CONTROLLED VOCABULARIES IN THE DIGITAL AGE:

ARE THEY STILL RELEVANT?

William Andrew Baker

Dissertation Prepared for the Degree of

DOCTOR OF PHILOSOPHY

UNIVERSITY OF NORTH TEXAS

August 2017

APPROVED:

Guillermo Oyarce, Committee Chair
Brian O'Connor, Committee Member
Will Senn, Committee Member
Suliman Hawamdeh, Chair of the Department of Information Science
Kinshuk, Dean of the College of Information
Victor Prybutok, Dean of the Toulouse Graduate School

Baker, William Andrew. *Controlled Vocabularies in the Digital Age: Are They Still Relevant?* Doctor of Philosophy (Information Science), August 2017, 147 pp., 22 tables, 7 figures, references, 105 titles.

Keyword searching and controlled vocabularies such as Library of Congress subject headings (LCSH) proved to work well together in automated technologies and the two systems have been considered complimentary. When the Internet burst onto the information landscape, users embraced the simplicity of keyword searching of this resource while researchers and scholars seemed unable to agree on how best to make use of controlled vocabularies in this huge database. This research looked at a controlled vocabulary, LCSH, in the context of keyword searching of a full text database. The Internet and probably its most used search engine, Google, seemed to have set a standard that users have embraced: a keyword-searchable single search box on an uncluttered web page. Libraries have even introduced federated single search boxes to their web pages, another testimony to the influence of Google. UNT's Thesis and Dissertation digital database was used to compile quantitative data with the results input into an EXCEL spreadsheet. Both Library of Congress subject headings (LCSH) and author-assigned keywords were analyzed within selected dissertations and both systems were compared. When the LCSH terms from the dissertations were quantified, the results showed that from a total of 788 words contained in the 207 LCSH terms assigned to 70 dissertations, 246 of 31% did not appear in the title or abstract while only 8, or about 1% from the total of 788, did not appear in the full text. When the author-assigned keywords were quantified, the results showed that from a total of 552 words from304 author-assigned keywords in 86 dissertations, 50 or 9% did not appear in the title or abstract while only one word from the total of 552 or .18% did not appear in the full text. Qualitatively, the LCSH terms showed a hierarchical construction that was clearly designed for a

print card catalog, seemingly unnecessary in a random access digital environment. While author-assigned keywords were important words and phrases, these words and phrases often appeared in the title, metadata, and full text of the dissertation, making them seemingly unnecessary in a keyword search environment as they added no additional access points. Authors cited in this research have tended to agree that controlled vocabularies such as LCSH are complicated to develop and implement and expensive to maintain. Most researchers have also tended to agree that LCSH needs to be simplified for large, full text databases such as the Internet. Some of the researchers have also called for some form of automation that seamlessly links LCSH to subject terms in a keyword search. This research tends to confirm that LCSH could benefit from simplification as well as automation and offers some suggestions for improvements in both areas.

ACKNOWLEDGMENTS

It is with both reverence and humility that I offer my sincere gratitude to the members of my dissertation committee since, without their generous guidance and direction, this project would have turned out much less effective than it did.

Thank you to Dr. Will Senn who graciously agreed to become a committee member at the eleventh hour yet still offered insights as well as reading materials that helped improve this research.

Thank you to Dr. Brian O'Connor who pointed me in a new direction with his additional readings as well as his suggestions for developing my topic.

A special thank you to Dr. Guillermo Oyarce, my major professor, was there for me from the very beginning, was always available for a meeting, kept me focused and on track while always maintaining a cordial yet professional attitude towards myself and this project.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

BACKGROUND

Introduction

Before the digital revolution and the introduction of fast, powerful online technologies, the principal access point for locating books in a library was the card catalog. Searching for an author or title of a work was straightforward; the card catalog was alphabetical. Searching for books by subject was also alphabetical although the terms on the cards were, in most libraries, not simple subject terms. These terms were most likely a specially designed set of words and phrases known as a controlled vocabulary which has been defined as "a list or database of subject terms in which all terms or phrases representing a concept are brought together" (Taylor & Joudrey, 2009, p. 334). Thus there was a "controlled" process that determined the creation and selection of those subject terms.

In the early 1980s, online public access catalogs (OPACs) began replacing printed card catalogs and with them came keyword search capabilities. Searchers no longer needed to know the exact title, or find the exact subject heading; instead, keywords could be typed into a search box and if those words appeared anywhere in the title, subject terms or other metadata, the record would be retrieved. And since a record in an OPAC had much more space available than small physical cards, more information could be added to each record such as abstracts, table of contents, and author-assigned keywords. These additions greatly improved the possibility of a successful keyword search which, then, began to call into question the necessity of controlled vocabularies (Gross & Taylor, 2005). With the increasing amount of full text databases, in particular the Internet, available for keyword searching, research on the future of controlled vocabularies is being called for, particularly in full text environments, as Tina Gross and her colleagues have asserted:

> In the long term, the ultimate test of the importance of controlled vocabulary will be its effect in full text environments. While most studies that have looked at the role of subject metadata in full text searching indicate that a controlled vocabulary is needed in full text environments, research in this area needs to continue and expand as the extent and accessibility of full text resources increases. (Gross, Taylor, & Joudrey, 2015, p. 30)

This research seeks to add to this body of knowledge by looking at a full text database and the relationship of Library of Congress subject headings (LCSH) to that database. Specifically, the research will try and ascertain if or how the role of a controlled vocabulary such as LCSH changes with the addition of full text keyword search capabilities.

The Digital Revolution

When OPAC's began offering keyword searching, it initially seemed to be a compliment rather than a replacement for controlled vocabularies. Searchers might use a generic subject term to locate a book then click on the controlled vocabulary term of the book to find other books on that subject. This complimentary relationship seemed to worked equally well with the introduction of online periodical databases. Later, full text databases began to emerge which enabled keyword searching, not only of metadata within an item's record, but also within the body of the item itself. Then the Internet burst onto the scene with its single box, keyword search capabilities. Libraries began to adopt "Googlized" single box, keyword searching of their online information resources. However, though these modern, single-search-box databases could return thousands of records to a searcher in seconds, that searcher would take months to view and consider so many items. As a result, an information seeker often considered the first few pages of results, ignoring the vast majority of returned items. This would certainly seem to make such powerful retrieval systems, in some ways, a waste of time. If over 95 percent of the

retrieved items will be ignored then how can these result be called relevant? And within those hundreds of pages there must surely be some relevant results that are buried too deeply to be found. So, unless the searcher is seeking a single precise item or very specific answer for his or her information needs, an important decision must be made: At what point can the searcher say, I know I have found the most relevant results, and, therefore, I can conclude my search. It would seem that the vast majority of searchers must certainly be finding relevant information within those first two or three pages of results, or search engines like Google and Yahoo would not be so popular. However, since these search engines do not make use of controlled vocabularies, when an appropriate item is located, there is not a way to click on a subject term located in the metadata of the record to help zero in on a specific subject, as there is in most OPACs or periodical indexes. If controlled vocabularies could be successfully transitioned to the Internet, it would surely help improve the search experience.

With the power and speed as well as the enormous size of these new databases and single box keyword search capabilities, manually assigning terms from a controlled vocabulary began to be questioned, especially LCSH since it was never designed for a digital environment (Chan, 2005). These and undoubtedly some other factors have led to what seem to be pertinent and important questions: Are Library of Congress subject headings (LCSH), a controlled vocabulary which has been extensively used by libraries across the United States for the majority of the twentieth century, still valid and necessary in today's twenty-first century digital environments? And if so, how can they be best modified for huge full text databases such as the Internet?

A Historical Foundation

There has been such an explosion of information in the second half of the twentieth century that continues into the twenty-first century, so much so that the current age has come to

be known as the Information Age (Headrick, 2000). However, what is called the age of information, according to Daniel Headrick, actually began in the early seventeen hundreds as better information systems began to be implemented. Then, as now, time was an important component of the information retrieval process. "The time it takes to obtain and use the relevant information puts a premium on the efficiency with which it is organized" (Headrick, 2000, p. 6). An important first step in organizing information is attaching unique names to unique items within specific groups:

> In order to capture a new piece of information and place it in the existing body of knowledge, one must identify it with precision, in other words, give it a distinct name. To avoid ambiguity and confusion, a one-to-one correspondence must exist between every term and the object it represents. (Headrick, 2000, p. 17)

These terms must not only indicate the precise name of objects but these names must also identify the relationships between individual terms. This could be accomplished with a nomenclature or system of names which were designed to express the underlying taxonomies or systems of things (Headrick, 2000). One of the first important nomenclatures was Linnaeus's method of classifying the plant kingdom by class, order, genus, and species which had its roots in the works of Aristotle: "The starting point of the nomenclature was the Aristotelian or Scholastic method of defining things *per genus et differentiam*, that is by naming the genus and describing, in a phrase, what differentiated one particular species from others in its genus" (Headrick, 2000, p. 23). Aristotle described ontological categories using language as a clue while a later philosopher, Kant, used concepts to approach categories of objects: "The goal of the Aristotelian and Kantian category system was to describe the categorical structure that the world would have according to human thought and language" (Almeida, 2013, p. 1687).

Another contribution to this early information age was the beginning of one of the first major reference works, the encyclopedia. Initially, encyclopedias were arranged thematically but because they would grow to be multivolume sets, specific information was not easily found. One solution that continues to the current day was the addition of an index at the end of the work. A second innovation was a replacement of the original thematic arrangement of entries with an alphabetical one which proved to be the triumph of an effective information system over a learning system: "The public favored works in alphabetical order designed for rapid reference, rather than didactic works arranged thematically. Erudition was being replaced by efficient information storage and retrieval systems" (Headrick, 2000, p. 172).

Paul Otlet, a pioneer of documentation in the first part of the twentieth century also sought to represent human knowledge in a systematic and unified way using an ontological framework (Ducheyne, 2009). He sought to provide a representation of the world using a notational system that would provide a scheme which could display the objective relationships between elements: "Ultimately, documentation had to become a 'cosmoscope' whereby all knowable elements of reality and the relations between them could be overseen, comprehended and contemplated" (Ducheyne, 2009, p. 234).

A Theoretical Framework

A controlled vocabulary is not only a tool to aid in the storage and retrieval of information but is also a subset of our natural language which is used for communication both with each other and with automated technologies. For this research, the theoretical framework is anchored with theory from both information science and linguistics. On the linguistic side, research from Noam Chomsky and H.P. Grice is featured, while within information science, the

focus is be on the research of Brenda Dervin, Christine Borgman, and Yan Zhang. This section will conclude with cognitive scientist Don Norman's three part mental model.

Noam Chomsky – Naming

Noam Chomsky has been called the most influential figure in linguistics during the last half century (Wardhaugh & Fuller, 2013). He theorized that "it is the linguist's task to characterize what speakers know about their language, that is, their competence, not what they do with their language, that is, their performance" (Wardhaugh & Fuller, 2013, p. 4). In one of Chomsky's major works, *Reflections on Language* (1975), he explained that "the place of the language faculty within cognitive capacity is a matter of discovery not stipulation" (p. 43). Rather than imposes a system for using language, Chomsky believed that the ways that languages are actually used must be studied, that is, the systems are discovered rather than prescribed. He further explained that our actual language may be the result of the interaction of several mental faculties rather than just language. And these interactions were not a simple process but were quite complex. Even naming and categorizing entities, according to Chomsky, were not as simplistic an endeavor as they might appear:

> Noting that an entity is named such-and-such, the hearer brings to bear a system of linguistic structure to place the name, and a system of conceptual relations and conditions, along with factual beliefs, to place the thing named. To understand "naming," we would have to understand these systems and the faculties of mind through which they arise. (Chomsky, 2007, p. 46)

For Chomsky, these factual beliefs, coupled with what he called "common-sense expectations" also played a role in whether a thing is capable of being name and categorized: "Or

to put it differently, we keep certain factual assumptions about the behavior of objects fixed when we categorize them and thus take them as eligible for naming" (Chomsky, 2007, p. 45).

H. P. Grice – Context and Rules

Grice was a British philosopher who theorized that in a conversation, a speaker's intentions were often quite different from what that speaker's words meant literally (Grice, 1957). Most all of us at one time or another say one thing yet mean something else. Communication was, for Grice, a complex process where context was of central importance to fully understanding meaning:

> Again, in cases where there is doubt, say, about which of two or more things an utterer intends to convey, we tend to refer to the context (linguistic or otherwise) of the utterance and ask which of the alternatives would be relevant to other things he is saying or doing, or which intention in a particular situation would fit in with some purpose he obviously has (e.g., a man who calls for a "pump" at a fire would not want a bicycle pump. (Grice, 1957, p. 387)

In a later work, *Logic and Conversation*, Grice argued that communication between individuals in conversation was an activity which, at best, should be guided by a set of rules. He explained that language, because of meaning problems, could benefit from rules which moved towards an ideal language:

> The proper course is to conceive and begin to construct an ideal language, incorporating the formal devices, the sentences of which will be clear, determinate in truth value, and certifiably free from metaphysical implications; the foundations of science will now be philosophically secure, since the statements of scientist will be expressible (though not necessarily actually expressed) within this ideal language. (Grice, 1975, p. 42)

He developed what he called a cooperative principle which consisted of four maxims:

(1) Quantity (Make your contributions as informative as is required.)

(2) Quality (Do not say what you believe to be false, do not say that for which you lack adequate evidence.)

(3) Relation (Be relevant)

(4) Manner (Avoid obscurity of expression and ambiguity. Be brief (avoid unnecessary prolixity and be orderly) (p. 45-46)

Grice believed that adherence to such rules would result in a more successful and more meaningful information exchange. And his work is still relevant today, as Jens-Erik Mai (2013) explained: "Grice's understanding of meaning and communication provides a solid framework for understanding the production, organization, retrieval, and use of information as creation, generation, and exchange of meaning" (p. 685). Mai also suggested that these principles were not only theoretical but practical as well, acknowledging Grice's striving for an ideal language while at the same time admitting the difficulty of the task: "The maxims acknowledge the messiness of real language, and its inability to capture and represent the world as it actually is, even though this would be the goal of an ideal language" (p. 685).

For Chomsky and Grice, the better our understanding of language, the better our exchange and use of information will be. This should also apply to human-computer interaction:

> To get computers to understand one another, we can program them to communicate unambiguously: but the ultimate goal for a spoken dialogue system is to be able to accommodate all the ambiguity and uncertainty of normal human discourse. (Cummins & Ruiter, 2014, p. 135)

Brenda Dervin – Sense-Making

Brenda Dervin has been developing her theories of sense-making since 1972 and they have evolved into a generalized communication-based methodology useful for the study of sense-making in any context (Savolainen, 2006). Rather than viewing information search and retrieval as a singular activity such as a user inputting terms into a search box and then selecting items deemed appropriate, Dervin's sense-making takes a more holistic approach. That user might be a student writing a paper on a controversial topic and the retrieved information is not only used to complete an assignment, but also to help make sense of the topic. Based on the retrieved information, that student may then engage in a more meaningful conversation with others, and this should help all concerned come to a better understanding of the topic. The process, according to Dervin (2000), is not a series of isolated search sessions but rather an ongoing process of understanding:

> The central idea here is that information is made and unmade in communication—intrapersonal, interpersonal, social, organizational, national, and global. With this view of information, information design cannot treat information as a mere thing to be economically and effectively packaged for distribution. Rather, it insists that information design is, in effect, metadesign: design about design, design to assist people to make and unmake their own informations, their own sense. (p. 43)

Dervin's sense-making also touched on a question which is important to information use studies: Information in context:

> Similarly, while sense-making focuses on the human individual, it does not rest on an individualistic theory of human action. Rather, it assumes that structure, culture, community, are created, maintained, reified, challenged, changed, resisted, and destroyed

in communication, and can only be understood by focusing on the individual-in-context, including the social context. (Dervin, 2000, p. 46)

When viewed through the lens of Dervin's sense-making theories, a controlled vocabulary becomes something more than simply a set of specified terms to help locate appropriate books and articles. It instead becomes a tool to assist in the process of locating the most appropriate information to not only resolve a current information need but also to instruct us "about the nature of the world we live in: its history, its future, its functioning, our place in it, our possible actions, and the potential consequences of those actions" (Dervin, 2000, p. 35).

Christine Borgman and Yan Zhang – The Mental Model

Christine Borgman (1999) presented a mental model of a bibliographic database to some students and user guides of that same database to other students. Her research was based on the mental model theory which proposes that "people can be trained to develop a 'mental model' or a qualitative simulation of a system which will aid in generating methods for interacting with the system, debugging errors, and keeping track of one's place in the system" (p. 435). She found that the students who had been given the mental model did better on complex tasks than those with only user guides. Yan Zhang (2008) did a study of undergraduates' mental models of the Web. The study concluded that three sources contributed to the construction of mental models: personal observation, communication with others, and class instruction.

Don Norman—The Three-Part Mental Model

While Borgman and Zhang have presented effective research within mental model theory, cognitive scientist Don Norman's (2013) three-part mental model seems to be the more appropriate choice for this research project.

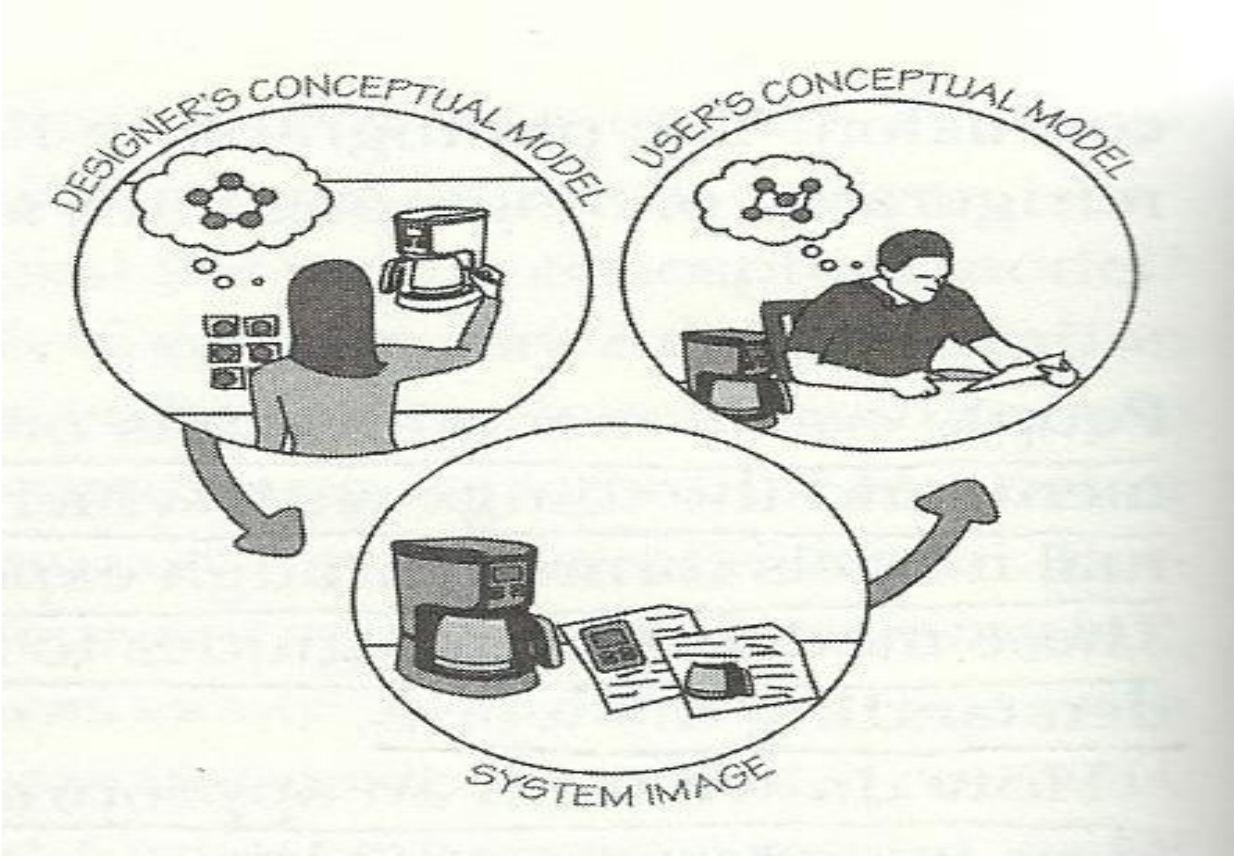


Figure 1.1. Don Norman's three-part mental model diagram.

Norman explained his definition of a mental model which seems to include elements of Dervin's sense-making:

> People create mental models of themselves, others, the environment, and the things with which they interact. These are conceptual models formed through experience, training, and instruction. These models serve as guides to help achieve our goals and in understanding the world. (Norman, 2013, p. 31)

Norman's model is a good example of a user-centered design (Russel-Rose & Tate, 2013). The model has only three components: the designer's conception, the system image, and the user's conceptual model, as Norman explained:

- The designer's conceptual model is the designer's conception of the product, occupying the vertex of the triangle.
- The system image is what can be perceived from the physical structure that has been built (including documentation, instructions, signifiers, and any information available from websites and help lines.
- The user's conceptual model comes from the system image, through interaction with the product, reading, searching for online information, and from whatever manuals are provided. (Norman, 2013, p. 31)

Norman recognized that the burden was on the designer to construct an intuitive system with the user clearly in mind during the design and construction process since the designer would not be available himself to assist the user: "The designer expects the user's model to be identical to the design model, but because designers cannot communicate directly with users, the entire burden of communication is on the system image" (Norman, 2013, p. 31).

In a similar way, catalogers working with controlled vocabularies such as LCSH should also have a mental model of the user when assigning metadata to information resources. As with Norman's model, ideally both the user's mental model and the cataloger's mental model meet when an appropriate LCSH subject terms is selected, resulting in a relevant item being located.

Chapter Summary

Most information professionals would probably agree that an ideal search is one that will retrieve the exact information, in the desired form, within a timely manner, and with a system that requires no special skills or training. This research endeavors to consider the role of a controlled vocabulary such as LCSH within the scope of this online search process. Since users have in a sense "voted with their keyboards" by embracing keyword searches within single

search boxes based on the Google model (Woods, 2010), the high cost and labor-intensive nature of LCSH make this resource seem ever-more prohibitive, especially considering the sheer volume of information resources available in an every-expanding World Wide Web. However, since this venerable resource has helped link information seekers with needed resources for over a century, it would seem unlikely that LCSH would be completely dismantled. But will it need an extensive overhaul or only minor to medium revisions to perform equally well within online environments? Or will it be replaced by a more effective and robust system. This hope is that this research will make a contribution to the literature currently addressing the future of controlled vocabularies such as LCSH.

CHAPTER 2

PROBLEM STATEMENT, RESEARCH QUESTIONS, AND LITERATURE REVIEW

Introduction

Since this research centers on the use of Library of Congress subject headings (LCSH) within the context of keyword searching of a full text database, the primary focus of the literature review is on studies of LCSH and its role in online database searching. This literature review is divided into four sections which will follow the problem statement and research questions. Section 1 deals with ontologies, thesauri, and controlled vocabularies, as well as some individuals who have made important contributions to information storage and retrieval, setting a foundation for section 2 which looks specifically at LCSH, both past and present. Section 3 features articles which deal with LCSH in the digital age. The final section contains articles focusing on the controlled vocabulary versus keyword search debate. This literature review also attempts to create a snapshot of LCSH from its beginnings in the early part of the twentieth century to its current status within an online information infrastructure, the Internet, and other digitization projects, that continue to grow at an astronomical rate.

The Problem Statement

In the late seventies and early eighties, online public access catalogs (OPAC) began appearing in libraries as did automated periodical indexes. However, these new systems were somewhat difficult to use since there were many vendors offering these products, which meant that, although they were similar, each had a unique platform and interface which often required a bit of a learning curve. The Internet, with its powerful search engines, in particular Google, seemed to provide a standard that users embraced: A single search box with keyword search capabilities. Libraries began imitating Google by offering single search boxes for their OPACs

and online periodical databases. The successes of the Internet and Google as well as the popularity of keyword searching have called into question the role of controlled vocabularies such as LCSH. Specifically: Are Library of Congress subject headings, a controlled vocabulary which has been extensively used by libraries across the United States for the majority of the twentieth century, still valid and effective in today's twenty-first century digital environments?

The Research Questions

This research considers Library of Congress subject headings (LCSH) within a full text search environment with the goal of determining the relationship of LCSH to the search process. The study also compares LCSH terms to user-generated keywords and phrases within a full text searching environment.

There are three general questions that this research addresses:

- If a database has the ability to keyword search the title, abstract, and other metadata, as well as the full text of a record, would the role of a controlled vocabulary such as LCSH change because of the addition of full text keyword search capabilities?
- If this research suggest that LCSH should be retained but revised or modified for current online technologies, what types of revisions or modifications are being suggested? And, are there any that seem most promising?
- If a modified or revised version of LCSH still proves to be deficient in modern full text search and retrieval databases, are there systems that can replace its subject search capabilities? Can such systems be incorporated into modern search engines such that they are either seamless or simple enough so that users can use them intuitively?

The Literature Review

Ontology

While Linnaeus was concerned with systematically classifying living organisms, ontologies have been used in modern information science, co-occurring with words such as information retrieval, knowledge management, indexing, and search and navigation (Gilchrist, 2003). It is still a case of grouping similar objects using natural language, an endeavor common to ontologies: "If you have a number of objects, it is possible to arrange them into groups and apply labels to the groups--librarians have traditionally done this in classifying books for arrangement on shelves" (Gilchrist, 2003, p. 15).

Most might agree that ontology is concerned with what kind or category of things can exist. Almeida (2013) suggested that ontological theories should specify category systems structured according to hierarchical levels. He also believed that most followers of Aristotle's writings thought that he was suggesting that a category system should provide an inventory of things that exist. Almeida further explained that the philosopher Kant also had developed categorical systems "to describe the categorical structure that the world would have according to human thought and language" (p. 1687). Category systems are also used to represent document contents for retrieval: "In information science, ontological principles may be used to support the building of categorical structures for representation of the content of documents" (Almeida, 2013, p. 1691). From an application point of view, the most prevalent ontological activity is "developing static ontologies such as taxonomies or controlled vocabularies" (Jurisica, Mylopoulos, & Yu, 2004, p. 393). Ontologies can also constitute a shared vocabulary within a specific subject domain: "Ontology provides a vocabulary for representing knowledge about a domain and describing specific situations therein" (Bhat, 2013, p. 41). Ontology has also been

defined as “a written, formal description of a set of concepts and relationships in a domain of interest” (Boonyoung & Mingkhwan, 2014, p. 371).

