional-requirements-for-subject-authority-data/frsad-final-report.pdf

IFLA. (2013). *Functional Requirements for Authority Data: A Conceptual Model*. Retrieved from: https://www.ifla.org/files/assets/cataloguing/frad/frad_2013.pdf

IFLA. (2017). *IFLA Library Reference Model: A Conceptual Model for Bibliographic Information*. Retrieved from: https://www.ifla.org/files/assets/cataloguing/frbr-lrm/ifla-lrm-august-2017.pdf

Intner, S.S. (1989). Much ado about nothing: OCLC and RLIN cataloging quality. *Library Journal 114*(2), 38-40.

IT. (2016). A Dictionary of Physics (7 ed.) In *OXFORD Reference*. Retrieved from: http://www.oxfordreference.com/view/10.1093/acref/9780198714743.001.0001/acref-9780198714743-e-1592?rskey=zkGtwY&result=1

Jackson, S. (1958). *Catalog Use Study: Director's Report*. Chicago: American Library Association.

Joudrey, D. N., Taylor, A. G., & Miller, D. P. (2015). *Introduction to cataloging and classification*. (11th ed.). Santa Barbara, CA: Libraries Unlimited.

Kaplan, B., & Duchon, D. (1988). Combining qualitative and quantitative methods in information systems research: A case study. *MIS Quarterly, 12*(4), 571-586.

Krathwohl, D. R. (2009). *Methods of Educational and Social Science Research: the logic of methods* (3rd ed.). Long Grove, IL: Waveland Press, Inc.

Krejcie, R. V., & Morgan, D. W. (1970). Determining sample size for research activities. *Educational and Psychological Measurement, 30,* 607-610.

Krikelas, J. (1972). Catalog use studies and their implications. *Advances in Librarianship, 3*, 195-220.

Kuhn, T.S. (1962). *The structure of scientific revolutions* (4th ed.). Chicago, IL: The University of Chicago Press, Ltd.

LaFrance, M. (1989). The quality of expertise: Implications of expert-novice differences for knowledge acquisition. *SIGART Newsletter, 108*, 6-14.

Laney, D. (2001). 3D Data Management: Controlling Data Volume, Velocity and Variety In *Application Delivery Strategies, file 949*. Retrieved from: http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

Langridge, D.W. (1989). *Subject analysis: principles and procedures*. London: Bowker-Saur.

Larson, R. (1991). Between Scylla and Charybdis: Subject searching in online catalogs. *Advances in Librarianship, 15*, 175-236.

Lasswell's model in communication models. (2010). In *Communication Theory*. Retrieved from: http://communicationtheory.org/lasswells-model/

Library of Congress. (no date). *What's New in BIBFRAME 2.0*. Retrieved from: https://www.loc.gov/bibframe/docs/bibframe2-whatsnew.html

Library of Congress. (2020a). *BIBCO Standard Record (BSR) Metadata Application Profiles (MAPs).* Retrieved from: http://www.loc.gov/aba/pcc/bibco/bsr-maps.html

Library of Congress (2020b). *Library of Congress Subject Headings Manual*. Retrieved from: https://www.loc.gov/aba/publications/FreeSHM/freeshm.html

Library of Congress. (2019a). *MARC 21 to BIBFRAME 2.0 Conversion Specifications*. Retrieved from: http://www.loc.gov/bibframe/mtbf/

Library of Congress. (2019b). *BIBFRAME Editor and BIBFRAME Database Manual*. Retrieved from: https://www.loc.gov/aba/pcc/bibframe/BIBFRAME-Manual-Final-2019-07-12.pdf

Library of Congress Working Group on the Future of Bibliographic Control. (2008). *On the record: report*. Retrieved from: https://www.loc.gov/bibliographic-future/news/lcwg-ontherecord-jan08-final.pdf

Looking for information: a survey of research on information seeking, needs, and behavior. (2012). (3rd ed.) Ed., D. Case. London, UK: Emerald.

