8-2018

AN EVOLUTIONARY APPROACH TO BIBLIOGRAPHIC CLASSIFICATION

David Linn Sims
University of Tennessee

To the Graduate Council:

I am submitting herewith a dissertation written by David Linn Sims entitled "AN EVOLUTIONARY APPROACH TO BIBLIOGRAPHIC CLASSIFICATION." I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Communication and Information.

Suzanne L. Allard, Major Professor

We have read this dissertation and recommend its acceptance:

David G. Anderson, Bradley Wade Bishop, Stuart N. Brotman

Accepted for the Council:
Dixie L. Thompson

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

AN EVOLUTIONARY APPROACH

TO BIBLIOGRAPHIC CLASSIFICATION

A Dissertation Presented for the

Doctor of Philosophy

Degree

The University of Tennessee, Knoxville

David Linn Sims

August 2018

ACKNOWLEDGEMENTS

If it had not been for my dissertation chair, Suzie Allard, I may not have completed this degree. She had the understanding of those of us who pursue a PhD while working full time and having a family. She also had incredible patience as I would constantly under estimate the time it would take to do all that was needed to complete a dissertation.

But I am also grateful to so many others: my dissertation committee members David Anderson, Wade Bishop, and Stuart Brotman for their patience through comprehensive finals, and valuable advice in both the dissertation proposal and final dissertation; Robert Patton at Oak Ridge National Laboratory (ORNL) for advice related to text-mining and assistance with term extraction for the dissertation research project; Jim Malone, Priyanki Sinha, and Todd Suomela, my closest cohort colleagues at The University of Tennessee, Knoxville, for interactions that provided early encouragement for pursuing this degree; Jenni Sinclaire with John Wiley & Sons publishing company for help with source material for the dissertation research project; Mike Paulus, Jennifer Caldwell, and Cynthia Manley at ORNL for ongoing encouragement and support; Kay Hedly, my mother-in-law for her support in the final three months of the dissertation project; Lonnie Crosby, Tabitha Samuel, and Gary Rogers, The University of Tennessee staff for help with the acquisition of the source material for the dissertation project; and UT-Battelle, LLC (M&O contractor for the Department of Energy's Oak Ridge National Laboratory) for educational assistance.

And of course, I am very grateful for having a mother, father, and sister who always supported and encouraged my education goals and impressed upon me the importance of education, each in their own way.

But I am most grateful for the support, understanding, and tolerance of my spouse, Beka Hedly, and my children, Ava Grace and Sean. For nearly as long as I have been married, and for the entire lives of my young children, I have been working on this degree. Countless times, Beka has taken care of the children while I studied and worked towards this end goal. Countless times my children have asked me to play only to hear my standard response: “I need to do my school work today,” even though it was the weekend or a vacation day I was using. Most major holidays a laptop accompanied me on trips. This was not the life Beka had envisioned when we were married nor was I the husband and father she had envisioned. Needless to say, much time and attention is owed to my family—and this is one debt I am looking forward to repaying!

ABSTRACT

This dissertation is research in the domain of information science and specifically, the organization and representation of information. The research has implications for classification of scientific books, especially as dissemination of information becomes more rapid and science becomes more diverse due to increases in multi-, inter-, trans-disciplinary research, which focus on *phenomena*, in contrast to traditional library classification schemes based on *disciplines*.

The literature review indicates 1) human socio-cultural groups have many of the same properties as biological species, 2) output from human socio-cultural groups can be and has been the subject of evolutionary relationship analyses (i.e., phylogenetics), 3) library and information science theorists believe the most favorable and scientific classification for information packages is one based on common origin, but 4) library and information science classification researchers have not demonstrated a book classification based on evolutionary relationships of common origin.

The research project supports the assertion that a sensible book classification method can be developed using a contemporary biological classification approach based on common origin, which has not been applied to a collection of books until now. Using a sample from a collection of earth-science digitized books, the method developed includes a text-mining step to extract important terms, which were converted into a dataset for input into the second step—the phylogenetic analysis. Three classification trees were produced and are discussed. Parsimony analysis, in contrast to distance and likelihood analyses, produced a sensible book classification tree. Also included is a comparison with a classification tree based on a well-known contemporary library classification scheme (the Library of Congress Classification).

Final discussions connect this research with knowledge organization and information retrieval, information needs beyond science, and this type of research in context of a unified science of cultural evolution.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1 – INTRODUCTION & GENERAL INFORMATION

INTRODUCTION

This dissertation is about an evolutionary classification approach to a perceived information organization and representation need. Keeping with the evolutionary theme of the dissertation, in areas of the Background of the Subject section (and in more detail in chapter 2), I include some examples of early evidence of human behavior that eventually led to information organization and representation. This 'deep' background is not usually discussed in courses or textbooks related to organization and representation of information, which is one of the pillars of information science. This seems a noteworthy omission because the story of organization and representation of information by humans does not begin with the advent of libraries or even the written record. Furthermore, for a holistic view of a subject at the graduate level, one should at least be aware of research that expands understanding of a subject as well as extends understanding as far back in time as the scholarly evidence permits. Finally, this evolutionary approach also communicates the range of research to which I have been exposed during my study at The University of Tennessee's College of Communication and Information.

An organizing animal

Organizing things—tangibles or intangibles, consciously or unconsciously, or under formal or informal structures—is a common behavior of humans. We organize to make sense of our world, for efficiency of time, and for psychological comfort. Markets and retail organizations have categories of items to sell (e.g., vegetables, meat, clothing, automotive, etc.) placed in specific, identifiable areas. Educational institutions organize academic departments

and courses. In our homes, we organize rooms, books, clothing, dishes, tools, etc. News organizations categorize content according to subjects, for example geographic regions (e.g., world, nations, and local areas) or sociological categories (e.g., art, business, economics, entertainment, health, politics, sports, etc.). Science categorizes around characteristics and properties of a natural phenomena. Our art, music, and literature are often categorized according to formal or informal genres. Libraries, museums, and knowledge management departments within organizations categorize with pre-defined classification schemes or other knowledge organization schemes such as taxonomies or ontologies. These user communities, along with their patrons, will benefit most from this dissertation.

Whether referred broadly as *categorization* or more structured as *classification*, our organizing behavior is "central to any understanding of how we think and how we function, and therefore central to an understanding of what makes us human" (Lakoff, 1987, p. 6); it is "arguably one of the most central and generic of all our conceptual exercises… [and] the foundation not only for conceptualization, language, and speech, but also for mathematics, statistics, and data analysis in general" (Bailey, 1994, p. 1). Much like other aspects of our human socio-cultural environment, the end result of organization in general, and categorization in particular, can be seen in every aspect of our life, though we rarely ever give thought to the processes related to these activities.

BACKGROUND OF THE SUBJECT

Any classification problem is ultimately related to which category a particular 'thing' belongs. The human ability to *categorize* aspects of our environment would need to be a cognitive ability well before notions of formal classifications emerged. Indeed, categorization is

considered a core cognitive process in humans. It is not difficult to imagine that humans (as well as other animals) have been categorizing observable phenomena since the emergence of our species—categorizing predator and prey, beneficial and toxic plants, and friend and foe.

As background to the subject of classification, this section is subdivided into two parts for the dissertation's introductory foundation to the subject. First, because categorization is a fundamental cognitive trait in humans, a general understanding of categorization from a cognitive science perspective is presented. Second, due to this fundamental cognitive trait in humans, a summary of categorization's prehistoric evidence from archaeological science is presented (more information about each of these topics is presented in the literature review).

Cognitive science research in categorization

Cognitive science research in categorization emerged in the mid-1900s. "Cognitive science views categories and categorization as the main way that we make sense of experiences… As we move through our lives, we automatically categorize people, animals, and things into categories" (Abbas, 2010, p. 35).

Two categories of research help us understand this core cognitive function in humans. Behavioral studies have resulted in a large body of cognitive science research from both human and non-human animals regarding the ability to categorize. Evolutionarily, the ability to categorize has provided vertebrates with survival and reproduction advantages over their evolutionary history (Smith, Zakrzewski, Johnson, Valleau, & Church, 2016). Ethnological studies of extant traditional human societies also enable us to gain insights into human socio-cultural behavior before knowledge was stored and communicated in written form. Studies related to colors and kinship words have demonstrated the influence of culture on category formation

and ethnobiological studies of folk taxonomies suggest humans have the ability to describe a basic level for naming things, which includes a generic (genus) level of plants and animals (Taylor, 2004, pp. 299-300).

Cognitive science traditionally understood categories solely by the properties shared by members of a category (Abbas, 2010, p. 35) and held to the principle that "categories should be clearly defined, mutually exclusive and collectively exhaustive" (Larsson, 2016, p. 132). This was the prevailing approach for knowledge organization until the mid-1900s when Ludwig Wittgenstein suggested some things do not easily fit into categories. He used games to support his assertion that a common set of parameters that all games must share under a classical categorization view are difficult to define (Abbas, 2010, p. 35; Taylor, 2004, p. 298). For example: Is a game for mental development or physical development? Is a game for recreation or for a career? Is a game played by one person or by groups of people? Other concepts that undermined the classical view of categories were that 1) categories could contain subjects that do not have common properties (the example provided by Taylor of *ball*, *bat*, and *umpire* can belong to the category *baseball*, 2) some categories could be subjective (e.g., *tall, short*), which lead to *fuzzy set theory* to deal with *graded terms*, and 3) *basic level categories* with studies of children's first learning level of categories (Taylor, 2004, pp. 298-299).

In the 1970s, experimental work by Eleanor Rosch and colleagues demonstrated the utility of the above concepts, which eventually led to Rosch's development of *prototype theory*. "Prototype theory proposes that human categorization is based on both human experience (of perception, motor activity, and culture) and imagination (of metaphor, metonymy, and mental imagery)" (Abbas, 2010, p. 36). Rosch's basic reasoning was that categories have 'best'

examples (i.e., prototypes), which was in conflict with classical cognitive theory, which suggested that categories based on similar features would not have a 'best' representative of the category. Her research illustrated how humans can create 'best' representatives for categories based on their experiences (e.g., *robin* for a bird category over *ostrich*; a 5-foot person believes there to be more tall people in the world that does a 6-foot person; creation of *ad hoc* categories such as "things to take camping" when needed), which create challenges for humans who classify library materials (Taylor, 2004, p. 300).

Prehistory of information organization and representation

Symbolic communication—the use of symbols represented in ways that transmit meaning to a group—is a fundamental component for information organization. Evidence of symbolic communication is considered a modern human trait, and personal ornaments (e.g., shell beads, modified teeth and bones) are considered early evidence of human use of symbolic communication. Some of the oldest human personal ornaments date back to about 75,000 years ago (Henshilwood, d'Errico, Vanhaeren, Van Niekerk, & Jacobs, 2004).

The earliest symbolic communication innovations directly related to information organization and representation include symbolic memory (d'Errico, 1998) and storage (Donald, 1991) systems. Though stone etchings dating back over 70,000 years ago have been discovered (Henshilwood et al., 2002), which may one day be considered early information storage, there is little scientific evidence for these memory/storage systems before 50,000 years ago. From ethnographic research, d'Errico (1998) suggests this lack of evidence may be due to using organic materials (e.g., wood, skin, vegetal matter) for early information systems. Cave art from

Spain dating over 40,000 years ago (Pike et al., 2012) is the most concrete evidence of early information organization and representation.

The cultural evolutionary innovation that would enable the most complex symbolic storage and memory systems was the emergence of writing systems. The first proto-writing systems emerged in Mesopotamia and Egypt over 5,000 years ago and in Mesoamerica sometime between 1140-400 BC. These early systems used pictograms (representations of physical objects) and ideograms (representations of concepts) but were not capturing actual speech such as phrases or sentences and not representing information in any grammatical order (Trigger, 2004). True writing systems—recording both phonetic information (word sounds) and semantic information (word meanings)—seem to have emerged independently in Sumer and Egypt (around 2500 BC), China (1200 BC), and Mesoamerica (around 250 AD).

From an evolutionary perspective, a connection can be made between the background of the subject and background of the problem. Archaeological researchers note that technical complexity increases over time for both memory devices and cave art. This increase in complexity is also seen in the evolution of writing systems—from use of simple pictures of objects and concepts to more complexity as systems emerged that included word sounds and meanings. This 'simple-to-complex' characteristic of evolution is also observed by biological, socio-cultural, and library/information science theorists over the last 100+ years. Support of this assertion is presented in detail in the literature review of chapter 2.

BACKGROUND OF THE PROBLEM

This dissertation explores an information organization problem and one possible solution utilizes what could be termed a natural approach to book classification. Prior to formal

science, humans needed to catalog, organize, and classify information objects and natural objects that were being collected. “Natural classification… is what botanists, zoologists, and geologists did before they had explanations of phenomena. It is what psychology did before it had etiologies of pathological conditions. It is an underappreciated aspect of understanding how science develops” (Wilkins & Ebach, 2014, p. 2). “In the discipline of natural history, researchers systematically study natural objects (animals, plants, minerals)--naming, describing, classifying, and uncovering their overall order. They do this because such work is an essential first step before other, more complex analysis can be undertaken.” (Farber, 2000, p. 2). Thus, *uncovering the ‘natural order’ is a key component of a ‘natural’ classification*.

Though most people might not consider classification of books as being a classification of natural objects in the way animals, plants, and rocks are considered natural objects, books are culturally generated objects and culture has become the natural environment of humans. This latter statement is relatively easy to comprehend if we consider that it is now possible for many humans to live an entire existence and never step foot on the dirt of the Earth. In other words, an entire existence can now be lived within human-created physical and social structures. Consider that in 2008, 54% of the world’s population lived in urban areas rather than rural areas, and by 2050, that percentage is estimated to be 66% (United Nations Department of Economic and Social Affairs, 2017).

A book is a recording, or data storage, of human thought at a point in time. There are research approaches (e.g., hermeneutics) for interpreting texts that incorporate historical, social, and cultural criteria into the analysis process. This suggests the content of books—whether fiction or non-fiction—is influenced by culture, and as such, book content changes

over time as culture changes over time. The classification connected to the most scientifically-based theory for change over time in natural phenomena—specifically, changes in living systems over time—is the classification for biological species. Therefore, a biological approach to classifications of cultural phenomena is one potential solution for a book classification problem.

The literature review provides substantial evidence of scholarly connections between biological and cultural phenomena. These connections have been made numerous times for well-over 100 years from scholars within both socio-cultural and biological disciplines. Though the literature review provides copious detail, below are the main points from this section of the review:

From socio-cultural science contributions, the assertions are:

- Culture is an offshoot of biological life and is composed of living systems just as biological ecosystems are composed of living systems. Socio-cultural phenomena are as real as biological organisms and can be studied with biological methods.
- Natural selection evolutionary theory is the only scientific theory that explains non-random change over time in living things.
- Culture possesses key attributes of natural selection theory: heritable variation and competitive selection. And because culture possesses the equivalent of organisms, biological evolutionary theory is applicable to cultural change.
- Cultural forms change over time by becoming more complex and emerging into new cultural forms, but still retain elements of smaller, less complex cultural forms.

- Packages of information are being transmitted—whether as biological genes or as cultural memes (and by their nature, communication and information sciences are positioned for research pertaining to packaging and transmission of cultural information).

From biological contributions, the assertions are:

- Principles of natural selection theory can be applied to cultural phenomena.
- Similarities/parallels between cultural phenomena and their biological analogs have been made (albeit somewhat inaccurately in my opinion). In particular, the identification and transmission (reproduction and inheritance) of the cultural analog of the biological gene has been made.
- Phenotypes[1] can extend beyond a physical body.

From my contributions, the assertions are:

- James Miller's living systems theory can be used to justify any human collective as the cultural analog to a biological organism.
- The process of cultural organism replication (i.e., inheritance) can be illustrated with a relatively simple mathematical model.
- Richard Dawkins' extended phenotype concept can be employed to justify the use of output (books) from a cultural organism (a book publishing company) for a biological evolutionary-based analysis.

[1] The phenotype is the observable anatomical, physiological, and behavioral characteristics of an organism's genotype.

The literature review also provides, in detail, solid evidence for the connection to evolutionary-based library classifications, which also extends back over 100 years. In summary, the information organization theorists uphold the importance of a *common origin* principle for information organization, but none view books as phenotypic expressions of living systems. In other words, none are looking at books in the way an evolutionary biologist would look at books. Though information organization theorists have developed evolutionary classification principles, there is no evidence from library and information science literature (or from several other disciplines) of the use of analytical techniques employed in biology to uncover evolutionary relatedness of books or even a test of an evolutionary book classification using the contemporary classification tools of evolutionary biology.

If such a classification were possible, it would seem to be a more scientific approach to book classification (based on comments by information organization theorists) than what has been used in the past—it would be a classification built on the overarching theory of natural selection's decent with modification and the more specific philosophical assumptions of phylogenetics. If common origin is truly an important principle for information organization, then it seems a methodology that will support that principle should be tested.

Cladistic taxonomy – an evolutionary classification method

Introduced in the mid-1900s by German biologist Willi Hennig[2] was the cladistic method of inferring evolutionary relationships among biological species. "The development of Hennig's cladistic method..., i.e. a purely phylogenetic approach which defines classes by the possession of the same evolutive characteristics of living beings and gives up all other criteria, changed

[2] Willi Hennig is considered the founder of phylogenetic systematics.

drastically biologists' view of taxonomy" (Parrochia & Neuville, 2013, p. 9). Currently most evolutionary biologists believe cladistics is the best method for inferring the relationships of organisms, and subsequently, for classifying them.

Though more detail is presented in chapters 3 and 4, in short, the research project follows the basic steps of the cladistic process using the PAUP* phylogenetic software (Swofford, 2002):

1. Select sample from a population of taxa (the things to be classified);
2. Obtain characters from the taxa (the traits of the taxa to be analyzed; the things that change);
3. Create a character list and code it;
4. Create a character matrix to include each taxon and its coding for the characters;
5. Insert data from the matrix into phylogenetic algorithms;
6. Analyze the results—i.e., produce the classification.

Applications of phylogenetics to cultural phenomena. There has been growth over the last three decades in the use of biological evolutionary analytical methods applied to cultural phenomena. One methodology has been termed a phylogenetic approach to understanding cultural diversity (a clear reference is Mace, Holden, & Shennan, 2005). In biology, phylogenetics is the study of the evolutionary relationships of organisms. Phylogenetic analysis is conducted to infer the relationships of organisms based on changes in characters (i.e., genetic features in populations of organisms that change over periods of time). The cladistics mentioned above is an example of a phylogenetic analytical method. The representation of a phylogenetic analysis of organisms is usually represented in a branching (or tree) structure

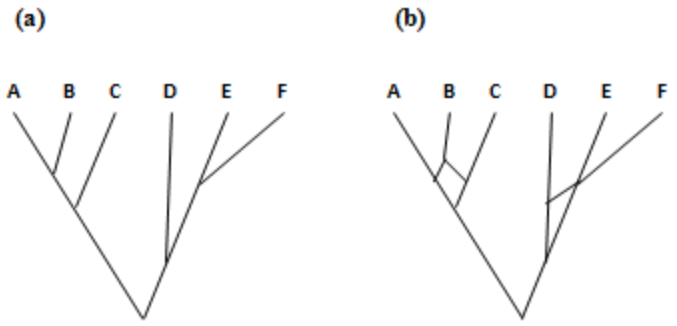


Figure 1. Example of Phylogenetic Tree and Network.
Tree (a) and network (b). Letters represent different taxa such as biological species.

illustrating the evolution of species over time due to changes in genetic frequencies resulting from genes being transferred from parent to child (i.e., vertical gene transfer). A network structure is used, for example, when hybridization or horizontal gene transfer has occurred. Figure 1 provides examples of each structure. And in phylogenetic theory, the tree (or network) is considered the classification.

Contemporary biological phylogenetic analysis conducted on extant (existing) organisms uses genetic information as the characters for analysis. Currently, there does not seem to be either a cultural equivalent to a genome or even much research activity in this space, though there is some evidence of the concept on the Web:

- The *Human Memome Project* (The Center for Human Emergence, n.d.) has been on the web since at least December 2004, but does not seem to have made much progress (comparison made using the internet archive Wayback Machine).
- An idea for a *Human Cognome project* (Horn, 2002) was presented at a National Science Foundation Converging Technologies workshop in December 2001, but no evidence that

this has moved forward (comparison made using the internet archive Wayback Machine*).*

- *Mapping China's Cultural Genome* (Woodrow Wilson International Center for Scholars, n.d.) is a source for United States policy developers and 'mapping' and 'genome' are only metaphors in light of the type of research suggested in this dissertation.

But prior to genetic analysis, biological taxonomists used *phenotypic* morphological characters for evolutionary classifications, and paleontologists studying fossil remains will continue to rely on fossil phenotypes due to the absence of genetic material in fossils (Lee & Palci, 2015). Therefore, even though there is no cultural equivalents of a genotype, phenotypic expressions of cultural 'organisms' (i.e., observable output such as behaviors, materials, technologies, etc. from groups of people) can justifiably be used for phylogenetic analysis, especially those expressions critical for the cultural organisms' survival and reproduction, such as books for authors and book publishing organizations.

As the literature review (chapter 2) will confirm, it is sufficient here to say there is growing evidence of the use of biological concepts and tools such as cladistics to support the notion that cultural phenomena can be explored with approaches and even analytical techniques similar to those of the biological sciences. Therefore, an information science classification need might also benefit from theories and methodologies borrowed from biological science classification.

Theory and philosophical principles

Three theoretical/philosophical concepts underlie the proposed research project: evolutionary classification, natural selection theory, and phylogenetic philosophical principles.

Evolutionary classification. The library and information science theorists from the literature review claimed 1) an evolutionary ordering of books would be the most natural, objective, and/or scientifically-based type of book classification and 2) a common origin would be the most important principle to follow in an evolutionary classification (but it also must make sense to the users of the classification).

The classification approach used by most contemporary biological taxonomists is phylogenetics and the most used method is cladistics, which is grounded in Darwinian evolutionary theory. Classifications "based solely on the criterion of resemblance or similarity between specimens..." (Rivero, 2016, p. 52) is known as *phenetics*. This is what biological taxonomists used prior to genetic analysis, and is still used in paleontology taxonomy where fossils are the only evidence of past life. But "phenetics lacks a theoretical model and thus constitutes an empty taxonomic method that struggles to tell a story beyond the classification models themselves. In cladistics, in contrast, Darwinian evolution provides a theoretical framework for the Linnaean taxonomic system, where system and theory become interconnected" (Rivero, 2016, p. 53).

Natural selection theory. Because phylogenetics (and the cladistic method in particular) is grounded in Darwinian evolutionary theory, the use of phylogenetics and cladistics will support the fundamental principles of evolution by natural selection:

> Evolution occurs whenever the following conditions exist:
>
> 1) variation: there is a continuing abundance of different elements;
>
> 2) heredity or replication: the elements have the capacity to create copies or replicas of themselves;

> 3) differential 'fitness': the number of copies of an element that are created in a given time varies, depending on interactions between the features of that element and features of the environment in which it persists (Dennett, 1995, p. 343).

Thus, populations of entities that are evolving by natural selection must exhibit *variation*, *differential reproduction*, and *inheritance*.