Thesauri

While most people think of a book of synonyms when they hear the word, thesauri can also be described as a classified list of terms within a particular field, for use in indexing and information retrieval (Gilchrist, 2003). Thesauri have traditionally been an integral component of the indexing of documents: “Historically, the primary purpose of a thesaurus was to aid in the creation of a subject index to support access to documents in a collection” (Willis & Losee, 2013, p.1331).

A thesaurus can also be thought of as a complex hierarchy which includes an associative relationship between related terms such as *friction* and *wear* or *tolerance* and *prejudice*: “Because of the multiple term relationships they include, thesauri are the most complex controlled vocabularies to create and maintain” (Leise, 2008, p. 123).

Controlled Vocabulary

If a selected piece of information is written down, then it would be reasonable to assume that someone wrote it down, although that person might not have affixed his or her name to the information. It should also be a reasonable assumption that when something unique is written, it will be given a title, which, among other reasons, will help locate the information. And if that information is stored with other information in some organized manner such as in an alphabetical list, a user seeking that information should be able to find it quickly if he or she knows the title of the desired item. However, when a user is not seeking a unique item of information but is instead looking for whatever is available within a specific subject, finding appropriate information becomes more challenging, even for alphabetically arranged subject systems

because human language often has a number of words and synonyms to denote the myriad of subjects that have been, and continue to be, discovered, identified, or invented. Also, some words have multiple meanings for use in different contexts. Controlled vocabularies have been designed to cope with these language difficulties. Fred Leise (2008) defined a controlled vocabulary as a list of terms designed to:

1. Collect similar information,
2. Assist content authors in consistently tagging content, and
3. Enable users to find the information they need by translating their language into the language of the information store" (p. 121).

Terms in a controlled vocabulary can have a hierarchical relationship when one term is broader than another such as the relationship between *automobile* and *engine*, or when those narrower terms are examples of the broader term such as *buildings* and the *Sears Tower* (Leise, 2008). According to Patricia Harping (2013), a controlled vocabulary is "an information tool that contains standardized words and phrases used to refer to ideas, physical characteristics, people, places, events, subject matter, and many other concepts. "Controlled vocabularies allow for the categorization, indexing, and retrieval of information" (p. 1). Harping went on to describe the purpose of a controlled vocabulary, which she said was, "to organize information and to provide terminology to catalog and retrieve information. While capturing the richness of variant terms, controlled vocabularies also promote consistency in preferred terms and the assignment of the same terms to similar content" (p. 13). For Harping, the two most important functions of a controlled vocabulary are: "(1) to gather together variant terms and synonyms for concepts and (2) to link concepts in a logical order or sort them into categories" (p. 13). Another important quality of a controlled vocabulary is not just its array of preferred or alternate terms

but how well that vocabulary represents terms and concepts in the collection which it supports (Willis & Losee, 2013).

Callimachus of Cyrene and the Library of Alexandria

The library of Alexandria was arguably one of the great achievements of antiquity and possibly the most famous library in the ancient world, as Michael Harris (1999) explained: The Alexandrian Library flourished for several hundred years, and for at least 200 years it was of tremendous importance in the cultural development of the Hellenic world. It drew scholars from great distances and from almost all fields of knowledge. Thousands upon thousands of scrolls were bought, copied, stolen, and compiled for its shelves until it contained, according to some estimates, over 600,000 rolls. (p. 45)

Callimachus was associated with the Alexandria Library from 260 to 240 B.C. and though is not known whether he held a role such as head librarian or was simply a library assistant, what is certain is that Callimachus was a seminal figure because he compiled a catalog of the holdings of the great library known as the *Pinakes* or tables (Harris, 1999). The massive work's complete title is *Tables of Persons Eminent in Every Branch of Learning together with a List of Their Writings* (Casson, 2001). The *Pinakes* were divided into eight major subject categories: oratory, history, laws, philosophy, medicine, lyric poetry, tragedy, and miscellany (Harris, 1999). The last category was where cookbooks were listed (Casson, 2001). The work of Callimachus could be described as an early attempt to group similar items around a standard set of subject terms.

Frances Bacon and Jean Le Rond d'Alembert

A major figure in Renaissance thought was the influential English thinker, Frances Bacon, whose views helped bring about what is now known as the scientific method which emphasizes

careful thought and empirical observation (Battles, 2015). His is also famous for organizing knowledge into specific categories which he first presented in 1623:

Class I. History (Memory)

1. Natural History
2. Civil History
    a. Ecclesiastical
    b. Literary
    c. Civil, Proper

Class II. Philosophy (Reason)

1. Science of God
2. Science of Nature
    a. Primary Philosophy
    b. Physics
    c. Metaphysics
    d. Magic
    e. Natural Philosophy
3. Science of Man

Class III. Poetry (Imagination)

1. Narrative Poetry
2. Dramatic Poetry
3. Allegorical Poetry (Brown, 1898, p. 29)

Almost one hundred fifty years later, in 1767, the French philosopher Jean Le Rond d'Alembert would make some additions and modifications to Bacon's system, making it better suited for the state of science in d'Alembert's day (Brown, 1898). It would become known as the Bacon-d'Alembert system (Brown, 1898):

Class I. History

1. Sacred History
2. Ecclesiastical History
3. Civil History
4. Natural History

Class II. Philosophy

1. General Metaphysics or Ontology
2. Science of God

    a. Natural Religion
    b. Revealed Religion
    c. Science of Good and Evil
3. Science of Man
    a. Universal Pneumatology
    b. Arts of Thinking, Retaining, Communicating (= Logic, Writing, Printing, Declamation, Symbolism, Grammar, Rhetoric)
    c. Morals (= Ethics, Jurisprudence, Commerce)
4. Science of Nature
    a. Mathematics
    b. Physics

Class III. Poetry

1. Narrative Poetry
2. Dramatic Poetry
3. Allegorical Poetry
4. Music, Painting, Sculpture, Architecture, Engraving (Brown, 1898, p. 29)

These knowledge classification systems are representative examples of how great thinkers have endeavored to organize and categorized knowledge (Taylor & Joudrey, 2009). The two systems also seem to show how the world's knowledge had been expanded since the eight broad categories of Callimachus. However, for classification of library materials, these systems seem to lack any type of specific subdivisions within the major categories.

Thomas Jefferson's Classification System

When the original Library of Congress had its collection burned by a British army in 1814, Thomas Jefferson offered his collection of over 6,700 volumes, and in early 1815, Congress approved the purchase of this collection for $23,950 (Gilreath & Wilson, 1989). Along with his extensive collection of books, Jefferson also passed along his handwritten catalog which reflected his system for classification of the collection. It was adapted from Francis Bacon's book, *The Advancement of Learning* in which Bacon presented his system of knowledge (Gilreath & Wilson, 1989).

Jefferson's classification system was important because: "There was no standard method for organizing book catalogs at the time, though most printed library catalogs of the period arranged titles either according to the size of the volumes or in broad subject categories" (Gilreath & Wilson, 1989, p. 4). Jefferson made some changes to Bacon's categories and also added forty-four additional sections or "chapters" as he called them, for additional subjects (Gilreath & Wilson, 1989). George Watterston, the Librarian of Congress during this period, retained Jefferson's system and his chapter arrangement but, unlike Jefferson, alphabetized the books within each of the chapters (Gilreath & Wilson, 1989).

Antonio Panizzi

A key figure in the organization of information in the period after the middle ages was Antonio Panizzi. He began work at the British Museum in 1831 as a library assistant and would rise to the level of principle librarian, a post he would hold from 1856 to 1866 (Battles, 2015). His work is considered so important that some believe "modern librarianship begins with Sir Anthony Panizzi" (Koch, 1914, p. 256). During his thirty-five years at the British Museum he would revolutionize the organization of its vast collection (Glasgow, 2001). When Panizzi arrived at the British Museum, its original seven volume catalog, a simple alphabetical list of the books in the museum, had expanded to forty-eight volumes (Battles, 2015). Panizzi saw that a new catalog was needed and proceeded to develop 91 rules for cataloging the collection, with the majority dealing with the creator of the work: "In total there are fewer than twenty rules that do not pertain to authors, editors, translators, and others directly involved with the production of the work" (Smiraglia, Lee & Olson, 2011, p. 140). Another of his innovations, designed to give the patron more independence, was the addition of a pressmark to the entry of each book in the catalog:

> Like a call number on a modern library book, the gnomic pressmark indicated precisely the place where the book was to be found among the shelves of the library stacks (or "presses," as bookshelves were commonly called). Unlike call numbers, however, pressmarks referred not to a scheme of knowledge, but to a location; they are not classification, but only coordinates. (Battles, 2015, p. 132)

With Panizzi's work can be seen the beginnings of the modern library, a collection centered around a catalog that featured works with main access points by author or creator, with a designated number for locating book on the shelves. However, as Battles noted, what's missing was a more precise method of subject arrangement.

Melvil Dewey

It would seem hard to overstate the importance of the classification system that Melvil Dewey invented in 1876. Until then, libraries assigned books to a specific location on a shelf, but as new materials were always being added, catalogs had to be amended and updated, as Matthew Battles (2015) explained: "The old system, in which each book was assigned a fixed spot on a shelf, would no longer do; each new addition of books required an overhaul of the entire catalog" (p. 138). Arlene Taylor and Daniel Joudrey (2009) conveyed the simplicity of Dewey's classification system as well as its seemingly limitless scope:

> He divided all knowledge into ten main classes, with each of those divided again into ten divisions, and each of those divided into ten sections—giving 1,000 categories into which books could be classified. Like its predecessors, it was enumerative in that it listed specific categories one by one. In late editions he added decimals so that the 1,000 categories could be divided into 10,000 then 100,000 and so on. (p. 78)

The Dewey Decimal classification system is into its second century of use and the numbers attest to its incredible success: "It is used in 2,000,000 libraries in 135 countries including national biographies of 60 countries, and has been translated in over thirty languages" (Satija, 2013, p. 277).

Hans Peter Luhn and Keyword Searching

In 1958 at the International Conference on Scientific Information, Hans Peter Luhn, who at the time worked for IBM, presented a system he called "Keyword-in-Context" (KWIC) indexing (Williams, 2010). His work had an almost immediate impact, as Robert Williams (2010) explained:

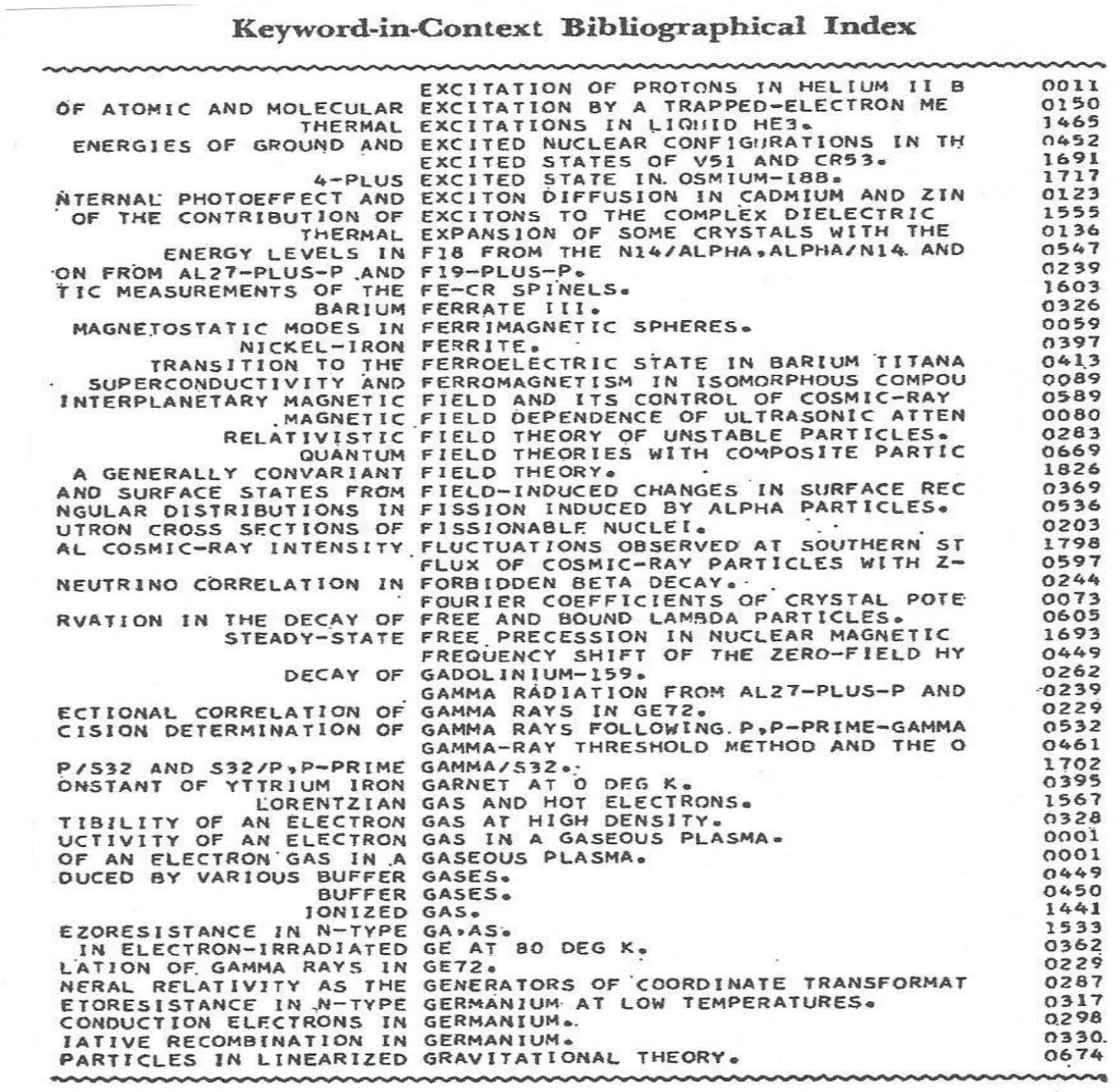
> Chemical Abstract Service (CAS) began using it to produce an index to new chemical publications. The first five monthly trial versions were published in 1960, and in1961, biweekly issues began to be issued. *Chemical Titles* became the first periodical to be organized, indexed, and composed almost completely by computer. (p. 845)

Luhn considered his KWIC system to be somewhat analogous to a concordance in that the keywords would include words in close proximity to them so that the context of the keyword could be considered, hence the name keyword in context:

> In dealing with a variety of subjects, as would be the case in the problem under discussion, the significance of such single keywords could, in most instances, be determined only by referring to the statement from which the keyword had been chosen. This somewhat tedious procedure may be alleviated to a significant degree by listing selected keywords together with surrounding words that act as modifiers pointing up the more specific sense in which a keyword has been applied. This method of indexing

words is well established in the process of compiling concordances of important works of literature in the past. (Luhn, 1966, p. 161)

Interestingly, Luhn defined a keyword somewhat circuitously, by considering words that were not considered significant: “Since significance is difficult to predict, it is more practical to isolate it by rejecting all obviously non-significant or ‘common’ words, with the risk of admitting certain words of questionable status” (Kuhn, 1966, p. 161). Luhn considered the words that were left as significant or “key” words. In his system, the keywords assumed a fixed position with a fixed number of surrounding words retained to the left and right of the keywords, as shown in the diagram below.



Keyword-in-Context Bibliographical Index

	EXCITATION OF PROTONS IN HELIUM II B	0011
OF ATOMIC AND MOLECULAR	EXCITATION BY A TRAPPED-ELECTRON ME	0150
THERMAL	EXCITATIONS IN LIQUID HE3.	1465
ENERGIES OF GROUND AND	EXCITED NUCLEAR CONFIGURATIONS IN TH	0452
	EXCITED STATES OF V51 AND CR53.	1691
4-PLUS	EXCITED STATE IN OSMIUM-188.	1717
NTERNAL PHOTOEFFECT AND	EXCITON DIFFUSION IN CADMIUM AND ZIN	0123
OF THE CONTRIBUTION OF	EXCITONS TO THE COMPLEX DIELECTRIC	1555
THERMAL	EXPANSION OF SOME CRYSTALS WITH THE	0136
ENERGY LEVELS IN	F18 FROM THE N14/ALPHA,ALPHA/N14 AND	0547
ON FROM AL27-PLUS-P AND	F19-PLUS-P.	0239
TIC MEASUREMENTS OF THE	FE-CR SPINELS.	1603
BARIUM	FERRATE III.	0326
MAGNETOSTATIC MODES IN	FERRIMAGNETIC SPHERES.	0059
NICKEL-IRON	FERRITE.	0397
TRANSITION TO THE	FERROELECTRIC STATE IN BARIUM TITANA	0413
SUPERCONDUCTIVITY AND	FERROMAGNETISM IN ISOMORPHOUS COMPOU	0089
INTERPLANETARY MAGNETIC	FIELD AND ITS CONTROL OF COSMIC-RAY	0589
MAGNETIC	FIELD DEPENDENCE OF ULTRASONIC ATTEN	0080
RELATIVISTIC	FIELD THEORY OF UNSTABLE PARTICLES.	0283
QUANTUM	FIELD THEORIES WITH COMPOSITE PARTIC	0669
A GENERALLY CONVARIANT	FIELD THEORY.	1826
AND SURFACE STATES FROM	FIELD-INDUCED CHANGES IN SURFACE REC	0369
NGULAR DISTRIBUTIONS IN	FISSION INDUCED BY ALPHA PARTICLES.	0536
UTRON CROSS SECTIONS OF	FISSIONABLE NUCLEI.	0203
AL COSMIC-RAY INTENSITY	FLUCTUATIONS OBSERVED AT SOUTHERN ST	1798
	FLUX OF COSMIC-RAY PARTICLES WITH Z-	0597
NEUTRINO CORRELATION IN	FORBIDDEN BETA DECAY.	0244
	FOURIER COEFFICIENTS OF CRYSTAL POTE	0073
RVATION IN THE DECAY OF	FREE AND BOUND LAMBDA PARTICLES.	0605
STEADY-STATE	FREE PRECESSION IN NUCLEAR MAGNETIC	1693
	FREQUENCY SHIFT OF THE ZERO-FIELD HY	0449
DECAY OF	GADOLINIUM-159.	0262
	GAMMA RADIATION FROM AL27-PLUS-P AND	0239
ECTIONAL CORRELATION OF	GAMMA RAYS IN GE72.	0229
CISION DETERMINATION OF	GAMMA RAYS FOLLOWING P,P-PRIME-GAMMA	0532
	GAMMA-RAY THRESHOLD METHOD AND THE O	0461
P/S32 AND S32/P,P-PRIME	GAMMA/S32.	1702
ONSTANT OF YTTRIUM IRON	GARNET AT 0 DEG K.	0395
LORENTZIAN	GAS AND HOT ELECTRONS.	1567
TIBILITY OF AN ELECTRON	GAS AT HIGH DENSITY.	0328
UCTIVITY OF AN ELECTRON	GAS IN A GASEOUS PLASMA.	0001
OF AN ELECTRON GAS IN A	GASEOUS PLASMA.	0001
DUCED BY VARIOUS BUFFER	GASES.	0449
BUFFER	GASES.	0450
IONIZED	GAS.	1441
EZORESISTANCE IN N-TYPE	GA,AS.	1533
IN ELECTRON-IRRADIATED	GE AT 80 DEG K.	0362
LATION OF GAMMA RAYS IN	GE72.	0229
NERAL RELATIVITY AS THE	GENERATORS OF COORDINATE TRANSFORMAT	0287
ETORESISTANCE IN N-TYPE	GERMANIUM AT LOW TEMPERATURES.	0317
CONDUCTION ELECTRONS IN	GERMANIUM.	0298
IATIVE RECOMBINATION IN	GERMANIUM.	0330
PARTICLES IN LINEARIZED	GRAVITATIONAL THEORY.	0674

Figure 2.1. Example of Luhn's keyword in context.

Luhn also seemed to anticipate future possibilities of keyword searching by suggesting that not only the title, but the abstract and even the text could be searched for keywords: "It will be a matter of experience as to whether KWIC indexing needs to be extended to include abstracts or even parts of the text in order to provide the degree of resolution required under given circumstances" (Luhn, 1966, p. 163). Still, Luhn might have raised an eyebrow if he would have been told how universal keyword searching would become in the years ahead.

The Problem of Relevance

The user of an ancient library would have probably been confronted with the same problem an Internet searcher faces today: How to select and retrieve relevant items from the myriad of materials that are available. In the library of Alexandria, that could have meant sifting through up to six hundred thousand scrolls (Harris, 1999). Today's Internet searcher can have a similarly daunting task since the Internet contains literally billions of sites (*Internet Live Stats* (http://www.internetlivestats.com/). Information seekers from ancient times to the modern era have wrestled with the problem of locating relevant information. W. S. Cooper (1971) considered relevance as a primary concept in information retrieval:

> "Relevance" is one of the most fundamental, if not *the* fundamental, concept encountered in the theory of information retrieval. The concept arises in this way: If a user of an information retrieval system has some definite information need, then it seems reasonable to say that some of the information stored in the system is "relevant" to his need, and that the rest is "irrelevant." (p. 19)

In the same article, Cooper identified a major challenge to locating relevant information: The system storing the information and the user requesting information both use natural language:

> The aim, then, is to define relevance as a relationship holding between pieces of stored information on the one hand and the user's information needs formulated as information need representation on the other. Moreover, both the stored information and the information need representations are assumed to be linguistic entities of some kind. (p. 22)

David Blair (2006), in his study of Ludwig Wittgenstein, related the philosopher's studies of language and meaning to modern information systems. Blair showed how Wittgenstein's work had a natural affinity to the field of information because, as Cooper (1971) noted above, both the user and the system use natural language in the information retrieval process:

> Specifically, ordinary language is the best medium for us to express our information needs, and any subset of ordinary language that may be used as an access language to the information system will be correspondingly less effective than ordinary language for searching. (Blair, 2006, p. 18)

For Blair, the closer an information system came to understanding natural language, the better that information system should be in retrieving relevant information for the user.

While our language is arguably quite sophisticated, it can often be vague or imprecise. Different searchers seeking the same information can often use different terms while, conversely, different indexers may assign different terms to the same document. M. E. Maron and J. L. Kuhns (1960) called this inexactness of language *semantic noise*:

> Just as the correspondence between the information content of a document and its set of indexes is not exact, so also the correspondence between a user's request, as formulated in terms of one or many index words, and his real need (intention) is not exact. Thus there is semantic noise in both the document indexes and in the request for information. (p. 219)

Inherent in this noisiness of language is the problem of context, as Maron and Kuhns further explained:

> That is to say, the meaning of a term in isolation is often quite different when it appears in an environment (sentence, paragraph, etc.) of other words. The grammatical type,

position and frequency of other words help to clarify and specify the meanings of a given term. (Marion & Kuhns, 1960, p. 219)

Controlled vocabularies were clearly an attempt to address the problem of language, meaning, and context by grouping similar resources around specially selected words and phrases. The obvious goal of this process was and still is a system to make it easier to locate relevant materials.

Library of Congress Subject Headings

It has been called the most comprehensive non-specialized controlled vocabulary in the English language (Chan, 2005). Library of Congress subject headings (LCSH) is also well known to library professionals: "People with library backgrounds tend to default to what they know. *LCSH* is a standard format familiar to librarians" (Walsh, 2011, p. 332). Initially designed for printed materials, LCSH is also the most popular controlled vocabulary for subject access in digital collections (Walsh, 2011).

LCSH: The Beginning

In 1898 the Library of Congress converted from an "author plus a classed-catalog to a dictionary catalog, which incorporated author, title, and subject entries into a single file" (Stone, 2000, p. 2). Lois Chan (2005) pointed out that the dictionary format for the card catalog was based on Charles A. Cutter's book, *Rules for a Dictionary Catalog* which had been published in 1876, noting that Cutter's statement of "objects" can still be used to describe the function of subject entries:

1. To enable a person to find a book of which . . . the subject is known [and]
2. To show what the library has . . . on a given subject [and] in a given kind of literature. (Chan, 2005, p. 9)

Chan further explained that while Cutter's setting was a library with books located by using a card catalog, his comments can be generalized to a wider milieu: "for 'book' read 'library material' or 'information resources'; for 'what library has' read 'what is available'" (Chan, 2005, p. 9). Elaine Svenonius (2000) presented an effective example of why Cutter was opposed to many of the classification catalogs of the day, which could meet the needs of scholars but seemed less effective with the ordinary public:

> How, for instance, could a non-scholar find something on the badger if it was necessary first to look under **Science**, then under **Natural History,** then under **Zoology**, then under **Vertebrates**, then under **Mammals**, then under **Monodelphi**, then under **Carnivora**. (p. 26)

That Cutter was concerned with finding aids useful to the general public is also reflected in a bit of advice he gave about the importance of simplicity: "The reader at first glance is frightened by the appearance of a system to be learned, and perversely regards it as a hindrance instead of an assistance" (Cutter, 1904, p. 123). His comment would be as appropriate in the modern digital age as in his own time.

A second significant event occurred in 1902 when the Library of Congress began a service specifically targeted at the library community, as Alva Stone (2000) explained:

> Meanwhile, with the ALA community clamoring for greater standardization as well as more cooperative cataloging, there arose a steady and appreciative market for the LC catalog Card Distribution Service, which began in 1902. To the larger and "outside" library world, then, that was when subject headings formulated and assigned by LC began to be noticed and utilized. (p. 2)

In the summer of 1909 the first printing of *Subject Headings Used in the Dictionary Catalogues of the Library of Congress* began which would later be titled *Library of Congress Subject Headings* (Stone, 2000). From these seemingly humble events, the Library of Congress subject headings would, in the course of the twentieth century, expand from a single subject access system designed for one library to become the most used subject retrieval tool for libraries not only in the United States but in other countries throughout the world: "Libraries that have adopted, translated, or adapted LCSH as the basis for their controlled vocabularies include those in Belgium, Brazil, Canada, the Czech Republic, France, Great Britain, Lithuania, Malaysia, and Portugal" (Chan & Hodges, 2000, p. 226). Making its cataloging records available to other institutions was a key factor in LCSH becoming so widely accepted, first as printed cards and later in online environments: "More recently, with the advent of the online age, MARC records have been distributed electronically. And since 1993, LC cataloging data and LCSH itself have also been accessible online through the Internet" (Chan & Hodges, 2000, p. 227).