Manby, A. (2014). Data Insights: About Big Data. In *IBM Smarter Business Summit*. Retrieved from: https://www-

01.ibm.com/events/wwe/grp/grp037.nsf/vLookupPDFs/Manby_Data_Insights/$file/Manby_Data_Insights.pdf

Mann, T. (1991). Cataloging quality, LC priorities, and models of the Library's future. *Opinion Papers, No.1*. Washington, D.C.: Library of Congress Cataloging Forum.

MARCEdit Development. (2013). *About MarcEdit*. Retrieved from: MARCEdit Development website: https://marcedit.reeset.net/about-marcedit

Marchionini, G., Dwiggins, S., Katz, A., & Lin, X. (1993). Information seeking in full-text end-user-oriented search systems: the roles of domain and search expertise. *Library and Information Science Research, 15*(1), 35-69.

Marchionini, G. (2004). From information retrieval to information interaction. In Sharon McDonald & John Tait (Eds.), *Advances in Information Retrieval*. New York, NY: Springer, 1-11.

Markey, K. (1984). *Subject Searching in Library Catalogs*. Dublin, Ohio: OCLC.

Markey, K., & Demeyer, A. (1986). *Dewey Decimal Online Classification Project*. Dublin, OH: OCLC.

Mayernik, M. (2009). The distributions of MARC fields in bibliographic records: A power law analysis. *Library Resources and Technical Services, 54*(1), 40-54.

Metadata. (2019). In *Merriam-Webster.com*. Retrieved from: https://www.merriam-webster.com/dictionary/metadata

Mierswa, I., & Klinkenberg, R. (2020). *RapidMiner Studio (9.6)* [Data science, machine learning, predictive analytics]. Retrieved from: https://rapidminer.com/

Miksa, F.L. (1983). *The subject in the dictionary catalog from Cutter to the present*. Chicago: American Library Association.

Milgram, S. (1967). The small world problem. *Psychology Today, 2*, 60–67.

Miller, E., Ogbuji, U., Mueller, V., & MacDougall, K. (2012). *Bibliographic Framework as a web of data: Linked Data model and supporting services*. Retrieved from: https://www.loc.gov/bibframe/pdf/marcld-report-11-21-2012.pdf

Miller, M. (2014). *The networked catalog. The New York Public Library*. Retrieved from: https://www.nypl.org/blog/2014/07/31/networked-catalog

Mixter, J., & Childress, E. (2013). *FAST (Faceted Application of Subject Terminology) Users: Summary and Case Studies*. Dublin, Ohio: OCLC Research. Retrieved from: http://www.oclc.org/content/dam/research/publications/library/2013/2013-04.pdf

Moen, W.E., & Benardino, P. (2003). Assessing metadata utilization: an analysis of MARC content designation use. In *Proceedings of the International Conference on Dublin Core and Metadata Applications (Seattle, WA, Sept. 28 - Oct.2, 2003).* Retrieved from: https://dcpapers.dublincore.org/pubs/article/view/745

Moen, W.E., Miksa, S.D., Eklund, A., Polyakov, S., & Snyder, G. (2006). Learning from artifacts: Metadata utilization analysis. In *Proceedings of the Joint Conference on Digital Libraries, June 11-15, 2006, Chapel Hill, NC.*

Neuendorf, K. (2002). *The Content Analysis handbook*. Thousand Oak, CA: Sage Publications.

Newman, I., & Benz, C.R. (1998). *Qualitative-quantitative Research Methodology: Exploring the Interactive Continuum*. Carbondale, IL: SIU Press.

Nowack, B. (2009). *The Semantic Web – Not a Piece of Cake…* Retrieved from: http://bnode.org/blog/2009/07/08/the-semantic-web-not-a-piece-of-cake

OCLC Research. (2020). *About FAST*. Retrieved from: https://www.oclc.org/research/areas/data-science/fast.html

Osborn, A. D. (1941). Crisis in cataloging. *Library Quarterly, 11*(4), 393-411.