Phylogenetic philosophical principles. Cladistics is a method to create phylogenetic classifications and will therefore support the underlying phylogenetic philosophical assumptions: (derived from Wiley & Lieberman, 2011, p.3).

- All species are linked genealogically (i.e., ancestor-descendant) and the genealogical relationships are discoverable and reconstructable.
- All characters can be useful in relationship discovery but some may be more useful than others depending on the analysis being conducted.
- Trees created by phylogenetic analysis are inferences of the genealogical relationships.
- Classifications are based on the trees produced.

Prediction. An evolutionary tree can be used to predict properties of organisms (e.g., if two species are closely related, then properties of one will likely be found in the other). Evolutionary trees can also predict 'missing links.' A very simple example that occurred in preparing this dissertation is visualized in Figure 2. I was thinking of the difficulty of placing a book about computational chemistry into a single slot within a traditional library classification scheme. I knew computational chemistry existed because a friend of mine obtained a PhD in the discipline. Visualizing chemistry at the bottom, I assumed there would need to be something akin to mathematical chemistry before computational chemistry, which could represent a transition between the other two disciplines. To test the assumption, I did a quick

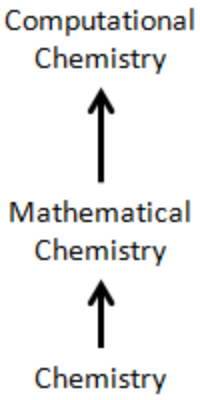


Figure 2. Simple Tree Illustrating "Mathematical Chemistry" Transition Subject.

Google search, and—maybe obviously to some people—quickly found evidence for mathematical chemistry's existence.

Books and phylogenetic classification

Can books be the units of analysis for a phylogenetic classification? The literature review chapter will provide the connection between 1) the book publishing organization as a living system and 2) the book as a phenotypic extension of that particular type of cultural 'organism'. But for the background in this introduction chapter, the short answer is 'yes.' I suggest that books are the most important outputs of the book publishing industry, for without sales of books, a book publisher cannot survive and reproduce (a book author could survive issuing free downloads if s/he has other resources for survival).

Can books be classified using phylogenetic methodology? Books can be viewed in various ways. High-level or 'coarse' views include the subject of the book and the general description of the book (usually provided by the publisher). A mid-level view would be a book review. Getting more granular, we find the book's metadata. And the most granular views are

found in research related to computer information systems (such as recommender systems) for search and retrieval that focus on words and phrases and their locations within the book (title, abstract, body of text, index). A phylogenetic view of a book would be in this latter space. And when viewed from this computer science perspective, a phylogenetic classification research project seems very reasonable.

In closing this section, from the literature review, Trigger (2004) suggested written recording systems could be a source of data for cultural evolution analysis due their development being "historically better documented than is that of many forms of material culture" (p. 39). Thus if early written recordings can provide data for evolutionary analysis of cultural phenomena, then *all* written recordings—including contemporary books—should have, at the least, some capability for evolutionary analysis.

Current classification issue: The increase in amount, diversity, and complexity of science and technical information

With the emergence of the Internet and World Wide Web and the increased computing power and digital storage capacity, most of us are very familiar with concepts of 'information explosion' and 'big data'. But scholarly publications have been increasing for quite some time. The number of peer reviewed, English-language journals have been increasing at ~3.5% average annual growth rate since about the year 1700 (Ware & Mabe, 2015). Using data provided by Tenopir and King (2009), the average annual growth rate of US-based journals was ~2.7% from 1965-1995 (p. 171, Table 6.2). Research reported from Ware and Mabe (2015) and Bornmann and Mutz (2015) based on the actual number of scientific article publications per year suggests a ~3% average annual growth rate. The number of articles from US-based journals increased

annually by ~4.5% average growth rate from 1965-1995 (Tenipor & King, 2009, p. 171, Table 6.2).

There have also been larger growth rates reported for scientific literature. For example, Bornmann and Mutz (2015) provided evidence of an 8-9% average annual growth rate in scientific article publications over roughly 70+ years prior to 2010. The researchers used Web of Science ("WoS"; Clarivate Analytics) and conducted citation research starting with the middle of the 17th century—which was a period of "development of the institutionalized structures of modern science with publication of the results of scientific work in journals and manuscripts undergoing critical peer review before publication" (Bornmann & Mutz, 2015, p. 2217; original Bornmann, 2011). This resulted in three distinct periods of growth: "from less than 1% up to the middle of the 18th century, to 2 to 3% up to the period between the two world wars and 8 to 9% to 2010" (Bornmann & Mutz, 2015, p. 2220). However, the data for these three phases is based primarily on citation analysis rather than the actual number of scientific article publications per year (which the authors discuss).

Bornmann and Mutz (2015) remind the reader "there is currently no literature database containing every publication since the beginning of modern science to today that can be used for statistical analysis" to track growth rates (p. 2220). But an 8-9% average annual growth rate (rather than the 3-4% from the rates in the first paragraph of this subsection) does fit a more intuitive growth rate in the 'information explosion' and 'big data' age. Additionally, by 2014, China had become the world's "third largest producer of research articles" (Morrison, 2014) and by 2016, "China's share of global science and engineering publications… [had] pulled within a percentage point of those from the United States" (Witze, 2016). This does not include India,

but it is safe to say that these countries will continue to increase in science and technical publications.

The global growth in science and technical research suggests that the specialization of scientific and technical disciplines and fields will continue to increase, leading to more rapid diversification of scientific literature. There is also a growing trend in the use of knowledge from different disciplines for scientific purposes, which can result in increases in both diversity and complexity[3] of scientific literature. One indication of this increase in complexity and diversity can be found with a simple WoS search of *multidisciplinary*, *interdisciplinary*, and *transdisciplinary* ("M-I-T-disciplinary") publications. Searching (*multidisciplin** OR *interdisciplin** OR *transdisciplin**) OR (*multi-disciplin** OR *inter-disciplin** OR *trans-disciplin**) generated 140,431 results. The amount per decade starting with the 1950s can be seen in Figure 3.

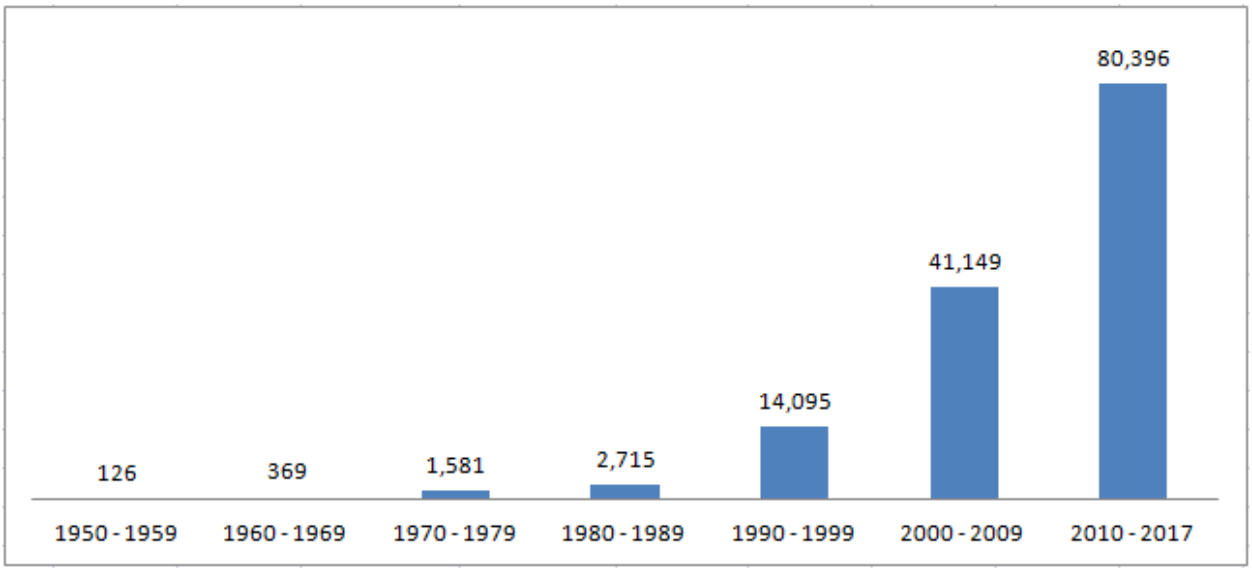


Figure 3. Web of Science Results for M-I-T-disciplinary Publications (2017 results through August 2017).

[3] Here, *complexity* is understood to be knowledge utilized from more than one discipline; this is analogous to a system's growth in complexity resulting from the integration of more parts.

STATEMENT OF THE PROBLEM

The global growth in science and technical research suggests that the specialization of scientific and technical disciplines and fields will continue to increase, leading to more complexity and diversity of information. A possible indicator of this is observed by the increases of the use of multi-, inter-, trans-disciplinary (M-I-T-disciplinary) words in a popular science and technical literature database, which suggests these types of sciences are increasing at even faster rates than the overall science and technical research publications.

M-I-T-disciplinary sciences increasingly address phenomena. Traditional library classifications schemes are based on disciplines rather than phenomena and are not well suited for M-I-T-disciplinary sciences.

The increase in the volume, diversity, and complexity of information packages in general, and in science and technical information specifically, will challenge traditional library classification schemes in the coming years as well as scholars attempting to find relevant literature.

PURPOSE OF THE STUDY

Primary purpose: To demonstrate a new evolutionary approach and methodology for a book classification that is designed for more complex and diverse information packages. In its most simple definition, evolution is change over time, so an evolution-based classification system is, by design, one that can incorporate taxa that change over time. The research project is a proof-of-concept test for an evolutionary book classification. If successful, the general methodology could be used in a much larger classification effort.

Secondary purpose: To continue expansion of cultural evolution classification. This project uses a phylogenetic method applied to cultural artifacts (books), extending cultural evolution research in general, and cultural phylogenetics research in particular.

RESEARCH QUESTIONS

Can a phylogenetic-based method be developed to produce a sensible classification of books with only words and/or phrases used as the characters for analysis?

If so:

- Does the classification have natural hierarchical groupings of books?
- Is this representation of books useful?
- Is there any insight regarding the emergence of a new discipline or field similar to the emergence of a new biological species?

If not:

- What methodological modifications should be considered to produce an evolutionary classification?

SIGNIFICANCE OF THE STUDY

As the literature review will suggest, there has been no research project like the one proposed for this dissertation, even though library and information science theorists' writings support such a research project. The project has the potential to improve library classification by demonstrating an approach to classification built on evolutionary relatedness of books in a collection. This is needed due to the presumed continued increase in more diverse scientific output from more M-I-T-disciplinary projects along with the continued increase in the amount

of information being produced. Without a more flexible classification system to better manage this growing diversity, the traditional library classification systems could become more of a liability to a library rather than an asset.

This research could also lead to a more intuitive approach to classification (i.e., nature-inspired), which could improve efficiency of information access and retrieval and even aid in learning. For example, an evolutionary classification of books may enable more efficient literature reviews and quicker understanding of history when a user can easily see the evolutionary history of a body of research. Such a classification may also enable the identification of both gaps in the literature and lesser-known works overlooked by scholars due to the common practice of citing well-known works. Therefore, the user communities—i.e., libraries in particular, but also museums and knowledge management departments within organizations, along with their patrons—would have the most interest in this research.

Printed books are store houses for human knowledge and span many centuries. Even fictional works can tell us something about the zeitgeist of various eras in human history. Deriving aboutness (i.e., the subject) of a book can be difficult and time consuming. This project could provide more support for the use of text as data sources for cultural evolution research.

The research project method is a workflow design utilizing a term-weighting technique from a text-mining tool, followed by the use of a phylogenetic analytical tool to create a tree-based classification. The workflow could potentially be automated and used with any text-based corpus. It may also be possible to incorporate more tools into the workflow for expanding into other text-based sources where evolution would be of interest (e.g., journals, fiction, etc.).

DEFINITIONS

***Automatic classification*:** a type of supervised machine learning that either assigns an information package to an existing classification scheme or creates a scheme, then makes the assignment.

Automatic indexing: a computer's use of a controlled vocabulary to index large amounts of documents.

Classification: a systematic way of grouping of things into categories based on some feature or features the things have in common.

Communication: the transmission of information between entities.

Character: a feature in a population of entities that is used as data in an analysis of evolutionary relatedness of a group of entities containing similar features.

Cladistics: a method to reconstruct evolutionary histories of taxa based on shared, derived characters from a common ancestor and assumes a branching (tree-like) pattern to the evolutionary history, as opposed to a network pattern.

***Clustering*:** a type of unsupervised machine learning whereby the computer algorithms create distance measures among the information packages and groups (i.e., clusters).

Culture: human beliefs, customs, technologies, etc. that are socially transmitted by imitation, teaching, or language.

Cultural evolution: change over time in tangible and intangible attributes of culture.

***Dendrogram*:** a branching, tree-like hierarchical classification used to explore phylogenies.

Entity: a tangible or intangible thing with an independent existence.

Evolution: change over time in the heritable characteristics of related entities.

Genotype: the complete gene set of an organism.

Heritable: capable of being transmitted, acquired, and utilized by others.

Infer**:** a conclusion obtained from evidence and reasoning.

Information: a symbol or a group of symbols that contain meaning.

Interdiscipline: connections between different disciplines to create a new disciplinary whole with modification of traditional boundaries (e.g., a unified science of cultural evolution).

Machine learning**:** the use of computer algorithms that enable a computer "to automatically learn and improve from experience without being explicitly programmed" (Jmila, Khedher, & El Yacoubi, 2017, p. 884).

Methodology**:** a collective of methods used in a project, discipline, or field.

Multidiscipline: use of different disciplines but usage remains within traditional boundaries of each discipline (e.g., use of several sciences to solve an environmental problem).

Phenotype: the observable anatomical, physiological, and behavioral characteristics of an organism's genotype.

Phylogeny: the evolutionary histories of a group of related entities.

Phylogenetics: the study of the evolutionary relationships and histories of entities.

Related: sharing a common characteristic or set of characteristics.

Supervised machine learning**:** use of external information (e.g., human feedback or labeled text) to guide machine learning algorithms.

Symbol: something that represents something else.

Symbolic communication: the use of symbols or groups of symbols organized and represent in ways that transmit an intended meaning within a socio-cultural group.

Synapomorphies: characteristics shared among a group of entities that are derived from a common ancestor.

Transdiscipline: integration of different disciplines that transcends each discipline's traditional boundaries to create new approaches or systems (e.g., holistic healthcare)

***Unsupervised machine learning*:** does not use external information (e.g., human feedback or labeled text) to guide machine learning algorithms.

ASSUMPTIONS, LIMITATIONS, DELIMITATIONS

Assumptions

- Something does not come from nothing; there is always a predecessor or ancestor that something comes from. When humans invent, we build on things or concepts that already exist. Thus, there is always some common origin to cultural creations.
- An evolutionary ordering of books would be a natural, objective, and scientifically-based type of book classification.
- Because phylogenetic methodology is grounded in Darwinian evolutionary theory, the use of the methodology will support the fundamental principles of evolution by natural selection (see 'Theory and philosophical principles' subsection above).
- Because cladistics is a method used to create phylogenetic classifications, the use of the method may support the underlying phylogenetic philosophical assumptions (see *Supported theory and philosophical principles* section above).
- Natural (cultural) selection is operating on books. From an author's perspective, a publisher selects books to be published. From a publisher's perspective, sales, citations, and downloads are the evidence.

- Books are collections of characters (words/phrases) and are the phenotypic (visible) expressions of the 'genotypes' (thoughts, processes) of the living systems that produce the books (authors and/or publishers).
- If enough characters (words/phrases) are selected, transmitted, and amassed in a population, a new species (discipline) may emerge.

Limitations

- Though this classification method could eventually be used, in whole or in part, as a classification methodology for all books, the research project is a proof-of-concept project limited to a small set of earth science books from Wiley publishing company's Online Books. The reason for this limitation is:
    - Science and technical disciplines and fields use very specific words and phrases that are often contained within disciplines/fields, which should increase the probability of success for the research project;
    - The books are presented in a consistent, structured, digital form, which is conducive to computational tools for data extraction;
    - And though The University of Tennessee's library had a subscription to the Online Books used to obtain the sample for the project's data source, staff from the publisher could provide assistance with content delivery to minimize acquisition time.
- Because this is a proof-of-concept study, there is no random sampling from an entire population (though a random sample was drawn from the books provided by the

publisher). Therefore, no inferences of a statistical nature can be drawn from this project's results and applied to a larger population of books.

- The literature review regarding automatic classification is primarily from information and library science literature rather than a review of computer science literature.
- The data was limited to the intellectual content of the book chapters that has been digitized by optical character recognition (OCR) technology, and therefore recognizable by automatic text-mining tools. The front matter, index, references, etc. were removed during a data cleaning process. Table data, formulas, and algorithms, are often OCR'd, but these were either 1) easily identifiable and removed from the term set or 2) were lower-weighted terms within the term set and did not contribute substantially to the overall outcome of the evolutionary analysis. Text on page headers and footers were not removed and possible issues are discussed in chapter 4 under the *Discussion of the initial terms* subsection.
- Images such as maps, graphs of experimental results, photos of geomorphic features, etc., (and any text included inside images) were excluded from phylogenetic analysis due to limits of OCR technology and the research objective of only using words and phrases extracted with a text-mining tool. Of course images may contain much information not found in the text, but in science and technical books, there are typically discussions and captions related to images that are available for phylogenetic analysis. In other words, the important information conveyed by an image would have likely been discussed in the text by the author(s).

- There may be biological phylogenetic, computer science, and/or mathematic principles that were violated of which I have no knowledge. For example, the project uses a method (cladistics) that assumes vertical transmission of the characters in the taxa, but written text in books is usually considered the result of non-vertical transmission. This particular limitation example is covered in detail in chapter 2 under the section titled *Non-vertical transmission criticism (Galton's Problem)*.

Delimitations

- This was not a phylogenetic test per se, but rather a book classification test using a phylogenetic approach. A proof-of-concept test was conducted to determine if books can be sensibly classified using a phylogenetic method.
- This was not a text-mining project, though it utilizes a text-mining technology to create the book datasets for the possible classification.
- The outcome of this project was not intended to be a new information retrieval method (it was intended to be new knowledge organization method), it may result in one.

CONCLUSION

The research project is a classification of books grounded in contemporary evolutionary theory, which in turn could eventually lead to a classification more conducive for expected increases in diverse information and more intuitive for search and retrieval of library catalogs. There is literature evidence that suggests a phylogenetic classification related to knowledge organization is desirable, but there is no evidence that an actual method has been developed and tested on a set of books.

Books are the storehouses of the human experience—ideas, concepts, processes, dreams, beliefs, stories, discoveries, inventions, etc.—aspects of culture that are often difficult to analyze the further back in time we go. With modern computational, digitization, and text-mining technologies, we have the technical ability to conduct large-scale analysis of text-based works. Using methodologies to analyze books for their evolutionary relatedness, as presented in this dissertation, continues the path of using books (and other information records) as a useful means of organizing human scholarship, past and present.

CHAPTER 2 – LITERATURE REVIEW

INTRODUCTION

This dissertation's research project is a test of an evolutionary book classification. Phylogenetics is the study of the evolutionary relationships among organisms and the evolutionary trees produced by cladistic analysis is the method preferred by most biological taxonomists to classify organisms. I was broadly familiar with some of the literature from previous course work, primarily phylogenetic studies related to cultural phenomena. As I suggested during my initial PhD orientation, I wanted to study culture through a biological lens, due to having an intuition that culture was somehow alive. During the early coursework, I began focusing on locating scholarly material that supported my emerging belief that aspects of culture could be classified in ways similar to biological species. This fit nicely with my information science interests, due to organization and representation of information being one of the two pillars of information science, and classification being within that pillar.

But it was not until this dissertation that the theoretical connections were clearly made to support this intuition: that human populations are composed of living systems (groups, organizations, communities, states, etc.); that every living system will have the cultural equivalent of a genotype (an organism's entire set of genes) and a phenotype (the expression of the genes); that some parts of the phenotype are more important to survival and reproduction; that specific phenotypic traits can be used to create an evolutionary classification by methodologies used in biology; and that library and information science scholars suggested

such a classification would be the most objective and scientific, with the literature from these scholars ranging from the late 1800s through 2017.

CONCEPTUAL FRAMEWORK

In keeping with the framework of this research, developing an evolutionary book classification, I have used somewhat of an evolutionary conceptual framework for this literature review. *Evolutionary* used in this conceptual framework context is a generic use of the word to be understood as change over time and with the use of *predecessors* in order to tell an evolutionary story. This is in contrast to *biological* evolution, which is understood as descent with modification from a common ancestor, a concept which will also be used in this dissertation with the classification project. In general, the flow of this literature review is:

- Fundamental human cognitive trait: categorization
- Emergence of human information organization/representation
- Emergence of writing systems
- Emergence of academic discipline of classification
- Scholarly evidence connecting biological and cultural phenomena
- Scholarly evidence connecting information organization to evolutionary classification
- Applications of phylogenetics to cultural phenomena
- Non-vertical transmission criticism (Galton's Problem)
- Automatic classification in library and information sciences
- A note about classification trees
- Detailed search descriptions

The first section begins with a review the most fundamental human cognitive trait related to classification, which is the ability to categorize things in our environment. This is followed by reviews of early human organization and representation of information with archaeo-cultural evidence that begins with earliest evidence of possible symbolic communication (body ornaments), emergence of extrasomatic information storage (rock etchings) and information representation (cave art), emergence of the most complex symbolic storage (writing systems). This section ends with the emergence of the academic discipline of classification.

Next is the section explores two foundational topics important to this dissertation's project:

- *Scholarly support for a connection between biological phenomena and cultural phenomena;*
- *Scholarly support for a connection between book classification and biological classification.*

The first topic is answered with a review of literature from socio-cultural scholars followed by a review of literature followed by biological scholars. The second topic is answered with a review of literature from information organization scholars, though a small section provides follow-up support from computer science. The chapter continues with a review of literature related to the phylogenetic approach to cultural phenomena, which is the methodology proposed for this dissertation's project. The chapter ends with a review of the literature related to automatic classification in library and information sciences due to the

possibility of an automated classification process emerging from the method developed for this project.

Search descriptions are included in some of the sections below as needed and a more extensive presentation of search descriptions for two sections is presented in the detailed search descriptions section at the end of this chapter.

REVIEW OF RESEARCH

Fundamental human cognitive trait: categorization and its deep roots

Categorization is a core cognitive process in humans and is the starting point for this dissertation's evolutionary perspective of classification. Cognitive and linguistics researcher George Lakoff (1987) asserts that "any time we either produce or understand any utterance of any reasonable length, we are employing dozens if not hundreds of categories: categories of speech sounds, or words, of phrases and clauses, as well as conceptual categories. Without the ability to categorize, we could not function at all, either in the physical world or in our social and intellectual lives" (p. 6).