LCSH and the Digital Age

As more and more information sources are digitized, a major weakness of LCSH becomes more apparent; it was designed to retrieve printed materials from physical libraries (Walsh, 2011). Even before the digital age, LCSH presented persistent difficulties, as Karen Fischer (2005) explained: "Six decades of literature demonstrate persistent complaints about LCSH: complicated syntax, inadequate syndetic structure, outdated terminology, lack of specificity in the list, and complicated, inconsistent application of subdivisions" (p. 65). Fischer also noted that: "Newer criticism targets the lack of adaptability and flexibility of LCSH in the online environment" (p. 74). LCSH was not designed to deal with digital information systems which can use different search and retrieval systems: "Unlike traditional libraries that use

Library of Congress Classification for organization and retrieval, digital libraries use metadata forms for organization and retrieval" (Walsh, 2011, p. 329). Another difficulty LCSH presents for transition from print to digital materials is its complexity: "Such a complex categorization scheme, developed for manual document classification, may not be suitable for automatic document classification" (Pong, Kwok, Lau, Hao, & Wong, 2008, p. 219). Another major problem with LCSH, one that is not confined to the digital age, is consistency. Library of Congress subject headings are not always consistently applied by human experts (Frank & Paynter, 2004). And because of its complicated syntax and intricate rules for application and implementation, LCSH requires highly trained personnel to appropriately and accurately assign its subject headings to documents (Chan, 2005). There simply aren't enough trained professionals to facilitate the catalog of all digital resources (Harper & Tillet, 2009). And as digital collections continue to grow in size, manually assigning controlled vocabulary terms will become prohibitive (Frank & Paynter, 2004). The LCSH headings can be quite complex, as Elaine Svenonius (2009) explained:

> An LCSH string begins with a main heading that focuses on the aboutness of the document to be described. This may or may not be followed by qualifying terms called *subdivisions.* The LCSH syntax rules specify when these subdivisions can be used and in what order. (p. 179)

Svenonius went on to give examples of the most common syntactic constructions used in LCSH with respect to the four facets that the system employs: Topic, Place, Time, and Form:

> Topical main heading—Place—Topic—Time—Form
>
> ex.: Art criticism—France—Paris—History—Nineteenth century—Bibliography

Topical main heading—Topic—Place—Time—Form

ex.: Art—Censorship—Europe—Twentieth Century—Exhibitions

Geographic main heading—Topic—Time—Form

ex.: France—Intellectual life—Sixteenth century—Periodicals (Svenonius, 2000, p. 179)

The three above examples reveal the complex structure of LCSH, which does, however, works quite well for moving forward or backward through a card catalog. Once an appropriate item is selected from the card catalog, the user would then go to the stacks, locate the item, and then also move forward or backward to see the actual books which would be similar in subject. But this system does not seem to be a good fit for digital environments which do not have a linear arrangement like a card catalog or books on a shelf. Lois Chan (2005) also finds LCSH lacking in regards to online environments, especially the web: "LCSH as it stands is too cumbersome, with too many intricate rules for forming subject strings, for it to be effective in dealing with the enormous scope of web resources" (p. 13).

In spite of these shortcomings there are those who strongly believe that LCSH can still play a prominent role in online information retrieval by complimenting the keyword search capabilities of most digital environments (Tillotson, 1995; Harper & Tillet, 2007). Research by Strader (2009) and a similar study using similar methodology by Schwing, McCutcheon, and Maurer (2012) also concluded that when used together, LCSH and keywords enhance access:

> Thus, while the keywords tend to represent more current, cutting edge ideas, as well as terms that are more specific within the sciences, LCSH, in contrast, tends to be more stable and to connect to broader subjects. This complimentary nature means that there is

value in the uniqueness of both keywords and LCSH in comparison to one another. (Schwing, McCutcheon, & Maurer, 2012, p. 924)

For these and other researchers who favor the use of LCSH in digital environments, there is still the question of how best to utilize it? In a recent study that considered the potential of LCSH for subject access to digital resources, the researchers suggested the need for more powerful algorithms which could go beyond simple word matching:

> In LCSH-based automated subject indexing, the development of sophisticated algorithms for linking LC subject headings to target vocabulary is crucial as current state-of-the-art automated algorithms rely primarily on simple word matching. Therefore, for terms that are not in LCSH, a novel algorithm or approach should be devised to link them to proper LC subject headings. (Yi & Chan, 2010, p. 685)

Research by Yi and Chan (2009) in another study, investigated the feasibility of creating a mechanism for matching relevant LCSH terms to folksonomies while at the same time highlighting the difficulty of mapping a controlled vocabulary to social, user-assigned tags. They presented LCSH terms as a tree structure in order to graphically represent the complexities of its design. Their results showed that roughly 61 percent of user-assigned tags could be found in LCSH, with an additional 10 percent of the remaining tags having potential matches. Yi and Chan concluded that matching LCSH to folksonomies could produce favorable results: "Implicitly, this study demonstrates the feasibility of using LCSH for the discovery of web resources" (Yi & Chan, 2009, p. 898).

Another problem becoming ever more prevalent as digital libraries grow exponentially in size is the unexpectedly large volume of search results which can be generated by current keyword search methods. When presented with voluminous pages of results, users often

abandon a search after viewing only a fraction of the items returned. Other researchers are calling for new methods for implementing and refining a search so that users are not overwhelmed by their results, which often happens with traditional keyword searches:

> Therefore, in order to facilitate precision search and discovery of archived materials which enables patrons to focus their exploration efforts on the most relevant items of interest and reduces the recall effort, i.e. the ratio of desired to examined, we need to go beyond the traditional keyword-based search methods currently employed. (Joorabchi & Mahdi, 2013, p. 726)

Joorabchi and Mahdi also believed that controlled vocabularies such as LCSH could play a role in the improvements they have called for.

LCSH vs. Keyword Searching

Sevim McCutcheon (2009) looked at some of the strengths and weaknesses of both keyword and controlled vocabulary searching. Focusing specifically on the Library of Congress subject headings, she admitted that this resource was originally designed and developed for print materials and that the powerful keyword searching as well as automated indexing capabilities of modern computerized databases have many questioning the usefulness of LCSH in our modern era. McCutcheon listed some of the strengths of keyword searching as:

1. Speed: There is virtually no delay in a word appearing online and its being keyword searchable
2. Versatile: Keyword searching works quite well for retrieving factual information from multiple sources
3. Ease of use: Keyword searching is both convenient and intuitive (p. 62)

McCutcheon noted that keyword searching usually returned a huge amount of items, while only a portion would pertain to the subject being searched. She also pointed out that current keyword searching cannot show relationships between terms and concepts, and that many search results have the correct search term but in the wrong context.

Her list of Library of Congress subject headings strengths were:

1. Concepts: The LC terms are not just words but often phrases that describe important concepts
2. Hierarchy; headings and subheadings move the searcher from general terms and concepts to more specific ones
3. Controlled vocabulary: The LCSH has a one hundred plus year history, is widely used, and is taken as a model around the world (p. 63)

The author considered complexity and cost to be the two main weaknesses of the LCSH scheme: "It is not as simple and intuitive as keyword searching and it is costly to build, to maintain, and to update" (McCutcheon, 2009, p. 63).

McCutcheon concluded her study by suggesting that keyword and controlled vocabulary were not equivalent and, therefore, not interchangeable. She considered the two search methods complimentary and should, she suggested, be used together.

Thomas Mann (2008) explained his views on the limitations of Google keyword searching, especially in the area of academic scholarship. When compared with the Library of Congress classification systems, he found Google's ability to search by conceptual categories, an important aspect of scholarly research, particularly lacking. He suggested that Google's search mechanisms and relevancy ranking methods worked best for quick information seeking rather

than scholarship. Mann also noted that Google's keyword search capabilities could not limit search results to the right word in the right context.

Jeffrey Beal (2008) described what he considered to be numerous problems associated with full text searching of online databases. He defined full text searching as: "The type of search a computer performs when it matches terms in a search query with terms in individual documents in a database and ranks the results algorithmically" (p. 438).

Some of the problems Beal associated with full text searching were:

- The Synonym Problem: The author considered this possibly the most pervasive weakness of full text searches. As he pointed out, a concept often has more than one word or phrase associated with it, and a search engine will only return results from the terms that have been entered. He used the example "leprosy" and "Hansen's disease."
- Variant Spellings: The search engine will only return results as the searcher spells them, but many terms have variant spellings. Examples given were "harbor/harbour" and "donut/doughnut." The author also pointed out that there are at least sixty different terms that all mean "Atlantic cod."
- Homonyms: Often a single word will have more than one meaning such as "cookies" for computers and "cookies" that are baked. The author suggested that homonyms could be the chief cause of low search precision.
- Disambiguation of Personal Names: The author described name disambiguation as a process in which the database makes each person's name unique. He noted that since this process necessarily involved adding metadata, most all full text documents would lack this feature.

- The Aboutness Problem: The author pointed out that individual words and even complete sentences do not always map directly to phenomena in the real world. Books and articles often have clever titles that fail to describe the work's content.
- Search Term not in a Resource: The author used an example of the golfer Arnold Palmer getting two holes in one on the same hole on consecutive days. An article describing this incredible feat never mentioned the word "golf."
- Searcher Doesn't Know a Term: As the author stated: "When a searcher does not know the correct term for a concept, it can be very difficult for the searcher to find desired information" (p. 442).

Karen Calhoun (2006) conducted interviews with library, academic, and information professionals between October and December 2005 for a report on the changing nature of the library catalog that was prepared for the Library of Congress. She reported that, "There were no strong endorsements for LCSH" (p. 33).

Some of the less critical comments from the interviews were:

- There is a need for subject cataloging in the context of clustering related content. LCSH is not ideal but it offers a readily available means of labeling clusters.
- For subject access, is the technology good enough so we can move away from manually assigned controlled vocabularies?
- It is hard to say how well LCSH serves as a source of controlled terms, but it is better than relying on keywords alone.

Some of the more critical comments from the interviews were:

- Now with the ability to search full text or even TOCs, do we need subject analysis for textual materials?

- LCSH requires too much behind-the-scenes understanding to be useful.
- There is a real question whether LCSH is cost effective. (Calhoun, 2006, p.33)

Tina Gross (2005) and Arlene Taylor wanted to find out what proportion of records returned by a keyword search would be lost without a keyword appearing in the subject headings. They took 3,397 transaction logs of keyword searches from a university library catalog. They discovered that 35.9% of successful keyword searches would be lost without subject headings.

Gross (2015) and her colleagues followed up the 2005 article with a more comprehensive study, questioning whether keyword searching has made the use of controlled, subject vocabularies, with their inherent high cost to produce and maintain, obsolete. They first looked at the current body of literature centered on this debate and then did a research study of their own. Some article comments in favor of discontinuing controlled vocabularies in favor of keywords were:

- "Using controlled vocabularies such as LCSH and MSH (Medical Subject Headings) for topical subjects is no longer as necessary or valuable" (p. 6).
- "Abandon the attempt to do comprehensive subject analysis manually with LCSH in favor of subject keywords: urge LC to dismantle LCSH" (p. 7).
- "While it is recognized as a powerful tool for collocating topical information, LCSH suffers, however, from a structure that is cumbersome from both administrative and automation points of view" (p. 7).
- "Clay Shirky, in a blog posting about ontologies in 2005, asserted that categorization belongs to a world where things are placed on shelves, not in the digital world" (p. 9).

- "Overall, the research from both transaction log analysis and user-response studies shows that subject searching is difficult for patrons, unlikely to be very successful, and becoming less frequent as patrons' behavior is shaped by keyword search engines such as Google" (p. 9).

Some article comments in favor of retaining controlled vocabularies were:

- "We need to be able to explain and defend the added value of subject thesauri in the databases for which we pay a considerable percentage of our materials budget" (p. 5).
- "For a quick, cursory search, keyword searching is promising even on the Web; but for more in-depth or extensive searches, the limitations of keyword searching, such as the lack of control over synonyms and the need for content to make the words more specific, will result in many irrelevant items for the searcher to wade through" (p. 8).
- "Controlled vocabulary offers the benefits of consistency, accuracy, and control . . . which are often lacking in the free-text approach" (p. 9).
- "Keyword searching 'cannot segregate the appearance of the right words in conceptual contexts apart from the appearance of the same words in the wrong contexts" (p. 10).
- "For keyword searching of bibliographic records, including those that have been given tags by users of the system, most studies show that controlled vocabularies cannot be replaced by keyword searching for in-depth, scholarly work" (p. 23).

Gross and Taylor's 2015 research looked at university library transaction logs of keyword searches from an online catalog. A sample size of 227 searches was selected for the study. When the results of all searches were aggregated, the authors concluded that 27.7% of search results would be lost without subject headings.

While the majority of articles mentioned so far in this review have pointed out the complementary nature if not the necessity of using controlled vocabularies within keyword searches, a 2007 article by Bradley Hemming (2007) and colleagues suggested that controlled vocabularies may no longer be necessary within a full text domain. They used the biomedical terms *Arabidopsis* and *schizophrenia* to search in full text journals containing articles that would include these terms. Both a metadata search and a full text search were performed and with both terms more articles were discovered by full text searching than with metadata searching. The researchers did concluded that metadata searches, while having a smaller recall rate when compared to full text searching, had a higher precision rate. They also suggested that a relevance ranking feature that could filter articles based on their usefulness to the searcher could help improve full text relevance. The researchers ranked the results from each search by judging each retrieved article on a five-point scale from *Definitely Useful* to *Definitely Not Useful*. They found a correlation between the number of hits of the search term in the full text article and the article's usefulness ranking. Hemming and his colleagues concluded that keyword searching within a full text database my render metadata such as a controlled vocabulary unnecessary: "This suggests that rather than accepting metadata searching as a surrogate for full text searching, it may be time to make the transition to direct full text searching as the standard" (Hemminger, Saelim, Sullivan, & Vision, 2007, p. 2350).

Jeffrey Garrett (2007) also did a study of subject headings within a full text environment. He looked at the feasibility of adding subject headings to an eighteenth century online collection. Conversely, he found that keyword searches could miss an enormous amount of relevant information because the words in use today to describe a historical topic may be entirely different than those used in the eighteenth century. He pointed out that words such as *hygiene*

and *prostitution* are used far more frequently today than in that time period. He concluded that adding subject terms to items in the collection was beneficial: "The fact is that the assignment of descriptive language in the subject heading fields frequently attaches important terms and concepts to a bibliographic record that the record will not otherwise contain" (p. 74). Garret did, however, conceded that more powerful search algorithms could make manually-assigned subject terms unnecessary: "Another direction for further research could be to investigate whether 'smart' relevance-determining algorithms run against full text can produce distillations of content, replacing the need for manually assigned subject headings" (p. 75).

Chapter Summary

Most of the reviewed articles addressing controlled vocabularies either stated or strongly suggested that a controlled vocabulary used in conjunction with keyword searching was still the best method for retrieving relevant information within an online environment. The majority of the articles also asserted the importance of the Library of Congress subject headings as arguably the most popular controlled vocabulary in use today. However, there were still those researchers who thought controlled vocabularies might be of less importance or even unnecessary in digital environments, especially full text environments. A few of the articles in this literature review also suggested that an appropriate automated ranking algorithm within a keyword/full text searching environment could make controlled vocabularies less relevant.

CHAPTER 3

METHODOLOGY

Introduction

This research looks at Library of Congress subject headings (LCSH) within a full text keyword search environment. LCSH subject terms are also compared to author-assigned keywords and phrases within both a metadata and a full text searching environment to try and ascertain the relationship between these two metadata items. The primary goal of this study is to try and determine whether the roll of a controlled vocabulary such as LCSH changes with the addition of full text search capabilities. A second goal of this research is to try and determine the importance of an LCSH subject term or word in an LCSH term if it is unique to a record, that is, the term or individual word appears nowhere else in the metadata or the full text of an item, but is only in the LCSH field of the metadata.

Research Design

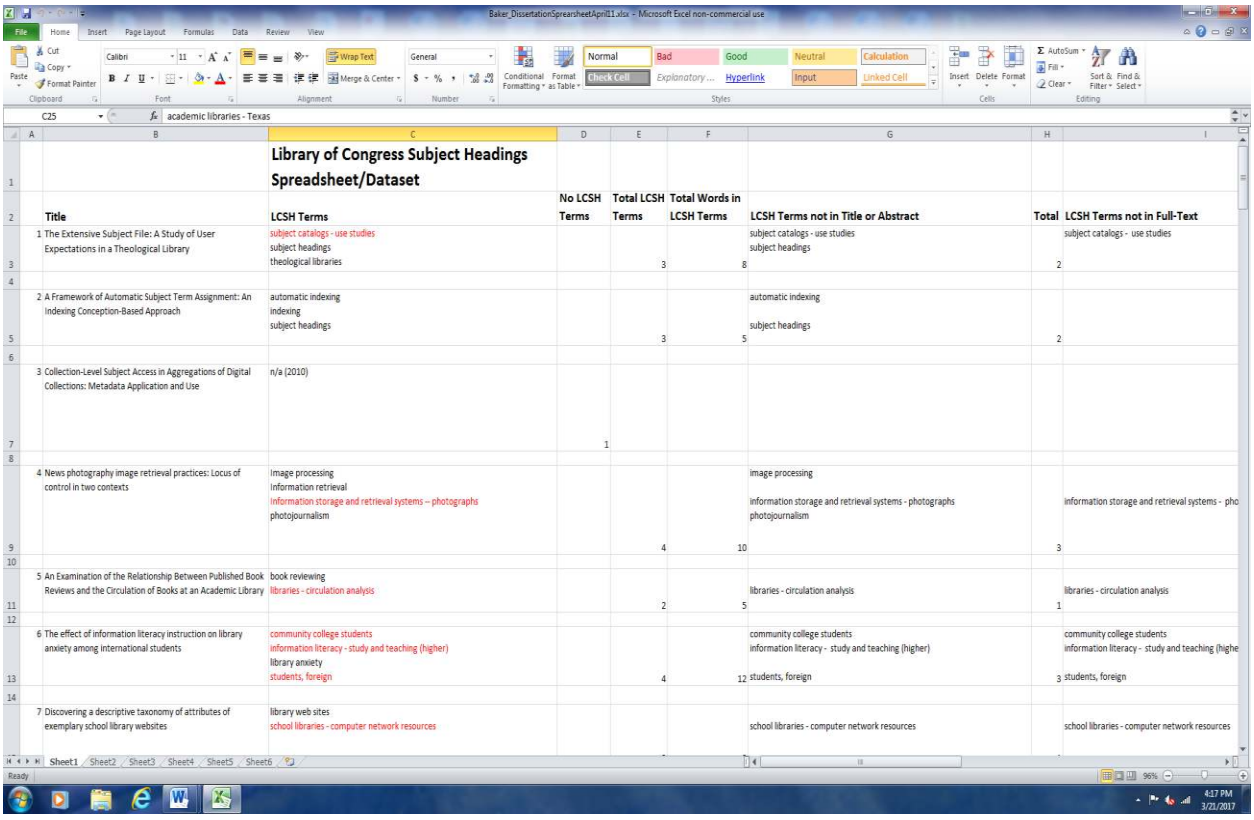
UNT Theses and Dissertations digital database was selected to build a dataset (https://digital.library.unt.edu/explore/collections/UNTETD/) to create the quantitative data for this research since it has a metadata page with both LCSH terms and author-supplied keywords. The database also contains the full text of each thesis and dissertation which can also be searched by keyword. While the pages which contain the title, abstract, and metadata as well as the full text can all be searched simultaneously, the metadata and the full text of the article can each be searched separately.

The terms *library of congress subject headings* are used for the initial search. The objective is to return any dissertation that contained at least one of these terms. With this research, the goal is not to look at a single set of comprehensive search results but rather an

eclectic selection of dissertations so that the relationship of the LCSH terms in each dissertation could be compared to both the metadata and the full text. A preliminary search, limited to dissertations only, yields eighty-six records. *Microsoft Excel 2010* is used to build a dataset which includes the dissertation titles in retrieval order, the LCSH terms for each title, how many of these terms are contained in the title, abstract, and full text of each dissertation, including both exact words or phrases and partial words or phrases. This data is first studied quantitatively and then a more qualitative study is undertaken.

The Spreadsheet/Dataset

The *Excel* spreadsheet consists of eighty-six rows, one for each dissertation. There are twenty-eight columns. Below is a screenshot showing the first nine columns.



	Library of Congress Subject Headings Spreadsheet/Dataset						
	Title	LCSH Terms	No LCSH Terms	Total LCSH Terms	Total Words in LCSH Terms	LCSH Terms not in Title or Abstract	Total LCSH Terms not in Full-Text
1	The Extensive Subject File: A Study of User Expectations in a Theological Library	subject catalogs - use studies subject headings theological libraries		3	8	subject catalogs - use studies subject headings	2
2	A Framework of Automatic Subject Term Assignment: An Indexing Conception-Based Approach	automatic indexing indexing subject headings		3	5	automatic indexing subject headings	2
3	Collection-Level Subject Access in Aggregations of Digital Collections: Metadata Application and Use	n/a (2010)	1				
4	News photography image retrieval practices: Locus of control in two contexts	Image processing Information retrieval Information storage and retrieval systems – photographs photojournalism		4	10	image processing information storage and retrieval systems - photographs photojournalism	3
5	An Examination of the Relationship Between Published Book Reviews and the Circulation of Books at an Academic Library	book reviewing libraries - circulation analysis		2	5	libraries - circulation analysis	1
6	The effect of information literacy instruction on library anxiety among international students	community college students information literacy - study and teaching (higher) library anxiety students, foreign		4	12	community college students information literacy - study and teaching (higher) students, foreign	3
7	Discovering a descriptive taxonomy of attributes of exemplary school library websites	library web sites school libraries - computer network resources				school libraries - computer network resources	

Figure 3.1. Screen shot of first nine columns of spreadsheet/dataset.

Quantitative Analysis

The goal of this study is to analyze LCSH terms assigned to a dissertation to try and determine how necessary the terms are when full text search capabilities are available as opposed to only metadata. For each of the records that have been assigned LCSH subject headings, both the abstract and the full text are searched with each individual subject heading. This first step is to determine if an exact match occurs either in the abstract or the full text. Next, the abstract and full text are again searched with the subject term, this time to determine if a partial match, i.e. any word or phrase from the subject term, is present. Lastly, the author-assigned keywords and key phrases are compared to the subject headings to also find exact or partial matches. The results are then quantitatively compiled to determine what percentage of subject headings are unique to each record and what are the partial-match percentages. There is also an attempt to find a correlation between the cataloger-assigned LSCH terms and the author-assigned keywords.

Qualitative Analysis

After the data is compiled quantitatively, each of the unique subject terms, LCSH terms that did not appear in the title, abstract, or the full text, is compared to the content of the dissertation to which it was attached. Knowing that a subject cataloger assigned the term based on what he or she felt was appropriate or necessary even though the author assigned different terms, what is the relationship between the term, the dissertation, and relevancy to a prospective searcher? This process is repeated for each dissertation containing a unique LCSH term with a goal of establishing a correlation between the term and the dissertation. A key point of the research is to try and determine the importance of a term that has been laboriously assigned by cataloger or subject specialist to a dissertation when that term appears nowhere else in the

metadata, the abstract, the author-assigned keywords, or the full text.

A second part of this qualitative study is a comparison of these unique terms as they relate to the other metadata of the record such as the title and the abstract. These results are compared to a similar qualitative study for each term as it relates to the full text that the article is associated with. The goal here is to try and ascertain, in a more qualitative way, the contribution the LSCH term makes to the relevance of the article that it is attached to as well as the relationship of the author-assigned keywords to the LCSH terms.

The 28 Excel Spreadsheet/Dataset Columns

From the eighty-six dissertations, LCSH headings and author assigned keywords for each dissertation were extracted and input into an Excel spreadsheet along with quantitative data for both. The final layout included twenty-eight columns and eighty-six rows. Below is the title of each of the twenty-eight columns along with the first two to five rows of information. (All 86 dissertation titles are listed in the appendix)

Table 3.1

Title

	Title
1	The Extensive Subject File: A Study of User Expectations in a Theological Library
2	A Framework of Automatic Subject Term Assignment: An Indexing Conception-Based Approach
3	Collection-Level Subject Access in Aggregations of Digital Collections: Metadata Application and Use

- Number: The first column lists the number of the dissertation. They are listed in the order that they were retrieved, from 1 to 86.
- Title: The title of each dissertation.

Table 3.2

LCSH terms

LCSH Terms	No LCSH Terms	Total LCSH Terms	Total Words in LCSH Terms
subject catalogs - use studies subject headings theological libraries		3	8
automatic indexing indexing subject headings		3	5
n/a (2010)	1		

- LCSH Terms: The Library of Congress subject headings assigned to each dissertation. If the dissertation did not contain LCSH terms, an n/a will appear in the box along with the year the dissertation was added to the database.
- No LCSH Terms: If the dissertation did not have LCSH terms, this box will have a 1 so that the total number of dissertations without LCSH terms can be tabulated at the bottom of the spreadsheet.
- Total LCSH Terms: The total number of LCSH terms for each dissertation, with the total from all dissertations at the bottom of the spreadsheet.

- Total Words in LCSH Terms: The total number of individual words from all the terms in each dissertation, with the total from all dissertations at the bottom of the spreadsheet.

Table 3.3

LCSH Terms not in Title, Abstract, or Full Text

LCSH Terms not in Title or Abstract	Total	LCSH Terms not in Full text	Total
subject catalogs - use studies subject headings	2	Subject catalogs - use studies	1
automatic indexing subject headings	2	Automatic indexing	1
Image processing information storage and retrieval systems – photographs photojournalism		Information storage and retrieval systems - photographs	

- LCSH Terms not in Title or Abstract: The LCSH terms that did not appear in the title or abstract as the exact word, words, or phrase.
- Total**:** The number from the adjacent column, with the total from all dissertations at the bottom of the spreadsheet.
- LCSH Terms not in Full Text: The LCSH term did not appear in the full text of the dissertation as the exact word, words, or phrase.
- Total: The number from the adjacent column, with the total for all dissertations at the bottom of the spreadsheet.

Table 3.4

LCSH Words not in Title, Abstract, or Full Text

LCSH Words not in Title or Abstract	Total	LCSH Words not in Full Text	Total
catalogs, use, studies headings	4		
headings	1		
processing storage photojournalism	3		
Libraries analysis	2		
counseling secondary education	3	counseling	1

- LCSH Words not in Title or Abstract: Individual words from the LCSH terms that did not appear in the title or abstract of the dissertation.
- Total: The number from the adjacent column, with the total at the bottom of the spreadsheet.
- LCSH words not in the full text: Individual words from the LCSH terms that did not appear in the full text of the dissertation.
- Total: The number from the adjacent column, with the total at the bottom of the spreadsheet.