Palmer, C. (1996). Information work at the boundaries of science: Linking library services to research practices. *Library Trends, 44*(2), 165-191.

Park, J., & Brenza, A. (2015). Evaluation of semi-automatic metadata generation tools: A survey of the current state of the art. *Information Technology and Libraries, 34*(3), 22-42. Retrieved from: http://ejournals.bc.edu/ojs/index.php/ital/article/view/5889

Pennanen, M., Serola, S., & Vakkari, P. (2003). Changes of search terms and tactics while writing a research proposal: A longitudinal case study. *Information Processing and Management, 39*(3), 445-463.

Phillips, M.E., Tarver, H., & Zavalina, O.L. (2019). Using metadata record graphs to understand controlled vocabulary and keyword usage for subject representation in the UNT Theses and Dissertations Collection. *Proceedings of the 22nd International Symposium on Electronic Theses and Dissertations*. Retrieved from: http://etd2019.upt.pt/etd-2019-program/

Phillips, M.E., Zavalina, O.L., & Tarver, H. (2019a). Exploring the utility of metadata record graphs and network analysis for metadata quality evaluation and augmentation. *International Journal of Metadata, Semantics, and Ontologies, 7*(3), 1-13.

Phillips, M.E., Zavalina, O.L., & Tarver, H. (2019b). Using metadata record graphs to understand digital library metadata. *Proceedings of the International Conference and Workshop on Dublin Core and Metadata Applications*. Retrieved from: https://www.dublincore.org/conferences/2019/abstracts/#14

Phillips, M.E. (2020). *Exploring the Use of Metadata Record Graphs for Metadata Assessment.* [Unpublished doctoral dissertation]. University of North Texas, Denton, United States.

Pierson, H. W. (1934). The forest of pencils. *Library Quarterly, 4*(2), 306–313.

Prell, C. (2012). *Social network analysis: history, theory and methodology*. Los Angeles, CA: Sage.

Price, D. J. D.S. (1963). *Little science, big science*. New York, NY: Columbia University Press.

Price, D. J. D.S. (1965). Networks of scientific papers. *Science, 149*(3683), 510-515. doi:10.1126/science.149.3683.510

Price, D. J. D.S. (1975). *Science since Babylon*. New Haven: Yale University Press.

Riva, P., & Zumer, M. (2017). The IFLA Library Reference Model, a step toward the Semantic Web. Paper presented at: IFLA WLIC 2017 – Wrocław, Poland – Libraries. Solidarity. Society. In *Session 78 - Standards Committee*. Retrieved from: http://library.ifla.org/1763/1/078-riva-en.pdf

RDA Steering Committee. (2010). *Resource Description and Access*. Chicago: American Library Association; Ottawa: Canadian Library Association; London: Chartered Institute of Library and Information Professionals (CILIP). Retrieved from: www.rdatoolkit.org

RDA Co-Publishers (2010). *RDA Toolkit*. Chicago: American Library Association; Ottawa: Canadian Federation of Library Associations; London: Facet Publishing. Retrieved from: www.rdatoolkit.org

Salo, D. (2017). *Retooling libraries for the data challenge*. Retrieved form: https://minds.wisconsin.edu/bitstream/handle/1793/46142/DataChallenge.pdf

Saracevic, T. (2009). Information science. In: Marcia J. Bates and Mary Niles Maack (Eds.), *Encyclopedia of Library and Information Science*. New York: Taylor & Francis. pp. 2570-2586. Retrieved from: https://comminfo.rutgers.edu/~tefko/SaracevicInformationScienceELIS2009.pdf

Šauperl, A. (2002). *Subject determination during the cataloging process: observation*. Lanham, MD: Scarecrow Press.