Evolutionary cognitive science suggests that the human brain has cognitive components that developed deep in evolutionary time. The ability to categorize "has conferred fitness advantages on vertebrates for hundreds of millions of years" and a large body of cognitive science research exists from both human and non-human animals regarding the ability to categorize (Smith et al., 2016). Insights into the past can be obtained by studying phenomena in the present.[4] For example, categorization research from our closest living relatives has provided

[4] This concept is known as the *principle of uniformitarianism* in geology and summarized by the phrase 'the present is the key to the past.' In short, the concept is understood as the natural forces at work today in the earth are the

evidence of the importance of prefrontal cortex in primates' abilities to form mental categories (Freedman, Riesenhuber, Poggio, & Miller, 2002) and researchers have observed chimpanzees' abilities to categorize fruits into distinct species in the wild (Janmaat, Ban, & Boesch, 2013). In humans, research suggests our ability to categorize begins at a very early stage of human development. Perszyk and Waxman (2016) provided evidence for a connection between a human infant's ability to form object categories and human (and non-human primate) vocalization sounds.

Ethnological studies of extant traditional human societies enable us to gain insights into human socio-cultural behavior thousands of years before knowledge was stored and communicated in written form. For example, Floyd Lounsbury illustrated how categories differ between cultures with examples from Native Americans' use of kinship words to refer to relatives of the mother and different words for relatives of the father. Similarly, Brent Berlin and Paul Kay's work with color categorization also showed the influences of culture. Ranges included two to eleven primary color categories, though some cultures that distinguish eleven colors may use the same word for more than one color. Later, Paul Kay and Chad McDaniel demonstrated that all cultures with less than eleven primary colors categorized green, blue, and black as 'cold' colors and red, orange, yellow, and white as 'warm' colors (Taylor, 2004, p. 299).

More specific to classification, the subject of this dissertation, from research in ethnobiological classification (more specifically, folk taxonomy or folk classification), Brent Berlin and colleagues discovered "that a single level of classification, the genus level, was the

---

same forces active in the past, though not necessarily at the same rates or intensities. This principle enables geologists to use today's Earth processes to explain past environments captured in rock structures. This general concept, even if not explicitly stated, has also been applied to understand past human behavior—e.g., by studying both present-day humans and non-human primates.

psychologically basic level at which [the study group] named the plants and animals in their region. This level, called the '*folk-generic level*' by Berlin, was in the 'middle' of the folk classification hierarchy…" (Abbas, 2010, pp. 37-38), which supported the psychologically basic level mentioned above. Berlin's folk classification hierarchy included:

- Unique Beginner (plant, animal)
- Life Form (tree, bush, bird)
- Intermediate (leaf-bearing tree, needle-bearing tree)
- Genus (oak, maple)
- Species (sugar maple, white oak maple)
- Variety (cutleaf staghorn sumac)

(Abbas, 2010, p. 38).

And not only did the Linnaean biological classification evolve from folk classifications, but Linnaeus also believed the *genus* to be the level at which humans studying taxonomic biology could most likely find agreement (Abbas, 2010, p. 38). Thus, Linnaeus' biological classification also supports a psychologically basic level category.

Emergence of human information organization/representation

Organization and representation of information (also known as *knowledge organization*) is one of two pillars of information science (the other pillar being information access and retrieval, also known as *information retrieval*). For communication to occur, information must be organized and represented by a sender in such a way that a receiver can obtain or construct meaning. "The ability to communicate detailed, concrete information as well as abstract concepts allowed early humans to cooperate and plan for the future in ways unique to our

species, thus enhancing their survival during rough times and boosting their reproductive success in good times" (Balter, 2009, p. 711).

At its most fundamental level, *information* can be understood as a symbol or a group of symbols that contain meaning. A *symbol* is something that represents something else. "Symbols can be anything experienced by human beings through their sensory perceptions of sight, sound, touch, smell and taste" (Edwards, 2005, p. 90). Humans give meaning to symbols, meanings are formed within human socio-cultural groups, and "agreement about the meanings enables symbolic communication" (Edwards, 2005, p. 90). Thus, communication is the transmission of information between entities, and symbolic communication is the use of symbols or groups of symbols represented in ways that convey an intended meaning within a socio-cultural group.

There seems to be general agreement in the archaeological community that "archaeological evidence of abstract or depictional images indicates modern human behavior" (Henshilwood et al., 2002, p. 1279). Personal ornaments (e.g., shell beads, modified teeth and bones) are considered early evidence of human use of symbolic communication. Some of the oldest human personal ornaments have been found in the South Africa dated about 75,000 years ago (Henshilwood et al., 2004) and personal ornaments have been found in the presence of Neanderthal remains with date ranges from 30,000 – 55,000 years ago (Caron, d'Errico, Del Moral, Santos, & Zilhão, 2011). Though there is debate regarding this connection to Neanderthals, most recent evidence suggests Neanderthal use of mineral pigments and marine shells 115,000 years ago in South Africa (Hoffmann, Angelucci, Villaverde, Zapata, & Zilhão, 2018).

The earliest symbolic communication innovation directly related to information organization and representation is the extrasomatic symbolic storage and memory systems, formally referred to as *external symbolic storage* (Donald, 1991) or *artificial memory systems* (d'Errico, 1998). Etchings on pieces of red ochre from South Africa date to about 77,000 years ago and may be the oldest evidence of recorded, organized information (Henshilwood et al., 2002). However, that is speculation at best because there is little evidence of any storage/memory devices in the Lower and Middle Paleolithic[5], but that could also be due to the material used for the devices. Based on ethnographic records, many memory storage devices were likely made of incised wood or bamboo or knotted strings of wool, skin, or vegetal matter (d'Errico, 1998)—materials that will degrade faster than stone or hardened clay, so it is not unreasonable to assume that more devices were used than what the concrete evidence suggests.

The most concrete evidence for early human for symbolic information representation is human creations of pictures in caves. Evidence found in several parts of the world ranging from about 25,000 to 40,000 years ago. For example, Pike et al. (2012) presented evidence of cave art from Spain ranging from roughly 41,000 - 22,000 years ago, which contained graphical representations that include anthropomorphic figures, animals, human hand stencils, and discs (Pike et al., 2012). And most recent evidence suggests Neanderthal cave paintings possibly older than 64,000 years ago in Spain (Hoffmann et al., 2018). Keep in mind that pictograms would later be precursors to some of the earliest writing systems that emerged over 5,000 years ago.

[5] In general, 50,000 years ago is considered the end of the Middle Paleolithic and the beginning of the Upper Paleolithic, though demarcations can vary depending on geographic location.

From an evolutionary perspective, it is interesting to note that in both the memory devices and cave art, researchers cite increased complexity over time as a trend—for memory devices: "the amount of stored information and the miniaturization of the marks used to store it" (d'Errico et al., 2003, p. 33); for cave art: the "technological and graphic complexity… and increase in figurative images" (Pike et al., 2012, p. 1412). Though this may be common sense, it seems interesting because socio-cultural theorists also observe that human organizations grow in complexity over time similar to biological organizational levels and library theorists suggest *simple-to-complex* over time is a principle to follow in an evolutionary classification (see *Social and cultural research contributions* and *Information organization and evolutionary classification* sub-sections below).

Emergence of writing systems

Writing systems enable the most complex symbolic storage and artificial memory systems and have had the most impact on human societies (d'Errico, 1998, p. 20). Literacy, the ability to read and write (i.e., the ability to organize and represent symbols that create information that can be communicated over time), is closely associated with the emergence of large, complex groups. "Many social theorists rank the behavioural innovation of literacy next to the advent of agriculture as one of the most consequential changes that humans have experienced during the Holocene" (Mullins, Whitehouse, & Atkinson, 2013, p. S143). The authors also suggest that both writing and record keeping were required for the mega-scale 'empires' to emerge.

The precursors to the earliest writing systems were symbols representing words or

Figure 4. Contemporary Logogram.
Communicates smoking is not allowed within an area.

phrases (but not syntax) consisting of pictograms (representations of physical objects) and ideograms (representations of concepts). For a modern example, a 'no smoking' sign (Figure 4) contains both a pictogram (cigar or cigarette) and an ideogram (not allowed).

Proto-writing systems emerged over 5,000 years ago, first appearing in southern Iraq (i.e., southern Mesopotamia, Sumer) about 3400 BC (termed *proto-cuneiform*) and in Egypt about 3300 BC. Itemizing goods, names, and quantities are some of the earliest uses of proto-cuneiform (Trigger, 2004, p. 47). Trigger suggests Mesoamerican proto-writing emerged with the Olmec culture sometime during 1140-400 BC (p.48). These early systems were all similar in that they were not capturing actual speech such as phrases or sentences and not representing information in any grammatical order (p.47).

Scholars assert that a true writing system must be able to record both phonetic information (word sounds) and semantic information (word meanings). Many scholars believe that true writing systems emerged independently in Sumer, Egypt, China, and Mesoamerica. Characteristics of recorded spoken language were visible around 2500 BC in Sumer and Egypt (Trigger, 2004, p. 62) and evidence suggests these two earliest writing systems evolved from

their proto forms. However, there is no evidence to suggest that the writing system that emerged in China, appearing around 1200 BC, emerged from a proto-writing system. Furthermore, though the Mayan writing system in Mesoamerica is visible by around 250 AD, Trigger states there is no evidence any proto-writing system in Mesoamerica was evolving into a true writing system prior to Spanish contact (pp. 48-49).

I end this section with insight from Trigger that is supportive of the research project presented in this dissertation: the study of writing systems "offers a useful way to evaluate evolutionary approaches to understanding change in cultural phenomena. Because of writing's role as a recording device, its development is historically better documented than is that of many forms of material culture" (Trigger, 2004, p. 39). Though Trigger specifically cites the overall *development* of writing systems, any part of development analysis would include *what* was being recorded using these writing systems—i.e., symbols and words/phrases. And though it may be difficult or impossible to understand exactly *what* was being communicated with the symbols, having that knowledge would obviously be of value to researchers studying these systems. Therefore, evolutionary analytical methods of the "what" that is being communicated in today's written communications could assist researchers in studying the "what" of the earliest writing systems.

Though there is much more literature that could be reviewed, this section ends because the emergence of writing is typically where an introductory information organization book would begin. In the footnote[6] below, I have included some historical points to give an example

[6] Short historical points for information organization and classification from 2000 BCE through the 1800s:

- A book list of 62 titles on a Sumerian tablet found in Nippur (about 2000 BCE); bibliographic information on tablets from the Hittites (about 1500 BCE).

of this type of progression, which was derived from Arlene Taylor's *The Organization of Information* (Taylor, 2004, pp. 49-50, 301).

Emergence of academic discipline of classification

Keeping with the evolutionary theme of this dissertation, tracing back to the early evidence for an academic study of classification and its general philosophical history is justified.

The historical evidence suggests Plato (428-347 BCE) was one of the earliest Western philosophers to think about groupings of things that have similar properties. "We customarily hypothesize a single form in connection with each of the many things to which we apply the same name. ... For example, there are many beds and tables. ... But there are only two forms of such furniture, one of the bed and one of the table" (Grube, 1992, p. 265). However, as an academic endeavor, most researchers credit Aristotle (384–322 BCE), a student of Plato, as the founder of one of the bedrocks of information organization: ontology, or the study of what we know.

Aristotle's *Categories* serves as a foundation for philosophical ontology. Ontology is the study of being or existence and is associated with questions that identify, differentiate, and establish relationships between categories of the things that exist. Examples of the general

---

- Callimachus, often referenced as the first bibliographer, who created the first known library catalog titled *Pinakes* at the ancient library at Alexandria, Egypt (third century BCE).
- Emergence of European monastery and university libraries and eventually the printing press (Middle Ages).
- Universities of the late Middle Ages divided their books into seven (7) subject classes based on fields that were taught: Trivium (Grammar, Rhetoric, Logic) and Quadrivium (Arithmetic, Music, Geometry, Astronomy), which had fixed shelf locations within the classes.
- Increase in libraries and expansion of collections, more complex systems for organizing collections emerged, author and subject indexes emerged (1500s-1700s). Many classifications were based on philosopher Francis Bacon's early seventeenth century categories, which were history [natural, civil, literary, ecclesiastical); philosophy (including theology); works of imagination (poetry, fables, etc.]
- Modern library classifications in the 1800s (Taylor, 2004, pp. 49-50, 301).

types of questions asked include: What exists? What are the characteristics of the things that exist? What are the relationships among the things that exist? Whether an information science classification or a biological classification, the most basic components of each can be rooted in ontology: i.e., *identification of entities, description of entities, and placement of entities into a classification structure based on relationships among the entities*.

Arlene Taylor (2004, p. 298) traces the roots of *classification* back to Aristotle's "classical theory of categories" or what June Abbas (2010) terms "classical classification"— which was based on common aspects shared by the members of a category and arranged in a hierarchical order. "The most widely used classification schemes in the United States are based upon the classical theory of categories" (Taylor, 2004, p. 300). Aristotle's detailed observation, communication of both similar and dissimilar characters, and groupings based on inherent properties found within his *Historia Animalium* (350 BCE) provided the foundations to Linnaean biological taxonomy and classification as well as classification schemes used in library and information science. His approach was deductive: "Through observation one can define the inherent properties of a plant or animal such as form, habits, and habitat. Using deduction, one could posit that if animals share the same properties or characteristics, they are similar to each other and can be grouped into categories" (Abbas, 2010, p. 29).

Wilkins and Ebach (2014) also acknowledge that scientists and philosophers often credit Aristotle with beginning classification. The source of this is Aristotle's defining of *kinds* by their *essences*—i.e., by their most salient attributes that make things distinct (although Aristotle never used a singular word that could be interpreted as "essence"). But the authors suggest that reevaluations of Aristotle's work on "kinds" indicate that he was more concerned with 1)

logical classification of words and 2) functional classes (i.e., what we might today consider "models") than what resembles biological or scientific classifications today (Wilkins & Ebach, 2014, p. 31). However, the authors do acknowledge that Aristotle did begin a tradition of "ten topics"—that is, "all concepts could be reduced, or rather generalized, to ten ultimate concepts" (p. 31); these included: what-it-is, quantity, quality, relation, location, time, position, possession, doing, and undergoing. But it was not clear whether Aristotle meant these to be actual real things or simply concepts of the mind (pp. 31-32).

Regardless of new insights, evidence from scholars of both information and biological sciences suggests that Aristotle has been the earliest and most influential Western source for classification thinkers, so rooting the history of Western classification with Aristotle seems appropriate.

Scholarly evidence connecting biological and cultural phenomena

This section explores scholarly support for a connection between biological phenomena and cultural phenomena. Both socio-cultural and biological researchers have contributed to the creation of a body of work that connects biological phenomena with cultural phenomena. Below are brief examples to illustrate the depth and breadth of these connections.

Social and cultural research contributions. Some of the earliest and most prominent social science researchers made connections between biological theories and cultural phenomena. Herbert Spencer (1820-1903), one of the founders of sociology (and philosopher of biology and psychology disciplines and scholarly contributor to the fields of astronomy and education, among others), was possibly the first to apply Darwin's natural selection theory to human groups.

> In popular opinion it is the name Charles Darwin that is most often associated with the idea of evolution. But Darwin applied it only to organic life. It was Spencer—whom Darwin himself called 'the great expounder of the principle of Evolution'—who extended the principle to include all of nature (Carneiro, 1967, p. lvi; original, Darwin, 1890, p. 10).

Today, any complementary combination of "Herbert Spencer" and "evolution" must be spoken softly in some academic circles due to this combination often associated with the pejorative phrase 'social Darwinism'—that Darwin's natural selection theory ("survival of the fittest" in Spencer's words (Carneiro, 1967, p. 78)) was used to justify colonialism and concentrations of power and wealth (the most fit) at the expense of the weak and poor (i.e., the least fit) as well as the Eugenics Movement of the early 1900s and associated atrocities of Nazi Germany and the Holocaust. But this focus of Spencer is disrespectful (at the least) because the broadening of evolutionary theory to science in general, and the application of evolutionary theory in particular, was remarkable and should not be omitted from background sections of research that applies aspects of biological evolutionary theory to cultural phenomena. As Carneiro asserts:

> No other thinker before or since [Spencer] has known so large a proportion of the scientific knowledge of his day, or has pieced it together into so all-embracing and rigorous a system. Details have been added and here and there conceptions and interpretations have been changed; but by and large the picture of the cosmos we have today, in which evolution is seen as giving rise successively to inorganic, organic, and superorganic phenomena, was first presented to the world by Herbert Spencer (Carneiro, 1967, pp. lvi-lvii).

Willey and Sabloff (1974) suggest the early theorists of anthropology also connected Darwinian evolutionary theory with archaeological data:

> From biological evolution the idea of progress was extended to the history of human societies and culture; and two of the founders of anthropology, E. B. Tylor (1832-1917) and L. H. Morgan (1818-81), saw in this principle of cultural evolution, and in the findings of archaeology... the data from which to construct a model of the human social and cultural past" (Dunnell, 1980, p. 35; original: Willey & Sabloff, 1974, p. 14).

A.R. Radcliffe-Brown (social anthropologist, founder of structural functionalism theory) suggested social phenomena were a class of natural phenomena and the result of social structures that unite humans. He believed social structures were as real as biological organisms (1940, p.3) and suggested social phenomena could be studied using biological methods:

> I conceive of social anthropology as the theoretical natural science of human society, that is, the investigation of social phenomena by methods essentially similar to those used in the physical and biological sciences (p.2).

Leslie White (anthropologist, cultural evolution theorist) perceived culture as integrated organic systems analogous to biological organisms: the dynamic aspects of culture—technology, social organization, philosophy, and sentimental/attitudes—are "kinds of behavior of the cultural system as an organic whole—as breathing, metabolizing, procreating, etc., are processes carried on by a biological organism as a whole" (1959, p. 19). And like any other system, a cultural system "tends to establish and maintain an equilibrium, even though this be a moving equilibrium" (1959, p. 27).

Similar to White, Julian Steward (anthropologist, cultural evolution theorist, founder of cultural ecology) also perceived culture as an integrated organic system, with a focus on the organization of cultural phenomena for study:

> In the growth continuum of any culture, there is a succession of organizational types which are not only increasingly complex but which represent new emergent forms… The concept is fairly

> similar to that of organizational levels in biology[7]. In culture, simple forms such as those represented by the family or band, do not wholly disappear when a more complex stage of development is reached, nor do they merely survive fossil-like, as the concepts of folkways and mores formerly assumed. They gradually become modified as specialized, dependent parts of new kinds of total configurations (Steward, 1963, p. 51).

Another connection that harkens back to Spencer comes from anthropologists Sahlins and Service (1960) who suggested both biological and cultural evolution "can be embraced within one total view of evolution" whereby "cultural evolution can be considered… a continuation, on a new line, of *the* evolutionary process" (p.8)—a process observable in "both life and its offshoot, culture…" (p.9). And with the "continuation on a new line" and "offshoot" concepts implanted in our minds, it is not much of a stretch to envision cultural entities—i.e., groups of humans bounded together in some way—being organized and represented in a branch-like tree similar to that of biological classification.

The connection between biological lifeforms and bounded human groups is presented in *living systems theory* by James Miller (psychologist, behavioral scientist). Miller suggests groups eight(8) categories of living systems exist on Earth (see Figure 5)—including human organizations, communities, and even societies—which all share the same twenty (20) life-processing subsystems as those of biological systems (Miller, 1978; Miller & Miller, 1990). The theory also shares the view of many cultural evolution researchers that organizational types *may* increase in complexity over time, ultimately leading to new organizational forms. The word "may" is stressed to make sure the reader understands that there is no requirement for any population of a living system to become more complex or even change radically over time.

[7] I.e., organelle, cells, tissues, organs, organ systems, organisms, populations, communities, ecosystem, biosphere

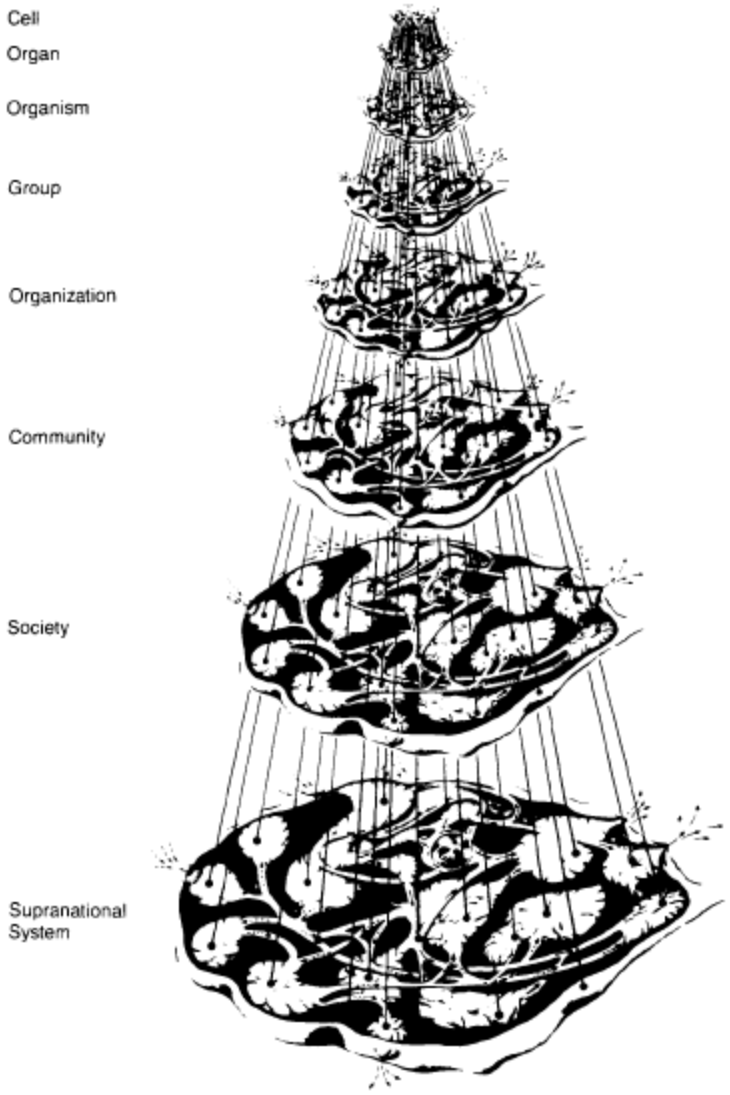


Figure 5. Categories of living systems on Earth.

Note. From "The nature of living systems," by J. G. Miller and L. J. Miller, 1990, *Behavioral Science*, 35, p.158. Copyright © 1990 John Wiley & Sons, Ltd. Reprinted with permission.

Consider that the Ginkgo Biloba tree has remained anatomically unchanged for about 200 my and the horseshoe crab for about 500 my and single-celled organisms have existed on Earth for almost 3.5 billion years).

Contemporary archaeologists and sociologists have also advocated for evolutionary theory based on biological evolution, while recognizing "the mechanisms of heritable variation and competitive selection are quite different in biological, cultural, and social evolution" (Runciman, 2009, p. 3). There is recognition that biological evolution is the strongest theory that explains change in living systems. Robert Dunnell (2000) sums it up nicely: scientific evolution is "a theory in which the form and diversity of life is explained by a set of mechanisms operating on the transmission of variability between individuals" (p. 190). Others include:

> Evolution is, in one version or another, the only scientific theory that explains *change* in living things... (O'Brien & Dunnell, 1996, p. vii).
>
> Evolutionary theory explains the origin and differential persistence of traits in living forms, i.e., change. Evolutionary theory, therefore, is the only scientific theory that explains change (why rather than how) (Lipo, 2001, p. 5).