Table 3.5

Partial LCSH Terms in Title, Abstract, or Full Text

Partial LCSH Term(s) in Title or Abstract	Partial LCSH Term(s) in Full text
	"use studies" subject, catalogs
automatic, indexing subject	automatic, indexing
image "retrieval systems" information, photographs	"information storage" "retrieval systems" photographs

- Partial LCSH Term(s) in the Title or Abstract: Partial words or phrases from the LCSH subject headings that appear in the title or abstract of the dissertation.
- Partial LCSH Term(s) in the Full Text: Partial words or phrases from the LCSH subject headings that appear in the full text of the dissertation.

The remaining rows of the spreadsheet deal with the author-assigned KEYWORDS from each of the dissertations. While some of the dissertations did not have LCSH terms, all had author-assigned keywords and phrases in the metadata.

Table 3.6

Keywords and Phrases

Keywords and Phrases	Total Keywords and Phrases	Total Words
subject card files user expectations	2	5
subject indexing processes text categorization (TC) automatic subject term assignment subject indexing approaches	3	13
metadata subject access digital libraries information access content analysis information seeking behavior	6	12

- Keywords and Phrases: The author-assigned keywords and phrases for each dissertation.
- Total Keywords and Phrases: The number from the adjacent column, with the total at the bottom of the spreadsheet.
- Total Words: The total number of words for all keywords and phrases for each dissertation, with the total at the bottom of the spreadsheet.

Table 3.7

Keywords or Phrases not in Title, Abstract, or Full Text

Keywords or Phrases Not in Title or Abstract	Total	Keywords or Phrases Not in Full Text	Total
digital libraries information access information seeking behavior	4	information access	1
indexing photojournalism information seeking in context tagging subject analysis	5		

- Keywords or Phrases not in Title or Abstract: The list of keywords or phrases that do not appear in the title or abstract of each dissertation.
- Total: The number from the adjacent column, with the total at the bottom of the spreadsheet.
- Keywords of Phrases not in Title or Abstract: The list of keywords or phrases that do not appear in the full text of each dissertation.
- Total: The number from the adjacent column, with the total at the bottom of the spreadsheet.

Table 3.8

Single-Word Keywords not in Title, Abstract, or Full Text

Single-Word Keywords	Words not in Title or Abstract	Total	Words not in Full Text	Total
1	seeking, behavior	2		
3	indexing photojournalism information, seeking tagging subject	6		

- Single-word Keywords: The number of single-word keywords for each dissertation, with the total at the bottom of the spreadsheet.
- Words not in the Title or Abstract: The list of individual words from the keywords and phrases that did not appear in the title or abstract of the dissertation.
- Total: The number from the adjacent column, with the total at the bottom of the spreadsheet.
- Words not in Full text: The list of individual words from the keywords and phrases that did not appear in the full text of the dissertation.
- Total: The number from the adjacent column, with the total at the bottom of the spreadsheet.

Chapter Summary

This study used the UNT library's theses and dissertations electronic database to study LCSH subject terms in the context of a full text environment. The study also looks at the relationship between author-assigned keywords and LCSH. An EXCEL spreadsheet was used to create the study's dataset. The spreadsheet made it easy to tabulate the quantitative findings as well as to present the LCSH terms and author-assigned keywords in such a way as to make qualitative findings easier to visualize. The results from the spreadsheet/dataset are presented in the next chapter.

CHAPTER 4

COMPILING THE DATA

Introduction

Since the purpose of this study is to try and ascertain the current and future role of controlled vocabularies, specifically the Library of Congress subject headings, the assembled spreadsheet/dataset is analyzed using both quantitative and qualitative methodologies. While the quantitative information i.e. the totals, percentages, and statistics, should provide definitive numerical information from which to make accurate and appropriate judgments, predictions, and conclusions, most anyone would certainly agree that numbers alone rarely tell the complete story. Therefore, the data will also be studied using qualitative methods in an attempt to get a more accurate assessment of the information contained within the data.

Breaking Down the Spreadsheet/Dataset

Although the spreadsheet/dataset is one page consisting of eighty-six rows and twenty-eight columns, it can be considered in two parts. The first part deals with the LCSH terms in each of the dissertations while the second part deals with the author-assigned keyword terms from each of the dissertations. The quantitative results are grouped separately so that they can be compared and contrasted, with the LCSH quantitative findings listed first.

LCSH Terms

- Number of Dissertation Titles: 86
- Number of Dissertation Titles without LCSH Terms 16
- Number of Dissertation Titles with LCSH Terms 70
- Number of LCSH Terms 207
- Number of Words in LCSH Terms 788
- Average number of Words per LCSH Term 3.8

- Number of single-word LCSH Terms 9 or 4%
- Number of LCSH Terms (exact phrase) not in title or abstract 185 or 89%
- Number of LCSH terms (exact phrase) not in full text 155 or 74%
- Number of LCSH words not in title or abstract 246 or 31%
- Number of LCSH words not in title, abstract, or full text 8 or 1%

Analyzing the Numbers

Searching UNT's digital Theses and Dissertations database with the terms *library of congress subject headings* and limiting the collection to dissertation produced 86 dissertation titles. The goal was not a comprehensive search of the database, but rather a search that produced dissertations with all or any of these words in the metadata or full text. The result was a list of dissertations with a variety of different topics. This was intended since the goal was to look at each individual set of LCSH terms for each dissertation and make observations, judgements, and inferences about the relationship between the LCSH terms and their dissertation, particularly within a full text environment. Therefore, the goal was an eclectic group of dissertations rather than a similar group that shared a specific subject. From the group of 86, there were 16 dissertations that were not assigned LCSH subject terms. From this group of 16, there was one from 1984 and one from 1987. The other fourteen were between 2010 and 2014. From the 70 dissertations that contained LCSH terms, there were a total of 207 individual terms. The LCSH terms ranged from just one (word or phrase) in a dissertation up to eight. The chart below graphs the breakdown. There were 9 dissertations with just one LCSH term (16%), 26 with two terms (37%), 10 with three terms (14%), 11 with four terms (16%), 6 with five terms (9%), 3 with six terms (4%), 2 with seven terms (3%), and 1with eight terms (1%).

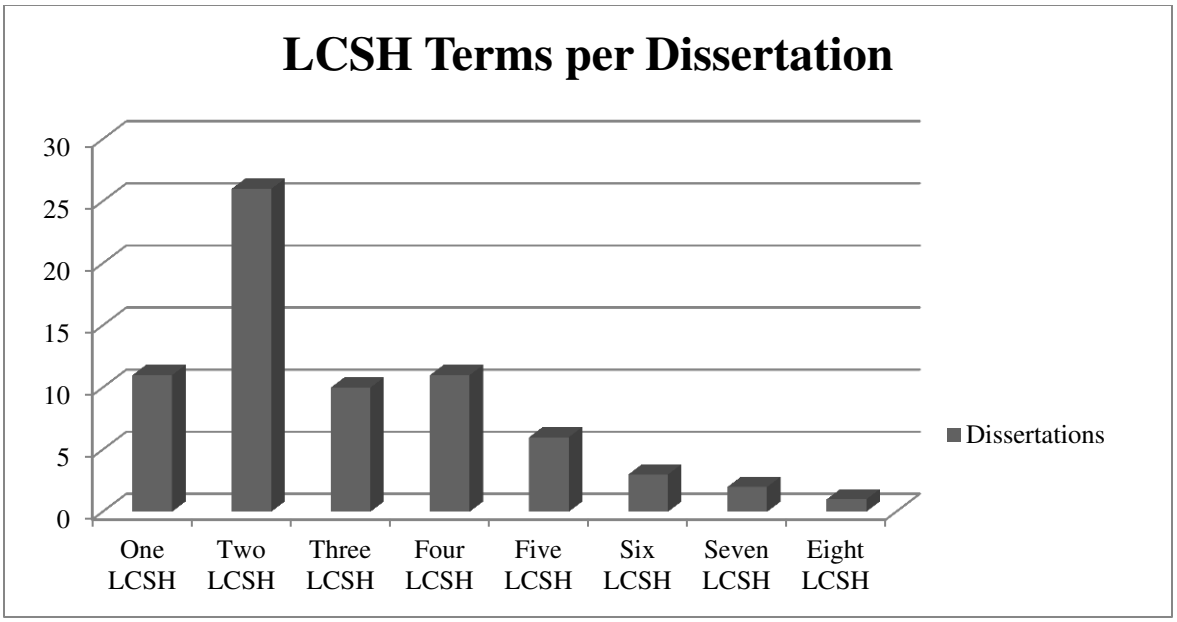


Figure 4.1. LCSH Terms per Dissertation.

Over half of the dissertations, 37, had only one or two terms, while there were only 12 with five, six, seven, and eight terms. This raises the question as to why so many dissertations had so few LCSH terms and, conversely, why so few dissertations had so many terms. Conventional wisdom would lead to the conclusion that the more subject terms, the more likely the item is to be retrieved. It may also be such that the cataloger believed that the number of LCSH terms assigned was sufficient, however many terms were assigned. The average number of words per LCSH terms was 3.8. Only 9 out of the 207 LCSH terms were single-word, about 4%.

From the 207 LCSH terms, 185 or 89% did not appear in the title or abstract as an exact phrase. For the full text, 155 LCSH terms, or 74%, did not appear in the full text as an exact phrase. This would appear to be a high number until considering the hierarchical nature of LCSH which was designed for use with a card catalog rather than a full text/keyword search environment. Another statistic that would seem to confirm LCSH's hierarchical nature is the small number of single-word LCSH terms.

From the 207 LCSH terms there were a total of 788 individual words. This is a significant statistic since most modern search engines will search for individual words unless the search is limited to an exact phrase. From this total of 788 words, 247 did not appear in the title or abstract of the dissertations. This is one of the more significant statistics as it means that without these 247 words, 31% of the dissertations would not be found if users entered these terms and the LCSH terms containing them were missing from the metadata.

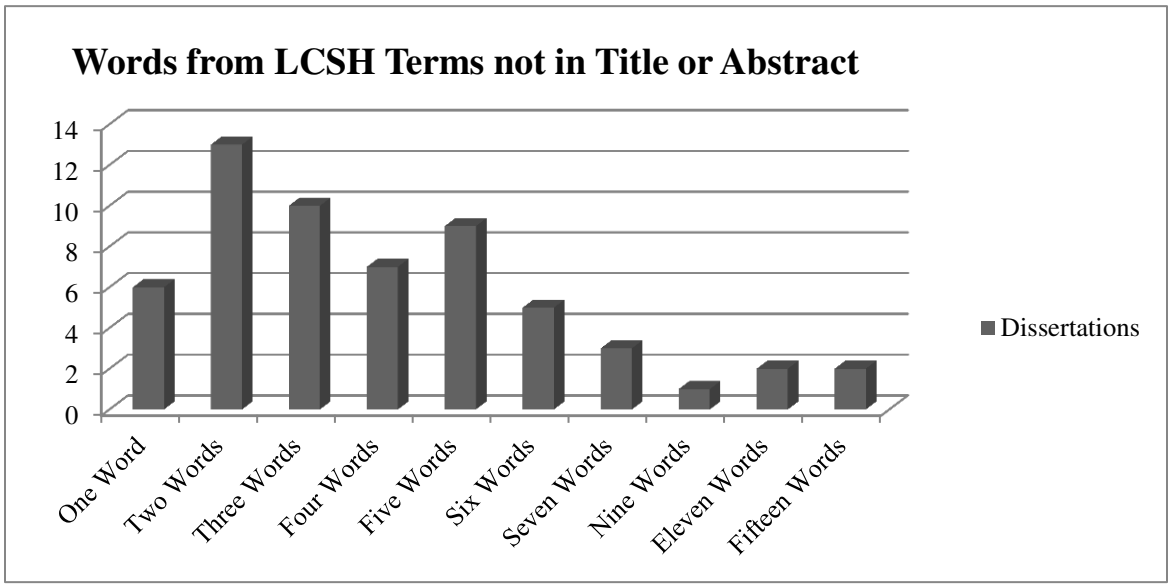


Figure 4.2. Words from LCSH Terms not in Title or Abstract

From the 70 dissertations, 58 had words from their LCSH terms that did not appear in the title or abstract. The above chart shows the breakdown: 6 had one term absent from the title or abstract, 13 had two words absent, 10 had three words absent, 7 had four words absent, 9 had five words absent, 5 had six words absent, 3 had seven words absent, 1 had nine words absent, 2 had eleven words absent, and 2 had fifteen words absent. From these 58 dissertations, 39 had between two and five words that did not appear in the title or abstract, over 67%. This would

also seem to confirm that without the LCSH terms a portion of these dissertations might not be retrieved, if full text searching was not available.

However, when full text search capabilities are available, this number drops considerably. From the total of 786 words in the LCSH terms, only 8 or a little more than 1% did not appear in the title, abstract, or full text of the dissertations. This last statistic raises an important question: "If 8 words from the LCSH terms did not appear in the title, abstract, or full text of the dissertations, then how relevant are they? Would they be important to the relevance and recall of the dissertation, or would they be less-relevant terms that might place the dissertation so far down on a search list that they would be deemed irrelevant by a searcher? This question is addressed in the qualitative section of this chapter.

Author-Assigned Keyword Terms

- Number of dissertations with author-assigned keywords 86
- Number of dissertations without author-assigned keywords 0
- Number of author-assigned Keyword terms 304
- Number of words in Keyword terms 552
- Average number of words per Keyword Term 1.8
- Number of single-word Keyword Terms 112 or 37%
- Number of Keywords Terms (exact phrase) not in title or abstract 70 or 23%
- Number of Keywords Terms (exact phrase) not in full text 9 or 2%
- Number of Keyword words not in title or abstract 50 or 9%
- Number of Keyword words not in title, abstract, or full text 1 or .18%

Analyzing the Numbers

Though this study centers on the Library of Congress subject headings as a controlled vocabulary, the author-assigned keywords were also tabulated quantitatively so that they might be compared to the LCSH terms. The obvious consideration is that author-assigned keywords are much simpler to assign than LCSH terms. While some of the dissertations did not contain LCSH terms, all 86 included author-assigned keyword terms, and there was a total of 302 keyword terms. The range of terms was between two and ten, as depicted in the graph below. There were no dissertations with only one Keyword term (0), 25 with two Keyword terms (29%), 30 with three Keyword terms (35%), 15 with four Keyword terms (17%), 4 with five Keyword terms (5%), 6 with six Keyword terms (7%), 4 with seven Keyword terms (5%), none with eight or nine Keyword terms and 2 with ten Keyword terms (2%).

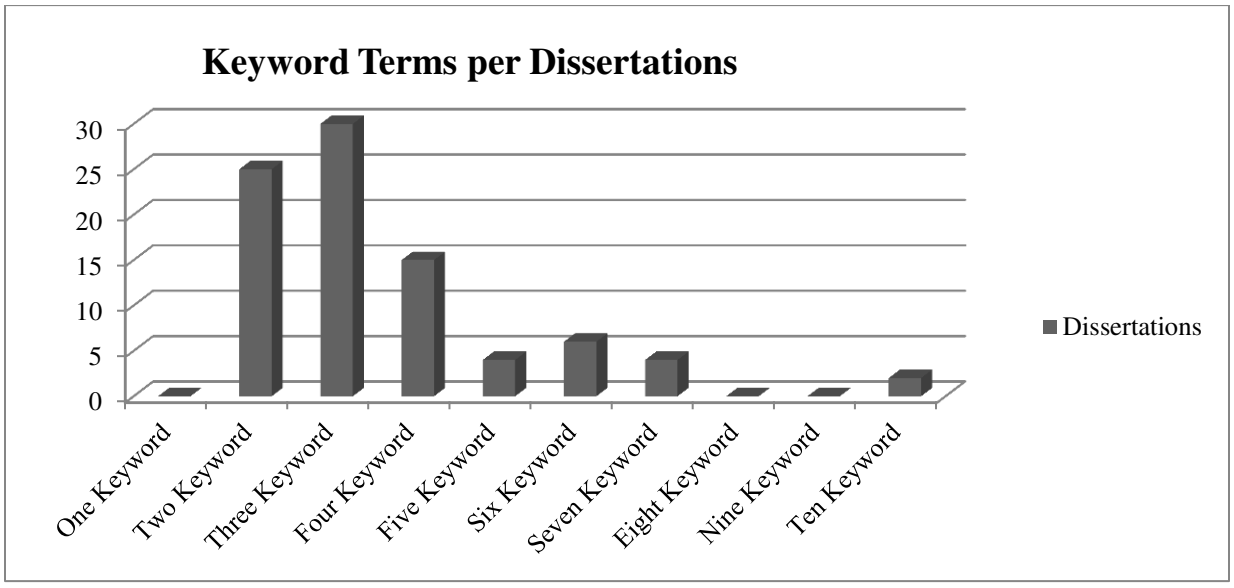


Figure 4.3. Keyword Terms per Dissertation.

Almost two-thirds of the dissertations had either two or three author-assigned keyword terms, and, when four is added, the total from one to four terms is 81%, very consistent with the

LCSH one-to-four terms which was 83%. Three was the most often-assigned number of Keyword terms (35%) while for LCSH, two terms were assigned most often (37%). The average number of words per Keyword terms was 1.8, and 112 of the 304 Keyword terms were single word, about 37%. Not only do authors assign shorter phrases than are assigned with LCSH, 3.8 versus 1.8, but they are much more likely to assign single-word terms to a dissertation, 37% versus 5%.

There were a total of 548 individual words from the 304 Keyword terms as opposed to 786 words in the LCSH terms, meaning there were over two and a half times more words in the LCSH terms than in the Keyword terms. This would seem to be significant because, as has been mentioned, generally the more words assigned to an item's metadata, the better the chance that a variety of searchers using a variety of terms will retrieve that item.

The number of Keyword terms (exact word or phrase) not in the title or abstract was 70 or about 23%, a much lower figure than the 70% for LCSH terms. The number of Keyword terms (exact word or phrase) not in the full text was 9 or about 2%, substantially lower than the 70% for LCSH terms. This would seem to suggest that authors tend to assign words and phrases that appear in the title, abstract, or text of their dissertations whereas LCSH terms are selected from existing schedules or tables that do not tend to appear in the title, abstract, or full text.

From the total of 548 words in the Keyword terms, 50 or about 9% did not appear in the title or abstract contrasted with about 30% of LCSH terms which did not appear in the title or abstract. And only one of the 552 words from the Keyword terms did not appear in the title, abstract, or full text, a little less than .2%. Clearly, the ability to find any word, whether that word be from an LCSH term or an author-assigned Keyword, is greatly enhanced when the retrieval system has full text search capabilities.

Highlights from the Quantitative Analysis

- Of the 788 words in 207 LCSH terms assigned to the 70 dissertations, 246 words or 31% do not appear in the title or abstract. When full text is added this total goes from 246 to 8, or just over 1%.
- Of the 207 LCSH terms only 9 or 4% were single word, while of the 207 keywords or phrases, 112 or 37% were single word.
- There were a total of 304 author-assigned keywords and phrases containing a total of 552 words.
- From the total of 552 words, 50 or 9% did not appear in the title or abstract. When full text is added, the total goes from 50 to 1 or .18%.

Qualitative Analysis

In this section selected dissertations and their LCSH and Keyword terms are analyzed in a more qualitative manner. The chosen dissertations are those having certain representative characteristics. They are selected because they best represent specific LCSH terms or, conversely, specific author-assigned keyword terms. The purpose is to try and deduce the effectiveness of the terms assigned whether cataloger-assigned LCSH terms or author-assigned Keyword terms. The overall goal, again, is to determine the role of LCSH within a modern automated database, particularly a full text database.

Table 4.1

Title/LCSH Terms

Title	**LCSH Terms**
Media Agenda-Building Effect: Analysis of American Public Apartheid Activities, Congressional and Presidential Policies on South Africa, 1976-1988	apartheid - public opinion mass media - Unites States - influence public opinion - United States South Africa - foreign relations - United States United States - foreign relations - South Africa
The Anglo-American Council on Productivity: 1948-1952 British Productivity and the Marshall Plan	British productivity council economic assistance, American - Great Britain Great Britain - economic conditions - 1945 - 1964 Great Britain - Foreign economic relations - United States industries - Great Britain - history - 20th century Marshall Plan reconstruction (1939 -1951) - Great Britain United States - foreign economic relations - Great Britain

The above chart and its two dissertations are excellent representations of the hierarchical nature of LCSH terms. The majority of the terms are obviously taken from a list that moves from a general term to a more specific one. From the list of thirteen assigned LCSH terms from the two dissertations, only two are non-hierarchical: *Marshall Plan* and *British productivity council***,** both from the second dissertation. This type of construction seems a good fit for a linear card catalog but inappropriate for the random nature of keyword searching of modern digital databases.

Table 4.2

LCSH Terms/Keywords and Phrases

LCSH Terms	**Keywords and Phrases**
apartheid - public opinion mass media - Unites States - influence public opinion - United States South Africa - foreign relations - United States United States - foreign relations - South Africa	media agenda apartheid congressional policies presidential policies
British productivity council economic assistance, American - Great Britain Great Britain - economic conditions - 1945 - 1964 Great Britain - Foreign economic relations - United States industries - Great Britain - history - 20th century Marshall Plan reconstruction (1939 -1951) - Great Britain United States - foreign economic relations - Great Britain	Anglo-American council on productivity industry Marshall plan history

The above chart contrasts the hierarchy of LCSH with the author-assigned keywords and phrases for the same two dissertations. When the author assigns terms to his or her work, especially a scholarly work such as a dissertation, it could be assumed that he or she was an authority on the subject, while a cataloger is an authority on LCSH terms, not necessarily the topic of a dissertation. However, two terms in the Keywords and Phrases section of the second dissertation are worth noting. The terms *industry* and *history* are such simple words and can be found in a myriad of dissertations that have nothing to do with this topic, it would seem to make them poor choices for metadata terms. Conversely, the LCSH terms are all much more specifically related to the dissertation topic.

Table 4.3

LCSH Terms/LCSH Words or Phrases in Title, Abstract, or Full Text

LCSH Terms	**LCSH Words or Phrases in Title, Abstract, or Full text**
apartheid - public opinion mass media - Unites States - influence public opinion - United States South Africa - foreign relations - United States United States - foreign relations - South Africa	apartheid, "public opinion" "mass media" "United States" influence "public opinion" "United States" "South Africa" "United States" foreign, relations "United States" "South Africa" foreign, relations
British productivity council economic assistance, American - Great Britain Great Britain - economic conditions - 1945 - 1964 Great Britain - Foreign economic relations - United States industries - Great Britain - history - 20th century Marshall Plan reconstruction (1939 -1951) - Great Britain United States - foreign economic relations - Great Britain	British productivity council economic, assistance, American, "Great Britain" "Great Britain" "economic conditions" 1945, 1954 Great Britain, foreign, "economic relations" "United States" industries, history, "Great Britain" 20th, century Marshall Plan reconstruction, "1939 - 1951" "Great Britain" foreign "United States" "economic relations", Great Britain

The next chart contrasts the LCSH terms (exact phrases) of the two dissertations with any words or phrases from these terms as they appear in the title, abstract, or full text of the dissertation. As is shown, all words from the LSCH terms are present in either the title, abstract or full text of these two dissertations, and this was the case for 59 of the 70 dissertations that had LCSH terms. One might even argue that, in the case of these two dissertations, the LCSH terms assigned to them are unnecessary, unless they are also used in conjunction with a ranking algorithm since every word for all LCSH terms appears in the title, abstract, of full text of the dissertations. The above two sections again show that the LCSH system was clearly designed for

a card catalog and has simply been transferred to online environments. If a searcher in the past went to a card catalog wanting information on relations between the United States and South Africa during the apartheid era, he or she could have looked up either *United States* or *South Africa*, kept thumbing through the cards until the subtopic *foreign relations* was found then either *United States – foreign relations – South Africa* or *South Africa – foreign relations - United States* was found. However, in a digital environment the searcher can type in any form of these words and phrases and the desired information can be found.

Table 4.4

Title, LCSH Terms, Keywords and Phrases

Title	**LCSH Terms**	**Keywords and Phrases**
Keywords in the mist: Automated keyword extraction for very large documents and back of the book indexing	automatic indexing	construction-integration keyword extraction back of the book indexing
A Philosophy for Two-year Occupational Programs in Public Junior College Curricula	business education	occupational programs public junior colleges business curricula
Alternative Funding Models for Public School Finance in Texas	education -- Texas – finance	alternative funding model public school finance property tax power equalization percentage equalization foundation school program

The above table lists three dissertation titles with their accompanying LCSH terms and author-assigned keyword terms. The first title deals with an automated keyword extraction method and was assigned the LCSH term, *automatic indexing*. The cataloger evidently thought

this would be sufficient, as no other terms were assigned. The author assigned two terms that were already in the title, *keyword extraction* and *back of the book indexing*, so they might be considered redundant except for use by a ranking algorithm. However, the third term, *construction-integration* seems to be more unique. The term does not appear in the title or abstract, yet it appears 33 times in the body of the dissertation, so it would seem to be very important to the author and his work. Without full text search capabilities, this information would be lost to a searcher if the author had not included it in the metadata. And since the dissertation certainly deals with automatic indexing, this is a case where LCSH and keywords seem complimentary.

The second dissertation works in a similar way. As the title, LCSH terms, and keywords attest, the subject deals with business curricula in junior colleges. The author's terms, *occupational programs* and *public junior college(s)*, appear in the title so, again, they seem redundant. However, the author's term, *business curricula*, appears in the abstract once and in the body of the dissertation 95 times, so in this case the author-assigned keywords would seem unnecessary in this full text keyword searchable database.

The last title seems to clearly explain the subject of this dissertation, and the LCSH terms and the keywords tend to confirm the topic. But, as can be seen, the LCSH words *charter schools*, *students*, and *disabilities* appear in the title, again making them redundant in a keyword search environment. And here again the first two author-assigned key phrases, *alternative finding model* and *public school finance* were taken directly from the title, while the last four, *property tax, power equalization, percentage equalization,* and *foundation school program* appeared in the abstract as well as the body of the dissertation which seems to call into question their usefulness.