Schieber, P. (1987). The wit and wisdom of Grace Hopper. *OCLC Newsletter, 167* (March/April 1987), 9.

Schöch, C. (2013). Big? Smart? Clean? Messy? Data in the humanities. *Journal for Digital Humanities, 2*(3). Retrieved from: http://journalofdigitalhumanities.org/2-3/big-smart-clean-messy-data-in-the-humanities/

Schniderman, S. (2006). *Statement of Saul Schniderman Representing the Library of Congress Professional Guild, AFSCME Local 2910, before the Committee on House Administration Concerning the World Digital Library. July 27*. Retrieved from: http://www.guild2910.org/.

Schreur, P. (2018). *The evolution of BIBFRAME: from MARC surrogate to Web conformant data model*: Paper presented at. IFLA WLIC 2018 in Session 141 - Cataloguing. Retrieved from: http://library.ifla.org/2202/1/141-schreur-en.pdf.

Schultz-Jones, B. (2009). Examining information behavior through social networks: An interdisciplinary review. *Journal of Documentation, 65*(4), 592-631.

Schultz-Jones, B., Snow, K., Miksa, S., & Hasenyager, R. (2012). Historical and current implications of cataloging quality for next generation catalogs. *Library Trends, 61*(1), 49-82. doi:10.1353/lib.2012.0028

Shieh, J., & Reese, T. (2015). The importance of identifiers in the new Web environment and using the Uniform Resource Identifier (URI) in subfield zero ($0): A small step that is actually a big step. *Journal of Library Metadata, 15* (3/4), 208-226.

Shreeves, S. L., Knutson, E. M., Stvilia, B., Palmer, C. L., Twidale, M. B., & Cole, T. W. (2005). Is quality metadata shareable metadata? the implications of local metadata practices for federated collections. In H. A. Thompson (Ed.), *Proceedings of the twelfth national conference of the association of college and research libraries* (pp. 223-237). Chicago, IL: Association of College and Research Libraries.

Sliffe, B.D., & Williams, R.N. (1995). *What's behind the research? Discovering hidden assumptions in the behavioral sciences.* Thousand Oaks, CA: Sage.

Smith, M., Ceni A., Milic-Frayling, N., Shneiderman, B., Mendes Rodrigues, E., Leskovec, J., & Dunne, C. (2010). *NodeXL: a free and open network overview, discovery and exploration add-in for Excel 2007/2010/2013/2016, from the Social Media Research Foundation*. Retrieved from: https://www.smrfoundation.org

Smith-Yoshimura, K., Argus, C., Dickey, T.J., Naun, C.C., Ortiz, L., & Taylor, H. (2010). *Implications of MARC Tag Usage on Library Metadata Practices: Report produced by OCLC Research in support of the RLG Partnership*. Retrieved from: www.oclc.org/research/publications/library/2010/2010-06.pdf.

Snow, K. (2011). *A Study of The Perception of Cataloging Quality Among Catalogers in Academic Libraries*. [Unpublished doctoral dissertation]. University of North Texas, Denton, United States.

Sockel, A. (June 21, 2016). *BISAC subject headings have been added to OverDrive Marketplace*. Retrieved from: https://company.overdrive.com/2016/06/21/bisac-subject-headings-have-been-added-to-overdrive-marketplace/

Soergel, D. (2009). Digital libraries and knowledge organization. In S. R. Kruk & B. McDaniel (Eds.), *Semantic Digital Libraries*, (pp. 9-39). Berlin: Springer.

Soglasnova, L. (2018). Dealing with false friends to avoid errors in subject analysis in Slavic cataloging: an overview of resources and strategies. *Cataloging & Classification Quarterly, 56*(5/6), 404-421.