Dunnell (2000) goes as far as suggesting archaeology as a science must embrace biological evolutionary theory if the discipline is to continue to be viewed as a science:

> Since evolution is the only scientific theory that explains change, and since the explanation of change remains at the core of archaeology, the fate of scientific evolution in archaeology would seem to rest on the degree to which archaeology continues to construe itself as science (Dunnell, 2000, p.192).

Sociologist W.G. Runciman (2009) suggests Darwin's natural selection theory combined two ideas that were already present during his time—*heritable variation* and *competitive selection* (p.2), which had profound implications for socio-cultural sciences:

> Not only does natural selection explain more about human behaviour than the overwhelming majority of twentieth-century sociologists were willing to concede, but the heritable variation and competitive selection of information which affects behaviour in the phenotype is a process which operates also at both the cultural level, where the information is encoded in *memes*[8] – that is, items or **packages of information** transmitted from mind to mind by imitation or learning – and the social level, where it is encoded in the rule-governed *practices* which define mutually interacting institutional roles (pp. 2-3; bold emphasis mine).

Thus information packages, whether biological or cultural, are *traits* that vary within a population and can be *inherited* by others in the population, and some traits get selected for transmission over time and others do not. And, of course, books are a source of organized cultural information packages—a collection of cultural traits much like an organism is a collection of biological traits; some of the book traits are the same for all books of certain types (e.g., title page, table of contents, chapters, index), some traits are similar (e.g., mysterious events of most all mystery novels are solved at the end) and some are very different (e.g., the content of an information science text book and a fantasy novel), but no two books that are different original works of authorship are the same[9].

Biological research contributions. Researchers making biological connections with cultural phenomena are not limited to socio-cultural researchers. Biological philosopher and cognitive scientist Daniel Dennett (1995) presents a generic version of the theory of evolution by natural selection applicable for a science of cultural evolution:

> Evolution occurs whenever the following conditions exist:
>
> 1) Variation: there is a continuing abundance of different elements;

[8] The term "meme" was coined by evolutionary biologist Richard Dawkins. See next section for more about this.
[9] This should go without saying, but to remove any doubt, mass produced copies of one book are obviously identical but each copy would not be different original works of authorship.

> 2) Heredity or replication: the elements have the capacity to create copies or replicas of themselves;
>
> 3) Differential 'fitness': the number of copies of an element that are created in a given time varies, depending on interactions between the features of that element and features of the environment in which it persists (p. 343).

In other words, if a population of entities (biological or cultural) has variations of traits/characteristics, has pathways (e.g., inheritance, communications) to pass those traits/characteristics to other members of the population, and has differential survival and reproduction (not all entities will survive to reproduce/replicate)—then evolution by natural selection occurs. That is, those traits/characteristics that enable more entities to survive and reproduce will increase in frequency within the entity population (i.e., the traits/characteristics will be 'selected') over time. For example, in the population of 4-wheeled transportation apparatuses, after the invention of the internal combustion engine, the number of horses powering the apparatuses diminished and the number of internal combustion engines powering the apparatuses increased.

In Dennett's words: "there is no denying that there is cultural evolution, in the Darwin-neutral sense that cultures change over time, accumulating and losing features, while also manifesting features from earlier ages" (p. 345). So for Dennett, the only question left for debate is how closely cultural evolution is to "genetic evolution, the process that Darwinian theory explains so well…" (p. 345).

Regarding a cultural analog to the gene, evolutionary biologist Richard Dawkins used *meme* in his book *The Selfish Gene* to illustrate the cultural analog to the biological gene, the latter of which he suggests is basically a replicator of information:

> I think that a new kind of replicator has recently emerged on this very planet. It is staring us in the face. It is still in its infancy, still drifting clumsily about in its primeval soup, but already it is achieving evolutionary change at a rate that leaves the old gene panting far behind. The new soup is the soup of human culture. We need a name for the new replicator, a noun that conveys the idea of a unit of cultural transmission, or a unit of imitation. 'Mimeme' comes from a suitable Greek root, but I want a monosyllable that sounds a bit like 'gene'. I hope my classicist friends will forgive me if I abbreviate mimeme to meme. If it is any consolation, it could alternatively be thought of as being related to 'memory', or to the French word meme. It should be pronounced to rhyme with 'cream'. Examples of memes are tunes, ideas, catch-phrases, clothes fashions, ways of making pots or of building arches. Just as genes propagate themselves in the gene pool by leaping from body to body via sperms or eggs, so memes propagate themselves in the meme pool by leaping from brain to brain via a process which, in the broad sense, can be called imitation (Dawkins, 1989, p. 192)

Regarding the cultural analog of genetic transmission (i.e., replication/inheritance), population biologists L.L. Cavalli-Sforza and M.W. Feldman (1981) developed a mathematical theory of cultural transmission and tested the theory with models using data from sociology, archaeology, and epidemiology. Regarding cultural constituents such as human thought, speech, behavior, and artifacts, the researchers suggest these “cultural entities” (p.10) share one thing in common:

> They are capable of being transmitted culturally from one individual to another. Transmission may imply copying (or imitation); copying carries with it the chance of error. Thus we have in cultural transmission the analogs to reproduction and mutation in biological entities. Ideas, languages, values, behavior, and technologies, when transmitted, undergo “reproduction,” and when there is a difference between the subsequently transmitted version of the original entity, and the original entity itself, “mutation” has occurred (Cavalli-Sforza & Feldman, p.10)

The authors continue by stating if only the analogs of reproduction and mutation are involved in cultural evolution, there would be only random change. Without something analogous to the natural selection process, there is no *adaptive* significance—i.e., no traits that enhance survival and reproduction that can be 'selected' by members of a population and ultimately spread throughout that population (what biologists refer to as *fitness*). And without an adaptive significance, the entire premise of a cultural evolution process similar to that of biological evolution collapses (and also reduces the strength of any argument that cultural researchers can legitimately apply biological theory and analytical techniques to cultural phenomena).

There is no question that we humans select cultural traits that can increase our survival and reproduction success (i.e., fitness)—such as obtaining a job or a college degree, membership in a religious organization, participation in online dating, controllable characteristics of speech, etc. Cultural traits have given humans the ability to inhabit every part of the Earth, as well as outer space. But culture to increase *human* fitness is not the view of culture that this dissertation explores. Rather, this dissertation is guided by the assumption that there are cultural analogous of biological species—i.e., cultural *organisms*— and therefore, biological evolutionary theory operates similarly within populations of cultural organisms.

The cultural organism. One possible support for the above 'cultural organism' assertion that comes from the biological researchers is from Cavalli-Sforza and Feldman's assertion that humans create *cultural objects*, which are the cultural equivalent of biological organisms, and humans enable those cultural organisms to emerge, survive, and reproduce. Examples given by the authors are the productions of jets, cars, washing machines, and violins. The authors

considered these cultural objects to be "*second-order organisms*," which are created by humans, which are the *first-order organisms*. Traits within these cultural objects are selected by humans—first by the producers and second by the acquirers of the cultural objects (p.17-19). Though the humans who are acquiring the cultural organisms may be acquiring to increase human survival and reproduction (e.g., acquiring the latest clothing fashion to attract a mate), the humans also create the cultural environments (or better yet, the cultural ecosystems) for cultural organisms to flourish.

> Car models or makes of musical instruments that are selected because of some aesthetic and technical qualities that appeal most to prospective customers will prosper (p.18). Even language and its components (words, rules, and sounds) can be regarded as cultural 'objects,' and the cultural fitness (the appeal to the speaker) of various alternatives for its components, rules, and so on, determines the Darwinian fitness of those components of language in the sense that they can be considered as second-order organisms (Cavalli-Sforza & Feldman, p.19).

If we assume this is a correct analogy, books would also be what Cavalli-Sforza and Feldman term second-order organisms. But Cavalli-Sforza and Feldman's second-order organism is only defined by examples, and the suggestion seems to be that either the components of cultural objects or the cultural objects themselves are the cultural equivalents of biological organisms.[10] Reading Dawkins and Cavalli-Sforza and Feldman's excerpts (above) together, one might believe that Dawkins' memes are also considered to be cultural organisms. And indeed they are—cultural *viruses* as Dawkins clarifies:

[10] A search for the 'second-order organism' term within other literature from Cavalli-Sforza and/or Feldman could not be found using a Google Scholar search nor could records having variants of the term be found from *Web of Science* 'topic' searches, so the term does not seem to fit into a 'common knowledge' category.

> When you plant a fertile meme in my mind you literally parasitize my brain, turning it into a vehicle for the meme's propagation in just the way that a virus may parasitize the genetic mechanism of a host cell (Dawkins, 1989: 192).

Viruses may be organisms, but there is debate whether they are living or just some type of organic structure. Viruses can survive and reproduce, but only within a host. Viruses are smaller and less complex than, say, bacteria, some of which are considered the smallest living organisms. Viruses have DNA or RNA (typically not both) but no cell structure. A meme (a catchy phrase, an idea, a visual image) that is transmitted from brain to brain (the hosts) in a population does seem to be the cultural analog of a virus because memes are usually portions of a larger cultural works such as books, sound recordings, speeches, etc. For example, consider that one of the most used concepts from *The Selfish Gene* is the 'meme' concept. But the entire contents of the book cannot be a meme (a cultural virus) because the entire contents will likely never reside in any one person's brain (other than maybe Richard Dawkins' brain). This 'portion' characteristic of a meme also seems to fit with the 'less complex' nature of the biological virus (i.e., having DNA or RNA (typically not both), smaller size, and no cell structure compared to even the smallest truly living biological organisms (bacteria).

So if a meme is a virus and if a virus is not a truly 'living' organism, then exactly what is the cultural organism? This is an important question because the classification method proposed in this dissertation is a phylogenetic method, which indicates possible evolutionary histories and relatedness among *biological species*. So a cultural analog of the biological organism should be clearly established to justify the use of the proposed classification method.

Analogies of biological organisms using cultural objects seem inaccurate when applying Miller's living systems theory to this topic, which suggests the *cultural group*—rather than a

cultural object produced by the cultural group—is analogous to the biological organism. The cultural group is a group of humans bounded by some unifying socio-cultural force (e.g., practices per Giddens' structuration theory; see Giddens, 1984, p. 17). The cultural group as analogous to the biological organism seems more logical because, fundamentally, cultural objects cannot reproduce cultural objects—i.e., books cannot reproduce books. So how does a cultural group (i.e., a cultural organism) replicate itself? What exactly is being replicated or inherited?

Replication/inheritance in cultural organisms (e.g., a book publishing organization) is *not* accomplished by the production of cultural objects. The cultural objects (or services) that are produced by cultural organizations are the life sustaining attributes needed to keep the organizations alive. ***Replication/inheritance is found in the organization's ability to replicate when members leave and new members enter.*** [11] So unlike the reproduction ability of biological organisms, which create new organisms, cultural organisms pass on parts of their 'genetic' (or memetic, *sensu* Dawkins) code to a new person entering. This is due to the permeable boundaries of cultural organisms. Cells, organs, and organisms have physical boundaries but cultural organizations have conceptual boundaries. There may be a physical building where the organization members carry out the functions of the organization, but it is the people that create and sustain the invisible boundaries (i.e., the structures and functions) of an organization with their day-to-day activities.

[11] The three bold/italicized statements in this section are my assertions and I believe the overall concept they represent to be novel and important missing pieces in the contemporary cultural evolution literature. More literature review needs to be conducted to reference someone else's invention of this concept, or more writing on my part needs to be done to fully develop this concept, which is outside the scope of this dissertation.

Given enough time, an entirely new group of people will have replaced all previous people in the organization, and thus, the organization will have replicated itself. In other words, because people create the structures and functions of a cultural organism, the only way an organization can pass on its genotype (or memotype, *sensu* Dawkins) to the 'next generation' is for new people (i.e., the next generation or the 'offspring') to replicate the structures and functions of the pervious people (i.e., the previous generation, or the 'parents'). The replication is ongoing, i.e., not marked with clear boundaries as with biological parents and their children. Obviously, this is because the cultural organism is maintained with people who flow in and out at varying points in time. Most likely, ***the parent generation's organization will be different than the offspring generation's organization, so there is descent with modification over time.*** So for cultural organisms, again, using Miller's living systems theory, we need to align our thinking toward collectives of people rather than the outputs of collectives of people. A relatively simple mathematical model might be:

Offspring organization = parent organization members – member outflow + member inflow.

The above equation needs a *time* component, and an *environmental force* component and maybe differentiation for the types of members (e.g., decision makers verses laborers). And maybe Miller's twenty (20) life-processing subsystems need to be part of the model. Regardless, from a fully developed model based on the above, various scenarios could be calculated and debate could begin for what constitutes an offspring organization and when an offspring organization has emerged.

Alignment with natural selection theory. Here is an example to clearly align the above cultural organism concept with natural selection theory. Within the population of book

publishing organisms, there are variations of traits (e.g., in the internal processes that produce the books, in the subject matter of the books produced, in the authors who are under contract with the publishing company, etc.); some traits will be similar (or the same) across all publishing companies and some will vary. Some traits (e.g., a new process that is made public in trade journals or a new subject of books being produced) provide a survival/reproduction advantage and those traits can be inherited by other publishing organisms. Not all organisms (companies) in the book publishing population will acquire necessary adaptive advantages (traits) and will eventually cease to exist and the adaptive traits that enable the remaining publishing organisms to continue to survive will be more prevalent in book publishing population.

Cultural objects. So what about the cultural objects (or services) produced by the cultural organisms, such as a publishing company's books? If they are not the cultural analogs to biological organisms, what are they? More importantly for this dissertation, why should they be classified using a biological classification method if they are not the cultural equivalent of biological organisms?

As with organisms, organizations have visible manifestations of their internal processes. The *genotype* is the genetic code of an organism and it will create traits of the organism's anatomy, physiology, and (to a large degree in most animals) behavior. Regarding the latter, in the 1930s, Konrad Lorenz asserted that behavior could "be analyzed in much the same ways as anatomical and physiological properties have long been studied" (Smith, 1980: 196). Lorenz himself believed this to be his greatest contribution to science. He was one of the founders of the discipline/field of ethology (the study of animal behavior).

Richard Dawkins (1982) took behavior one step further by suggesting the phenotype (the expression of the genotype) does not need to be limited to a physical body. Examples of 'extended' phenotypes include bird's nests, beaver dams, caddisfly cases, and spider webs.

> Nobody has any trouble understanding the idea of genetic control of morphological differences. Nowadays few people have trouble understanding that there is, in principle, no difference between genetic control of morphology and genetic control of behaviour... And if we decide to allow that both morphology and behaviour may be inherited, we cannot reasonably at the same time object to calling caddis house colour and spider web shape inherited. The extra step from behaviour to extended phenotype, in this case the stone house or the web, is as conceptually negligible as the step from morphology to behaviour. From the viewpoint of this book an animal artefact, like any other phenotypic product whose variation is influenced by a gene, can be regarded as a phenotypic tool by which that gene could potentially lever itself into the next generation (p. 199).

To be clear, what Lorenz and Dawkins are referring to is *innate* animal behaviors, not learned as in the case with humans and learned aspects of culture. But others have extended this to human culture. Vinicius (2010) is one example:

> A limitation of the original [extended phenotype] definition is that not all extended phenotypes are encoded in genes; for example, some may depend on information stored and transmitted by animal brains (i.e. they may be coded by memes rather than genes). Nest building in birds or dam building in beavers do seem to be effects of genes, as those are innate actions involving no learning or teaching from other individuals. However, dam building in humans is clearly a learned process, and such extended phenotypes that need to be learned only evolved due to the existence of the animal brain as an extended carrier of information (p.196).

Assuming the extended phenotype concept can be extended to cultural organisms, we can understand or view the book as an extension of the book publishing organism and that the

book is just as important to the book publishing organism for survival and reproduction as a web is to a spider.

Summary. From the above socio-cultural science contributions, the assertions are:

- Culture is an offshoot of biological life and is composed of living systems just as biological ecosystems are composed of living systems. Socio-cultural phenomena are as real as biological organisms and can be studied with biological methods.
- Natural selection evolutionary theory is the only scientific theory that explains non-random change over time in living things.
- Culture possesses key attributes of natural selection theory: heritable variation and competitive selection. And because culture possesses the equivalent of organisms, biological evolutionary theory is applicable to cultural change.
- Cultural forms change over time by becoming more complex and emerging into new cultural forms, but still retain elements of smaller cultural forms.
- Packages of information are being transmitted—whether as biological genes or as cultural memes (and by their nature, communication and information sciences are positioned for research pertaining to packaging and transmission of cultural information).

From the biological contributions, the assertions are:

- Principles of natural selection theory can be applied to cultural phenomena.
- Cultural phenomena and biological analogs have been made (albeit somewhat inaccurate). In particular, the identification and transmission (reproduction and inheritance) of the cultural analog of the biological gene has been made.

- Phenotypes (expressions of the genotype) can extend beyond a physical body.

From my contributions, the assertions are:

- James Miller's living systems theory can be used to justify any 'human collective' as the cultural analog to a biological organism.
- The process of cultural organism replication (i.e., inheritance) can be illustrated with a relatively simple mathematical model.
- Richard Dawkins' extended phenotype concept can be employed to justify the use of output (books) from a cultural organism (a book publishing company) for a biological evolutionary-based analysis.

Of course, all of the above is only a very small sample of general literature that supports the assertion that cultural phenomena can be studied using evolutionary biology's theories, concepts, and methodologies. And as one might assume, there is also no shortage of criticism for attempting to link cultural research with biological theories. And, of course, there are differences between biological evolution and cultural evolution. If this was the focus of the dissertation, the literature review chapter would need to provide much more evidence supporting the various connections with a biological view of culture. But the point to this section of the dissertation is to briefly communicate to a communications and information science community that a strong scholarly foundation exists in linking biological and cultural phenomena. Even Thomas Kuhn (1996) suggested science could be illustrated using an evolutionary tree based on common origin:

> Imagine an evolutionary tree representing the development of the modern scientific specialties from their common origins in, say, primitive natural philosophy and the crafts. A line drawn up

> that tree, never doubling back, from the trunk to the tip of some branch would trace a succession of theories related by descent (p.205).

This should be enough to justify the use of a biological classification method as a possible solution to an information science problem that this dissertation explores.

This section concludes with a graphic that helps to mentally solidify the connections between biological evolution and cultural evolution. One research team with backgrounds in anthropology, psychology, neuroscience, and biology has made arguments for an approach to a science of cultural evolution including characteristics, techniques, and framework analogous to evolutionary biology (Mesoudi, Whiten, & Laland, 2004, 2006). The proposed academic framework is presented in Figure 6.

Scholarly evidence connecting information organization to evolutionary classification

This section explores scholarly support for a connection between book classification and biological classification. This section's literature search strategy is presented in the detailed search descriptions section at the end of this chapter.

Information organization and evolutionary classification. Unknown to many library and information science professionals is the connection to evolutionary-based library classifications bibliographer, and J. S. Lesley, an American geologist and librarian, created book classification schemes based on evolutionary order in the 1840s and 1860s respectively. But Dousa suggests the word "evolution" was implied but not explicitly stated in these writings[12] and the first library theorists to explicitly introduce evolutionary order as the foundation for a library

[12] Dousa suggested these schemes deserved a closer comparison to the schemes of Cutter and Richardson but was beyond the scope of the 2011 paper.

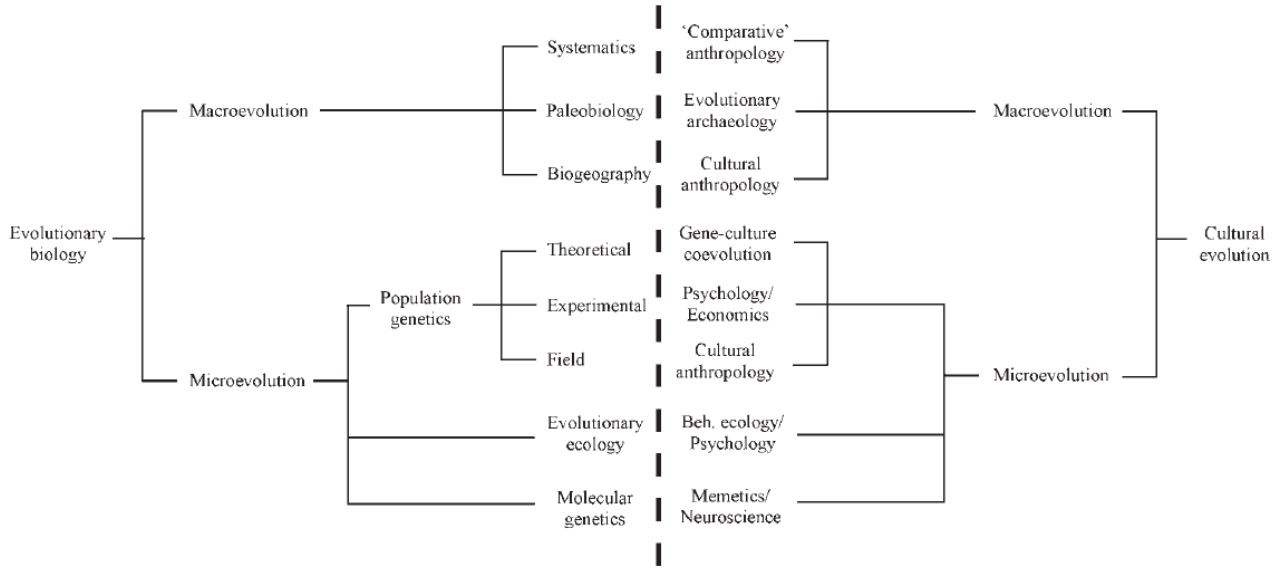


Figure 6. Proposed Academic Discipline Framework for a Science of Cultural Evolution. Left side includes disciplines of biological evolution. Right side includes the analogs for cultural evolution.

Note. From "Towards a unified science of cultural evolution," by A. Mesoudi, A. Whiten, and K. Laland, 2006, *Behavioral and Brain Sciences*, 29, p. 331. Copyright © 2006 by Cambridge University Press. Reprinted with permission.

classification scheme were Charles Cutter (1837–1903) and Ernest Richardson (1860–1939) (Dousa, 2011, p. 77).

To set the stage, Dousa (2011) suggests there were at least three cultural influences at work on both Cutter and Richardson. One was the 1860 publication in the United States of Darwin's *Origin of Species*. By the end of that century, the book had stimulated much debate and was widely read beyond scientists and religious leaders (p. 78). Secondly, as mentioned previously, English philosopher Herbert Spencer had extended Darwin's concept of evolution to encompass all of nature, suggesting evolution was a universal law "that provided an explanatory key for phenomena as diverse as the formation of the solar system from the nebular mass, the embryological development of animals within the womb, the development of more complex organisms from simpler ones, and the development of larger, pluralistic societies from simpler, homogeneous ones" (Dousa, 2011, p. 78; original Copleston 1994, p. 128). Finally, there was a movement in both Europe and the United States during the 19th century to classify the sciences. The French philosopher Auguste Comte (founder of positivism) and Herbert Spencer both created classifications. And though both outcomes were similar, Compte's philosophical basis for ordering the sciences was based on increasing complexity over time (Dousa, 2011, p. 78; original Shera, 1965, p. 81), which tends to be a principle also found in both Cutter and Richardson.