Table 4.5

Title, LCSH Terms, Keywords and Phrases

Title	**LCSH Terms**	**Keywords and Phrases**
Integration of Students with Disabilities into a Contemporary Technology Education Program: A Case Study	mainstreaming in education technical education	children autism learning disabilities technology education
A multi-state political process analysis of the anti-testing movement	educational accountability - political aspects - United States educational tests and measurements - political aspects - United States	political process model high-stakes testing social movement framing processes mobilizing structures political opportunity accountability
Choice for All? Charter Schools and Students with Disabilities	charter schools - Texas students with disabilities - education - Texas	charter schools students with disabilities PEIMS IDEA

The above chart lists three dissertations titles from the twenty-six dissertations that contained two LCSH terms. The first LCSH term for the first dissertation is *mainstreaming in education*. The term does not appear in the title or abstract but appears in the full text fourteen times. Again, with the addition of full text keyword searching, the importance of LCSH seems to diminish. The keyword phrase *learning disabilities* appears once in the abstract, as *learning disabled*, and six times in the dissertation, three times as *learning disabilities* and three times as *learning disabled*, again calling into question the usefulness of these additional words to the metadata.

Neither of the LCSH terms from the second dissertation, *educational accountability – political aspects – United States* and *educational tests and measurements – political aspects – United States*, appear in the title or abstract as the exact phrase; however, all words and some phrases from both LCSH terms appear in either the title, abstract, or full text—except the word *measurements* which does not appear in the title, abstract or full text. Here again is an example of the somewhat cumbersome nature of the hierarchical form of LCSH. It also might be suggested that the word *measurements* is not essential to the metadata of this dissertation.

The two LCSH terms from the final dissertation, *charter schools – Texas* and *students with disabilities – education – Texas* do not appear in the title, abstract, or full text as the complete phrase, but all the words or phrases from each term do appear in the title, abstract, or full text. Again, the obvious conclusion here is that in a database with full text search capabilities, these LCSH terms appear not be necessary, especially as they are constructed.

Although the sample size is very small, the above examples do seem to suggest that assigning LCSH terms as they are currently constructed to be used in a modern keyword search database, especially one with full text search capabilities, is not cost effective and possibly not practical. Conversely, when authors assign keywords and phrases that are already in the metadata or full text of the document, this too seems to be inappropriate in a keyword search environment.

The next section will look at the nine words from the LCSH terms that did not appear in the title, abstract, or full text. The goal here is to try and assess the importance of these missing words to the search and retrieval process.

Table 4.6

Title, LCSH Terms, Keywords and Phrases

Title	**LCSH Terms**	**Keywords and Phrases**
Evaluating e-Training for public library staff: A quasi-experimental investigation	"distance education" "library education" ("continuing education") library employees - training of - computer-assisted instruction "public librarians" - - training of - computer-assisted instruction "web-based" instruction	online training public library staff evaluation methodologies e-training

Although the word *assisted* did not appear in the metadata or full text of this dissertation, the word *assist* appears five times. A sophisticated search engine should retrieve the various common tenses of verbs, meaning that this may in fact not be a word that would be lost. In the text, instead of using the phrase *computer assisted*, the author used *computer mediated* which appears four times. This would seem to suggest that the word *assisted* is an important synonym, especially when used in a phrase such as *computer assisted,* a term that most searchers would probably use instead of the phrase *computer mediated*. It is also worth noting that the LCSH phrase *distance education* would seem to be a synonym for *online training* while *web-based instruction* could be considered a synonym for *e-training*. This would seem to be an example of the complementary nature of LCSH and keywords, even in a full text environment as both do seem important components of the metadata.

Table 4.7

Title, LCSH Terms, Keywords and Phrases

Title	LCSH Terms	Keywords and Phrases
A multi-state political process analysis of the anti-testing movement	"educational accountability" - political aspects – "United States" educational tests and measurements - political aspects –"United States"	political process model high-stakes testing social movement framing processes mobilizing structures political opportunity accountability

While the word *measurements* did not appear in the metadata or full text, the singular form, *measurement* appears three times, so this word might also not be one that would be considered missing since modern search engines should return the plural of keywords.

Table 4.8

Title, LCSH Terms, Keywords and Phrases

Title	LCSH Terms	Keywords and Phrases
Improving Recall of Browsing Sets in Image Retrieval from a Semiotics Perspective	image files - abstracting and indexing "image processing" "information retrieval" semiotics	image retrieval semiotics connotations

The term *files* is another example of a plural form of a noun that does not appear in the text, but the singular form, *file*, does appear, albeit only once. Here again, the word would probably not be considered missing. Also of note here is that the LCSH term *image processing* would seem to be a synonym for the key phrase *image retrieval* and both terms could be

considered important for the retrieval of this dissertation, again verifying the complementary nature of LCSH terms and keyword terms.

Table 4.9

Title, LCSH Terms, Keywords and Phrases

Title	**LCSH Terms**	**Keywords and Phrases**
The History of Speech and Drama Education in the Dallas Public Schools (1884-1970)	Dallas (Tex.) drama - study and teaching - Texas - Dallas - history education - Texas - Dallas - history "oral communication" - study and teaching - Dallas - Texas - history "public schools" - Texas - Dallas - history	speech and drama education Dallas public schools

This abbreviation *Tex.* may also be somewhat meaningless since *Texas* appears 86 times in the text. As an aside, it seems somewhat curious that in the first LCSH term the abbreviation for Texas is used and in the four other LCSH terms, Texas is spelled out, an obvious inconsistency. The use of the abbreviation instead of the complete word by a searcher would seem to be remote. Again, the thought here is that the search engine would retrieve items with either the abbreviated form of a state as well as the complete spelling.

Table 4.10

Title, LCSH Terms, Keywords and Phrases

Title	LCSH Terms	Keywords and Phrases
Problem-Based Learning for Training Teachers of Students with Behavioral Disorders in Hong Kong	"behavior disorders" in children - China – "Hong Kong" problem children - education - China – "Hong Kong" "problem-based learning" - China – "Hong Kong" teachers of problem children - training of - China – "Hong Kong"	problem-based learning teachers behavioral disorders

Any study dealing with students and/or teachers in Hong Kong would certainly benefit from having the word *China* as a search term since it would broaden the search to any city in China. It would seem reasonable to assume that most searchers interested in this topic would want similar studies from anywhere in China, so this LCSH term would seem to be an important addition to the metadata of this dissertation. However, the complex construction of the LCSH terms just to get one synonym, again, seems inappropriate.

Table 4.11

Title, LCSH Terms, Keywords and Phrases Title, LCSH Terms, Keywords and Phrases

Title	LCSH Terms	Keywords and Phrases
An examination of music for trumpet and marimba and the Wilder Duo with analyses of three selected works by Gordon Stout, Paul Turok, and Alec Wilder	Stout, Gordon. Duet, trumpet, marimba "Trumpet and percussion music" - history and criticism Turok, Paul, 1929- "concert variations," trumpet, marimba op. 51. no 3 Wilder, Alec. Suites, trumpet, marimba	trumpet Wilder Stout Turok percussion duo marimba

The term *criticism* would seem to be an important addition since the work could be considered a critical study of a specific type of music and musicians, and a percentage of searchers wanting such studies would most likely use the term by itself or to construct phrases. But it seems to be an expensive way to include an extra synonym. And all but one of the author-assigned keywords (*percussion*) is contained in the title, again, making these terms somewhat redundant as metadata.

Table 4.12

Title, LCSH Terms, Keywords and Phrases

Title	LCSH Terms	Keywords and Phrases
Government and Private Funding of Nonprofit Visual Arts Organizations in the State of Texas: An Analysis	art and state - Texas art "fund raising" - Texas "nonprofit organizations" - Texas - finance	funds distribution arts funding arts administration

The term *finance* would seem an appropriate synonym for a study with the word *funding* in the title and in the keywords. A searcher might use other terms such as *institution*s in conjunction with *finance* and still retrieve this study, so this LCSH word does seem an important addition, yet again, an expensive synonym.

Table 4.13

Title, LCSH Terms, Keywords and Phrases

Title	LCSH Terms	Keywords and Phrases
The Search for Order and Liberty : The British Police, the Suffragettes, and the Unions, 1906-1912	Democracy – "Great Britain" -- History Demonstrations – "Great Britain" -- History Great Britain -- Politics and government -- 1901-1936 "labor movement" – "Great Britain" -- History -- 20th century "labor unions" – "Great Britain" -- History -- 20th century Police – "Great Britain" -- History Suffragists – "Great Britain" -- History -- 20th century	British police suffragettes unions

The term *20th* when used with *century* may be significant if the search engine does not search both the number *20th* and the word *twentieth.* The phrase *twentieth century* appears twice in the full text of this dissertation. Here again it should be the case that a sophisticated algorithm would retrieve documents with both forms when either is used as a search term.

Highlights from the Qualitative Analysis

While the LCSH terms rarely appear exactly in the title, abstract, or full text, most all the words and phrases which make up the LCSH terms do appear in the title, abstract, or full text. The hierarchical nature of LCSH, which makes it such a good system for a print card catalog, seems to be too complex for a keyword search/full text environment.

- Of the 788 words from all LCSH terms, 8 did not appear anywhere else in the metadata or in the full text Though this is a very small sample, it does suggests that just over 1% of items in a full text environment might be lost without LCSH terms. However, looking specifically at these eight words, some of them do seem unimportant or insignificant.

- Authors often assigned Keywords as metadata that appeared both in the metadata as well as the full text, making these terms redundant in a keyword search environment. It would seem that authors could benefit from some guidance when adding keyword metadata.

Chapter Summary

The quantitative findings in this chapter suggest that around one-third of items in a database might be lost without LCSH subject terms, a figure consistent with the findings of other researchers (Gross & Taylor, 2005). However, with the addition of full text search capabilities, the findings in this research suggest that number drops substantially.

When the LCSH terms are examined in a more qualitative manner, their hierarchical design, while well-suited for the linear nature of a card catalog that points to books on a linear shelf, both of which move from the general to the specific, makes them seem too complex for a keyword search/full text environment. A full text database is not linear, and it never presents the same materials in the same order because searchers rarely search with the exact same terms, and searchers are always modifying their searches by adding and deleting terms, so the list of items presented to them is always changing. Could LCSH be improved to work better within this less static and more fluid environment? Keyword searches of extremely large databases such as the Internet often return an exorbitant amount of items, sometimes in the millions, much more than any searcher could consider. Can controlled vocabularies such as LCSH be modified so that, as they have done so well with books and texts in print, they help improve the relevance of digital items found with keyword searching? Also, could LCSH be modified so that it works more as a partnership with authors as they assign keywords rather than as the two systems operate now, separate but complimentary? The final chapter addresses these issues.

CHAPTER 5

FINDINGS AND DISCUSSION

Introduction

This research has been undertaken with the goal of contributing to the literature dealing with controlled vocabularies, particularly within full text environments. According to Tina Gross (2015) and her colleagues: "In the long run, the ultimate test of the importance of controlled vocabulary will be its effect in full text environments" (p.30). Specifically, this project hopes to contribute to the literature which seeks to answer the question: Are controlled vocabularies still valid and effective in today's twenty-first century digital environments, particularly full text environments? Put another way, can LCSH be adjusted, modified, or revised, so that it works as effectively in a full text digital keyword search environment as it did in a printed card catalog environment to help searchers find the most relevant items?

Research Question 1

If a database has the ability to search the title, abstract, and other metadata, as well as the full text of a record, would this affect the role of a controlled vocabulary in full text keyword search environments? If so, does this finding suggest a change to the importance of a controlled vocabulary such as LCSH?

LCSH in a Full Text Environment

Quantitative data in this research confirmed what other research had discovered (Gross, T., Taylor, A. G., & Joudrey, D. N., 2015): One third of searches of OPACs would fail if LCSH terms were eliminated, clearly a substantial number. Research from this dissertation found that when full text search capabilities are added to a keyword search environment, only about one

percent of searches would be lost without LCSH terms in the record. The sample size was only eighty six dissertations, clearly not large enough to make any definitive conclusions.

Daniel Alemneh, Mark Phillips, Laura Waugh, and Hanna Tarver (2015) conducted a much larger study of the same Thesis and Dissertations database at the University North Texas. These researchers looked at transaction logs to compile a dataset containing over 43,000 unique query results. They listed the percentage of each search query that was found in either the full text or the metadata of each record. Their metadata consisted of four fields: Title, Subject, Agent (both creator and contributor), and Description. They found that almost 96 percent of queries had at least one term listed in the full text of a record, with only 4 percent of query terms appearing only in the metadata. Their study, along with the spreadsheet/dataset results from chapter four, suggest that when full text capabilities are added to keyword search environments, retrieving material that contains some or all of a user's search terms increases dramatically. This could indicate that the user may be less willing to use the LCSH terms in the metadata to help locate more relevant items, being more or less satisfied with the search results.

Controlled Vocabulary as Keyword Search Aid

As has been repeatedly stated in this research, LCSH is quite complex. Rebecca Dean (2009) succinctly detailed the complexity of LSCH by first explaining what the controlled vocabulary system was—and was not:

> LCSH is not a true thesaurus in the sense that it is not a comprehensive list of all valid subject headings. Rather LCSH combines authorities, now five volumes in their printed form, with a four-volume manual of rules detailing the requirements for creating headings that are not established in the authority file and for the further subdivision of the established headings. (p. 334)

Dean then presented a revealing example of just how challenging the assigning of LCSH subject terms could be:

> For example, **Burns and scalds–Patients–Family relationships** is a valid heading formed by adding two pattern subdivisions to the established heading **Burns and scalds**. The subdivision **Patients** is one of several hundred subdivisions that can be used with headings for diseases and other medical conditions. Therefore, it can be used to subdivide **Burns and scalds**. However, the addition of **Patients** changes the meaning of the heading from a medical condition to a class of persons. Now, since **Family relationships** is authorized under the pattern for classes of persons, it can also be added to complete the heading. (Dean, 2009, p. 334)

Obviously, creating as well as assigning LCSH terms to a work is very complex. Clearly, something simpler seems needed in large digital databases, especially the Internet.

Researchers who have argued that without LCSH terms in metadata, a percentage of searches would fail without the keywords the LCSH contain seem to be arguing more for additional keywords than for the value of a controlled vocabulary. To insist that a controlled vocabulary such as LSCH should be retained because of the additional keywords that the system provides would seem, in one sense, to be an argument against the need for controlled vocabularies. LCSH, a system designed for a physical card catalog, has become very large and complex, very labor intensive, and arguably quite expensive. Some of the LCSH headings from the 80 selected dissertations in this research's dataset contained up to eight words, and these words were arranged according to strict and quite specific guidelines, as noted in the example above. Below are some examples of LCSH that appeared in a dissertation's metadata, taken

from the dataset. The terms in red are words that only appear in the subject heading but nowhere else in the metadata or the full text of the dissertation.

- library employees - training of - computer-assisted instruction
- image files - abstracting and indexing
- teachers of problem children - training of - China – "Hong Kong"
- "Trumpet and percussion music" - history and criticism
- "nonprofit organizations" - Texas – finance

To retain a multi-word subject heading created using a complex and time consuming set of instructions simply because it contained one word that didn't appear anywhere else in a record simply does not seem efficient. Something simpler and more attuned to the keyword search environment seems more appropriate.

The Complexity of Library Search and Retrieval Tools

It is quite possible and even arguable that when writings began appearing on clay tablets, the technology would have been considered state of the art. This technology must also have been used within some type of efficient system for storage and retrieval of the information contained on the tablets, otherwise, like today or any other time, the information would have been useless if it could not be readily retrieved when needed. A later technology, papyrus, was used for the recoding of written information which must have been lighter and therefore made the information contained within the papyrus more easily stored and retrieved. A further advancement would have been the use of scrolls which probably meant that much more information could be contained in one long document which could "scroll" up and down to help find the desired information. These scrolls must certainly have had the title of the work as well as its author on the ends or somewhere on the outside, and they were probably stored in a more

efficient storage and retrieval system, possibly stored in pots or chests of drawers within some type of subject or author arrangement. It would be millennia before paper, the printing press, and Guttenberg's all-important moveable type ensured that writing stored on sheets of bound paper, or what would become universally known as the *book*, became the state of the art technology for the storage and retrieval of information. Later would come improvements such as page numbering, chapter subdivisions, indexes, and tables of contents, and eventually most all book publishers would adhere to virtually all of these innovations. And as the printing of books proliferated, newer, larger and more efficient libraries with ever more efficient search and retrieval systems would be designed, developed, and implemented to facilitate the storage and retrieval of the ever increasing quantities of available books and the valuable or even invaluable information they contained. Automated computerized storage and retrieval systems would allow more efficient indexing of the huge amounts of information being generated by technologically advanced societies as they helped broaden the frontiers of human knowledge. The preferred method, and arguably the standard, for locating and retrieving information in written form would be a public, academic, or special library maintained by a well trained staff of information professionals, perhaps the state of the art for the twentieth century. However, in these modern libraries, with innovation came complexity, and with complexity came frustration. Many if not most of the search and retrieval systems, including online technologies, were difficult and therefore frustrating for a percentage of users.

Indexing: The Human Factor

From the ancient libraries right up to our huge digital collections, one concern could be considered the most overriding: How can searchers retrieve only the information that they are seeking when the number of documents in the collection being searched is too large to peruse all

or most of the items? Elaine Svenonius (2009) noted the importance of this question to the problems of indexing:

> The quest for precision in retrieval is so paramount and of such long duration that it would be possible to frame the history of indexing over the last hundred and fifty years in terms of the successive means devised to deal with it. (p. 148)

The Dewey decimal classification system coupled with the Library of Congress subject headings has enabled the retrieval of very specific items from very large library collections. These two systems as well as the Library of Congress classification system are of major importance in the organization of library materials, as Lois Chan (1990) explained:

> In subject analysis and access, the three major tools—the *Library of Congress Subject Headings (LCSH)*, the *Library of Congress Classification (LCC*), and the *Dewey Decimal Classification*—have been the mainstay in our effort to organize, represent, and arrange library materials. (p. 258)

It would seem hard to argue that a controlled vocabulary such as LCSH when used in conjunction with a decimal classification system such as LCC or Dewey has greatly increased the ability to locate relevant materials, especially when compared to earlier systems. However, these systems are quite complex, meaning that those who assign call numbers and subject heading to library materials, the indexers, do require special training. And even with extensive training as well as years or even decades of experience, these indexers still must make judgments that might be considered more intuitive than scientific, as Brian O'Connor (1996) succinctly explained:

> The rules for extraction generally are not explicit in the process of most human indexing. There is, in saying this, no value judgment of the quality of the representation made by

> the indexer. It is simply important to note that it may be difficult or impossible for a human indexer to specify the exact mechanism he or she used for highlighting. It may well be that personal knowledge of the types of users of the system will enable very good representation. However, we are left without any method of addressing issues of consistency across time and across settings. (p. 108)

Clearly, catalogers are people of varied backgrounds with different undergraduate and /or secondary degrees and therefore have different areas of expertise, so there will seemingly always be a problem with consistency. A question that could be asked at this point is whether, with the ever-advancing search technologies, algorithms may be just on the horizon that will make the assigning of index terms to a document more automated and therefore more consistent. Elaine Svenonius (2009) believed optimistically in that possibility:

> The algorithms used in search engines today are still fairly primitive; many are based on keyword searching alone. As yet such systems have not been able to deal with the scatter and clog of information caused by the synonymy and homonymy of natural language, nor can they provide semantically useful displays of bibliographic data. But they have the potential to do so, transforming the theory of bibliographic description into a theory of bibliographic searching. (p. 66)

An automated system of bibliographic description might not only reduce or eliminate inconsistencies of human cataloging, but should also greatly reduce the time it takes to assign index terms to individual items. This would seem to be essential if individual records in large databases and eventually even the billions of items on the Internet are to be more effectively cataloged and classified so that precision can be modestly if not greatly improved.

The Internet

In 1958 a government agency was created in the Department of Defense called the Advanced Research Projects Agency or ARPA as it would become known (Isaacson, 2014). The agency: "Marked an extension of the defense-oriented military-university collaborations that began in World War II" (Isaacson, 2014, p. 229). In October of 1969 ARPA selected four research centers as the foundation for a system to connect distant computers to share information: UCLA, Stanford Research Institute, University of Utah, and the University of California at Santa Barbara (Isaacson, 2014). This shared information system would become known as ARPANET and by the early 1980's would evolve into the Internet (Isaacson, 2014). The prescient designers of this shared information system insisted on one ingredient that would help pave the way for the modern Internet, as Walter Isaacson (2014) pointed out:

> These academic researchers of the late 1960s, many of whom associated with the antiwar counterculture, created a system that resisted centralized command. It would route around any damage from a nuclear attack but also around any attempt to impose control. (p. 251)

The World Wide Web was launched in 1991 based on HTTP protocol designed by Tim Berners-Lee who also gave this new version of the Internet its name (Isaacson, 2014). Using HTTP, "computer scientists around the world began making the Internet easier to navigate with point and click programs" (Hafner & Lyon, 1996, p. 258). In 1994 Lycos, one of the first Internet search engines was developed, and it was quickly followed by Excite, Infoseek, AltaVista, and Google (Isaacson, 2014).

The growth of the Internet, and its essential upgrade, the World Wide Web, in the last couple of decades has been nothing short of spectacular. It has been estimated that in 1993

fifteen million people in fifty countries were using the internet, with just over one hundred Web sites available (Auletta, 2010). By 2010 the number of Internet searches was three billion per day (Auletta, 2010), and according to a website called *Internet Live Stats* (http://www.internetlivestats.com/) there are currently over one billion web sites. Clearly, a tremendous amount of information is available on the Net, and it is growing astronomically.

The Influence of Google

Near the end of the twentieth century most modern academic and public libraries would most likely have an OPAC or online public access catalog located within the library's Web Page which contained the holdings of the library, and which was searchable by author, title, or subject. The OPAC was an automated version of the card catalog which it replaced. These modern libraries would also have automated periodical databases for locating periodical articles by author, title, or subject. They were online versions of print indexes such as *Reader's Guide, Eric, Journal of Chemical Abstracts*, and *Psych Abstracts*, to name only a few. The process for finding periodical articles by subject was also difficult and often frustrating for searches, especially those whose experience consisted of searching the Web with the Google search engine:

> The current process of finding an article—Choose a database from an alphabetical or subject list; Search the databases by appropriate keywords; Choose one or more citations based on the title and abstract; Switch to the library's catalog; Search the catalog by journal title; Interpret the holdings display to determine location and availability; Go to the selves or follow a link to a journal Web site (where you have to browse to a date or issue)—is time-consuming, complicated, and not intuitive to students raised in a point-and-click Google World. (Ponsford & vanDuinkerken, 2007, p. 160)

Dennis Warren (2007) also believed that students were frustrated with the large number of databases with dissimilar search interfaces presented to them when these students sought periodical articles: "One could well ask why we persist in presenting our users with alphabetical lists of databases, especially since many of the names of the databases tell us nothing about the content" (p. 262). Yongming Wang and Jia Mi (2012) echoed these same sentiments:

> Many academic libraries offer database A-Z lists as a directional guide. However, users are presented with too many choices and have little knowledge of where to begin. Currently, most libraries also provide 'Databases by Subject' help; however, users still have difficulty identifying appropriate databases. (p. 230).

Wang and Mi also succinctly delineated the ubiquitous influence of the Google search engine: "Today's users form their information seeking behavior by using Google and other Internet search engines. As a result, they expect library systems to work the same way as Google does: one simple search box, intuitive and instant results" (Wang & Mi, 2012, p. 229). Other researchers have verified Google's effect on libraries as well as society in general:

- "It is nearly impossible to discuss search and discovery in libraries without mentioning Google" (Lown, Sierra, & Boyer, 2013, p. 227).
- "It is difficult to dispute that Google has become the central search tool in society" (Swanson & Green, 2011, p. 222).
- Roberta Woods (2010) seems to have effectively summed up what she called the "Googlized" library patron: "The advent of Google made one box searching easy, with result sets that seemed precisely what the searcher had in mind. Thus the 'Googlized' library patron was born. This patron – our patrons – will no longer tolerate anything more complex than a single search box and a single, integrated result set" (p. 141).

Ponsford and vanDuinkerken (2007) noted in their research: "In other words, users expect to take their Google searching skills and apply them to find library resources" (p. 162). Such is both the influence as well as the pervasiveness of the Google search engine. It is also a reminder that searchers want search tools that are simple, easy to understand, and therefore easy to use.

Google's PageRank Search Algorithm

Google was founded by Sergey Brin and Larry Page, who met while both were pursuing PhDs. in computer science at Stanford University (Vaidhyanathan, 2009). In a technical report published in 1999, Page, Brin, and their colleagues described a method for automatically ranking web pages which they called PageRank based on academic citation analysis: "It is obvious to try to apply citation analysis techniques to the Web's hypertextual citation structure. One can simply think of every link as being an academic citation" (Page, Brin, Motwani & Winograd, 1999, p. 2). Page and Brin wanted to gauge the importance of web pages in the same way academic papers are ranked, by the number of times they are cited. The more a paper is cited, the more important it is considered. In a similar way, they reasoned, the more times a Web site has been viewed or linked to, the more valuable it should be to searchers:

> Using PageRank, we are able to order search results so that more important and central Web pages are given preference. In experiments, this turns out to provide higher quality search results to users. The intuition behind PageRank is that it uses information which is external to the Web pages themselves – their backlinks, which provide a kind of peer review. (Page, Brin, Motwani & Winograd, 1999, p. 15)

Siva Vaidhyanathan (2009) provided a simple yet concise example of how PageRank works: "Let's say you type 'shoe store' into a Google search box. Google's PageRank algorithm

sorts through Web pages containing the phrase 'shoe store.' It ranks these pages based on the number of other pages that link to those pages" (p. 66). In September of 1998 Google Inc. was formed, and by 2010 the company was taking in over $22 billion in sales a year, a remarkable achievement (Carr, 2011).

While Google's PageRank algorithm is undoubtedly the key to the company's incredible success, most users might agree that the overall simplicity of the Google search page, a single search box on an uncluttered page, has also contributed to the positive experience of using the search engine. Typing words and phrases into a single search box on a page with virtually no distractions may be about as intuitive as is possible for an Internet search engine or, in fact, any automated technology.

The Federated or Single Search Box

Google has not only influenced library patrons but library services as well. Many libraries have added single search box capabilities, or federated searching, to their list of online search capabilities. Roberta Woods (2010) explained that while many university libraries have an extensive list of databases, some of these resources were not being consistently used to justify their expense, paving the way for federated search options:

> Current library budgets cannot continue to maintain resources no one uses in the hope that one day users will seek out the content contained in them, thus the need for a federated system for searching across various forms of digitized content had its genesis. (p. 142)

By 2009 more than 2,500 libraries had added commercial federated search services to their resources (Wang & Mi, 2012). As Wang and Mi explained: "Although the concept of federated search or discovery services is not new, it appears in the library world as an answer and

an alternative to Google" (p. 230). Wang and Mi also provided an effective definition: "Federated search, also called metasearch, parallel search, or broadcast search, is defined as searching different resources at the same time and then presenting the search results in a unified way" (p. 229).