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

Sugimoto, C., Ding, Y., & Thelwall, M. (2012). Library and information science in the Big Data era: Funding, projects, and future: [a panel proposal]. *Proceedings of the Association for Information Science and Technology, 49* (1), 1-3. Retrieved from: http://onlinelibrary.wiley.com/doi/10.1002/meet.14504901187/full

Tagliacozzo, R., & Kochen, M. (1970). Information-seeking behavior of catalog users. *Information Storage and Retrieval, 6*, 363-381.

Taniguchi, S. (2017). Examining BIBFRAME 2.0 from the viewpoint of RDA metadata schema. *Cataloging & Classification Quarterly, 55*(6), 387-412.

Tarver, H.S., Phillips, M., Zavalina, O.L., & Kizhakkethil, P. (2015). An exploratory analysis of subject metadata in the Digital Public Library of America. *Proceedings of the International Conference and Workshop on Dublin Core and Metadata Applications, São Paulo, Brazil*. Retrieved from: https://dcpapers.dublincore.org/pubs/article/view/3761

Taube, M. (1953). *Studies in Coordinate Indexing*. Washington D.C.: Documentation Incorporated.

Taxonomy. (2017). Taxonomy. In *Online Etymology Dictionary*. Retrieved from: http://www.etymonline.com/index.php?term=taxonomy

Taylor, A.G., & Simpson, C.W. (1986). Accuracy of LC copy: A comparison between copy that began as CIP and other LC cataloging. *Library Resources & Technical Services, 30*(4), 375-387.

Tennant, R. (2002). MARC must die. *Library Journal, 127*(17), 26–27.

Thomale, J. (2010). Interpreting MARC: Where's the Bibliographic Data? *Code4Lib, 11*. Retrieved from: https://journal.code4lib.org/articles/3832

Thomas, S.E. (1996, Winter). Quality in bibliographic control. *Library Trends, 44*(3), 491-506.

Thornburg, G., & Oskins, M. (2007). Misinformation and bias in metadata processing: Matching in large databases. *Information Technology and Libraries, 26*(2), 15-26.

Tillett, B. (2004). *What is FRBR? A conceptual model for the bibliographic universe*. Washington, DC: Library of Congress Cataloging Distribution Service. Retrieved from: https://www.loc.gov/cds/downloads/FRBR.PDF

Topham, K. (2018). Of Python and Pandas: Using Programming to Improve Discovery and Access. In *BLOGGERS! The blog of SAA's electronic records section*. Retrieved from: https://saaers.wordpress.com/2018/10/09/of-python-and-pandas-using-programming-to-improve-discovery-and-access/

Uzzi, B. (1996). The sources and consequences of embeddedness for the economic performance of organizations: the network effect. *American Sociological Review, 61*(4), 674-698. Retrieved http://www.kellogg.northwestern.edu/faculty/uzzi/ftp/sources.pdf

Varian, H. (2008). Hal Varian answers your questions. In *Freakonomics*. Retrieved from: http://freakonomics.com/2008/02/25/hal-varian-answers-your-questions/

Varma, C. (2019). CISO Guide: Surface Web, Deep Web and Dark Web - Are they different? In *CISO platform*. Retrieved from: https://www.cisoplatform.com/profiles/blogs/surface-web-deep-web-and-dark-web-are-they-different

W3C. (2007). RDF and SPARQL: Using Semantic Web Technology to Integrate the World's Data. In *W3C*. Retrieved from: https://www.w3.org/2007/03/VLDB/

W3C. (2013). Semantic Web Activity Statement. In *W3C*. Retrieved from: https://www.w3.org/2001/sw/Activity

Weare, C., & Lin, W.-Y. (2000). Content analysis of the World Wide Web: opportunities and challenges. *Social Science Computer Review, 18*(3), 272-292.

Weinberg, A.M. (1961, July 21). Impact of large-scale science on the United States. *Science, 134*(3473), 117.

Weinberg, B. (1995). Why postcoordination fails the researcher. *The Indexer, 19*, 155-159.