Cutter's evolutionary order. According to Dousa, Cutter asserted "'the expansive classification follows the evolutionary idea throughout', claiming this as a point of superiority over the Dewey Decimal Classification, whose sequence of classes he deemed to be less 'scientific'" (Dousa, 2011, p. 81). In his Expansive Classification (EC) for books, the ordering for

the natural sciences was described by Cutter to be "'general' before 'special,' 'past' before 'present,' 'dynamical' or theoretical before 'statical' or descriptive" (Dousa 2011: 81). In practical terms for natural sciences, this meant matter to life; within life sciences, botany to zoology; and within zoology, monera (single-celled animals) to primates (Dousa 2011: 81). In other words, "the order of classes for the natural sciences given in the EC was based on the classical scale of nature (i.e., mathematical entities before physical entities, inorganic entities before organic entities, plants before animals, brute animals before human beings)—an order that, *mutatis mutandis*, finds numerous parallels in the natural science classes of other late 19th-century bibliographical classifications" (Dousa 2011: 81). Cutter believed this structure represented an evolutionary order based on both complexity and chronology (Dousa 2011: 81). But for Cutter, outside of the natural sciences, the basic evolutionary principles used (at the time) of "general to special, past to present, and abstract to concrete" had to be augmented with experience of the classificationist to make connections among classes based on cultural logic (Dousa, 2011: 82).

Though Cutter claimed to be producing an evolutionary classification, Dousa concluded that Cutter's EC "did not reflect any single, consistently applied notion of evolutionary order: rather, sequences of classes based on the development from simplicity to complexity and the movement from generality to speciality [*sic.*] were intermingled with sequences based on an idealized chronological order or what Cutter took to be a 'natural' or 'logical' order" (p.82). Cutter's results mirrored that of the classical scale of nature that was common at the time. Additionally, evolution based on descent from a common ancestor would have been a stronger

principle to employ—Darwin's *Origin of Species* had been published for 20-40 years prior to Cutter's evolutionary classification writings referenced by Dousa.

Richardson's evolutionary order. Whereas Cutter's approach sought a practical application for an evolutionary-based classification, Dousa states that Richardson's (1901) approach to an evolutionary-based classification was "the first full-scale theoretical treatment of the topic within the library community…" (p. 83). According to Dousa, Richardson believed there was a natural order to both material objects and ideas and that classifications, at least in theory, should be based on this natural order. Richardson suggested three general laws on which a natural classification should be developed: "(1) the law of likeness, according to which 'all things are organized according to their likeness'; (2) the historical law, according to which 'the progress of things in time is also in general a genetic progress in complexity'; and (3) the law of evolution, according to which 'the law of historical progress from the simple to the complex holds good of all things which tend toward continued existence'" (Dousa, 2011, p. 83). The laws formed the basis of Richardson's evolutionary classification, which Richardson stated as a "classification according to the order of likeness from the simplest to the most complex" (p. 83). He applied this to a theoretical ordering of the sciences, which, as with Cutter, greatly resembled "the schemes of Comte and Spencer" (p. 84).

Dousa omitted what I believe to be the strongest evolutionary classification principle made by Richardson. In Richardson's 'law of likeness,' he provides several 'sub-laws': "(a) The law that things like the same things are like each other, (b) The law that like draws like, (c) The law that like begets like, and (d) The law that *true likeness points to a common ancestry*" (Richardson, 1901, p. 15, emphasis mine). Given that relatedness by descent from a common

ancestor is a fundamental assumption of contemporary biological evolution, 'common ancestry' should be included in any classification endeavor claiming to be evolutionary. Richardson's 1901 writing may have been the first introduction of this concept to library classification scholars.

Though Richardson believed the classification principles should be the same under both theory and practice, he also asserted that ordering books based on the needs of the library user should be given priority over ordering books based on strict theory (Dousa, 2011, p. 84). As stated by Richardson, "the main fact about the classification of books is in brief the fact that it is an art not science" (Richardson, 1901, p. 49). Dousa stated that an analysis of the classification that Richardson designed for the Princeton University Libraries was clear evidence of many deviations from theory Richardson made in practice.

Classification Research Group's evolutionary order. The use of evolutionary ordering related to library and information science would emerge again in the 1950s with the English-based Classification Research Group, and with it, a primary principle used by Cutter and Richardson also emerged again: "the world of entities evolves from the simple towards the complex," but this time with an added principle that included "by an accumulation of properties or influences from the environment;—a process resulting in progressively higher levels of organization" (Dousa, 2011, p. 76). This 'accumulation of properties' could be interpreted as the result of 'selection' forces within the environment, at least for living systems. The 'simple to complex' concept is a theoretical principle underlying classification by integrative levels, which is most visible in the works of the International Society for Knowledge Organization in Italy (http://www.iskoi.org/ilc/index.php).

Gnoli's evolutionary order. The literature review revealed that Claudio Gnoli is by far the most relevant contemporary researcher in the library and information science literature pertaining to the type of evolutionary classification presented in this dissertation. Gnoli is a prolific knowledge organization researcher and writer (both theoretical and applied) and the breadth of his source subject matter and examples are admirable. He has been a practicing librarian since 1994, most recently as a librarian at the University of Pavia in Pavia, Lombardy, Italy—first in the Mathematics department and currently in the Science and Technology Library. Gnoli's work in knowledge organization is broadly related to classification of phenomena and he is a primary developer of the Integrative Levels Classification.

Though integrative levels is a main area of Gnoli's work, I am most interested in his views on classification by common origin rather than views of hierarchical levels. This is due to my broad interest in classifying cultural 'species' and Gnoli's broad interest classifying human knowledge—the former approaches classification as a biological taxonomist and the latter approaches classification as a librarian/information science classifier. I will briefly link Gnoli's integrative levels and my work in this dissertation, and then focus on his classification views of common origin.

Integrative levels. In his work on integrative levels, Gnoli continues the evolutionary tradition of emergence from simple-to-complex whereby the lower levels evolve into complex levels, which then retain parts from the lower level but have novel properties not found in the lower levels (Gnoli, 2006, p. 140). Similarly, other library and information science scholars—Tom Stonier (biologist/information theorist), Marcia Bates (information scientist), David Bawden (information scientist), and Wolfgang Hofkirchner (information theorist)—

independently identified "three levels of information organization – material, living and social – connected to each other by an evolutionary process of emergence, from the more simple to the more complex" (Gnoli & Ridi, 2014, p. 446). As mentioned previously, I subscribe to a view of culture based on James Miller's living systems theory: culture is composed of living systems of small and large 'bounded' human groups and as such, we can identify, describe, and classify cultural 'species' in the way biological taxonomists identify, describe, and classify biological species.[13] But I want to at least bring to the reader's attention the similarities between integrative levels and living systems:

Gnoli (2017) suggests phenomena can be broadly represented by:

> At least six major levels, each one representing patterns of the previous one in networks of a novel nature: *forms, matter, life, mind, society, culture* (Gnoli, 2017, p. 46).
>
> Each level of phenomena, though being made with parts from the lower levels, forms into a new whole, having emergent properties not present in the lower levels (Gnoli, 2006, p. 140).

Miller & Miller (1990) suggest there are:

> Eight levels of living systems: *cell, organ, organism, group, organization, community, society, supranational system* (p. 158).
>
> At each higher level of living systems there are important similarities to the lower levels, but there are also differences. Higher-level systems have emergent structures and processes that are not present at lower levels. Emergents are novel processes, made possible because higher level systems have a greater number of components with more

[13] I referenced Miller & Millier's work previously to make the connection between biological lifeforms and *bounded* human groups to support the assertion that human groups could be considered living entities and thus could be studied using analytical methodologies from biology.

> complicated relationships among them. This increased complexity makes the whole system greater than the simple sum of its parts, and gives it more capability (p. 163).

Though the levels are somewhat different due to the subjects of phenomena (Gnoli) verses living systems (Miller & Miller), the descriptions of these simple-to-complex levels are practically the same. Thus, *Gnoli and I are each using the same foundational theory to support, in part, our classifications. We also share much of the same principles of classification as the next section will illustrate*.

Classification by common origin. Gnoli stresses the need for phenomena to be "grouped into classes, based on both their similarity (morphology) and their common origin (phylogeny)" (Gnoli, 2017, p. 50). Classifications based on both of these two principles are "are more deep and informative, and in this sense more objective" (Gnoli, 2006, p. 145). In many cases, the more related two entities are, the more similar they will be, but common origin is more powerful of the two due to the potential for prediction:

> Common origin often has a bigger explanatory power of the nature of phenomena than has shape similarity alone. Once we know that two objects are historically related, we understand their structure in deeper ways, and on this basis we can also predict further characters not manifest at initial inspection: knowing that whales are mammals allows us to predict that they breathe by lungs and suckle their offspring, without need of checking this directly for every new whale individual that is discovered (Gnoli, 2017, p. 48).

Gnoli (2006) suggests a phylogenetic approach, similar in concept to that used by evolutionary biologists, can be used in classification of phenomena:

> Phylogenesis is a most informative source to classify phenomena at many levels. We just have to define phylogenesis in a broader sense: not just the inheritance of DNA variants by descent and modifications from biological ancestors, but any derivation of a phenomenon from pre-existing

> phenomena through a path of increasing logical depth. Indeed, [a biologist] was able to apply to cornets a method designed to classify organisms (p. 148).

The primary analytical method for phylogenetic analysis is cladistics. Gnoli acknowledges the strength of cladistic analysis is due to "a rigorous analysis of the characters[14] actually shared by organisms with their common ancestors" (Gnoli, 2006, p. 142). However, his main opposition to this is the potential conflict with similarity:

> Cladograms [the evolutionary "trees" that result from cladistic analyses] may seem to be the ultimate solution in terms of evolutionary biology. However, they produce some oddities. A sensational example is that, according to cladistics, because all birds are originated from a sub-group of reptiles, birds should not form anymore a sister class of reptiles, as in traditional and common sense systematics; rather, birds are now a subclass of reptiles! (p. 143).

Thus, Gnoli prefers Ernst Mayr's[15] view of an evolutionary classification, which Mayr termed *evolutionary taxonomy* and is based on both evolutionary relationships and morphological similarities. Therefore, birds would remain a separate class "by virtue of their remarkable differentiation from their ancestors" (p. 143). However, Gnoli stresses that "origin is more relevant, as it allows for more generalizations than naive classes based on similarity" (p. 144). He provides as examples dolphins and sharks, which look similar but are genetically far apart.

Discussion. Both Cutter and Richardson, library theoreticians who were formalizing their evolutionary classification ideas in the late 1800s, believed that, at least in principle or theory,

[14] Character: a feature in a population of entities used as data in an analysis of evolutionary relatedness of entities containing similar features. For this dissertation project, the characters are the words and phrases extracted from books.

[15] Ernst Mayr was one of the most prominent evolutionary scientists of the twentieth century. His work, along with others, resulted in evolutionary biology's 'modern synthesis': Darwinian natural selection, Mendelian heredity, and population genetics.

an evolutionary ordering of books was the most natural or scientifically-based type of book classification; but neither could achieve what could be considered a truly evolutionary-based book classification in practice. Richardson made the connection between similarity and a common ancestor, and over 100 years later, Gnoli would also suggest both similarity and common origin should be primary principles for evolutionary classification theory and that such a classification would be more objective and scientific. But Gnoli further emphasizes the common origin principle by suggesting it to be *the* most important principle for classification, at least for human knowledge and biological classifications. Similarity may be used if common origin is not possible, because as Richardson asserted, true similarity suggests a common origin. And though Gnoli devotes very little writing to the use of phylogenetic analytical techniques that have appeared in cultural studies over the last 20 years, he did reference one example (analysis of cornets by a biologist) to support his discussion of the need to broaden phylogenesis beyond the biological definition to make it useful for any phenomenon classification.

Cutter, Richardson, the Classification Research Group (CRG), and Gnoli all seem to agree that simple-to-complex is a fundamental principle in evolutionary classification theory but the latter two (CRG and Gnoli) understand that complexity results from the accumulation of properties from the environmental interactions (which I interpret as the result of Darwinian 'selection pressures' for living systems). And Cutter, Richardson, and Gnoli (and presumably the CRG) also believe that if given a decision choice between a practical classification that benefits

the user of information and a theoretical classification that would be less useful for the user, then the former should take priority over the latter.[16]

Because there is evidence from library/information science scholars (and, of course, evolutionary biology scholars) that suggests common origin to be a primary principle in any proposed evolutionary classification, and if a classification based on a principle of common origin is more informative, more intuitive/natural, more objective, more scientific, and has the potential for more generalizations and more explanatory power, it seems the use of methodology designed to uncover common origin should be at least attempted if one is truly interested in creating a true evolutionary classification of books. Cutter and Richardson did not have the phylogenetic methodologies (or even modern text-mining technologies for that matter) because cladistics was not invented until the mid-1900s. Gnoli has access to phylogenetic methodologies, but his scholarly work seems to be focused on philosophical principles and applying those in the development of the levels and notation system for the Integrative Levels Classification rather than conducting analyses utilizing phylogenetic methods from evolutionary biology to uncover evolutionary relationships of the taxa being classified.

Conclusion. Though information organization theorists uphold the importance of the 'common origin' principle for information organization, none view books as phenotypic expressions of a living system. In other words, none are looking at books in the way an evolutionary biologist would look at books. Though information organization theorists have developed evolutionary classification principles, none are using the analytical techniques employed in biology to uncover evolutionary relatedness. Simply put, the literature review has

[16] With today's computational power and technologies, it is possible that multiple classification schemes could co-exist.

not produced evidence of a test of an evolutionary book classification based on common origin using the contemporary classification tools of evolutionary biology. More specifically, there is no evidence of a dataset of information (e.g., books) being analyzed by computational phylogenetic algorithms to produce an example evolutionary classification for information organization. If such a classification were possible, it would seem to be a more 'scientific' approach to book classification than what has been used in the past—a classification that is built on the overarching theory of natural selection's decent with modification and the more specific phylogenetic philosophical assumptions. If common origin is the most important information organization principle, then it seems we should first develop a method that will satisfy that principle. More broadly, if culture is comprised of various living systems each with the equivalent of genotypes and phenotypes, then a phylogenetic approach (producing cladograms or networks) should be theoretically possible for *any* identified cultural taxa.

Computer science connection to biological evolution

There is often overlap with information science and computer science—i.e., it is not uncommon to find some level of computer science background in information science professionals. However, this is not the case with this dissertation's author[17], but because of the potential overlap, I wanted to at least take a quick look to make sure I was not missing something obvious and detrimental to my proposed project. Of course, the algorithms created for *any* phylogenetic analysis are ultimately written to run on a computer. But if phylogenetic-

[17] Due to the lack of computer science background, even a cursory analysis of any computer science methodology and its applicability to the research project proposed in this dissertation is beyond the scope of dissertation. I focused on the use of existing computational applications specific to phylogenetic analysis that others have used in the past but have not been applied specifically to book classification.

based text-mining was an active area of general computer science, it should be observable in a general literature such as text books.

In computer science, evolutionary computation is a branch that is directly influenced by Darwinian evolutionary theory. Noman and Iba (2016) state:

> Perhaps, the largest natural information processing system that we have studied most widely and understand reasonably is evolution. Evolution refers to the scientific theory that explains how biological hierarchy of DNA, cells, individuals, and populations slowly change over time and give rise to the fantastic diversity that we see around us. Through the evolutionary process, the changes taking place in an organism's genotypes give rise to optimized phenotypic behaviors. Therefore, evolution can be considered as a process capable of finding optimized, albeit not optimal, solutions for problems. (p. 4).

The authors categorize the types of evolutionary computation as genetic algorithms, genetic programming, evolutionary strategy, and evolutionary programming (p. 4). It is interesting to note that words typically associated with phylogenetic analysis in cultural projects (e.g. phylogenetics, cladistics, likelihood, neighbor joining, and PAUP[18]) do not appear in Noman & Iba's book chapter. The closest is genetic programming, which includes aspects of programming being represented by tree structures but does not seem to be directly related to the algorithms in cultural phylogenetic projects. This may be because of the difference in the objectives of the research projects: evolutionary computation projects are concerned with *optimization solutions* to problems (as noted in the last sentence of the inset referenced above) and phylogenetic projects are concerned with *inferring phylogenies* (i.e., the evolutionary histories of taxa) for information about transmission of (genetic or cultural) traits.

[18] The word 'parsimony' is in the chapter, but it is used only to let the reader know 'parsimony' is the term used for a control process to a known inefficiency in genetic programming.

In *Data Mining and Knowledge Discovery Handbook*, 2nd ed., Rokach and Maimon's (2009) one chapter is titled Classification Trees. Scanning through the chapter, as well as checking for truncations mentioned in the paragraph above (which are also terms associated with phylogenetic tree structures), provided no information about phylogenetic classifications. Granted, this reference is a bit dated, but cladistics emerged in biology in the mid-1900s, and the phylogenetic tree is now the standard for biological classification of organisms, and the title of the chapter is Classification *Trees* (emphasis mine).

However, the lack of evidence in general computer science technical books may be due to phylogenetic algorithms only recently being tested in non-biological, computer science applications. For example, Fiorini, *et al.* (2016) used a phylogenetic algorithm in an information retrieval project, which will be discussed in the *Automatic classification in library and information sciences* section below (it was classified in the Web of Science ("WoS"; Clarivate Analytics) category 'computer science artificial intelligence').

Conclusion. From this section's very simple review, there are two main points. First is that computer science suggests nature's evolutionary process results in one of the most efficient ways to approach a computing problem (and organization and representation of information is, at least in part, a computing problem). And second, though there is relatively new evidence for the use of phylogenetic algorithms for text-based concept categorization within the computer science domain, there remains no indication that any sort of test of a phylogenetic classification for books based on common origin has been conducted.

Applications of phylogenetics to cultural phenomena

Phylogenetic techniques borrowed from evolutionary biology have been used in cultural research studies from the fields of archaeology, cultural anthropology, linguistics, and textual criticism. The literature review indicates these types of studies have been conducted at least since the 1990s in cultural studies. But as far back at 1977, Platnick & Cameron suggested historical linguistics and textual criticism could benefit from the cladistic method used in phylogenetic investigations.

Examples of projects in the anthropological fields in the 2000s include marriage and residence customs (Fortunato, 2011; Fortunato, 2011; Fortunato, *et al*., 2006; Mulder, *et al.* 2001; Cowlishaw & Mace, 1996); lithic technologies, (Prentiss, *et al.*, 2015; Buchanan & Collard, 2007; Lycett, 2007; O'Brien, *et al.,* 2001); basketry, cradles, ceremonial dress, and earth lodges (Jordan & Shennan, 2009); cloth and textiles traditions (Larsen 2011; Tehrani & Collard, 2009; Tehrani & Collard, 2002); musical instruments (Tëmkin & Eldredge 2007), contemporary utensils (Schillinger, *et al.*, 2016), and folktales (Tehrani, *et al.*, 2016; Stubbersfield & Tehrani, 2013).

Of the anthropological studies, the folktale studies are from text sources as books or similar to books, so there would be some relevance to my proposed research project using text from scientific and technical books. These will be discussed in the *phylogenetic research projects specific to <u>text</u> analysis* section below.

Phylogenetic research projects specific to historical linguistics analysis. To reduce the literature to phylogenetic (or cladistic) studies related directly to *literature*, the following search was performed in July 2017 using WoS:

- Search description: strategy: *phylogen**

- To include historical linguistics, the search was limited to the following WoS categories:
    - linguistics, language linguistics
- Limited to: article, proceedings paper, book chapter
- Limiting to: English language
- Results: 140 records

Scanning thorough some of the first results, I learned that phylogenetic analysis in *language and historical linguistic* studies has been mostly concerned with inferring histories of language families. The characters that have been used in the analysis of change over time have been lexical (words), morphological (word formation and relationship to other words), phonological (speech sounds), or syntactical (rules of sentence and phrase formation) features or combinations thereof (Cabrera, 2017, pp. 68-69; Dipper & Schrader, 2008, p. 39). Of these characters, only the lexical studies would be potentially relevant to my research due to my focus on words and phrases found in books. Searching 140 records for *lex** reduced the results to 34 records.

However, after scanning through the first results, I learned that lexical phylogenetic studies are predominately studies of cognates (i.e., variations of words that have common origins). To support this claim, de Schryver, et al. (2015) state:

> To account for lexical variation we predominantly proceeded as is customary in lexicostatistic studies, that is, on the level of cognates or words having a common etymological origin (p. 129)

With scientific and technical books published in the English language from 1956-2017, and for my dataset, I do not expect to have words changing over time, though there will likely be hypernyms (a more general word of a specific word), synonyms (different word, same

meaning), and possibly homographs (spelled same, different meaning depending on the academic domain). There will also be new words to emerge over time, such as new words introduced from other disciplines or fields, rather than variants of an existing word. Therefore, the historical linguistic studies using text are not directly relevant to my research project.

To cover my bases, I reran and returned to the original 140 records and further limited to include only WoS category: literature. Of the six (6) results, all were either not relevant or had been retrieved from other searching.

Phylogenetic research projects specific to text analysis. This search was presumed to be the most relevant for my proposed research project. To reduce the literature to phylogenetic (or cladistic) studies related directly to *text*, the following search was performed in July 2017 using WoS:

- Search description: strategy: (*phylogen** OR *clad**) & *text*
- To include humanities and social science subjects and exclude biological subjects, the search was limited to the following WoS categories:
    - computer science interdisciplinary applications, linguistics, language linguistics, history philosophy of science, literature, humanities multidisciplinary, computer science information systems, social sciences interdisciplinary, music, medieval renaissance studies, literature romance, information science library science, folklore
- Limited to: article, proceedings paper, book chapter
- Limited to: English language
- Results: 43 records

I learned from this search that phylogenetic text analysis is found most in stemmatic studies from the field of textual criticism, which are typically analyses of either copying or printing variations visible in extant medieval manuscripts. Of the 10 records associated with actual productions of phylogenetic trees, five (5) were stemmatic studies. Of course, as with the historical linguistics studies, I will not be analyzing copying or printing variations in the science and technical monographs. But the search uncovered two (2) folktale phylogenetic studies that will be useful (the other three were two linguistics studies and one historical biological study). Also important in the results are the *types* of phylogenetic analyses conducted and the software used for the analyses.

Stemmatics

Bergel, Howe, & Windram (2015) used phylogenetic analysis as part of a stemmatics project of variations in surviving copies of a printed English ballad (ca. 1450–1800). Maximum parsimony analysis using PAUP* (Swofford, 2003) software and network analysis using NeighborNet algorithm (Bryant & Moulton, 2004) in SplitsTree4 software (Huson & Bryant, 2005).

Windram, Charlston, & Howe (2014) used phylogenetic analysis to investigate copying variations in surviving copies of printed musical text (i.e., sheet music). Phylogenetic tree analysis with maximum parsimony using PAUP* software and phylogenetic network analysis using NeighborNet algorithm in SplitsTree4 software.