The library user that has been conditioned to think only of entering terms into a single search box because of extensive experience with the Google search engine may seem less likely to spend time learning various segments of a library's resources, as Troy Swanson and Jeremy Green explained: "Users do not appear to be very aware of differences between databases, catalogs, and other tools. They search whatever search box is readily available" (Swanson & Green., 2011, p. 227).

In a study of a federated search tool known as *One Search*, implemented at New Mexico State University in 2008, Sarah Baker and Alisa Gonzales (2012) concluded that the majority of students surveyed found the service useful: "Federated searching appeals to students because they do not have to choose resources, learn controlled vocabularies, or know specialized database features to search many different resources" (p. 15). Baker and Gonzales concluded that students responded favorably to the federated service because it saved them time: "Students felt that searching individual databases was very inefficient and saw *One Search* as a way to save time in doing research or in completing an assignment" (p. 28).

Cory Lown, Tito Sierra, and Josh Boyer (2013) examined two semesters of transaction logs, nearly 1.4 million transactions, in order to determine user patterns using a federated search tool known as QuickSearch at North Carolina State University. One of their more interesting findings was that about 23 percent of searches were not focused on the catalog or articles modules: ". . . indicating that NCSH Library users attempt to access a wide range of information

from the single search box" (p. 240). This would tend to confirm that users view most any search box as they view the Google search box, as a "one box fits all searches" environment. However, these researchers also concluded what might be considered the major consensus of federated search capabilities: "A single search box communicates confidence to users that our search tools can meet their information needs from a single point of entry" (Lown, Sierra, & Boyer, p. 240).

The qualitative research in this study would tend to confirm that in a full text environment such as UNT's Thesis and Dissertation database, users seeking items by subject can enter keywords into a search box and select from the results list, without using or possibly even needing a controlled vocabulary such as LCSH. In a non-full text environment, up to one third of searches could fail without the LCSH terms, but when full text capabilities are added, that number drops substantially. Put another way, one might say that 66% of searchers in a non-full text environment could be successful without the need of a controlled vocabulary. And when full text is added, this research suggests that that number may go as high as 99%. In order to maintain its usefulness, a controlled vocabulary such as LCSH would appear to need to adapt to a full text search environment as well as the search habits of its current users.

The Amazon Books Model

Before the digital age, looking for information often meant a trip to a public, academic, or special library where traditional search and retrieval aids such as subject headings and decimal call numbers would be used. As has been discussed, these systems could seem difficult and confusing to users. Unfortunately, these tools were often the only aids available for search and retrieval of library resources. The modern digital age is now presenting information seekers with alternatives to these traditional systems, and they are much simpler and much more intuitive.

Library of Congress subject headings and Dewey or LC classification system are conspicuously absent from the Internet, and users have wholeheartedly embraced its simple keyword search capabilities. Users seeking information on today's Internet are undoubtedly quite similar to users who entered brick and mortar libraries in the past: They are searchers looking for information to satisfy simple questions or more complex problems. Today's modern libraries are still major repositories for information seekers, and those seeking information for more in-depth research can still borrow items for weeks or months at a time to be used in the comfort of their homes or offices. However, an alternative to borrowing and returning library materials is to purchase books from online retailers such as Barns and Noble and Amazon.

It could be argued that Amazon.com is as important to online retailing as Google is to the Internet; the giant one-stop online retailer cleared 61 billion dollars in sales in 2012 (Stone, 2013). An important sector of that sales figure are, of course, books which was what the company was initially based on when it began in 1995 (Stone, 1995). And like the Internet, the Amazon search system is also both simple and intuitive.

The success of Amazon books can be attributed to many factors, low prices being perhaps the most obvious as well as also the convenience of online shopping. However, two features have helped to make searching for books by subject both easier and more efficient for searchers of the Amazon site: *Look Inside the Book* and *Search Inside This Book* (Stone, 1995).

Obviously, an immediate problem for an online book retailer is that a potential buyer can't pick up an online book and peruse it as in a traditional bookstore. Amazon was able to solve this problem with their *Look Inside this Book* feature which allowed users to view a book's front and back covers as well as the index and tables of content for thousands of titles (O'Leary, 2004). In 2003 the company gave this feature a major upgrade when it introduced *Search Inside*

This Book which allowed keyword searching of every page of more than 120,000 titles as well as limited page viewing (Marinaro, 2004). When a searcher selects a book and clicks on *Look Inside!*, a box will appear that says *Search Inside This Book* if the feature is available for that particular work. The book can then be searched by keyword with the results listed below the box. This keyword full text search ability seems virtually the same as an Internet search or even a library database search, as Mick O'Leary (2004) explained:

> The SITB collection is useful for all sorts of reference, including personal, consumer, professional, and academic from high school to grad school. And for reference use—as opposed to reading—the five-page excerpts are often quite sufficient. In other words, SITB is also an alternative—and possibly a threat—to fee-based e-book research collections like netLibrary and ebrary. (p. 43)

Like the Internet, Amazon is proving that a large informational database can provide effective keyword full text searching without using traditional library tools such as decimal classification systems or controlled vocabularies. And based on the company's sales figures listed in the introduction to this section, users are embracing Amazon's intuitive search simplicity, just as users have embraced the Internet.

Research Question 2

If this research suggests that LCSH should be retained but revised or modified for current online technologies, what types of revisions or modifications are being suggested? If so, are there any that seem most promising?

The Future of LCSH

As LCSH moves well into its second century of service, this venerable tool has shown amazing adaptability as it has moved from being a system for finding books by subject for the

Library of Congress, to being one of the most widely used controlled vocabularies not only in the United States but in many foreign countries. Lois May Chan and Theodora Hodges (2000) explained that:

> In the course of the twentieth century, *Library of Congress Subject Headings (LCSH)* grew from a subject access system designed for a single library to become the main subject retrieval tool for libraries throughout the United States and in many other countries around the world. With its current size at approximately a quarter million terms, it is now the most comprehensive non-specialized controlled vocabulary in the English language, and, in addition, has become the *de facto* standard for subject cataloging and indexing in circumstances far beyond those for which it was originally designed. (p. 226)

Chan and Hodges (2000) gave two reasons for LCSH's phenomenal growth. The first is that: "Throughout the 20th century, the Library of Congress has made its cataloging records available to other institutions" (p. 227). The second reason is: "Almost from the beginning the Library took responsibility for giving other libraries an account of its own cataloging policies and practices" (p. 227). These researchers also pointed out how well LCSH has responded to both changing audiences as well as changing environments as libraries have moved from book and card catalogs to OPACs and other online environments. However they do wonder if LCSH can continue its dominant role into the future: "The question for us now is, can it continue to do as well in the future?" (Chan & Hodges, 2000, p. 228).

Karen Fischer (2005) compiled a study of research articles dealing with LCSH from 1990 to 2001. She admitted that while the library community is coming to terms with the inadequacies of LCSH, it is still apparent that it is a very rich and comprehensive list and with

nearly 270,000 terms: “No other controlled vocabulary comes close to the depth and breadth of LCSH” (p. 64), and that “few authors suggest that LCSH should be abandoned” (p. 75). However, Fischer did concede that LCSH has, over the years, had its detractors: “Six decades of literature demonstrate persistent complaints about LCSH: complicated syntax, inadequate syndetic structure, outdated terminology, lack of specificity in the list, and complicated, inconsistent application of subdivisions” (Fischer, 2005, p. 65). She also pointed out that newer literature criticism also listed a lack of adaptability and flexibility in LCSH within the online environment as weaknesses that needed to be addressed. Fischer concluded that the literature from 1990 to 2001 clearly indicated that LCSH must become more flexible, efficient, and easier to use: “LCSH has great potential, but its structure is based on the card catalog model. In order for the list to remain a viable tool in the digital age LC must endorse important changes to increase its adaptability in the online environment” (Fischer, 2005, p. 103).

At the close of the twentieth century it was clear that information storage and retrieval was vastly different from what it was at the start of the century, and the beginning of LCSH. Automated technologies have had a major impact on the changing face of information, not the least of which is the World Wide Web which has changed not only the way information is stored and retrieved, but on the behavior and expectations of the user:

> But the World Wide Web environment differs in many respects from that of a traditional library. Its store of resources is vast, and access to those resources is apparently easy even though retrieval results are not always satisfactory; as a result, information seekers have changed not only their behavior but their expectations. (Chan & Hodges, 2000, p. 229)

These researchers expressed their belief that LCSH, as currently applied: "Comes up somewhat short as an effective tool for subject access in the Web environment" (Chan & Hodges, 2000, p. 230). They concluded that there were three options for LCSH:

- creating a totally new controlled vocabulary covering all subject areas for subject access on the web; or,
- applying *LCSH* as it is currently done, and accepting the fact that its usefulness and effectiveness will be limited largely to the OPAC environment; or,
- retaining LCSH and applying it with a flexible and scalable syntax. (Chan & Hodges, 2000, p. 230)

Perhaps not surprisingly, Chan and Hodges suggested that while the first option might require too much time and effort, and the second option was not a very good one, the third offered real promise. They concluded their research by calling on the information profession to retain LCSH by developing and implementing the appropriate changes:

> *LCSH* contains an enormously rich vocabulary and it offers great potential for successful adaptation to the electronic environment. It is up to the information profession to determine whether it can fulfill that potential, and, if chances are judged good, what measures should be adopted toward that end. (Chan & Hodges, p. 233)

Subject searching in the Web environment is not the only reason researchers have called for improvements to LCSH. The OPACs are also automated tools that use LCSH as a controlled vocabulary and are often considered lacking in subject search capabilities. Pauline Cochrane (2000) believed that in spite of improvements in the design of OPACs over the years, their subject searching capabilities still need to be improved:

> Keyword searching is often touted as the best feature of OPACs, but any subject searcher knows that this mode of searching puts the burden on the user with little or no help provided to track down synonyms, homonyms, or related terms. Surely we can do better than that after twenty years of automated catalogs and indexers? (p. 75)

Cochrane called for improvements in three areas: (1) Notes in LCSH, (2) the cross reference structure of LCSH, and (3) the link between the LCSH headings and the Library of Congress Classification numbers.

A Faceted Syntax Approach

The advantages of a faceted syntax for LCSH are presented by James Anderson and Melissa Hofmann (2006) in their study, *A Fully Faceted Syntax for Library of Congress Subject Headings*. They defined *facets* as, "fundamental categories, aspects, or 'faces' of phenomena similar to the journalist's 'who, what, where, when, why.' Facets represent fundamental characteristics by which any documentary topic or form can be analyzed and described" (p. 8).

As has been previously mentioned, the LCSH system can be quite complex. Anderson and Hofmann point out how this complexity can be problematic for a user who is simply seeking an overview of what is available on a specific subject:

> Homosexuality is not a large topic in the Rutgers University Libraries catalog, yet there are nearly 200 separate unique headings beginning with the term "Homosexuality." For a user to scan through this list, it would take 10 separate screens, if 20 headings were displayed on each screen. Few users would have the patience for such a task, and it is even worse for large topics such as "Women" or "United States. (Anderson & Hoffman, 2006. P. 14).

Anderson and Hoffman defined *faceted syntax* as, "syntax (rules for ordering words in natural language or terms and descriptors in indexing languages) based on their facets" (p. 8). For one example, they used the work, *Black Baptists and African Missions: The Origins of a Movement, 1880-1915,* by Sandy D. Martin (Macon, GA: Mercer, c1989). Below are the LCSH subject terms associated with this work:

West Africans – Missions -- History – 19th Century.

African American Baptists – Southern States – History -- 19th Century.

National Baptist Conventions of the United States of America – Missions – Africa, West – History – 19th century.

Carey, Lott – Contributions to missions.

Anderson and Hoffman propose using the same LCSH headings and subdivisions but arranging them into a single string using faceted syntax:

West Africans. Missions <to> . <by> African American Baptists: National Baptist Convention of the U.S.A.; Carey, Lott. <from> Southern states; <to> Africa, West. 19th century. History.

These researchers noted that inserting words within angle brackets is consistent with standard practices in faceted indexing: "Natural language role indicators may be inserted within angle brackets to clarify the relationship between descriptors" (Anderson & Hoffman, 2006, p. 13). This example does seem to be easier to understand than the initial list of LCSH terms. The fact that it has more of a natural language flow would also seem to be a plus. Another advantage is the simplicity of the authors' syntax: "Single concept terms, plus any needed or helpful natural language role indicators, are placed into facets, in the order listed. These terms are combined into strings of terms for display" (Anderson & Hoffman, 2006. P. 19). For these researchers,

combining the LCSH terms into a single string should give the searcher a better overall view of the work than the separate LCSH terms.

Fast (Faceted Application of Subject Terminology)

Lois Mai Chan and her colleagues along with the OCLC Office of Research, with support from the Library of Congress, have developed their own faceted schema which they have named FAST (Faceted Application of Subject Terminology) (Fischer, 2005). FAST uses LCSH's rich vocabulary but also considerably simplifies the syntax of LCSH, and is designed to serve as a subject vocabulary for use on the Web. "The 'schema' is a controlled vocabulary built on the terminology and relationships already established in LCSH but structured with a different syntax (Fischer, 2005, p. 85). According to Qiang Jin (2008), the individual headings in FAST are not as lengthy as the LCSH strings: "The Library of Congress Subject Headings are longer strings in prescribed order while FAST breaks most strings apart" (p. 92).

Chan and her colleague Edward O'Neill, also a member of the FAST development team, reiterated that one of the main weaknesses of LCSH was that it was designed for a card catalog environment: "While LCSH has served libraries and their patrons well for over a century, its complexity greatly restricts its use beyond the traditional cataloging environment. It was designed for card catalogs and excelled in that environment" (O'Neill & Chan, 2003, p. 337). The goal of FAST was a subject vocabulary which was suitable for the web environment that would have the following characteristics:

- It should be simple in structure (i.e., easy to assign and use) and easy to maintain;
- It should provide optimal access points;

- It should be flexible and interoperable across disciplines and in various knowledge discovery and access environments across disciplines, not the least among which is the OPAC (Chan, Childress, Dean, O'Neill, & Vizine-Goetz, 2001, p. 39)

O'Neill and Chan explained that while the new system retained the hierarchical structure of LCSH as well as the use of subdivisions, FAST differed from LCSH in one major way:

> Its major difference from LCSH is that, in a particular FAST heading, subdivisions must belong to the same facet as the main heading. Topical headings can be subdivided by other topicals, geographic headings by other geographics, etc. That is, a particular main heading may not be subdivided by subdivisions from a different fact. (O'Neill & Chan, 2003, p. 338)

Rebecca Dean (2009), who was also a member of the FAST development team, defined both *simplicity* and *interoperability* as the terms were used in relation to the FAST project: "Simplicity refers to the usability by non-catalogers. Interoperability enables users to search across both discipline boundaries and across information and storage systems" (p. 332).

Assigning a subject term to an item, according to Rick Bennet, Edward O'Neill, and Kerre Kammerer (2014), is essentially a three part process:

1. The first phase is intellectual—reviewing the material and determining its topic.
2. The second phase is more mechanical—identifying the correct subject heading(s)
3. The final phase is retyping or cutting and pasting the heading(s) into the catalog interface along with any diacritics, and potentially correcting formatting and subfield coding. (p. 34).

Bennet, O'Neill, and Kammerer (2014) believed that faceting will make the task of assigning subject terms easier: "Without the complex rules for combining the separate

subdivisions to form an LCSH heading, only the selection of the proper heading is necessary" (p. 34).

According to Chan (2001) and her colleagues, the designers of FAST not only wanted a schema based on the terminology and relationships contained in LCSH but also with policies and procedures more inclined towards post-coordination. They considered pre vs. post-coordination as a key point when dealing with a controlled vocabulary:

> The central issue involving syntax of a controlled vocabulary is pre-coordination vs. post-coordination. Both have precedence in cataloging and indexing practice. Subject vocabularies used in MARC records are typically pre-coordinated subject heading strings, while controlled vocabularies used in online databases are mostly single-concept descriptors, relying on post-coordination for complex subjects. (Chan, Childress, Dean, O'Neill, & Vizine-Goetz, 2001, p. 42)

Precoordinate indexing has been defined as: "The assigning of subject terms to surrogate records in such a way that some concepts, subconcepts, place names, time periods, and form concepts are put together in subject strings, and searchers of the system do not have to coordinate these particular terms themselves" (Taylor & Joudrey, 2009, p. 467). Postcoordinate indexing has been defined as: "The assigning of single concept terms from a controlled vocabulary to surrogate records so that the searcher of the system is required to coordinate the terms through such techniques as Boolean searching" (Taylor & Joudrey, 2009, p. 467).

O'Neill and Chan outlined the basics of the FAST schema in six parts:

1. a controlled vocabulary with all headings established in the authority file, with the exception of headings containing numeric values only
2. based on the LCSH vocabulary

3. designed for an online environment
4. a post-coordinated faceted vocabulary
5. useable by people with minimal training and experience
6. amenable to automated authority control (O'Neill & Chan, 2003, p. 338)

Another feature of the FAST system is that the selection of a particular subject heading is be based on its usage in WorldCat:

> The establishment of a particular heading is determined by its usage in WorldCat, which also includes all of the headings assigned by the Library of Congress. Headings that have never been assigned in WorldCat will not be established in FAST even though they may be valid. (O'Neill & Chan, 2003, p. 338)

FAST will consist of eight distinct facets: topical, geographic (place), personal name, corporate name, form (type, genre), chronological (time, period), title, and meeting name (Childress & Chan, 2003). As of 2013 there were "approximately 1.7 million headings across all facets" (Mixter & Childress, 2013, p. 7).

Jeffrey Mixter and Eric Childress (2013) published a report presenting the results of interviews with sixteen institutions that had contacted OCLC regarding FAST. From this group of sixteen, nine had adopted FAST while seven had decided against adoption. Below is a list of the adopters.

- Boodleian Library, University of Oxford, United Kingdom
- Databib.org, Perdue University
- National Library of New Zealand
- RMIT Publishing, Australia
- Sterling and Francine Clark Art Institute

- Theodore Roosevelt Center Digital Library, Dickinson State University
- University of Amsterdam, the Netherlands
- University of Illinois at Chicago
- University of Chicago

According to Mixter and Childress (2013), the most often-cited positive attribute of FAST was its simple syntax. According to Mixter and Childress, some comments from the adopters were:

- "Prior to using FAST, Databib.org was using LCSH for subject cataloging, but it decided that the complexity of the vocabulary was too frustrating, and it took too much time to assign terms to a record." (p. 20)
- "It allows inexperienced catalogers to add headings quickly, and also allows users to easily discover materials by using facets." (p. 20)
- "The team selected FAST as their cataloging vocabulary because of its simplicity and ease of use. Since the majority of the cataloging was being done by library interns with little cataloging experience, there was a great interest in using a vocabulary that was easy to use and did not have the complexity typical of LCSH." (p. 34)
- "FAST was chosen because it has a simple syntax but still retains the semantic richness of LCSH." (p. 36)
- Unlike LCSH there is no need to string together multiple terms in order to form a valid subject term." (p.40)

Some of the reasons offered for not adopting FAST were: "An absence of customer support and concerns about OCLC's commitment to FAST going forward" (Mixter & Childress, 2013, p. 12). The University of Amsterdam, which had adopted FAST was, however, concerned

that no other large research university that they were aware of was currently using FAST (Mixter & Childress, 2013).. Mixter and Childress also admitted that FAST was not yet a true OCLC product but was still a research project, which might help to partially explain way it had not yet been widely embraced by the larger library community.

Arash Joorabchi and Abdulhussain Mahdi (2013) reported on a system they had developed for the automatic classification and subject indexing of scientific documents in digital libraries and repositories. They used FAST subject terms along with the Dewey decimal classification system as well as key concepts identified in Wikipedia. Joorabchi and Mahdi acknowledged that automated machine learning-based document classification systems should work in conjunction with traditional controlled vocabularies since so much time and effort had been put into developing controlled vocabularies: "These systems aim to combine the power of ML-based text classification methods with the enormous intellectual effort that has been put into developing library controlled vocabularies over the last century" (p. 727). Joorabchi and Mahdi called their system *concept matching approach* (CMA). After all key Wikipedia concepts are identified in a document that is to be classified and indexed, WorldCat is queried for MARC records containing the documents key concepts, retrieving the most relevant records. With this group of MARC records, each corresponding to a key concept identified in the document, the semantic relatedness of each MARC record to the document is measured and the most popular DDC class and FAST subjects for the work represented by each MARC record is identified. Joorabchi and Mahdi selected FAST because "it is a simplified version of the well-known Library of Congress Subject Headings schema (LCSH), designed to retain the rich vocabulary of LCSH while making it easier to understand and use" (p. 732).

Considering the recommendations for the simplification of LCSH, then considering the huge size of the Web, it still seems like these recommendations and products are still a bit too complex. An enormous amount of time and effort would surely still be needed if one of these systems is to apply a controlled vocabulary to virtually every individual web page. A simpler and more automated system still seems to be necessary.

Tagging, Folksonomies, Forms, and Author-Assigned Keywords

Since the introduction of Web 2.0 technologies, the Internet user has moved from a passive searcher to a collaborative participant (Lu & Kipp, 2014). User application of keyword terms or *tagging* has been defined as:

> A populist approach to subject classification. It is a process by which a distributed mass of users applies keywords to various types of Web-based resources for the purposes of collaborative information organization and retrieval. Tagging allows individual users to group similar resources together by using their own terms or labels, with few or no restrictions. Also referred to as *user tagging, social tagging,* and *social indexing.* (Taylor & Joudrey, 2009, p. 474)

This ability to tag web resources by individual users or groups of users has led to a term known as *folksonomies* which has been defined as: "The aggregation of tags created by a large number of individual users. The term is a blend of folks and taxonomy" (Taylor & Joudrey, 2009, p. 456). This user-generated tagging would seem to be a viable answer to the immense amount of material available on the web since they seem to be an inexpensive alternative to traditional labor intensive cataloging processes:

> Folksonomies differ from traditional approaches in that they employ user-generated tags applied by users instead of controlled-vocabulary keywords assigned by trained

professionals. In this manner, tagging takes advantage of the ability to produce low-cost metadata that are critical for current attempts to provide subject access to the enormous number of electronic information resources. (Lu & Kipp, 2014, p. 483)

While this form of user application of metadata to web information resources is more cost effective than more expensive and labor intensive methods of traditional cataloging, an obvious question seems to be, is it as effective? One the one hand, a user assigning tags to a resource would surely be someone who either participated in the creation of the information resource or was familiar with its content. In either case this user would seem knowledgeable enough to apply appropriate terms so that the resource could be found by those seeking the information contained in the resource. But could these taggers really be as effective as catalogers who have had years of experience prefaced with years of training? And could theses folksonomies be as effective as controlled vocabularies? Danielle Lee and Titus Schleyer (2012) compared medical subject headings (MeSH) to CiteULike tags and concluded that the tags were not comparable to the subject headings: "Although MeSH is a tightly curated controlled vocabulary whose terms are highly standardized in content and format, the same does not apply to social tags" (p. 1755)

In their study of mental models of taggers versus experts, Ya-Ning Chen and Hao-Ren Ke (2014) also found the tagging process lacking: "In practice, tags are a set of knowledge representations of a user's cognitive understanding of information resources and their content. "It seems that the mental models of article indexing of taggers are not in line with those of experts" (Chen & Ke, 2014, p. 1675). Chen and Ke also found that more than one third of tags were "identical, synonymous, or variant forms of title keywords of journal articles" (p. 1692).

The author of scholarly research such as a journal article or dissertation often assigns keywords to the work. Though not known specifically as social tagging since these works are

generally not considered web resources, the distinction between authors of scholarly works and social taggers does seem somewhat blurred. Authors of journal articles and dissertations are without question experts on the subject matter they have written about. But, again, does their expert knowledge of the subject translate to the ability to assign metadata as well as experts? In a study of electronic theses and dissertations (ETDs), Sevim McCutcheon (2011) discovered that in a set of 95 ETDs, two authors repeated the title as a keyword entry and many more repeated words in the title and/or abstract. She observed that: "If student authors select keywords that repeat acronyms and words in the title and abstract then no new access points are added" (p. 66). McCutcheon succinctly explained her belief in the superiority of a controlled vocabulary such as LCSH:

> First, when relying on author-supplied metadata, the onus is on the user to think up all possible vocabulary the author might have used to describe the work, a task which could be done successfully on a consistent basis only by a mind-reader. In contrast, LCSH relies on recognition of pre-coordinated strings of words. Once found via a subject browse search, or when a LCSH link within a bibliographic record is clicked, all the works containing that same subject heading are retrieved. (McCutcheon, 2011, p. 66)

However, the question still remains as to how best to achieve the superiority of controlled vocabulary with the cost effectiveness of tagging and author-assigned keywords. One solution suggested by Rebecca Lubas (2009) was to transfer some of the expertise of the cataloger to the taggers and authors in the form of additional training.

A study by Jane Greenberg, Maria Pattuelli, Bijan Parsia, and W. Robertson (2001), found that authors were, in fact, good candidates for assigning metadata to their work: "Resource creators are intimate with their work, they want their work to be discovered and consulted, and

they know their audience and can thus describe their resources appropriately" (p. 38). These researchers concluded that guidance could be provided by development of a form: "Finally, the study shows that the design of a simple form with selective use features may be the best means for author-generated metadata" (p. 44). The form they suggested is included below.