Wildemuth, B. (2004). The effects of domain knowledge on search tactic formulation. *Journal of the American Society for Information Science and Technology, 55*(3), 246-258.

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

Wilson, P. (1983). The Catalog as Access Mechanism: Background and Concepts". *Library Resources & Technical Services 27*(1), 4-17.

White, T. (2015). *Hadoop: The Definitive Guide* (4th ed.). Sebastopol: O'Reilly. Retrieved from: http://javaarm.com/file/apache/Hadoop/books/Hadoop-The.Definitive.Guide_4.edition_a_Tom.White_April-2015.pdf

WorldCat. (2020). WorldCat. Connect to the world's collected knowledge. Retrieved from: OCLC website: https://www.oclc.org/en/worldcat.html

Xie, Z, & Fox, E. A. (2017). Advancing library cyberinfrastructure for Big Data sharing and reuse. *Information Sciences and Use, 37*(3), 319-323.

Yang, C., Chen, H., & Hong, K. (2003). Visualization of large category map for Internet browsing. *Decision Support Systems, 35*, 89-102. Retrieved from: http://www.cis.drexel.edu/faculty/cyang/papers/yang2003f.pdf

Zavalina, O. (2007). Collection-level user searches in federated digital resource environment. In *Proceedings of the American Society for Information Science and Technology, 44*(1), 1-11.

Zavalina, O.L. (2012). Subject access: Conceptual models, functional requirements, and empirical data. *Journal of Library Metadata, 12* (2/3), 140-163.

Zavalina, O.L., Shakeri, S., & Kizhakkethil, P. (2016). Editing of library metadata records and its effect on subject access: An Empirical Investigation. *Proceedings of the International Federation of Library Associations World Library and Information Congress Satellite Conference "Subject Access: Unlimited Opportunities"*, Columbus, Ohio, August 11-12, 2016. Retrieved from: https://pdfs.semanticscholar.org/4989/d3c7ba96a173117264e8f3e1310d0f4dcf0f.pdf

Zavalina, O., L., & Zavalin V. (2018). Evaluation of metadata change in authority data over time: An effect of a standard evolution. *Proceedings of the Association for Information Science and Technology, 55*(1), 593-597.

Zavalina, O.L., Zavalin, V., & Miksa, S. D. (2016). Quality over time: A longitudinal quantitative analysis of metadata change in RDA-based MARC bibliographic records representing video resources*. Proceedings of the Association for Information Science and Technology, 53*(1), 1-5.

Zavalina, O.L., Zavalin, V., Shakeri, S., & Kizhakkethil, P. (2016). Developing an empirically-based framework of metadata change and exploring relation between metadata change and metadata quality in MARC library metadata. *Procedia Computer Science, 99*, 50-63.

Zeng, M.L. (2016). Subject Access, Smart Data, and Digital Humanities – Finding Unlimited Opportunities through their Intersections. In *Kent State University Keynote at IFLA Classification & Indexing Satellite Conference 2016*. August 11-12, Columbus, OH, USA. Retrieved from: https://www.slideshare.net/MarciaZeng/zeng-marcia-iflasubjectaccesssmartdatadh

Zeng, M., & Salaba, A. (2005). Toward an international sharing and use of subject authority data. *FRBR Workshop, OCLC, 2005*. Retrieved from: http://www.oclc.org/research/events/frbr-workshop/presentations/zeng/Zeng_Salaba.ppt

Zhang, X., Anghelescu, H., & Yuan, X. (2005). Domain knowledge, search behavior, and search effectiveness of engineering and science students: An exploratory study. *Information Research, 10*(2), 217.

Zhang X., Liu J., & Cole M. (2013). Task topic knowledge vs. background domain knowledge: impact of two types of knowledge on user search performance. In Rocha Á., Correia A., Wilson T., Stroetmann K. (eds.), *Advances in Information Systems and Technologies. Advances in Intelligent Systems and Computing*, 206. Springer, Berlin, Heidelberg.