Roos and Heikkilä (2009) conducted a comparison experiment of existing thirteen (13) computer-assisted stemmatology techniques applied to three artificial, hand-written, manuscript datasets. The two phylogenetic software packages used were PAUP* (tree-based)

and SplitsTree4 (network-based). A method developed by the authors and the PAUP* maximum parsimony method was found to be consistently better in reconstruction of the stemma across the three datasets.

Windram, Shaw, Robinson, & Howe (2008) used phylogenetic methods to analyze extensively studied manuscripts and compare results to traditional stemmatology results. The researchers found that the stemma produced by phylogenetic methods were comparable to traditional stemmatology results. Phylogenetic tree analysis with maximum parsimony (MP) using PAUP* software, phylogenetic network analyses using NeighborNet algorithm and Supernetwork algorithm (Huson, Dezulian, Klopper, & Steel, 2004), the latter was within the SplitsTree4 software.

Eagleton and Spencer (2006) used phylogenetic analysis of medieval text variants to better understand conflation of the texts. The phylogenetic network was produced using the NeighborNet algorithm.

Folktales

Stubbersfield and Tehrani (2013) used phylogenetic analysis to investigate psychological biases in the transmission of a contemporary legend. Phylogenetic tree analysis with maximum parsimony using PAUP* software.

Tehrani, Nguyen, & Roos (2016) used phylogenetic analysis to investigate oral or literary origin of a fairy tale. Phylogenetic tree analysis with Maximum Parsimony using PAUP* software and phylogenetic network analysis using NeighborNet and T-Rex (Boc, Diallo, & Makarenkov, 2012).

Linguistics

von Waldenfels (2017) used phylogenetic network analysis to analyze use of prepositions within a corpus of texts in Slavic languages. Network analysis using Neighbor-Net algorithm/diagrams in SplitsTree software.

Dipper and Schrader (2008) conducted, in part, a phylogenetic analysis of German dialect data from medieval text variants. The researchers concluded that the methods used correctly illustrated the known distinction between two German language families. Phylogenetic network analysis using the Neighbor Joining method (Saitou & Nei, 1987).

Historical Biology

von Lieven and Humar (2008) used a phylogenetic method to ascertain if Aristotle's correlation of animal descriptions in his *Historia animalium* were based on a prior (unknown) classification. Though this was an analysis of a book, the character matrix used for the species was populated using anatomy and development descriptions in *Historia animalium*. Therefore this was more of a biological phylogenetic study than any other type of study. The researchers concluded a relatively consistent classification underlies Aristotle's work. Phylogenetic tree analysis with maximum parsimony using PAUP* software.

Final phylogenetic search. Finally, in July 2017, another search was conducted identical to the phylogenetic text analysis search presented above, with the exception of replacing 'text' with 'book' OR 'monograph'. To include humanities and social science subjects and exclude biological subjects, the search was limited search to the following WoS categories: history philosophy of science, anthropology, social sciences interdisciplinary, women's studies, sociology, philosophy, linguistics, humanities multidisciplinary, classics, art, archaeology. Of the 16 records, two (2) were relevant but had already been retrieved from the 'text' search above.

Conclusion. From this section's search, there is no evidence within anthropology, language and linguistics, library and information science, or textual criticism of a phylogenetic research project using an actual dataset of books printed after modern printing methods were developed, with the objective of classifying the books based on common origin.

One of the primary arguments against applying phylogenetic cladistic (tree) methods (as opposed to phylogenetic network methods) to cultural phenomena is that cladistics methods assume the characters changing over time in a dataset happened by vertical transmission (transmission from parent to offspring). But culture is transmitted vertically, horizontally (by peers), and obliquely (from an older generation). So any cultural dataset used in phylogenic analysis contains the possibility that non-vertical transmission has occurred.

Non-vertical transmission is also present in biological datasets. Rivero (2016) states "horizontal transmission between genomes and bacteria... and high rates of horizontal transmission between many species and families of plants and animals" is known in the literature (p. 57). So horizontal transmission is not limited to cultural phenomena, but it is much more prevalent in cultural living systems due to humans 1) controlling the cultural living system processes and 2) moving in and out of human groups, organizations, communities, etc. Biology has responded with the creation of phylogenetic network applications, which have also entered into cultural phylogenetic studies (as noted in the 'phylogenetic research projects specific to text analysis' section above).

In anthropology, the topic of non-vertical transmission is often mentioned in the same context as 'Galton's problem'. In the late 1800s, Francis Galton (Darwin's cousin) raised this issue in a response to anthropologist E. B. Tylor's discussion of correlations of cultural

phenomena across societies. Galton suggested “any functional explanation for why two traits are correlated across a number is vulnerable to the possibility of that those societies may not be independent, because they may share a common history” (Mesoudi, 2011, p. 95).

Cultural phylogenetic researchers have dealt with this in at least two ways. Some have specifically sought a cultural group and/or a specific cultural phenomenon for phylogenetic study in which the group and/or phenomenon is very likely to have been isolated from other cultures for that particular phenomenon. For example, Larsen’s (2011) phylogenetic analysis of Polynesian bark cloth production in Polynesia, Pohnpei, Fiji, and Indonesia. Others have used ‘mapping’ to overlay a cultural phenomenon onto a recognized genetic tree of a group of people. For example, Cowlishaw and Mace’s (1996) phylogenetic analysis of cultural groups with known marriage and wealth customs, which were then mapped onto a previous, accepted language phylogenetic trees due to evidence that suggests “language phylogenies of human populations do correspond broadly to those produced on the basis of genetic evidence” (p. 89). And contemporary cultural phylogenetic studies provide phylogenetic network analyses and some even provide both a tree and network from the same dataset (several examples of providing both are found in the ‘phylogenetic research projects specific to text analysis’ section above; also see Tehrani, 2013).

Collard, Shennan, & Tehrani (2006) provides an interesting finding related to cultural phylogenetic datasets. The researchers obtained 21 datasets of biological data (i.e., animal mtDNA, morphology, behavior) and 21 datasets of cultural data (i.e., material culture/artifacts, practices, beliefs). The team then used the PAUP* 4 phylogenetic software package to analyze each dataset’s Retention Index (RI), which is a measure of the fit between the data and the

tree. “An RI of 1 indicates no homoplasies and a perfectly treelike evolutionary pattern, with lower RI values increasingly less treelike” (Mesoudi, 2011, p. 101); in other words, “an RI of 1 indicates that all similarities can be interpreted as shared derived traits, without requiring additional explanations, such as losses, independent evolution or borrowing” (Tehrani, 2013, p. 9). Surprisingly, the average biological RIs was 0.61 and the average cultural RIs was 0.59, which indicated that the cultural datasets were just as likely to be bifurcating trees as the biological datasets. Other high RI values from cultural datasets have also been cited. For example, Tehrani (2013) reported an RI of 0.72 for a folktale dataset.

Conclusion. For my proposed project, there is no question that Galton’s problem should be an issue: scientific and technical books have numerous citations—clear evidence of non-vertical transmission. I am not concerned about non-vertical transmission because for this project, the objective is to provide evidence that an evolutionary book classification is possible; the objective is *not* to reconstruct a literal phylogeny of books. Rather, it is to use a phylogenetic approach to learn if 1) a set of important words/phrases in books 2) can be considered input into a phylogenetic software package, 3) which was developed on the underlying principle of decent from common origin) 4) which will create a logical classification, preferably tree-like in structure.

Automatic classification in library and information sciences

A successful research project may also demonstrate a proof-of-concept of a novel automatic classification method for books (and possibly other text-based documents). As with phylogeneticists’ use of trees, the tree produced in this project would be considered the classification. Therefore, a review of the literature automatic classification methods is needed

to distinguish the contributions of this information science dissertation project to the overall discipline/field. This type of review is also needed due to the importance of automatic processes that 1) enable information researchers and professionals to be more efficient and effective and 2) enable information users to obtain the desired information. In other words, automatic information organization methods are commonly used in library and information sciences. Desale and Kumbhar, 2013 state:

> Many library professionals believe that automatic classification will help in classifying more effectively, quickly, and accurately. Due to the information explosion… classification schemes are becoming bulky and thereby expensive and unmanageable. Library professionals have invested their time in designing automatic document classification schemes as they help in standardizing the classification procedure. Standardization of classification helps in constructing uniform class numbers, which further helps in locating pinpointed information and documents (p.295).

And with the global growth and diversification of science and technical research suggested in chapter 1, the need for automatic classifications will likely continue to increase, and there may even be the emergence of a *variety* of automatic library classifications, depending on the needs of information users. As with the above scholarly evidence connecting information organization to evolutionary classification section, most of this section's literature search strategy will also be presented in the chapter Appendix due to the importance of the topic for this dissertation.

General types of automatic categorization techniques. There are four general types of automatic categorization of text-based information packages. Smiraglia and Cai (2017) claim many of the techniques used in knowledge/information organization are from the computer science domain:

> A very important extension of the traditional domain of knowledge organization… arises from attempts to incorporate techniques devised in the computer science domain for automatic concept extraction and for grouping, categorizing, clustering and otherwise organizing knowledge using mechanical means (p.215).

The authors conducted a review of these automated techniques most used in the knowledge organization domain, which includes automatic indexing, machine learning, automatic classification, and clustering (Smiraglia & Cai, 2017, p. 216). *Automatic indexing* is a computer's use of a controlled vocabulary to index large amounts of documents. One way to understand automatic indexing is as a precursor to the more advanced machine learning techniques used today in knowledge organization. In general, automatic indexing creates a list of most frequent words (i.e., descriptors) in single or multiple information packages. This can be accomplished with oversight by a professional or by fully automatic methods. *Machine learning* is the use of computer algorithms that enable a computer "to automatically learn and improve from experience without being explicitly programmed" (Jmila, Khedher, & El Yacoubi, 2017, p. 884). Supervised machine learning uses external information (e.g., human feedback or labeled text) to guide the 'learning' algorithms and unsupervised machine learning does not use external information. *Automatic classification* is a type of supervised machine learning that either assigns an information package to an existing classification scheme or creates a scheme, then makes the assignment. *Clustering* is a type of unsupervised machine learning whereby the computer algorithms create distance measures among the information packages and groups (i.e., clusters). To clarify, methods using unsupervised machine learning is what most would consider fully automatic categorization.

Most of the research projects related to automatic book classifications utilize machine learning methods (both supervised and unsupervised), which have become powerful in information organization and representation. Ibekwe-Sanjuan and Bowker (2017) state that machine learning algorithms have become very proficient at providing humans with relevant information due to the big data phenomenon and the response by various organizations to extract meaningful information from very large amounts of data. Specific examples the authors provide included, in part, are the learning recommender systems developed by Amazon and Netflix. "While not infallible, these [types of] algorithms have attained a level of performance that is acceptable to humans... [and now] provide users with suggestions and recommendations that can rival those of a human librarian or knowledge organization specialist" (p.188).

Much of the automatic categorization using text-only (rather than metadata) is to automatically identify subjects/topics and assign documents to pre-defined classes or categories using the textual content of the document. Machine learning algorithms are commonly used and typically require manual classification of a corpus (i.e., selection and manual labeling of documents), which is used to train an automated classifier, which is then used to classify/categorize unlabeled (i.e., new) documents for inclusion into the corpus (Joorabchi & Mahdi, 2011, p. 2; Busagala, Ohyama, Wakabayashi, & Kimura, 2012, p. 43; Desale & Kumbhar, 2013, p. 297).

Automatic classification. Specific to this dissertation's research project would be automatic book classifications, but the phrases "automatic book classification" and "automatic bibliographic classification" are virtually absent in scholarly literature. For example, a topic search in WoS conducted in November 2017 using the search strategy for the exact phrases

automatic bibliographic classification OR *automatic book classification* produced only four (4) records. The years of publication include 2002, 2009, 2012, and 2013 and all were within the Information Science & Library Science WoS subject category. (These four results have either been cited or were considered for use in this section's review.)

However, "automatic classification" is a well-known phrase. Over 3800 records were retrieved using the exact phrase "automatic classification" in a November 2017 WoS topic search, much of which is from the computer sciences and engineering literature. Within the Information Science & Library Science WoS category, 117 records were retrieved, which represents only three percent (3%) of the total (which is a simple confirmation of Smiraglia and Cai's (2017) claim mentioned above that many of the techniques used in knowledge/information organization are from the computer science domain). Limiting the broader 'automatic classification' 3800 records to topics including *book** OR *monograph** reduced the number of records to 18 records. The following is a short summary of these 18 results.

Eight (8) records were in the Information Science & Library Science WoS category, with a date range of these publications being 1995–2014. Six (6) of these publications were related to assignment of existing library classification scheme codes using book metadata. The other two publications were related to analysis of bibliographic references and categorization of social media 'tags' given to books. Of the remaining 10 records (generally from computer sciences and engineering literature), the date range of these publications was 1996–2016 and only one was related to an automatic classification using text from books: a categorization method based on sentiment (i.e., feelings expressed in the books; Bisio, Meda, Gastaldo, Zunino, & Cambria,

2016). The other nine (9) were related, in whole or in part, to categorization of book review sentiments, creation of a list of categories based (in part) on an existing textbook's index, classification of web information based on pre-existing classes, classification of defect types in scanned documents, classification of mobile phone contacts, classification of web services, clustering based on titles of books, classification of web documents into user- or community-specific topics, and the organization of bookmarks (i.e., web URLs).

In summary, the above search suggests most of the research and development of automatic classifications for books is the automatic assignment of books to existing library classification scheme codes using existing book metadata rather than the entire text. The automatic sentiment classification method paper was published in 2016, so it is too early to determine if this research path for books will continue, but sentiment analysis is very interesting and could be valuable component in an evolutionary classification of fictional books.

Automatic classification using existing library classification schemes. Most of the well-known library classification schemes—Colon Classification (CC), Dewey Decimal Classification (DDC), Library of Congress Classification (LCC), and Universal Decimal Classification (UDC)—have been used in automated classification research projects. In the 1990s, the Online Computer Library Center (OCLC) was engaged in automated classification of electronic documents using the DDC (Thompson, Shafer, & Vizine-Goetz, 1997), though webpages for both the Scorpion project (Online Computer Library Center [OCLC], n.d.) and Automatic Classification Research at OCLC (OCLC, n.d.) are currently labeled as 'closed'. Similarly, the GERHARD (German Harvest Automated Retrieval and Directory) automatic classification technology used a version of UDC to automatically classify German webpages (Carstensen,

Diekmann, & Möller, 2000). And as with the SCORPION project, the webpage for the GERHARD project provided by the authors is no longer active. OCLC does have a currently active research classification prototype named Classify (OCLC, n.d.) that helps users to classify books, magazines, movies, and music using the DDC system or the LCC system. This is a recommender system that requires input of a standard number (ISBN, OCLC#, UPC, or ISSN), title, author, or subject heading from an OCLC controlled vocabulary. The service then provides information regarding the classification numbers that others have used. Experimenting with a few titles (e.g., Selfish Gene, Wonderful Life), it is interesting to see the various classes that are used, which clearly illustrates the fact that catalogers use different classification notations for the same book.

Frank and Paynter (2004) used a machine learning method using Library of Congress Subject Headings from metadata of a virtual library's records to automatically assign an information package to a LCC notation. Kim and Lee (2002) created an automatic classification for use with book titles to produce the five (5) facets of CC (i.e., personality, matter, energy, space, and time). A knowledge base was first designed to enable the automatic classification. Similarly, Panigrahi and Prasad (2007) created a method for identifying the position of each of the five CC facets for use with an automatic classification system using document titles and a pre-existing knowledge base. Wang (2009) used bibliographic metadata (title and subject) and a supervised machine learning approach experiments with automated assignment of DDC classes.

Automatic topic identification. Text-based (as opposed to metadata) automatic topic identification methods can often be found in machine-based discourse analysis research. In general, the automatic techniques used "inspect the content of text from syntactic or lexical

perspective and attempt to uncover the hidden information within discourse structure of text" (typically referred to as 'topic segmentation'), which utilizes "multiple sentences or paragraphs of text" (Guo, Wang, & Lai, 2015, p. 4). These studies typically use pre-classified books or documents to determine how well the automated method can classify based on some unique approach. For example, Guo et al. (2015) developed a method for automated classification by identifying discourse segments and subtopics within electronic books using supervised machine learning. The researchers used 125 books, which were pre-assigned to one of five classes: medical science, agriculture, animal, computer, and geography. Similarly, Demarest and Sugimoto (2015) also used machine learning in a text discourse analysis project aimed at distinguishing between dissertation abstracts in pre-assigned disciplines of philosophy, psychology, and physics. Osborne, Salatino, Birukou, & Motta (2015) discuss the Smart Topic Miner developed to assist editors at Springer Nature. The tool analyzes metadata (title, abstract, authors' keywords, section title, and book title) of publications in a collection and provides, among other analytics, a hierarchical taxonomy of topics, which can be used in classification of proceedings and other literature.

Automatic text classification research project utilizing phylogenetic algorithm. One project was identified that is the closest to this dissertation's project. Simply stated, Fiorini, Harispe, Ranwez, Montmain, & Ranwez (2016) use a phylogenetic algorithm to organize text-based information packages, the results of which were reportedly represented in a tree-based classification, which is the general plan for my project. That research provides recent scholarly work supporting this dissertation's type of project, but there are also very distinct differences between the two projects.

First, the fundamental objectives of the projects are different: Fiorini et al.'s (2016) objective is improved information retrieval and my objective is knowledge organization—i.e., a sensible classification based on common origin. Second, the researchers use existing conceptual annotations (also known as 'semantic indexing') as the term sets in their project. Semantically annotated documents are difficult to create, as the authors acknowledge:

> Semantic indexing is as tedious as complex: a synthetic and relevant semantic annotation requires a good understanding of the subject area the documents refer to, as well as a deep familiarity with the chosen knowledge representation (pp. 133-134).

In contrast, this dissertation project uses a relatively simple approach to term-set creation: a term-weighting technique (discussed in chapter 3), which means there is no need for subject matter experts to be involved with term-set creation for the individual information packages. Third, the researchers' use of semantic annotations as the proxy for information packages is valuable for an information retrieval (IR) project due to IR's underlying goals of high retrieval relevance. In contrast, for a text-based evolutionary classification, retention of the actual words/phrases of books without any conversions of terms (e.g., no conversion of synonyms among several information packages to a generic term) is valuable due to one of the goals of evolutionary analysis is change over time: the fact that words emerge, change, and go extinct based on the cultural environment, all of which is critical to any analysis of text-based cultural evolution. Finally, the researchers use a phylogenetic approach using a distance matrix method and this dissertation's project suggests a parsimony method. In general, distance methods attempt to "minimize the distortion between the matrix of observed distances and the set of distances that is induced when all [operational taxonomic units] are assigned to specific nodes in the tree…"and parsimony methods attempt "to minimize the amount of

evolutionary change that is needed to explain a particular tree" (Mushegian, 2007, p. 154). As will be seen in Chapter 4, parsimony analysis was a better choice for knowledge organization than distance-related analyses.

Issues with metadata only as data source. As the literature referenced above indicates, the use of only metadata to classify books is a popular choice for research projects related to automatic classifications. But this approach has some weaknesses if using existing classification schemes, such as LCC or DDC. For example, multi-, inter-, and trans-disciplinary information has the inherent characteristic of multiple scientific subjects rather than one. (This is where CC has an advantage over LCC and DDC due to having multiple facets to describe/classify an object, but even CC is not immune to subjectivity due to humans creating the facets.) Suominen and Toivanen (2015, p. 2466) and Toivanen and Suominen (2014, p. 557) also suggest automated classification methods using a proxy such as metadata, including predefined categories, to represent the entire information package are limited when anticipating scale-up beyond research projects (e.g., diminished representation accuracy).

Additionally, culture in general, and science in particular, is constantly evolving, more so now than at any other time in history due to the number of people involved in science and technical fields and the speed at which new ideas transmit. Attempting to fit new science into preexisting categories is contrary to the way science proceeds:

> Preexisting categories of science provide a finite definition of new knowledge, fitting knowledge that is by definition infinite and new to the world into preexisting categories… They are best at monitoring the behavior of known and defined bodies of knowledge, but lend themselves poorly—if at all—to correctly identifying the emergence of truly new epistemic bodies of

> knowledge. In short, human-assigned subject categories are akin to using a rearview mirror to predict where a fast-moving car is heading (Suominen & Toivanen, 2015, p. 2464).

Though Suominen and Toivanen (2015) were making the above statement within a science mapping project that utilized journal articles as the data source, both journal articles and books represent culturally defined areas of science and technical domains and both books and journal articles can be indicators of changes in science and technology over time. Therefore, the need for subject category flexibility applies to classifications of both books and journal articles.[19]

Advantages of text-mining. The developments in text-mining—the identification of words and phrases that summarize the content using unique techniques to analyze co-occurrence[20] of these words and phrases—"have made taking advantage of semantic text a practical approach… Novel text-mining methods create value by being able to create practical categories directly from semantic text, rather than using preordained categories, keywords, or citations" (Suominen & Toivanen, 2015, p. 2466). This dissertation's knowledge organization research project takes advantage of this value with the use of Oak Ridge National Laboratory's Piranha information retrieval text-mining technology (Klump et al., 2010)—specifically, the technology's word-/phrase-generating component—to reduce the semantic text in books to a weighted list of important words/phrases (i.e., terms). In short, the tool's weighting algorithms are based on occurrence frequencies both within each book and across the entire corpus. The

[19] Arguably, science and technical books represent an even stronger defined science or technical area due to books often representing a synthesis of recent literature based on trend perceptions by publishers and authors. If true, science and technical books may be able to provide a more accurate forecast of trends in contrast to cutting edge journal articles.

[20] For this dissertation's project, co-occurrence is understood to mean the occurrence of the same words/phrases in both a book's term set and a corpus term set.

terms generated by the text mining tool are then used as input into a phylogenetic software package to create the evolutionary tree classification.

Due to advances in text-mining tools and computational power that enable humans to analyze entire text corpora for topic discovery for large-scale classification needs, it seems using anything less that the entire text (i.e., metadata only) is not taking advantage of all that is available for automatic text analysis and classification. Even if the use of metadata produces similar results in a research project, there remains an intuitive aspect (at least with researchers) for using the most reasonably attainable data to produce the most accurate results for a scientific study. Consider this data dichotomy as analogous to someone conducting research with a survey: if an entire population of interest is small enough to be surveyed, why not use the entire population, even though a random sample would produce very similar results? Or, more applicable to this dissertation, consider a biologist using only morphological information about animals to produce an evolutionary classification when DNA is available for analysis—the morphological information may provide a very similar (or even the same) classification, but the use of DNA is considered more accurate, to the point that it is now considered standard use when available (e.g., there is no DNA with fossils). Of course, this is not to suggest that metadata cannot be used even if the full text is available. For example, the date of first publication could be used with a project similar to this dissertation's project.

Conclusion. Most of the automatic classification of books within the information and library science literature utilize metadata, existing library classification schemes, or combinations thereof. I could not locate any that use the entire text of a book for an unsupervised automatic classification of books. The closest was Guo et al. (2015) discourse

analysis using 125 books in a *supervised* automatic classification project using pre-assigned classes. However, text-mining technologies have recently been used in an *unsupervised* automatic classification project related to science documents (Suominen & Toivanen, 2015), and with today's computational power, text-mining-based automatic classifications are possible, even with large book collections. Finally, though a computer science automatic text classification research project was identified that used a phylogenetic algorithm to create a classification tree of semantically annotated information packages, distinct differences were discussed that makes this dissertation's project unique.