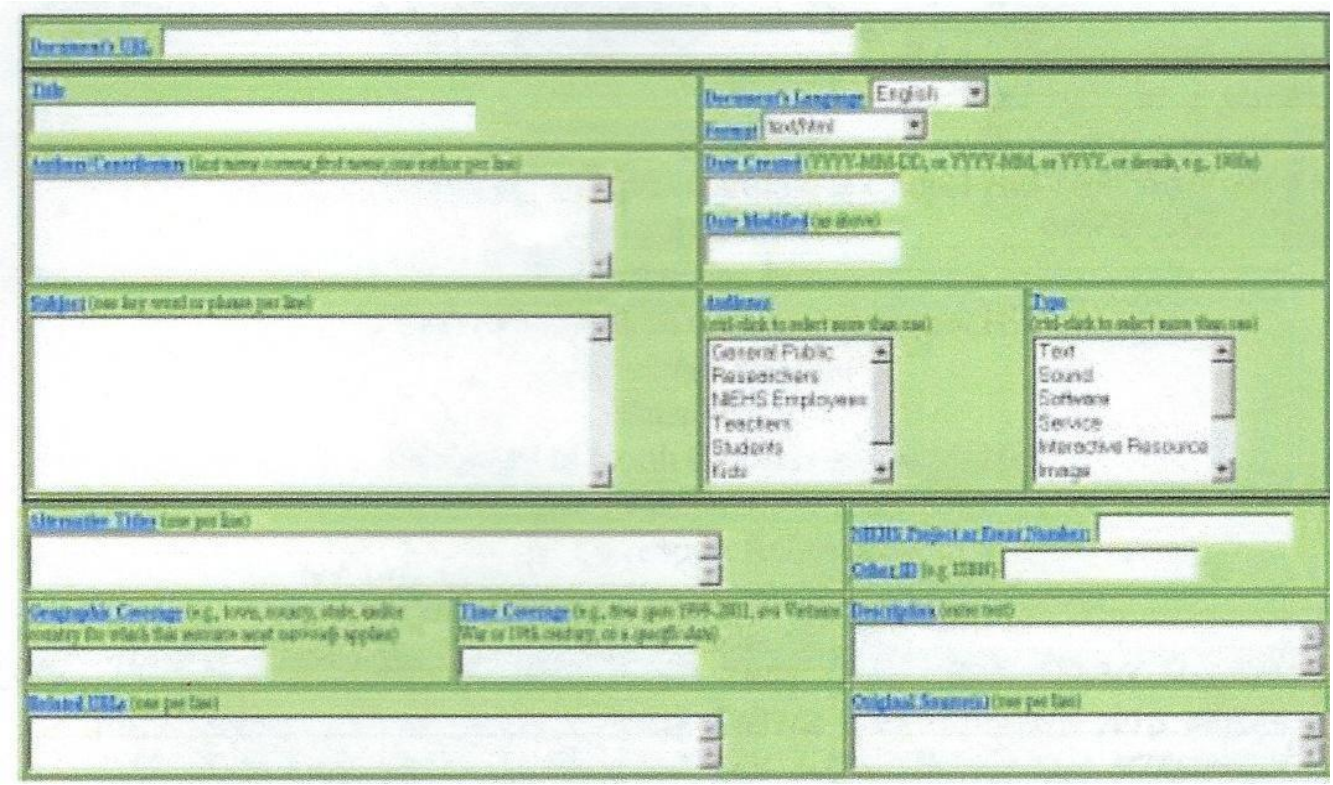


Fig. 5.1. Metadata Form.

Jung-Ran Park (2009) also noted that the enormous volume of online resources "makes (semi) automatic metadata generation an impending need" (p. 223). Park also called for visual aids to assist authors in adding metadata to their work: "A simplified version of metadata guidelines can be embedded within a web form or template in the form of a pop-up window or other forms, providing a benefit to catalogers or document authors in the creation of quality metadata" (p. 224).

As has been mentioned, when an author completes a work, he or she is considered quite knowledgeable in that subject area, especially authors of journal articles and dissertations. However, most would probably agree that expert knowledge of a subject is different than expert knowledge of the best metadata for that subject. Most would also probably agree that if authors of books, articles, and theses and dissertations were, in fact, given some assistance from professionals about assigning metadata to their works, the resulting collaboration might prove to be both efficient and effective.

Book Jackets: Overlooked Access Points?

A substantial portion of the articles and books cited in this research have argued that a controlled vocabulary such as LCSH is necessary because it often provides terms that are otherwise absent from the metadata or even the full text of the item. These terms provide additional access points to help users locate items that may be known by a variety of different terms and synonyms, and that the author may have omitted from the work itself. In today's huge full text digital keyword search environments, any additional terms that relate to a work should only increase the chances of that work being located and selected by a potential user.

An interesting source for additional access terms, one that has been seemingly overlooked by catalogers and others who assign metadata to informational items, is the book jacket. In a study titled *Book Jacket as Access Mechanism: An Attribute Rich Resource for Functional Access to Academic Books*, Brian and Mary O'Connor (1998) discovered that these enticing snippets of text can actually act as search aids to help link academic researchers to appropriate resources in a keyword search digital environment. These researchers believed that while traditional card catalogs as well as OPACs do have a number of helpful points of entry to desired materials, additional access points can still be desirable: "Reducing the number of

documents to be examined and reducing the time to examine each document are fundamental reasons for the construction of access mechanisms – some manner of representation of the documents" (O'Connor & O'Connor, 1998, p. 2).

After a preliminary search of a dozen book jackets, O'Connor and O'Connor came up with a set of 228 for final evaluation. The content analysis of the book jackets revealed seven attributes common to most of them. Below are the attributes with the number of book jackets out of the total of 228 that contained the attribute in parentheses:

- Subject Indicators (219)
- Evaluative Statements (183)
- Summary (227)
- Author Credentials (223)
- User Description (107)
- Reviews (157)
- Reviewer Credentials (144) (O'Connor & O'Connor, 1998, p. 7)

Clearly, the amount of helpful information contained within the book jacket can be surprisingly high. These authors concluded their research by reiterating their belief in the value of book jacket information in digital document environments: "The legacy of the book jacket as an enhanced representation palette can provide a substantial foundation for robust digital representations" (O'Connor & O'Connor, 1998, p. 11).

This research represents yet another example of the versatility of digital keyword search environments. The findings also demonstrate that in the modern digital document era, as with the Amazon example above, finding aids such as this may not render traditional library metadata obsolete, but they can certainly enhance them.

The examples above reveal that there are a number of solutions to be considered when deciding how best to modify controlled vocabularies for digital environments. The researchers of the articles seemed to agree that a controlled vocabulary such as LCSH will need to be simplified if it is to meet the needs of the immense amount of information available on the Internet.

Research Question 3

If a modified or revised version of LCSH still proves to be deficient in modern search and retrieval systems such as the Internet, are there systems that can replace its subject search capabilities? Can such systems be incorporated into modern search engines such that they are either seamless or simple enough so that users can use them intuitively?

Controlled Vocabularies and the Card Catalog

Before the current era of automated technologies, when a university student or library patron wanted a book or books on a certain subject, he or she went to the subject section of the library's card catalog and thumbed through the cards looking for that desired subject. If the student or patron was fortunate enough to find the desired subject listed, then he or she could thumb through the list of books filed under that subject, select an appropriate one, and head off to the stacks to find and peruse it. When the book was located, the student or patron would immediately notice that other books of the same subject were most often in the same area. In this instance both the classification system and the controlled vocabulary worked as designed; the controlled vocabulary term (probably LCSH) found the specific book and the classification system (probably Dewey or Library of Congress) located the book on the shelf and positioned it with books of similar subject in the vicinity, if not immediately to the left or right. This would prove to be a very effective system when the user hit on the right subject term. But what

happened when the user could not find his or her subject listed on any of the cards in the print catalog? In the physical library which housed the print card catalog as well as the print materials, the next step might have been to consult a library staff member for help and hopefully get referred to a knowledgeable reference librarian whose job it would be to convert his or her subject terms to controlled vocabulary terms used in the card catalog. This could also be done by consulting one or more of the big red books in the five volume set situated close to the print card catalogs: *Library of Congress Subject Headings*. The user could also have consulted the red books without asking for help, but they were complicated to use, especially without any prior training. But what if the user was reluctant to ask for help, and had no idea what those red books were for? At this point, however, there didn't seem to be a lot of options. One option would be to simply give up, which undoubtedly many, many library patrons must have done--an obvious instance where the library and its complicated systems for storage and retrieval of materials completely failed the user. However, if the user were a student needing support for an upcoming assignment, quitting was not an option. That student either tried to think of a synonym for the desired subject and hoped that term was in the catalog, or tried to think up a different topic for the assignment and hoped that topic would be in the card catalog. And if everything failed he or she may have simply started looking through the subject cards in the catalog or go directly to the stacks looking at titles hoping to find something appropriate for the assignment.

An excellent example of problems with finding subject terms in a controlled vocabulary using a print index would be a student writing a paper on the death penalty. When the student went to the print card catalog searching for the term, he or she would find nothing because in the LCSH controlled vocabulary system--used as subject entry for most libraries--the controlled vocabulary term selected for this subject was *capital punishment*. However, if the student did use

the term capital punishment then he or she could select an appropriate title from the card catalog, go to the stacks and find most all books on that subject located together. This example reveals both the advantage and major disadvantage of controlled vocabularies in the print environments: It could be somewhat of a hit or miss exercise.

OPACs and Keyword Searching

When online public access catalogs (OPACs) began appearing in libraries, they had one obvious advantage over print catalogs: Keyword searching of a record. Instead of thumbing through a printed card catalog hoping to find a subject term, users could now type their terms into a search box and the system would not only search the subject terms but also the title and any other information contained in the item's record. Using the death penalty/capital punishment example, if death penalty were input into the search box of an OPAC, it would still not be found in the subject terms, but if the title contained the words death penalty, as certainly many books on the subject would, the record for that book could be retrieved. In this instance, the student might scan the record and notice that in the subject terms the phrase *capital punishment* is listed, and when clicked on, more books on the subject immediately appeared. Or he or she could simply jot down the call number and go to the stacks to find that book and others on the topic circumventing the subject terms altogether. A subject search in the printed card catalogs meant that a user had to deal with controlled vocabulary terms as they were the only words printed on the subject cards. OPACs and keyword searching meant it was possible to find something using only keywords, although the search had a much better chance of succeeding if controlled vocabulary terms were input into the search box. However, as more information such as author-assigned keywords as well as table of contents and abstracts began appearing on the records of OPACs, a user's ability to find desired information without a controlled vocabulary

would continue to increase. And when the full text of the item itself, as well as the metadata, is available for keyword searching, as this research has shown, the necessity of a controlled vocabulary for finding desired information seems to be greatly reduced.

Google Books

Google began scanning books in 2002 with the intent of making them available and searchable to Internet users, as Anna Hoffman (2016) explained: "The project aims to do to the world's collection of printed books what the company has already done for web pages: index their contents, analyze their connections, and make them searchable" (p. 77). In 2004 Google announced that it had partnered with the Bodleian Library in Oxford, the New York Public Library, and the libraries of Harvard, Stanford, and the University of Michigan to digitize their collections of books (Green, 2010). Google has since partnered with other libraries, many in languages other than English (Green, 2010). The ultimate goal of the project is to digitize all the books that that have ever been published—*ever*—obviously a herculean task. Eric Schmidt, a former Google CEO, eloquently envisioned the lofty and yet seemingly altruistic goals of the Google Books Project:

> Imagine sitting at your computer and, in less than a second, searching the full text of every book ever written. . . Imagine the cultural impact of putting tens of millions of previously inaccessible volumes into one vast index, every word of which is searchable by anyone, rich and poor, urban and rural, First World and Third, *en toute langue* -- and all, of course, entirely for free. (Schmidt, 2005)

The Company has estimated that there are over 129 million books available for digitization worldwide and has set the ambitious goal of having them all digitized by the year 2020 (Jackson, 2010). In a New York Times article, Stephen Heyman (2015) reported that

Google had, up to that time, scanned 25 million books. While Google may not have all books digitized by 2020, it seems almost a certainty that within a few more decades the dream of digitally searching every word of every book available in all the world's libraries will be a reality. In a study comparing the contents of WorldCat, a database with over 200 million records, and Google Books, Xiaotian Chen (2012) noted that: "As of late 2010 and early 2011, there were hardly any WorldCat books that Google books could not retrieve" (p. 514). Edgar Jones (2009) suggested a few of the many advantages that digitized full texts have over printed books: "This indexing allows one to search both within individual volumes and across the entire collection, facilitating text-based research in general, but especially historical research and the comparison of variant texts" (p. 86).

It would be hard to overstate the importance of having virtually all books digitized and available to anyone without charge. The benefits of such an accomplishment to both individuals as well as societies are, of course, incalculable and would seem to be the culmination of what may have been the dream of ancient libraries: A central location housing the world's knowledge for all to use and enjoy. Digitizing virtually all available books worldwide could also have ramifications beyond the scope of current scholarship, as storage and retrieval systems continue to improve, as Andrew Green (2010) explained:

> In the future, as techniques for searching, text mining, document recognition and automated translation become more sophisticated, as the semantic web develops, and as 'cyberscholarship' becomes more common, these great reservoirs of text will yield new knowledge, and perhaps even generate new research fields. (p. 64).

The ability to search every word of every book ever written will, of course, be a monumental achievement. But for anyone in the world to conduct the search while sitting at a

computer terminal, or even with a table or Smartphone, with virtually no cost and requiring no special training other than the ability to type words into a search box still sounds more like science fiction than fact. However, it does seem to be just on the horizon.

While the many advantages of the Google book digitization project are obvious, and there are undoubtedly uses that have yet to be discovered, there will also be obstacles to overcome. One obvious problem, one that has plagued most large information databases that use keyword search capabilities as the primary means of access to the information they contain, is the problem of relevance. The ability to locate and retrieve all items containing specific words or phrases from a database with literally billions of records does not always translate into a successful search. There is a definite difference between matching search terms to words in a document as opposed to finding desired information on a specific subject, as Elaine Svenonius (2009) has pointed out:

> Bibliographic systems that rely for collocation on the automatic manipulation of character strings on documents, without attempting to interpret their meaning or to show relationships among them, are minimally featured systems. Keyword systems are of this type. While they are often useful for accessing information, they lack the retrieval power of systems in which bibliographic data are intelligently interpreted and organized through set formation and differentiation. (p. 25)

The Google search engine has conclusively demonstrated that most users, perhaps all at one time or another, will choose ease of use over a more perfect search if the latter involves an excessive amount of work. When the Google books digitization project is complete and all 129 million books can be searched simultaneously, and these books are available to anyone, anywhere, the problem of relevance inherent in keyword search systems such as Google will

probably still be problematic. And as with most Google searches, a few keywords input into a single search box could retrieve tens of thousands of books. Even with Google's successful PageRank retrieval system, it would seem that many relevant items may be overlooked because if they are not on the first couple of pages, the user may disregard them. Conversely, when a patron or student searches a library OPAC, although he or she can still be presented with thousands of books, when a promising book is located, that patron or student probably knows that when s selected book is located on the library shelf, other books of the same subject will most likely be in the same section. These books arrived at their locations through a decimal classification system, probably either Dewey of Library of Congress that had been designed, implemented, and improved upon by trained professionals. These classification systems are, arguably, the best and most precise methods of arranging books by subject that have ever been devised, including our present digital era. Of course they were designed to locate and retrieve physical books on a physical shelf in a physical library. Obviously, in the digital world of Google books, there will be no physical shelf to peruse. But with the currently available technologies, could it be possible to wed the intuitive simplicity of keyword searching to the ability to browse a section of books that have been brought together with the specificity of a classification system such as the Dewey or LC? The solution may already be in the modern OPAC just waiting to be discovered.

The Virtual Bookshelf: Digitizing Dewey

Obviously, the books that Google is digitizing are physical books, meaning that they exist somewhere. It probably also goes without saying that the vast majority of these selected books are in libraries around the world. As the third decade of the twenty-first century approaches, it also is a good bet that in virtually all libraries where these books are contained, there is an OPAC

with MARC records or something similarly digital for each one of them. In one of the fields of each record for each book there is a call number, which means that, with what must surely be the rarest of exceptions, every book being digitized has a call number. And since Google is retaining selected metadata for each book (Jones, 2009), the digitized call numbers should be available. Almost twenty-seven years ago Lois Chan suggested that a user's keyword search could be mapped to a classification system: "Term matching can also be accomplished through automatic switching, whereby the user's entry vocabulary is mapped to valid indexing vocabulary or class numbers without the user's being made aware of the switch" (Chan, 1990, p. 260). This mapping to the Dewey decimal classification system may, in fact, be not only possible but quite practical in the Google Books project.

In essentially the same way that users have been searching OPACs for books in the last forty years or so, users will type words that they think best describe their subject and wait to see the results. If nothing satisfactory appears on the first three or four pages, the searcher will probably go back to the search box and enter slightly different terms. When a promising book is found on the list, it will be selected for a more detailed inspection. But unlike the physical library, there is no shelf to walk to that contains more books of the same subject. However, there may be a way to replicate this process in the digital world. Imagine a small box within the screen of the book that had been selected. That box might say something similar to: "MORE LIKE THIS." When the box is clicked the search engine would immediately find the decimal call number for the book and, as Chan (1990) suggested, map the number to the appropriate decimal classification tables that have already been digitized, and return to the user the books closest in subject to the selected book based on the decimal classification system.

Since this is essentially dealing with ascending and descending numbers, albeit Library of Congress is alphanumeric, the necessary algorithm would seem to be remarkably simple. There would, of course, be interoperability issues to be worked out between classification systems as well as other possible difficulties, as there are with all new technologies, but, hopefully, the technology is available to solve them. And if this technology proves to be as doable as it seems, the need for a controlled vocabulary may, in fact, diminish. Perhaps most importantly, the process is automated and is therefore completely seamless to the user.

LCSH and Relevance Ranking

Most of the articles in the previous section have called for some form of simplification of controlled vocabularies such as LCSH so that it will be better suited for large, full text environments such as the Internet. Along with simplifying LCSH, some researchers, albeit a smaller chorus, have also been calling for some form of automation: (Yi & Chan, 2010), (Yi & Park, 2009), (Calhoun, 2006), (Garrett, 2007). The graphic below also seems to hint at not only how to simplify LCSH but to make it work in an automated system.

Table 5.1

LCSH Terms/LCSH Words or Phrases in Title, Abstract, or Full Text

LCSH Terms	LCSH Words or Phrases in Title, Abstract, or Full Text
apartheid - public opinion mass media - Unites States - influence public opinion - United States South Africa - foreign relations - United States United States - foreign relations - South Africa	apartheid, "public opinion" "mass media" "United States" influence "public opinion" "United States" "South Africa" "United States" foreign, relations "United States" "South Africa" foreign, relations
British productivity council economic assistance, American - Great Britain Great Britain - economic conditions - 1945 - 1954 Great Britain - Foreign economic relations - United States industries - Great Britain - history - 20th century Marshall Plan reconstruction (1939 -1951) - Great Britain United States - foreign economic relations - Great Britain	British productivity council economic, assistance, American, "Great Britain" "Great Britain" "economic conditions" 1945, 1954 Great Britain, foreign, "economic relations" "United States" industries, history, "Great Britain" 20th, century Marshall Plan reconstruction, "1939 - 1951" "Great Britain" foreign "United States" "economic relations", Great Britain

Each of the two rows on the left lists LCSH terms from a dissertation and in the two rows on the right exactly how words and phrases from these LCSH terms appeared in the title, abstract, or full text of the two individual dissertations. Creating such long strings as the LCSH terms on the left seems to be unnecessary in a keyword search environment where users simply input words and phrases. Breaking them up into simpler words and phrases would seem to be more effective since that is how they appear in both the metadata and the full text and would be in line with the simplification that has been called for in the earlier examples.

The table seems to graphically point to a way to help with relevancy rankings. When a searcher enters words and phrases into a search engine, documents containing those words are presented to the searcher with what ideally should be the most relevant documents first. However, documents containing less user search terms can sometimes be more relevant than those that have more of the terms. It seems possible to have the search engine look to the LCSH headings as well as the other metadata and full text, and if words or phrases are contained in the LCSH then those items would be listed first. This would seem to be similar to when a user searches a database such as an OPAC that uses LCSH terms. When an appropriate item is located, a click on the appropriate LCSH term brings up more items similar to the one selected. However, in this way the process is automated and requires no additional effort by the user. In the present era of users entering keywords into a single search box and sifting through the results, a more automated system for harnessing the power of a controlled vocabulary such as LCSH seems both appropriate and doable.

The Future: Totally Automated

A final question one might ask is whether, in the next few decades as the Google books project is being completed, the classification systems themselves along with the controlled vocabularies that accompany them could both be automated so that they might not only be converted to Google Books, but also to the Internet at large and its billions of documents. Deciding which specific words or phrases best describe the subject of a written work, along with developing and continually upgrading and modifying a classification system so that similar works are grouped together with as much precision as possible is not only expensive and highly labor intensive, it is also somewhat inconsistent because people can never be totally objective. Also, to decide on the subject of a work, particularly one that is difficult, especially esoteric, or

simply one that touches on several subjects, by simply looking at the title, glancing at the table of contents or index, and flipping through the pages would seem at best somewhat inadequate. We certainly have the technology to scan and count every word, phrase, and sentence contained in any work and can compare those words, phrases, and sentences to every other work in a digitized collection, something no person or group could do for even a few hundred books, much less 126 million. What is lacking, of course, is an algorithm or algorithms that can make sense of those words, sentences, and phrases is a way that would, like classification systems, decide what work is closest in subject to another. However, a system that can automatically group similar items, no matter how large the database, and do it as well or better than decimal classification systems and controlled vocabularies, while at the same time allowing searchers to input keywords into a single search box would seem to be the best marriage of simplicity and automation.

Limits of this Research Project

Though much can often be gleaned from a small sample, eighty-six dissertations from a database of thousands of records with hundreds of such full text digital databases on campuses not only around the country but around the world, the dataset for this research seems much too small to be considered an adequate sample. This research is also limited in that it uses quantitative data more than qualitative. To get a better understanding of the role of controlled vocabularies in the modern keyword search environment, it will be necessary to learn more about the user since that user's search habits should be the ultimate arbiter of whether controlled vocabularies such as LCSH can be effectively retained in future search engines.

Recommendations for Future Research

Future studies of this nature should consider the user and his or her relationship to controlled vocabularies such as LCSH. If users are not using them for what they were designed,

how can they be modified to best serve those users? Those using a Thesis and Dissertation database are most likely graduate students seeking support for their academic projects or faculty members working on their own research. Transaction logs have shown that UNT's Thesis and Dissertation database is heavily used meaning that a large pool of users is available for interviews or similar user-centered research projects (Alemneh, Phillips, Waugh, & Tarver, 2015).. Finding out what frustrations these users have and how best to use controlled vocabulary terms to address these frustrations would seem an appropriate supplement to this research.

Much of this study dealt with the influence of Google and its insistence on a single, simple, search box on an uncluttered page. While so much of the literature details as well as confirms Google's influence, the venerable search engine is approaching its twentieth birthday, a very long time for modern technologies. And yet, a controlled vocabulary has never been matched to Google. Research needs to continue to seek a marriage between the Google's simplicity and a controlled vocabulary's proven ability to improve relevance.

Chapter Summary

This dissertation research was put in motion by a sentence from the work of Tina Gross (2015) and her colleagues: "In the long run, the ultimate test of the importance of controlled vocabulary will be its effect in full text environments" (p. 30). As the literature review of this dissertation demonstrates, many research projects dealing with controlled vocabularies such as LCSH are based on online environments that do not have full text capabilities such as OPACs, however this research considered full text databases. Specifically, three full text databases were most heavily used as the focus for this research: (1) The Thesis and Dissertations Database of the University of North Texas, (2) The Internet, and (3) The Google Books project. Within the Thesis and Dissertations site, LCSH are a functioning part of the metadata, acting as hyperlinks

to all items with the same LCSH term or terms, functioning pretty much as a controlled vocabulary is designed to do—locate all material on the same subject. A controlled vocabulary such as LCSH is noticeably absent from the Internet, however, in the Google Books project LCSH may be made available in the metadata, as selected metadata for books being digitized is being retained (Jones, 2009). However, at present it is not clear how or if LCSH will be used in the Google Books project.

A controlled vocabulary such as LCSH has probably not been successfully converted to the Internet because when it was designed and implemented in the beginning of the twentieth century, a digitized full text database the size and scope of the Internet could not even have been imagined. LCSH was designed for use on small cards in narrow drawers of a printed card catalog, specifically to find most or all items on a selected subject.

One of the findings of this research is that improvements in technology have tended to lessen the role of a controlled vocabulary, particularly from the searcher's point of view. And the technology that seems to be most responsible for the diminishing role of controlled vocabularies like LCSH is keyword searching of digitized materials. When a user went to the printed card catalog looking for books by subject, the process tended to be all or nothing, either a card was found with the desired subject and books with that subject were listed, or nothing was found. It could be argued, therefore, that use of controlled vocabularies in *successful* subject searches of printed card catalogs was essentially 100 percent. There just wasn't another way to find materials by subject. And, of course, many subject searches failed because the user didn't know or couldn't find the correct subject term in the controlled vocabulary. When OPACs began replacing printed card catalogs, this equation changed. When users entered keywords into the search box, the title of the record was searched as well as the subject terms, meaning that the

search was no longer all or nothing. If a keyword search term or terms was in the title, the record was retrieved. And, unlike the print catalog where the subject terms was alphabetized by first word in the string, meaning that the first word had to be found or the rest of the terms were lost, if a keyword was found anywhere in the subject string, the record was retrieved. Successful searches were no longer 100 percent dependent on finding the exact subject term. As more information was added to the metadata of the records such as author-assigned keywords, abstracts, and table of contents, the number of successful keyword searches increased which meant that the number of successful subject searches that skipped the LCSH terms increased.

Clearly, as digital environments have improved, the need for a controlled vocabulary seems to be diminishing. The fact that the Google search engine has operated quite well without a controlled vocabulary for almost twenty years, in what is by far the world's largest full text database, would tend to confirm this finding.

The authors who have argued for simplifications and modifications to LCSH are, in a sense, admitting that LCSH, as it was initially designed and presently structured, is simply not a good fit for digital, keyword search environments. As the Google search engine has demonstrated and users have confirmed, inputting keywords and phrases into a single search box and choosing from a results list is preferable to having to learn how to use a more effective but difficult search aid such as a controlled vocabulary.

LCSH, like other controlled vocabularies was designed to make finding relevant information easier for the searcher, which it did well in print environments. However, it is still not completely clear if or how controlled vocabularies such as LCSH can be successfully transitioned to large, full text keyword search environments. Still, it would be hard not to argue

that some system that either modifies LCSH or replaces it is still needed to aid in sifting relevant items from the myriad of documents that make up the Internet.

A Final Thought

It has been suggested that written records exist for only about 10% of human history, and that 90% of man's history is lost because of the lack of written records (Wade, 2006). One might argue that rudimentary writing systems were available to early humans at some point, but no system was available to permanently retain the writings as their importance to future generations may not have yet been realized. It might also be argued that this did not change when a mature system of writing was devised or when more durable substances such as papyrus scrolls and clay tablets were invented. Writing probably became more permanent when efficient systems for storage and retrieval were devised. And these systems most likely came together with the advent of the first libraries, which appeared around 3000 BC (Harris, 1999).