Based on the above review, the notable aspect of this dissertation's research project that distinguishes it from the prior literature in automatic book classifications in information and library science includes:

- A classification based on contemporary evolutionary theory: descent with modification from a common origin.
- Unsupervised machine learning. The classification tree produced is a type of clustering machine learning technique. Therefore, there is no labeling of training documents and no subject matter experts are required.
- Relatively simple term-set creation. The term-weighting technique does not rely on complex semantic annotations of the information packages to be classified.
- The data source is the actual text of book chapters rather than using a proxy for the book such as title, abstract, and/or other metadata.
- Easily adaptable to any text-based science/technical collection. By utilizing a text-mining technique, there is no need for predefined classes and no need for preexisting,

controlled vocabulary because the text-mining technology creates its own taxonomy directly from the corpus.

A note about classification trees

The dissertation project will present a book classification in the form of a dendrogram (i.e., a branching 'tree'). A dendrogram is a diagram for presenting hierarchical categorizations and classifications. It is common in computer science literature and less common in information and library science literature. As mentioned previously, a chapter titled *Classification Trees* was observed in a data mining handbook (Rokach & Maimon, 2009). And from over 3,000 records retrieved from a November 2017 WoS topic search for the exact phrases "classification tree" OR "classification trees", only 13 were in the information science / library science category (the next closest WoS category to information and library science is computer science information systems category, which contained over 200 records).

An example of trees used in information and library sciences is Julien, Tirilly, Leide, & Guastavino (2012), who used the Library of Congress Subject Headings (LCSH) along with metadata (specifically, MARC bibliographic field 650: topical subject, added entry) from a collection within the McGill University Libraries to create a hierarchical tree of science and engineering books. Though the authors do use terms such as 'parent', 'child', and even 'common ancestor', these terms were used in the context of commonly known hierarchical tree structure rules rather than in an evolutionary or a phylogenetic context.

Of course, the idea of a tree to represent a book classification is not new. Figure 7 contains an image of a tree to illustrate the classification for the Library of Congress that Thomas Jefferson created (Library of Congress, 1900). But there is a very real difference

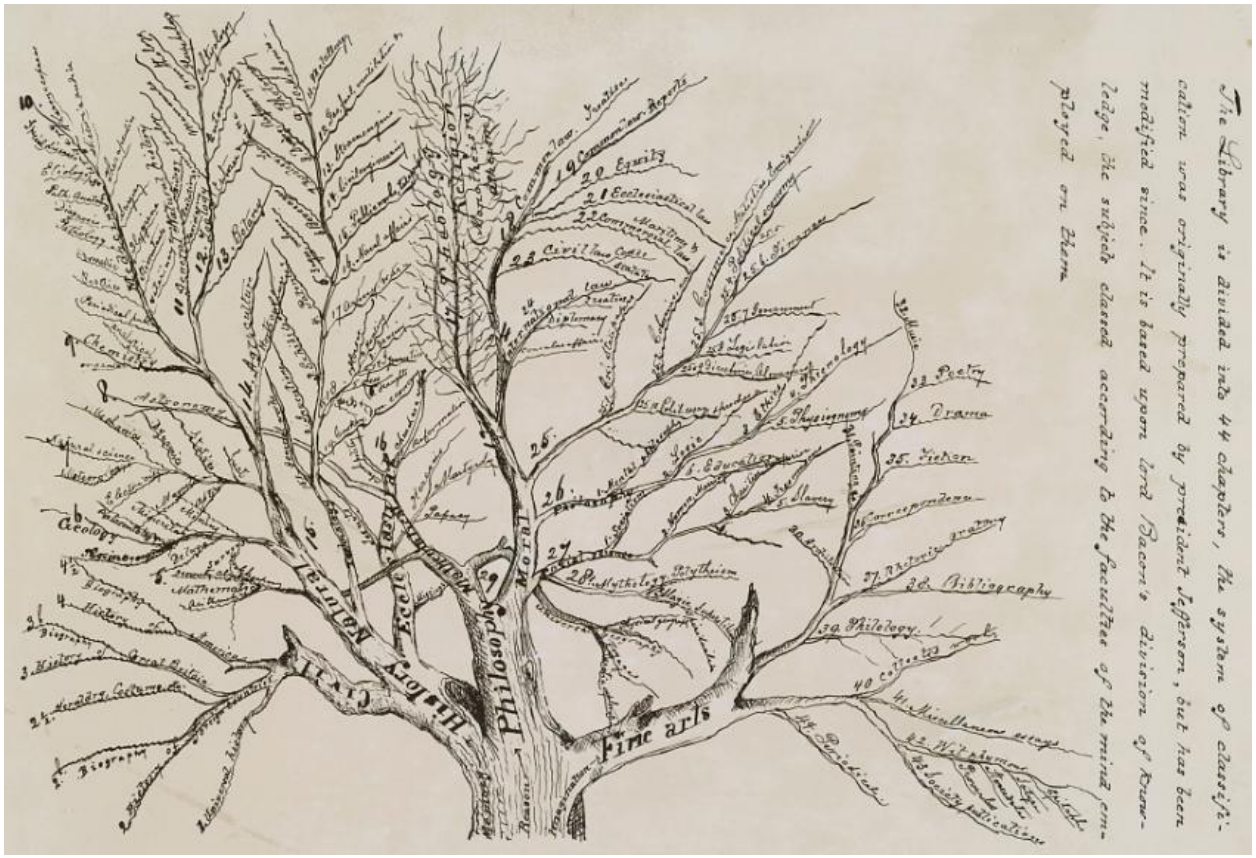


Figure 7. Classification Tree Drawing by President Thomas Jefferson to Illustrate Library of Congress Classification.

Photograph (circa 1900). The writing on the right side of the image is: “The Library is divided into 44 chapters, the system of classification was originally prepared by president Jefferson, but has been modified since. It is based upon Lord Bacon's division of knowledge, the subjects classed according to the faculties of the mind employed on them.” Courtesy of the Library of Congress, LC-DIG-ds-09241.

between the use of a branching pattern to represent a hierarchical structure of organized information packages and the use of a method to produce a branching pattern that is explicitly based on algorithms for use in analyzing possible common origins of information packages. The research project of this dissertation is an application of the latter.

DETAILED SEARCH DESCRIPTION

Detailed search description for tems "evolution" and "phylogenetics" in the library and information sciences literature. A broad Web of Science (WoS) search for *evolution** as a 'topic' resulted in almost 1.3 million records[21] as of July 12, 2017. Within research areas of 'information science & library science' there were 3,775 records or about 0.3% of the almost 1.3 records. The following show the number of results when limiting the 3,775 'information science library science' results:

- 252 records when limiting to *classif** OR *taxonom** OR *ontolog**
    - 9 records when limiting the 252 records to *biolog** (which would include biology and biological). One was relevant to the use of a biological evolutionary classification approach to literature classification.
    - 4 records when limiting the 252 results to *tree**. Only one was relevant to the use of a biological evolutionary classification approach to literature classification (Gnoli, 2006), though two records indicated the use of biological-based evolutionary computational techniques.

[21] Actual number: 1,293,014

    - 5 records when limiting the 252 results to *phylogen**. Four were relevant to a biological evolutionary classification approach to literature classification, the most relevant being Gnoli (2006).
    - 1 record when limiting the 252 results to *cladist**[22]: (Gnoli, 2006)
- 4 records when limiting the search to the exact terms *dewey decimal OR DDC*. No records of relevance—the use of 'evolution' is used only as a reference of change over time.
- 7 records when limiting the search to the exact terms *library of congress* OR *LCC* OR *LC*. No records of relevance—the use of 'evolution' is used only as a reference of change over time.
- 9 records when limiting the search to *phylogen** (which would include phylogenetic, phylogenetics, phylogeny, and phylogenies). Five were relevant to the use of biological evolution within the context of information or literature analysis, and Gnoli authored two (Gnoli, 2017, Gnoli, 2006) and co-authored one (Gnoli & Ridi, 2014).
- 1 record when limiting the search to *cladist**: Gnoli (2006).
- 5 records when limiting to *lineage**. Though none were relevant to biological evolutionary classification approach to literature classification, one article was interesting in that Georges (2017) investigated similarities and differences among classical music composers using existing data sources about the composers' personal musical influences coupled with 'ecological' characteristics such as "time period, geographical location, school association, instrumentation emphases, etc." (pp. 26-27).

[22] Cladistics is research approach used in biology to hypothesize about evolutionary relationships among species, based on shared, derived traits from a common ancestor.

This approach did not use either sound recording or musical scores in the analysis. The author considered the research to be early foundational work that could lead to an eventual phylogenetic classification of composers.

- 2 records when limiting to the exact phrase *cultural transmission*. One was tangentially relevant as a phylogenetic analysis of various stories obtained from internet web sources related to a particular type of legend (Stubbersfield & Tehrani, 2013).
- 0 records when limiting search to PAUP*, a popular phylogenetic tree analysis software package for inferring evolutionary relatedness.
- 12 records when limiting to the exact phrase *cultural evolution* and none were relevant.

In summary, there were only five (5) records relevant to the use of a biological evolutionary classification approach to literature classification that were labeled as research areas of 'information science & library science'. Of these, one was a phylogenetic classification project (Stubbersfield & Tehrani, 2013), which was published in the journal *Social Science Computer Review*. The other four (4) included conceptual discussions related to phylogenetic classifications.

More specific to this dissertations research project, a narrower WoS search for *phylogen** (which would include phylogeny, phylogenesis, phylogenetics) as a 'topic' yielded over 225,000[23] records as of July 12, 2017. Within research areas of 'information science & library science' there were 25 records:

- 17 results were related to a biological use of the *phylogen** variants;

[23] Exact number 225,696 as of July 16, 2017.

- 2 results were from the same letter to the editor of a journal whereby the authors were arguing specific points from their previous articles rather than presenting research.
- 1 result included the report of a research project related to representation and visualization for recommender systems. This is interesting due my proposed research project's use of Oak Ridge National Lab's *Piranha* technology, which is a recommender system. However, though the trees that were produced in the project gave the appearance of phylogenetic trees, the underlying algorithms were not true computational phylogenetic algorithms—the authors specifically state that trees created by recommender systems do not include the "inheritance relationship" aspect that true phylogenetic trees have (Hernando, Moya, Ortega, & Bobadilla, 2014, p. 98). Nonetheless, the authors demonstrate how a phylogenetic-type tree is useful for visually representing search results.
- The remaining five (5) were the same that were found in the broad *evolution** search.

This search resulted in one more relevant article (Hernando et al., 2014), for a total of six.

Of these six, a phylogenetic classification project pertaining to the cultural transmission of legends was reported by Stubbersfield & Tehrani (2013), which will be a reference source for my project, but it was published in the journal *Social Science Computer Review*, which is not a traditional information science journal so this is not a direct connection with library and information sciences.

Of the five traditional information science journals, the focus of Hernando et al. (2014) was representation and visualization for recommender systems. The graphs produce bifurcating

trees (which are easy to interpret), which give the impression of a phylogenetic classification. But the graphs utilize recommender system information, which is based on collaborative filtering that generates similarity metrics and measures between users or items and were not based on computational phylogenetic algorithms that identify evolutionary relationships between entities. Thus there is only the shape of a phylogenetic tree, rather than an actual cladistic-based phylogenetic tree.

Of the four published in library and information sciences sources**,** Dousa's (2010) paper presentation was a more philosophical treatment of E.C. Richardson's classification theory than Dousa's 2011 article that focused, in part, on Richardson's specific evolutionary ordering principles used in this dissertation due to the latter being more directly relevant. The remaining three publications were the most relevant library and information sciences sources—Gnoli 2017, Gnoli & Ridi, 2014, and Gnoli 2006). Clearly, Claudio Gnoli is the most relevant researcher in the library and information science literature pertaining to the type of classification approach (i.e., phylogenetic) presented in this dissertation.

Detailed search description for terms related to "automatic classification" in the library and information sciences literature. In October 2017, a series of WoS searches were conducted for the *Automatic classification in library and information sciences* section of the literature review, the results of which are presented below.

A search for the exact phrases *automatic book classification* OR *automatic bibliographic classification* resulted in only four (4) records, three (3) of which were used in the 'automatic classification' literature review section (one had outdated material and was too general, though the author was an author of a more updated article retrieved and used from this search).

Broadening the search strategy with (*automat** OR *unsupervised*) AND (*biblio** OR *book** OR *monograph**) AND *classif** produced 388 results. Limiting to WoS Categories *Library Science Information Science* and *Computer Science Information Systems* narrowed the results to 159, and further limiting to the last 10 years (2008 - 2017) reduced the results to 102. Of those, nine (9) had some relevance to automated classifications for books or documents, which included three (3) of the four (4) that were duplicates from the aforementioned search. Of the additional six (6) records, all were used in the 'automatic classification' literature review section.

Broadening the search even more to include text, documents, and categorization, I used the strategy (*automat** OR *unsupervised*) AND (*biblio** OR *document* OR *text**) AND (*classif** OR *categor**). This produced 13,024 records. Limiting the search to 'review' articles reduced the records to 191, and further limiting the search to WoS Categories *Library Science Information Science* and *Computer Science Information Systems* during the last 10 years (2008 - 2017) narrowed the results to 16 records. Of these, one was a duplicate from another search, and the remaining did not have any significant contribution to the discussion topics created from the literature obtained in the more narrow searches above.

Limiting the 13,024 results to those containing *phylogen** OR *clad** produced 19 records (none in *Library Science Information Science* or *Computer Science Information Systems* WoS Category). Of these, one was directly relevant to my project due to the researchers' use of a phylogenetic algorithm for text-organization represented in a tree-based classification (Fiorini, Harispe, Ranwez, Montmain, & Ranwez, 2016). This article is discussed in the *Automatic classification* literature review section. Of the non-relevant records retrieved, most were

biological phylogenetic articles not related to automated bibliographic, document, or text classification or categorization. Two projects used text-mining or text categorization techniques to assist with dataset building or classification, but not to create actual phylogenetic trees. Other non-relevant articles included 'automat' as a reference not relative to classification or categorization and the use of 'text' references from 'texture' (not words) or used as general reference to scholarly texts. And one used ‘phylogenetic’ in the context of ‘resembling’ a phylogenetic tree.

Finally, because ‘common origin’ is a fundamental principle mentioned by the library and information science classification theorists in the ‘Scholarly evidence connecting information organization to evolutionary classification’ section above, a search was conducted to explore any relevant ‘common origin’ literature. The first search strategy (*automat** OR *unsupervised*) AND (*biblio** OR *book** OR *monograph**) AND *classif** AND (“*common origin*” OR “*common ancestor*”) produced no records. The second search strategy (*automat** OR *unsupervised*) AND (*biblio** OR *document* OR *text**) AND (*classif** OR *categor**) AND (“*common origin*” OR “*common ancestor*”) produced one (1) result, but the use of ‘common ancestor’ was as a general comment in the opening sentence of the abstract and not relevant to any type of automated/unsupervised book, document, or text classification.

CHAPTER 3 – MATERIALS & METHODS

INTRODUCTION

My primary research question is: *Can a phylogenetic-based method be developed to produce a sensible classification of books with only words or phrases used as the characters for analysis?* As the literature review suggested, there does not seem to be an existing research project utilizing text extracted from books as input into a phylogenetic tool to create an evolutionary classification to test the common origin principle suggested by theorists. To answer the primary research question, a proof-of-concept test with a limited number of books is first needed to test the efficacy of this approach. Similar to Charles Cutter and Ernest Richardson's attempts at an evolutionary classification mentioned in the literature review, scientific books are good candidates for this project. The reason for this limitation is that science and technical disciplines/fields use very specific words and phrases that are often contained within the disciplines/fields, which should increase the probability of success for the project. Additionally, the books are presented in a consistent, structured, digital form, which is conducive to computational tools for data extraction.

However, Cutter and Richardson were relying on evolutionary concepts such as *general-to-specific* and *simple-to-complex* to form the foundations of their evolutionary classifications, which, as mentioned in chapter 2, could be successful with science. Though this dissertation's project also uses science books, the analysis is based not on *general-to-specific* and/or *simple-to-complex* but rather on the perceived evolutionary relationships based on the unique words

and phrase used in the science and technical books and produced by algorithms based, in part, on *principles of common origin*.

RESEARCH DESIGN

The design needed to answer the primary research question is a phylogenetic analysis producing a tree-like, hierarchical classification. The design is modeled after the cladistic phylogenetic analytical method is attributed to German biologist Willi Hennig (Hennig, 1966) and is an analysis of relatedness based on similar characters among different species. The cladistic method has been considered by some scholars to be a hypothetico-deductive science (Platnick & Cameron, 1977) whereby the resulting tree that emerges from a cladistic project is a hypothesis of the relationships of the taxa and from which true observations can then be made—therefore the trees can be falsified. But this claim of falsification is not without its critics (e.g., Vogt, 2007).

The classification tree produced by this project is considered a *dendrogram* rather than a *cladogram* or a *phylogram*[24]. All three types of trees suggest phylogenies (evolutionary histories of the taxa), but only the latter two are considered actual scientific *hypotheses* of a phylogeny and the former is more of an *exploratory* phylogeny. For example, using another text-mining tool to extract terms from the same set of books could result in a different classification tree.

The research project method broadly followed this sequence of steps: 1) select sample of books, 2) obtain important terms from each book, 3) convert the books and terms into a

[24] In addition to evolutionary histories, a phylogram also suggests evolutionary time between the related taxa and a cladogram does not.

matrix for input into the phylogenetic software, and 4) run the phylogenetic analysis. A diagram of the full method developed during the project is presented at the beginning of chapter 4.

POPULATION AND SAMPLE

Population

The *operational taxonomic unit (OTU) of analysis* for this phylogenetic classification project is individual books. The Wiley Online Books collection includes digitized science and technical books with Main Subject Categories as follows: Agriculture, Aquaculture & Food Science; Business, Economics, Finance & Accounting; Chemistry, Earth, Space & Environmental Sciences; Humanities; Life Sciences; Mathematics & Statistics; Medicine; Nursing, Dentistry & Healthcare; Physical Sciences & Engineering; Psychology; Social & Behavioral Sciences; and Veterinary Medicine. These are further subdivided with Specialized Subject Areas. Using a collection of previously digitized books obviously eliminates the need for digitization of printed books. The list provided to me in July 2017 from the publisher included 19,562 titles with print publication years 1936 - 2017.

A down-select was obtained for the Specialized Subject Area "Earth Sciences," which was within the broader Main Subject Category "Earth, Space & Environmental Sciences." Other than reducing the number of books to analyze, the primary reason for a down-select to include a specific sub-category was due to the assumption that if an evolutionary classification cannot be constructed from a collection of books that have already been categorized more narrowly twice (i.e., from the Wiley Online Books collection → Earth, Space & Environmental Sciences → Earth Sciences), then it would be difficult to argue an evolutionary classification for an even more diverse collection could be accomplished. Also, having 18 undergraduate hours in geology

was also a reason for selecting the earth-science books: by being more familiar with the subjects of the books, it could be easier to detect inconsistencies in a proposed classification.

Another reason for choosing "Earth Sciences" as the sample was due to the Specialized Subject Area being one of the largest specialized areas in the collection (939 initial titles), with a good range of years (1956 – 2017), and with few repetitive titles (e.g., series volumes, annual reviews compilations). I removed books with more than two (2) repetitive titles so as not to inadvertently skew the collection. The following records were removed:

- Biology of Antarctic Seas I-XXII
- Computational Seismology and Geodynamics 1-6 and selected papers
- Contributions of Space Geodesy to Geodynamics (3 books)
- Contributions to Antarctic Research I-IV
- Environmental Hydraulics (6 books)
- Fossils and Strata 59-61
- History of Geophysics 1-4

The final total of the population was 894 books with date ranges: 1956 – 2017.

Sample

Regarding the number of books needed for a proof-of-concept research project, from discussions with Dr. Robert Patton[25] (one of the developers of the text-mining technology used in this project), that number is generally whatever will achieve the objectives of the research project. For example, the number of documents for a proof-of-concept project with the objective of demonstrating a text-mining method can be scaled-up to run on high performance

[25] R. Patton, personal communication, February 2018.

computers will be much larger than the number of documents needed to demonstrate a sensible book classification using a phylogenetic approach. Therefore, because an evolutionary book classification project, as with the one presented in this dissertation, has not been conducted, the closest classification project using book text from the above literature review is Guo et al. (2015), who used a corpus of 125 books for an automatic e-book classification project.

However, expanding to provide examples of the number of sources used in actual cultural phylogenetic studies referenced in the above literature review, O'Brien et al. (2001) used 83 projectile point sources and Larsen (2011) used 71 ethnographic sources for a Polynesian bark cloth project. More specific to text-based phylogenetic studies includes, Windram et al. (2014), who used 16 surviving copies of printed musical text. Dipper & Schrader (2008) used five (5) texts to learn if various quantitative methods could "be sensibly applied to small text samples of historic German dialects" and stated further that their "next steps, in the context of a larger project, [would] be to expand [their] data to include up to 50 complete corpus samples" (p.50). Spencer et al. (2004) used 21 artificial texts in their research. And the closest research to my dissertation's project is Tehrani's (2013) phylogenetic folktale project in which the author extracted attributes from 58 folktale sources. Given the above, a book corpus within the range of 20 – 80 should be adequate for a proof-of-concept text-based phylogenetic classification project. In other words, if 20 – 80 books are sensibly classified using phylogenetic software to produce the classification tree, then that should be sufficient to convince experts in both the information science and automatic text categorization fields that the general method is feasible. Two random samples for a total of 85 books were drawn, but the final number of

books used was 51. Discussion of the winnowing-down of this set is presented in chapter 4 under the *Findings from data collection* section.

DATA COLLECTION

In a phylogenetic analysis, a character list obtained from the units of analysis is first constructed from the taxa to be classified. A *character* is a feature in a population of entities that is used as data in an analysis of evolutionary relatedness of a group of entities containing similar features. Characters are the data for phylogenetic analysis and are the traits that change in populations of entities over time. Therefore, using a phylogenetic approach to develop a classification of books with only words or phrases, words and phrases from the books will be used to create the character list.

Each book was analyzed by Oak Ridge National Laboratory's Piranha text-mining tool to produce the word/phrase character set for each book. One of the reasons Piranha was chosen due to its weighting method used to distinguish 'important' words and phrases (terms). This weighting method uses both the terms from the book and the terms from the entire corpus library to determine weighting (I view this as being analogous to the DNA of an individual and the DNA of a population of individuals). And if the proposed classification method in this project were to be applied in a library, the weighting method would also need to be utilized to update a collection each time a book enters or exits the collection.

Piranha functions

In general, Piranha is an information retrieval tool, a text-mining tool, or more commonly referred to by its creators as a document recommender system. A recommender

system provides suggestions to a user and, in cases of collections of books, journal articles, etc., would be analogous to a human librarian: the patron provides information to the librarian and the librarian searches, retrieves, and presents suggestions to a patron. Similarly, a recommender system takes input from user and performs search, retrieval, and presentation functions for suggestions of information package to the user.