While it may not yet be considered immanent, the completion of the Google Books project and its digitization of 126 million books seem tantalizingly close. If it is, then the digitization of every edition of every newspaper worldwide along with every issue of every magazine and journal article must also be within reach. To have such a vast store of information available from the search screen of a computer, tablet, or cell phone will surely mean that information will have reached a new state of democratization. When this all comes together, probably even sooner, more people will probably have cell phones than have access to a library. It seems that this single, vast, and all-encompassing information database will become the world's universal information library. But if the gateway to such an immense store of information is keywords input into a single, simple search box, which users should still undoubtedly demand, then what type of search engine will have been developed to seamlessly

retrieve the most relevant results from an almost infinitesimal amount of retrievable information? A good bet is that no matter what the final design of the algorithms, the goal of that state of the art search engine will be the same as it has been for all libraries from the ancient world to our modern era: To group similar items together in a system that makes finding the most relevant items as simple as possible for the user.

This research looked at the role of controlled vocabularies, specifically LCSH, in present and future digital databases. Improvements in technology have made this an important question; Improvements in technology will most likely provide a definitive answer. And, of course, while creating more questions along the way.

APPENDIX

THE 86 DISSERTATION TITLES FROM THE SPREADSHEET/DATASET

	Title/ Author
1	***The Extensive Subject File: A Study of User Expectations in a Theological Library*** **Cecil R. White**
2	***A Framework of Automatic Subject Term Assignment: An Indexing Conception-Based Approach*** **Eunkyung Chung**
3	***Collection-Level Subject Access in Aggregations of Digital Collections: Metadata Application and Use*** **Oksana Zavalina**
4	***News photography image retrieval practices: Locus of control in two contexts*** **Diane Rasmussen Neal**
5	***An Examination of the Relationship Between Published Book Reviews and the Circulation of Books at an Academic Library*** **Glenda A. Thornton**
6	***The effect of information literacy instruction on library anxiety among international students*** **Joel C. Battle**
7	***Discovering a descriptive taxonomy of attributes of exemplary school library websites*** **Joyce Kasman Valenza**
8	***Evaluating e-Training for public library staff: A quasi-experimental investigation*** **Teresa Dalston**
9	***Smoothing the information seeking path: Removing representational obstacles in the middle-school digital library.*** **June M. Abbas**
10	***User satisfaction in a government library: A case study of the Ministry of Foreign Affairs in Saudi Arabia*** **Jamal Abbas Tameem**
11	***Relation of Personal Characteristics To Type Of Position Among Bibliographic network coordinator's, Ex-coordinators, and Selected Library Department Heads*** **Lois Nicholson Upham**

12	***Evaluation by Korean Students of Major Online Public Access Catalogs in Selected Academic Libraries*** **Il-jong Park**
13	***Are Online Catalogs for Children Giving Them What They Need? Children's Cognitive Development and Information Seeking and Their Impact on Design*** **Stacy Creel**
14	***A Survey of All American History Textbooks Adopted for the Public High Schools of Texas from 1919 to 1970*** **Kenneth Reuben Durham**
15	***Korean studies in North America 1977-1996: A bibliometric study*** **Kyungmi Chun**
16	***Exploring Naming Behavior in Personal Digital Image Collections: The Iconology and Language Games of Pinterest*** **Tami Sutcliffe**
17	***Integration of Students with Disabilities into a Contemporary Technology Education Program: A Case Study*** **David Terrell Pullias**
18	***An Examination of Factors Contributing to Critical Thinking and Student Interest in an On-line College-level Art Criticism Course*** **Glenell McKinnon Beach**
19	***The Validity of Health Claims on the World Wide Web: A Case Study of the Herbal Remedy Opuntia*** **Michael A. Veronin**
20	***A multi-state political process analysis of the anti-testing movement*** **Carol DeMerle**
21	***A Study Of The Perception Of Cataloging Quality Among Catalogers In Academic Libraries*** **Karen Snow**
22	***Media Agenda-Building Effect: Analysis of American Public Apartheid Activities, Congressional and Presidential Policies on South Africa, 1976-1988*** **Ehikioya Agboaye**

35	***Research Information and Facilities Available to Graduate Art Students at Ninety European and North American Art Museums*** **Lois Swan Jones**
36	***The Development of a Program in Humanities for the Junior College Curriculum*** **Jacob Marshall Trieber**
37	***The Anglo-American Council on Productivity: 1948-1952 British Productivity and the Marshall Plan*** **Carl H. Gottwald**
38	***Assessment of the Current Status of Informatics in Colombia's Universities and Society*** ***Eusebio Jose Cabrales***
39	***An Investigation Into the Relationships Between the Technological Pedagogical Content Knowledge of University Teacher Education Faculty and Their Age, Rank, and Gender*** **Christina Hamilton**
40	***Jane McManus Storm Cazneau (1807-1878): a Biography*** **Linda Sybert Hudson**
41	***A Philosophy for Two-year Occupational Programs in Public Junior College Curricula*** **William Sterling McClung**
42	***The History of Speech and Drama Education in the Dallas Public Schools (1884-1970)*** **Rose-Mary Rumbley**
43	***A Study of Certain Variables and their Implication in the Vocational Rehabilitation Training of Veteran Trainees*** **Quinten Snow Mathews**
44	***A Case Study of Interpersonal Influences in a Band Music Setting: Bohumil Makovsky (1878-1950) and His Association with Selected Individuals Involved in Instrumental Music in the State of Oklahoma*** **Richard Charles Dugger**
45	***A Faculty Orientation and Design for Writing Across the Curriculum*** **Tahita N. Fulkerson**
46	***Preschool Teachers' Constructions of Early Reading*** **Karen Elledge Walker**

47	***User-Centered Evaluation of the Quality of Blogs*** **Sutthinan Chuenchom**
48	***Effectiveness of a Performance Contracting Program in Reading and Mathematics Relative to Educationally Deprived Secondary School Students*** **Peggy Joy Lloyd Kelley**
49	***Toward a Grounded Theory of Community Networking*** **Kathryn Masten-Cain**
50	***The Status of the Implementation of International Education in Texas Four-Year Colleges and Universities: a Comprehensive Study*** **Sarah Hodges**
51	***Problem-Based Learning for Training Teachers of Students with Behavioral Disorders in Hong Kong*** **Vivian Woon King Heung**
52	***The Organization, Implementation, and Evaluation of an Organized Guidance Program during the First Year in a Metropolitan High School*** **Morgan Clay Moses**
53	***Creating a Mythistory: Texas Historians in the Nineteenth Century*** **Laura Lyons McLemore**
54	***An Analysis of Certain Factors Associated with Financing Capital Outlay for Texas Public Schools*** **Allen James Herndon**
55	***Covering the Campus: The History of The Chronicle of Higher Education*** **Patricia L. Baldwin**
56	***Elizabeth Bishop in Brazil: an Ongoing Acculturation*** **Elizabeth Neely**
57	***Comparative Effects of Two Physical Conditioning Programs and Evaluation of Instruments for Measuring Physical Fitness*** **John Ray Montgomery**
58	***The Effect of Training in Test Item Writing on Test Performance of Junior High Students*** **Jeanne L. Tunks**

59	***The History of Alcoholism Treatment in the United States*** **Suzanne S. Brent**
60	***From reactionary to responsive: Applying the internal environmental scan protocol to lifelong learning strategic planning and operational model selection*** **David L. Downing**
61	***The implementation of international education in colleges and universities in the state of Texas: A follow-up study*** **Sara Hodges**
62	***An examination of music for trumpet and marimba and the Wilder Duo with analyses of three selected works by Gordon Stout, Paul Turok, and Alec Wilder*** **Christopher C. Foster**
63	***Government and Private Funding of Nonprofit Visual Arts Organizations in the State of Texas: An Analysis*** **Maurine C. Howard**
64	***The Search for Order and Liberty : The British Police, the Suffragettes, and the Unions, 1906-1912*** **Kung Tang**
65	***A Qualitative Study of Nine Elementary Principals Providing Inclusion for the Differently Abled*** **Cloyd L. Hastings**
66	***Issues for the Nineties: An Analysis of 14 State Master Plans for Higher Education*** **John Paul Thompson**
67	***André Malraux: the Anticolonial and Antifascist Years*** **Richard A. Cruz**
68	***Understanding the Motivation of Vietnamese International Students and Their Higher Education Experiences in the United States*** **Randy Scott Miller**
69	***Parents Of Children With High-functioning Autism: Experiences In Child-parent Relationship Therapy (Cprt)*** **Jeffrey M. Sullivan**

REFERENCES

Alemneh, D. G., Phillips, M. E., Waugh, L., & Tarver, H. (2015). *Understanding user discovery of ETD: Metadata or full text, how did they get there?* Paper presented at the United States Electronic Thesis and Dissertation Association annual conference, Austin, Texas, September 29 - October 1.

Almeida, M. B. (2013). Revisiting ontologies: A necessary clarification. *Journal of the American Society for Information Science & Technology, 64*(8), 1682-1693. doi:10.1002/asi.22861

Anderson, J. D., & Hoffmann, M. A. (2006). A fully faceted syntax for Library of Congress subject headings. *Cataloging & Classification Quarterly, 43*(1), 7-38. doi:10.1300/J104v43n01_03

Auletta, K. (2010). *Googled: The end of the world as we know it*. New York: Penguin Books.

Baker, S., & Gonzales, A., C. (2012). Graduate students and federated searching. *Internet Reference Services Quarterly, 17*(1), 13-31. doi:10.1080/10875301.2012.658738

Battles, M. (2015). *Library: An unquiet history*. New York: W. W. Norton & Company.

Beall, J. (2008). The weaknesses of full text searching. *Journal of Academic Librarianship, 34*(5), 438-444. doi:10.1016/j.acalib.2008.06.007

Bennett, R., O'Neill, E. T., & Kammerer, K. (2014). AssignFast: An autosuggest-based tool for FAST subject assignment. *Information Technology and Libraries, 33*(1), 34-43.

Bhat, M. H. (2013). Knowledge organization systems in digital environment. *Trends in Information Management, 9*(1), 38-53.

Blair, D. (2006). *Wittgenstein, language and information: Back to the rough ground*. The Netherlands: Springer.

Boonyoung, T., & Mingkhwan, A. (2014). Semantic search: Document ranking and clustering using computer science ontology and N-grams. *Journal of Digital Management, 12*(6), 369-378.

Borgman, C. L. (1999). The user's mental model of an information retrieval system: An experiment on a prototype online catalog. *International Journal of Human-Computer Studies, 51*(2), 435-452. doi:10.1006/ijhc.1985.0318

Brown, J. D. (1898). *Manual of Library Organization and Shelf Arrangement*. London: Library Supply Company.

Calhoun, K. (2006). *The changing nature of the catalog and its integration with other discovery tools: Final report*. Prepared for the Library of Congress. March 17, 2006. Library of Congress. Retrieved from https://www.loc.gov/catdir/calhoun-report-final.pdf

Carr, N. (2011). *The Shallows: What the Internet is doing to our brains.* New York: W. W. Norton.

Casson, L. (2001). *Libraries in the ancient world*. New Haven: Yale University Press.

Chan, L. M. (2005). *Library of Congress subject headings: Principles and applications* (4th ed.). Westport, Connecticut: Libraries Unlimited.

Chan, L. M. (1990). Subject analysis tools online: The challenge ahead. *Information Technology and Libraries, 9*(3), 258-262.

Chan, L. M., Childress, E., Dean, R., O'Neill, E. T., & Vizine-Goetz, D. (2001). A faceted approach to subject data in the Dublin core metadata record. *Journal of Internet Cataloging, 4*(1-2), 35-47. doi:10.1300/j141v04n01_05

Chan, L. M., & Hodges, T. (2000). Entering the new millennium: A new century for LCSH. *Cataloging & Classification Quarterly, 29*(1-2), 225-234. doi:10.1300/j104v29n01_16

Chen, X. (2012). Google books and WorldCat: A comparison of their content. *Online Information Review, 36*(4), 507-516. doi:10.1108/14684521211254031

Chen, Y., & Ke, H. (2014). A study on mental models of taggers and experts for article indexing based on analysis of keyword usage. *Journal of the Association for Information Science and Technology, 65*(8), 1675-1694. doi:10.1002/asi.23077

Chomsky, N. (2007). *On language: Chomsky's classic works language and responsibility and reflections on language*. (pp. 1-269). New York: The New Press.

Cochrane, P. A. (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. doi:10.1300/j104v29n01_05

Cooper, W. S. (1971). A definition of relevance for information retrieval. *Information Storage and Retrieval, 7*(1), 19-37.

Cummins, C., & Ruiter, J. P. (2014). Computational approaches to the pragmatics problem. *Language and Linguistics Compass, 8*(4), 133-143. doi:10.1111/lnc3.12072

Cutter, C. A. (1904). *Rules for a Dictionary Catalog* (4th ed.). Washington: Document Printing Office.

Dean, R. J. (2009). FAST: Development of simplified headings for metadata. *Cataloging & Classification Quarterly, 39*(1-2), 331-352. doi:10.1300/J104v39n01_03

Dervin, B. (2000). Chaos, order, and sense-making: A proposed theory for information design. In R. Jacobson (Ed.), *Information Design* (pp. 35-57). Cambridge, Massachusetts: MIT Press.

Ducheyne, S. (2009). "To treat of the world": Paul Otlet's ontology and epistemology and the circle of knowledge. *Journal of Documentation, 65*(2), 223-244. doi:10.1108/00220410910937598

Fischer, K. S. (2005). Critical views of LCSH, 1990-2001: The third bibliographic essay. *Cataloging & Classification Quarterly, 41*(1), 63-109. doi:10.1300j104v41n01_05

Frank, E., & Paynter, G. W. (2004). Predicting Library of Congress classifications from library of congress subject headings. *Journal of the American Society for Information Science & Technology, 53*(3), 214-227.

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

Gilchrist, A. (2003). Thesauri, taxonomies and ontologies - an etymological note. *Journal of Documentation, 59*(1), 7-18.

Gilreath, J. & Wilson, D. L. (Eds.). (1989). *Thomas Jefferson's library: A catalog with the entries in his own order*. Library of Congress. Retrieved from http://catdir.loc.gov/catdir/toc/becites/main/jefferson/88607928_intro.html

Glasgow, E. (2001). Sir Anthony Panizzi. *Library Review, 50*(5), 251-254. doi:10.1108/00242530110394529

Green, A. (2010). Big digitization: Origins, progress and prospects. *International Journal of Humanities and Arts Computing, 4*(1-2), 55-66. doi:10.3366/ijhac.2011.0007

Greenberg, J., Pattuelli, M. C., Parsia, B., & Robertson, W. D. (2001). Author-generated Dublin core metadata for web resources: A baseline study in an organization. *DCMI International Conference on Dublin Core and Metadata Applications,* Tokyo, Japan, 24-26, October 2001. 38-45.

Grice, H. P. (1957). Meaning. *The Philosophical Review, 66*(377-388).

Grice, H. P. (1975). Logic and conversation. In P. Cole, & J. L. Morgan (Eds.), *Syntax and semantics 3: Speech acts* (pp. 41-58). New York: Academic Press.

Gross, T., & Taylor, A. G. (2005). What have we got to lose? The effect of controlled vocabulary on keyword searching results. *College & Research Libraries, 66*(3), 212-230. doi:10.5860/crl.66.3.212

Gross, T., Taylor, A. G., & Joudrey, D. N. (2015). Still a lot to lose: The role of controlled vocabulary in keyword searching. *Cataloging & Classification Quarterly, 53*(1), 1-39. doi:10.1080/01639374.2014.917447

Hafner, K. & Lyon, M. (1996). *Where wizards stay up late: The origins of the Internet.* New York: Simon & Schuster.

Harper, C. A., & Tillett, B. B. (2007). Library of Congress controlled vocabularies and their application to the semantic web. *Cataloging & Classification Quarterly, 43*(3-4), 47-68. doi:10.1300/J104v43n03_04

Harping, P. (2013). *Introduction to controlled vocabularies: Terminology for art, architecture, and other cultural works*. Los Angeles, CA: Getty Research Institute.

Harris, M. H. (1999). *History of libraries in the western world* (4th ed.). Lanham, Maryland: The Scarecrow Press, Inc.

Headrick, D. R. (2000). *When information came of age: Technologies of knowledge in the age of reason and revolution, 1700-1850*. New York: Oxford University Press.

Hemminger, B. M., Saelim, B., Sullivan, P. F., & Vision, T. J. (2007). Comparison of full-text searching to metadata searching for genes in two biomedical literature cohorts. *Journal of the American Society for Information Science and Technology, 58*(14), 2341-2352. doi:10.1002/asi.20708

Heyman, S. (2015, October 28, 2015). Google books: A complex and controversial experiment. *New York Times*. Retrieved from http://www.nytimes.com/2015/10/29/arts/international/google-books-a-complex-and-controversial-experiment.html?_r=1

Hoffmann, A. L. (2016). Google books, libraries, and self-respect: Information justice beyond distributions. *The Library Quarterly: Information, Community, Policy, 86*(1), 76-92.

Isaacson, W. (2014). *The innovators: how a group of hackers, geniuses, and geeks created the digital revolution.* New York: Simon & Schuster.

Jackson, J. (2010, August 10). Google: 129 million different books have been published. *PC World,* August 10. Retrieved from http//www.pcworld.com/article/202803/google_129 million_different_books_have_been_published.html

Jin, Q. (2008). Is FAST the right direction for a new system of subject cataloging and metadata? *Cataloging and Classification Quarterly, 45*(3), 91-110.

Jones, E. (2009). Google books as a general research collection. *Library Resources & Technical Services, 54*(2), 77-89.

Joorabchi, A., & Mahdi, A. E. (2013). Classification of scientific publications according to library controlled vocabularies. *Library Hi Tech, 31*(4), 725-747. doi: 10.1108/LHT-03-2013-0030

Jurisica, I., Mylopoulos, J., & Yu, E. (2004). Ontologies for knowledge management: An information systems perspective. *Knowledge and Information Systems, 6*(4), 380-401. doi:10.1007/s10115-003-0135-4

Koch, T. W. (1914). Some old-time old-world librarians. *The North American Review, 200*(705), 244-259.

Lee, D. H., & Schleyer, T. (2012). Social tagging is no substitute for controlled indexing: A comparison of medical subject headings and CiteULike tags assigned to 231,388 papers. *Journal of the American Society for Information Science and Technology, 63*(9), 1747-1757. doi:10.1002/asi.22653

Leise, F. (2008). Controlled vocabularies: An introduction. *The Indexer, 26*(3), 121-126.

Lown, C., Sierra, T., & Boyer, J. (2013). How users search the library from a single search box. *College & Research Libraries, 74*(3), 227-241. doi:10.5860/crl-321

Lu, K., & Kipp, M. E. I. (2014). Understanding the retrieval effectiveness of collaborative tags and author keywords in different retrieval environments: An experimental study on medical collections. *Journal of the Association for Information Science and Technology, 65*(3), 483-500. doi:10.1002/asi.22985

Lubas, R. L. (2009). Defining best practices in electronic thesis and dissertation metadata. *Journal of Library Metadata, 9*(3-4), 252-263. doi:10.1080/19386380903405165

Luhn, H. P. (1966). Keyword-in-context index for technical literature (KWIC index). In D. G. Hays (Ed.), *Readings in automatic language processing* (pp. 159-167). New York: American Elsevier Publishing Company Inc.

Mai, J. (2013). The quality and qualities of information. *Journal of the American Society for Information Science & Technology, 64*(4), 675-688. doi:10.1002/asi.22783.

Mann, T. (2008). Will Google's keyword searching eliminate the need for LC cataloging and classification? *Journal of Library Metadata, 8*(2), 159-168. doi:10.1080/10911360802087366

Marinaro, T. S. (2004). Searching for profits inside Amazon—inside the book and in the margins. *Publisher's Research Quarterly*, *20*(3)*,* 3-8.

Maron, M. E., & Kuhns, J. L. (1960). On relevance, probabilistic indexing and information retrieval. *Journal of the Association for Computing Machinery, 7*(3), 216-244.

McCutcheon, S. (2009). Keyword vs controlled vocabulary searching: The one with the most tools wins *The Indexer, 27*(2), 62-65.

McCutcheon, S. (2011). Basic, fuller, fullest: Treatment options for electronic theses and dissertations. *Library Collections, Acquisitions, & Technical Services, 35*(2/3), 64-68. doi:10.1016/j.lcats.2011.03.019

Mixter, J. & Childress, E. R. (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.

Norman, D. (2013). *The design of everyday things*. New York: Basic Books.

O'Connor, B. C. (1996). *Explorations in indexing and abstracting: Pointing, virtue, and power.* Englewood, Colorado: Libraries Unlimited, Inc.

O'Connor, B. C. & O'Connor, M. K. (1998). Book jacket as access mechanism: An attribute rich resource for fundamental access to academic books. *First Monday, 3*(9), p. 1-12. Accessed at http://firstmonday.org/ojs/index.php/fm/article/viewArticle/616/573

O'Leary, M. (2004). Thinking outside the box. *Information Today*, *20*(2)*,* 39, 43.

O'Neill, E. T., & Chan, L. M. (2003). FAST (faceted application of subject terminology): A simplified vocabulary based on the Library of Congress subject headings. *IFLA Journal, 29*(4), 336-342.

Page, L., Brin, S., Motwani, R. & Winograd, T. (1999). The PageRank citation ranking: Bringing order to the Web. Technical Report: Stanford InfoLab. Stanford, Calif.: Stanford University. Retrieved from http://ilpubs.stanford.edu:8090/422

Park, J. R. (2009). Metadata quality in digital repositories: A survey of the current state of the art. *Cataloging & Classification Quarterly, 47*(3-4), 213-228. doi:10.1080/016393709027372

Pong, J. Y., Kwok, R. C., Lau, R. Y., Hao, J., & Wong, P. C. (2008). A comparative study of two automatic document classification methods in a library setting. *Journal of Information Science, 34*(2), 213-230. doi:10.1177/0165551507082592

Ponsford, B. C., & vanDuinkerken, W. (2007). User expectations in the time of google: Usability testing of federated searching. *Internet Reference Services Quarterly, 12*(1/2), 159-178. doi:10.1300/J136v12n01_08

Russell-Rose, T., & Tate, T. (2013). *Designing the search experience: The information architecture of discovery*. Waltham, MA: Elsevier.

Satija, M. P. (2013). Briefs on the 19th (1979) to the 23rd edition (2011) of Dewey decimal classification. *DESIDOC Journal of Library & Information Technology, 33*(4), 277-288.

Savolainen, R. (2006). Information use as gap-bridging: The viewpoint of sense-making methodology. *Journal of the American Society for Information Science and Technology, 57*(8), 1116-1125. doi:10.1002/asi.20400

Schmidt, E. (2005). Books of revelation. *Wall Street Journal*, October 18. Retrieved from http://www.wsj.com/articles/SB112958982689471238

Schwing, T., McCutcheon, S., & Maurer, M. B. (2012). Uniqueness matters: Author-supplied keywords and LCSH in the library catalog. *Cataloging & Classification Quarterly, 50*(8), 903-928. doi:10.1080/01639374.2012.703164

Smiraglia, R. P., Lee, H., & Olson, H. A. (2011). Epistemic presumptions of authorship. Proceedings of the 2011 iConference, February 8-11, Seattle, Washington.

Strader, C. R. (2009). Author-assigned keywords versus Library of Congress subject headings: Implications for the cataloging of Electronic Theses and Dissertations. *Library Resources & Technical Services*, *53*(4), 243-250.

Stone, A. T. (2000). The LCSH century: A brief history of the Library of Congress subject headings, and introduction to the centennial essays. *Cataloging & Classification Quarterly, 29*(1-2), 1-15. doi:10.1300/j104v29n01_01

Stone, B. (2013). *The everything store: Jeff Bezos and the age of Amazon*. New York: Back Bay Books, Little Brown and Company.

Svenonius, E. (2000). LCSH: Semantics, syntax and specificity. *Cataloging & Classification Quarterly, 29*(1-2), 17-30. doi:10.1300/j104v29n01_02

Svenonius, E. (2009). *The intellectual foundation of information organization*. Cambridge, Mass: MIT Press.

Swanson, T. A., & Green, J. (2011). Why we are not Google: Lessons from a library web site usability study. *The Journal of Academic Librarianship, 37*(3), 222-229.

Taylor, A. G., & Joudrey, D. N. (2009). *The Organization of Information* (3rd ed.). Westport, Connecticut: Libraries Unlimited.

Tillotson, J. (1995). Is keyword searching the answer? *College & Research Libraries, 56*(3), 199-206.

Vaidhyanathan, Siva. (2009). The Googlization of universities. *The NEA 2009 Almanac of Higher Education*. The National Education Association, 65-74. Retrieved from http://www.nea.org/assets/img/PubAlmanac/ALM_09_06.pdf

Wade, N. (2006). *Before the dawn: Recovering the lost history of our ancestors*. New York: Penguin Books.

Walsh, J. (2011). The use of *Library of Congress subject headings* in digital collections. *Library Review, 60*(4), 328-343. doi:10.1108/00242531111127875

Wang, Y., & Mi, J. (2012). Searchability and discoverability of library resources: Federated search and beyond. *College & Undergraduate Libraries, 19*(2), 229-245. doi:10.1080/10691316.2012.698944

Wardhaugh, R., & Fuller, J. M. (2013). *An introduction to sociolinguistics* (7th ed.). West Sussex, UK: John Wiley & Sons, Inc.

Warren, D. (2007). Lost in translation: The reality of federated searching. *Australian Academic & Research Libraries*, *38*(4), 258-269.

Williams, R. V. (2010). Hans Peter Luhn and Herbert M. Ohlman: Their roles in the origins of keyword-in-context/permutation automatic indexing. *Journal of the American Society for Information Science and Technology, 61*(4), 835-849. doi:10.1002/asi.21265

Willis, C., & Losee, R., M. (2013). A random walk on an ontology: Using thesaurus structure for automatic indexing. *Journal of the American Society for Information Science & Technology, 64*(7), 1330-1344. doi:10.1002/asi.22853.

Woods, R. F. (2010). From federated search to the universal search solution. *The Serials Librarian, 58*(1), 141-148. doi:10.1080/03615261003622957

Yi, K., & Chan, L. M. (2009). Linking folksonomy to Library of Congress subject headings: An exploratory study. *Journal of Documentation, 65*(6), 872-900. doi:10.1108/00220410910998906

Yi, K., & Chan, L. M. (2010). Revisiting the syntactical and structural analysis of Library of Congress subject headings for the digital environment. *Journal of the American Society for Information Science & Technology, 64*(1), 677-687. doi:10.1002/asi.21295

Zhang, Y. (2008). Undergraduate students' mental models of the Web as an information retrieval system. *Journal of the American Society for Information Science & Technology, 59*(13), 2087-2098. doi:10.1002/asi.20915