The Piranha technology has two main functions. The first is obtaining 'important' words/phrases (terms) from a corpus, which includes creating a library of important terms from the entire corpus, as well as creating lists of important terms from each individual source that makes up the corpus. Piranha uses a statistical machine learning approach using frequency counts to obtain the term lists and term library rather than deep learning (e.g., a word in close proximity to another word, where the word is located in the book, etc.). The term weighting technique is known as *term frequency-inverse corpus frequency*, and was created for faster term weighting of corpora than previous methods, while maintaining the same level of comparable quality of results (Reed et al., 2006; R. Patton, personal communication, September 2017). The weighting process that identifies the "important" terms will be discussed more in the *Corpus library creation* section below, but it is sufficient to say here that the identification is based on uniqueness (rather than similarity) and makes the terms important from an evolutionary perspective—i.e., evolutionary analysis is concerned with *change* among OTUs rather than *similarities* among the OTUs.

The second Piranha function is clustering of a corpus, which is accomplished by comparing the corpus term set with the individual source sets to form clusters of similar sources. The Piranha clustering function would not be used in this project because the

phylogenetic software will be used to organize the books based on algorithms that are used to suggest evolutionary relationships among the OTUs (i.e., among the individual books). Though clustering can decrease variance in a dataset, clustering can also increase the possibility of error. For example, the Piranha clusters would contain some, but not necessarily all, of the important terms of each book within a given cluster due to the weighting technique within Piranha. In other words, a cluster may not have all the important terms of a specific book, but would have enough terms for a book to be included in a cluster. This is analogous to resolution on a digital camera—if the resolution is reduced, the person viewing can still understand the imagery in the photo, but some digital information will be lost. Also, because clustering in Piranha is based on similarities of the information packages, the use of the clustering feature also increases the possibility that some books could be included together in a cluster that could be distant from an evolutionary-relatedness context. Therefore, it would be simpler (and thus more supportive of the parsimony principle and more scientifically elegant) if only the terms that characterize the books were enough for a phylogenetic analysis.[26]

From Piranha's two main functions, only the identification and extraction of important terms will be used.

Corpus term set creation

Piranha was used to scan a randomly generated sample set of books from the almost 900 books to create both the corpus' term set of important words/phrases (terms) as well as each individual book's list of terms. The individual book terms will ultimately be the 'characters' for the phylogenetic analysis. A *character* is a feature in a population of entities used as data in

[26] R. Patton, personal communication, September, 2017.

a phylogenetic analysis of evolutionary relatedness of a group of entities containing similar features. For organisms, early phylogenetic characters were morphological features. Later, molecular data emerged for phylogenetic analysis, including protein sequencing, which was followed by DNA sequencing (Brown, 2002). For this project, the characters are the terms generated by Piranha from a group of earth-science books.

The data source was limited to the intellectual content of book chapters that were digitized by optical character recognition (OCR) technology, and therefore recognizable by the automatic text-mining tool. The front matter, index, references, etc. were removed during data cleaning process. Images, along with any text within the images, were not included in the analysis due to those not being OCR'd. Image captions are typically OCR'd, but most science and technical captions use words also present in the text. Table data, formulas, and algorithms, are also often OCR'd, but most of these were lower-weighted terms, and did not contribute substantially to the overall outcome of the evolutionary analysis. These types of terms which were present were removed, which is discussed in greater detail in chapter 4.

To obtain the terms, the Piranha technology performs a statistical analysis based on frequency counting. More specifically, the technology uses term frequency counts and compares the counts 1) in each book and 2) across all books. Term identification is based on uniqueness of the word: if 'earth' is used 500 times in a book, 'earth' may not considered important; similarly, if 'earth' is found in every book in the corpus, it would likely not be considered important.

The general steps in the Piranha processing of the terms that characterize books include:

- Scan all text in each book of a corpus
- Remove stop terms such as *and*, *the*, *an*, etc.
- Assign a weight to the remaining terms by comparing terms in the individual books and across all books in the corpus.

In general, weighting is lower for terms that are more common to most books and higher for terms less common to most books. But even if a book contains a very common word, the word may be uncommon to the overall corpus terms. In this case, the book's common term would be a little higher weighted than if only the book was analyzed in isolation. The principle underlying the weighting is that uncommon words characterize a book and illustrates uniqueness when compared to other books. Table 1 conceptually illustrates a weighted list.

DATA ANALYSIS

Creating the character list and data matrix

Each book's character was compared with each other during the phylogenetic analysis. In preparation for the analysis, each book's character set was compared to the overall corpus term set and coded as either 'present' [1] or 'absent' [0] in the corpus term set. The exact number of terms in each book character set and the overall corpus set, along with how they were utilized is discussed in chapter 4.

A character data matrix was then created for each book by comparing each book's top set of terms with the overall corpus term set with coding being either absent [0] or present [1] (see concept in Table 2).

Table 1. Conceptual Example of Term Weighting.

Term Rank	Weight Rank
Term 1	Highest weighted
Term 2	Second highest weighted
Term 3	Third highest weighted
Term *n*	n^{th} highest weighted

Table 2. Conceptual Character Data Matrix Example.

Book Title	Corpus term 1	Corpus term 2	Corpus term 3	Corpus term *n*
Book 1	0	1	0	*n*
Book 2	1	1	0	*n*
Book 3	0	0	1	*n*
Book *n*	*n*	*n*	*n*	*n*

Once the character data matrix was completed, the data was entered into the PAUP* computational phylogenetics software, which was used to conduct the cladistic analysis to obtain the classification tree. “Phylogenetic Analysis Using Parsimony (PAUP)… [is] a widely used software package for the inference of evolutionary trees that supports a wide range of approaches to phylogenetic analysis and features relatively friendly input for data and output of results” (Stubbersfield & Tehrani, 2013, p. 94).

PAUP* 4 capacity

For almost 900 books, using 10 terms suggests over 9,000 term characters and using 20 terms suggests over 18,000 term characters. This dissertation project used considerably less than 900 books. However, the PAUP* 4 phylogenetic software can easily accommodate up to 900 books and many more characters (and character states) than the characters (and character

states) most likely needed for this project. The maximum number of taxa (e.g., books) the PAUP* 4 software can analyze is 16,384. The maximum number of characters (terms) that can be analyzed using a computer with a 32-bit processor is 1,073,741,824 (even if this project required 1,000 characters per book, a 900,000 character set would only be about 0.084% the software's character capacity). And the maximum number of character states for each character for a 32-bit processor is 32 (http://paup.phylosolutions.com/documentation/faq/). This dissertation project only used two character states (absent, present).

Rooting the tree

This project's tree was rooted. A *root* is the node on the tree that connects to all the other branches. The root is considered the last known common ancestor of all the taxa represented on the tree. If unknown, the root can also be a taxon that is known to be outside of the other taxa in the tree (i.e., the "outgroup"). With this project's earth-science book set—with the oldest book published in 1956—the use James Hutton's (1788) paper *Theory of the Earth*, served as the root. Hutton is considered the founder of modern geology and though this was a paper presented at the Royal Society of Edinburgh, it was almost 100 pages in length and was the foundational work for his more extensive treatment of the subject published in two book volumes in 1795. The book used as the outgroup to root the tree is not critically important and other more recent books could have been used with no change to the trees produced. If there was no root designated, the first book listed in the input into the PAUP* tool is the default outgroup if rooting is desired.

Reliability and validity

This research project follows a specific process and the process relies heavily on computer algorithms to 1) extract characters (i.e., terms) from the text of books, 2) rank order those characters according to a weighted 'importance' based on a type of frequency count, and 3) group together books according to their presumed evolutionary relationships. Given the same data and following the developed method, these algorithms should produce the same results if repeated. The manual human activities in the process—creating the corpus' character list, coding each book's characters as absent/present relative to the corpus (i.e., creating the character data matrix), and entering the character data matrix into the phylogenetic software—could possibly be automated with new computer algorithms. Determining the boundaries for the term set does introduce subjectivity, but with the developed method, this becomes repeatable). Therefore, if the overall method developed during this project is repeated using the same data and the same text-mining and phylogenetic tools, it would provide the same results, therefore making the method highly reliable.

The research project's validity—whether or not the method indicates what it is supposed to indicate—is less clear due to inherent subjectivity of this project's two primary research guides: theory and primary research question. First there is the theoretical classification principle of common origin on which the research project is based. Here, my research position is this: because the analytical tool to be used produces bifurcating trees of taxa and is ultimately based on Darwinian evolutionary theory of decent with modification from a common ancestor, the tool can be used to produce a classification tree of scientific and technical books based on common origin because those types of books are created from previous knowledge—i.e., common origin—which is indicated by the copious references

contained in scientific and technical books. But this research position is subjective because the fundamental idea that books can be evaluated in the same evolutionary-relatedness manner as biological organisms is debatable, at least at the current stage of human understanding.

The second component of subjectivity is the primary research question itself: *Can a phylogenetic-based method be developed to produce a sensible classification of books with only words or phrases used as the characters for analysis?* Of course, "sensible" is not quantifiable and highly subjective. However, assuming the phylogenetic classification tree of these books looks "sensible"' to me, one measure of validity of sensible could be using subject matter experts to further review the classification.

In summary, the project method will likely be highly reliable due to the use of algorithms and human methodological activity that could possibly be automated. Evidence to assist in validity of the project can be obtained from observation of the project's classification tree, but will have some level of subjectivity, which may reduce overall validity.

CHAPTER 4 – RESEARCH FINDINGS

INTRODUCTION

This chapter is divided two main sections: Findings from dataset development and findings from phylogenetic analysis (followed by a conclusion section). Most of the research effort with a phylogenetic project is assembling the dataset for input into the phylogenetic analysis tool. And because this project is the development of a potential automatic method to create a book classification, much attention is given to the research that led to the final dataset. Therefore, some method steps that are traditionally included in a methods and materials chapter are presented in this research findings chapter in order to more easily communicate the actual findings from my research. To better illustrate the importance of the dataset development to the overall project, Figure 8 presents the conceptual steps of the method.[27]

Findings include relevant discoveries from dataset development that could have

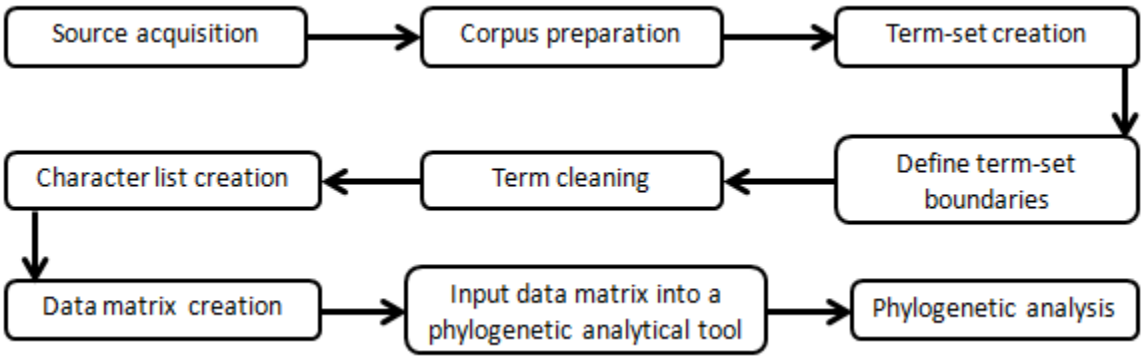


Figure 8. Research Project Method Conceptual Design.

[27] "Evolutionary Method of Organizing and Representing Text-Based Documents," UT-Battelle, LLC invention disclosure number 201804148; Department of Energy S-number S-138,815

negatively affected the phylogenetic analysis, such as issues with the optical character recognition (OCR) used on the PDF versions of the books that could (and did) affect the quality of the term extraction with the text-mining tool. Other discussions include conversion of the term set into input for the phylogenetic software, including the analysis that led to defining the term-set boundaries. Additional discussion is provided for term cleaning and reduction of the terms to a reasonable number. Findings from the phylogenetic analysis include considerations for analysis and tree display, followed by three analyses (parsimony, distance, and likelihood) and their associated trees. Comparison discussions of each tree are provided as well as a comparison of a tree to the Library of Congress Classification. Evidence suggests only the parsimony tree provides a sensible book classification.

The two main sections and subsections of chapter 4 include:

- Findings from Dataset Development
    - Book collection identification and acquisition
    - Corpus preparation
    - Term-set creation
    - Setting term-set boundaries
    - Term cleaning
    - Character list creation
    - Data matrix creation
- Findings from Phylogenetic Analysis
    - Data input
    - Analysis considerations

- Tree display considerations
- Creating trees from parsimony, distance, and likelihood analyses
- Tree comparisons
- Library of Congress classification comparison

FINDINGS FROM DATASET DEVELOPMENT

Book collection identification and acquisition

As discussed in chapter 3, a book corpus within the range of 20 – 80 books should be adequate for a proof-of-concept text-based phylogenetic classification project. Fifty (50) numbers were randomly generated from the base of 894 numbers (i.e., the total number of books in the collection) using a web-based random number generator (Hedges, 2018). To select, the book list was first sorted by the print publication year in ascending order, then by the print book thirteen-digit International Standard Book Number (ISBN13 number) in ascending order. Due to some books not being suitable for data collection (see 'Books removed' section below), another 35 books from the remaining 844 were pulled according to the same process as above.

The books arrived from the publisher in individual 'zip' folders with each book folder's title being the online book's ISBN13 number. Each book's folder contained separate subfolders that were titled as individual chapters (e.g., ch1, ch2… ch*n*) and other subfolders such as the book's front matter (i.e., title page, contents, forward, preface, acknowledgement, introduction, lists, etc.), appendices, indexes, plates, lists of notations/symbols, glossaries, etc. Some books had a 'references' subfolder, though most references were placed at the end of the chapters.

The most substantive sections of books—i.e., those containing the primary intellectual content—are the chapters. Therefore all subfolders that were not in a 'chpXX' subfolder were removed. The chapter subfolders contained a PDF document of the chapter and a Wireless Markup Language (.WML) metadata file. In other words, if a book contained 15 chapters, there would be 15 chapter subfolders, each containing a PDF document and a .WML file. Because of this, the *Piranha* text-mining tool would recognize each chapter as a separate document rather than a collection of documents within a single book. To enable *Piranha* to analyze the chapters as an entire book—thereby creating the individual book characters for phylogenetic analysis—all the PDF chapters had to be removed from their individual folders, placed into a single folder with the name of the online book's ISBN13 number, and all PDF chapters were then merged into a single document representing the entire book.

Corpus preparation

Preparing the corpus, creating the term-sets (i.e., the data), selecting which terms to use, and preparing the terms for use as 'characters' is the primary work of conducting the phylogenetic analysis (i.e., creating the classification 'tree'). Therefore, most of my research time was spent creating and understanding the data. One of the time consuming parts of this process was the removal of references.

In general, all text in the book chapters was removed after the final section of the main body—i.e., after the Discussion, Summary, or Conclusions section (or a combination thereof). Sections removed at the end of the chapters, other than the references, included acknowledgements, appendixes and lists of abbreviations and acronyms. Granted, the acknowledgements, appendixes, and lists would likely not have influenced the overall outcome

of the list of terms extracted by the *Piranha* text-mining tool (see discussions below for more information about the weighting of terms), but only a few extra seconds were required to remove them. However, if there was anything placed after the final section due to efficient use of space by the publishing company (e.g., a figure with a caption was placed after the final section), those were retained in the final book used for term extraction.

To reduce the time of manually removing the references, two automated citation extractors were suggested to me: *ParsCit* and *GROBID*. The developers of *ParsCit* were no longer supporting *ParsCit* and they recommended using *GROBID*. Neither of these tools actually removed the references, but a person who was familiar with both said a parsing script could likely be written for use with *GROBID* to remove the references, but the PDF documents would first need to be converted into text and the elimination of references from each chapter would probably still require review of the elimination because the tool was not 100% accurate in its normal citation extraction process. I chose not to use this for two reasons: 1) the modification of existing computer code is something I cannot do myself, and therefore falls outside the scope of the project and 2) the number of books required for this type of proof-of-concept project could be obtained with manual deletions using Adobe Acrobat Pro DC (see *Sample* subsection in chapter 3).

Considering an automatic classification system, the above citation extraction tools (or some equivalent) could be needed, though other sections of a chapter after the main body (i.e., acknowledgments, appendixes, lists, etc.) may also need to be eliminated, which would require more modifications of the citation extraction tools (or some other type of extraction tool). Of

course, if publishers provided files with only the text from the main body, that would eliminate the need for any extraction.

Books removed. Nineteen (19) books were eliminated from the dataset due to the inability to remove the references using Adobe Acrobat Pro DC (and executing the optical character recognition (OCR) feature did not enable the document to be edited). Though these could have been converted to Microsoft (MS) Word documents, it would not have been a perfect conversion. For example, a word hyphenated at the end of a sentence in the PDF is not recognized as a full word in MS Word (e.g., 'devel-oped' would be the text read by the text-mining tool in MS Word document but would be read by a human as 'developed' in the PDF document). I determined such a conversion was not necessary for a poof-of-concept test and time could be better used for other research activities. Another book was also removed from the list because several chapters only had abstracts (i.e., the main body was missing).

Other notable mentions. Though the preface was removed from the books for the final dataset, some books contained an introduction chapter or abstracts at the beginning of each chapter. Both remained in the final corpus because they were considered to be intellectual content. Also remained were the identifying aspects of authors (affiliated instructions, contact information, etc.) often present in the first page of a chapter. In-text citations also remained. The reason for including these (other than the overwhelming time consuming nature of removal of these types of text) is due to the limited appearance of these words and the low likelihood of them significantly impacting the terms to be used in producing the phylogenetic classification tree. In other words, these terms would likely not be weighted high enough by the text-mining tool to be included in the phylogenetic character list (see discussions below

pertaining to evaluation of proper nouns that were weighted high enough to influence the character list). Peer commentaries, which were only observed in one book, were removed due to being after the last section of the main body. However, a 'Reply' (i.e., a refute) to one of the chapters was given its own chapter by the publisher, so that peer commentary remained in the book.

Seven (7) chapters in online book ISBN13 number 9781118663615[28] had small acknowledgement sections but were kept due to the entire page of text would need to be removed to delete only the acknowledgement section. Chapter 17 of online book ISBN13 number 9781118665442 was incomplete but was included in the book. From the 50-book pull, online book ISBN13 numbers 9781118667422, 9781118667446, 9781118667842 had only one chapter, and from the 35-book pull, online book ISBN13 numbers 9781118667033 and 9781118667071 had only one chapter. These five (5) were kept as books in the corpus. All were geology field trip books except one.

On the final review of the books before term extraction processing with the text-mining tool, the following were missed in the initial cleaning but were removed during this final review: two books that still had acknowledgement sections, another had an acknowledgement section as well as a small section of references, and one had an appendix with an associated image.

Final corpus. The corpus' print publication years of the final 65 books selected for term extraction ranged from 1969 to 2017 including one (1) from the 1960s, three (3) from the 1970s, 14 from the 1980s, 16 from the 1990s, 13 from the 2000s, and 18 from the 2010s. From calculations made, it took about 13 minutes per book and about 50 seconds per chapter to

[28] The IBSN13 numbers referenced through this chapter can be found with their associated titles in APPENDIX A: Corpus Dataset Book List or in APPENDIX B: Term Set Removals.

clean the books, which translates to about 14 hours to manually clean a 65-book scientific and technical corpus.

Term-set creation

The terms extracted by the text-mining tool were saved in separate plain text (.txt) files for each of the 65 books in the project corpus, with the file titles being the online book ISBN13 number. These files contained the terms of each book along with each term's associated weight number, which represented the weight of a term's importance relative to both the book and the overall corpus. Each of these .txt files was converted into a Microsoft Excel file with a separate column for term and the weight, which enabled easier initial review. Table 3 is an example of the top 20 terms from the book one of the books.

Discussion of the initial terms. There were two notable observations during the initial review of the term sets: 1) the presence of fragmented words and variants of words and 2) the presence of highly weighted terms that were also observable in the titles of books and/or chapters printed in headers and footers of the books.

Regarding fragmented words, these seemed most likely the result of issues during the OCR process, editorial issues, oddities in text (e.g., figure captions, formulas, symbols, abbreviations, etc.) or combinations thereof, rather than issues with the text-mining tool (more about the OCR issues is discussed in the Issues with Quality of OCR Processing section below). Table 4 illustrates that 'convect' should have been 'convection'. But most fragments could not have been easily corrected as in this example. Fortunately, fragmented, highly ranked terms (i.e., terms that best distinguish one book from another) were not substantial: only nine (9) of the 65 term sets (14%) had more than one fragment in the top 20 terms, but the variations of the

Table 3. Top 20-ranked terms from *Geosimulation: Automata-based Modeling of Urban Phenomena*.

Rank	Term	Weight
1	geosimulation	71.36536506
2	geosimulations	71.36536506
3	land-uses	63.99997808
4	land-use	63.99997808
5	automata	63.06311841
6	automata-based	61.66918901
7	real-world	55.93533952
8	self-organization	47.71078314
9	self-organize	47.71078314
10	self-organized	47.71078314
11	self-organizes	47.71078314
12	self-organizing	47.71078314
13	agent-based	46.57713432
14	yaffo	45.53505324
15	portugali	44.99761237
16	1960s	44.65617885
17	benenson	44.61163919
18	non-fixed	44.30512622
19	1970s	43.57188274
20	multi-agent	42.89811282

Table 4. Fragment Example from Corpus Terms.

Term column	Additional letters due to fragmentation
situ…	
con-vection	
convect	ion
convectional	
convec	

same terms with the same or similar weights were present in the top 20 terms of all the book term sets. In general, understanding of these fragments and variations would also provide some degree of confidence in the use of this text extraction method within an automated classification process—i.e., fragments and variations could automatically be combined and/or removed if a weight was above a certain threshold.

The second notable observation was that many of the highly weighted terms were also observable in the titles of books and chapters. Of course the header/footer sections of each book's page should have been removed before conducting the keyword extraction from the corpus. But that would have taken an even longer amount of time for the data cleaning process without my ability to write some sort of script to automatically remove those. Fortunately, there is evidence to suggest that this may not be necessary or there was only a minimal consequence as a result of these header/footer terms being included in the term sets.

Consider the book titled *Geosimulation: Automata-based Modeling of Urban Phenomena*. This title was on (almost) every other page in the header section.[29] The text-mining recognized all words except the stop word "of." It is intuitive to believe that the repetition of that many words would have significant influence on the weight of the terms *geosimulation*, *automata-based*, *modeling*, *urban*, and *phenomena*. But keep in mind that the *Piranha* weighting algorithms utilize not only the terms from the individual book but also the terms from the entire corpus to determine the overall weight for each term in an individual book. Therefore, the text-mining tool's creation and utilization of the corpus term set can assist

[29] The 'cleaned' book included 254 pages, the number of rows in the spreadsheet corresponding to the number of individual terms extracted by Piranha was 6,532, and the weight range of the terms was from 71.36536506 to 0.55879068.

in increasing or decreasing a single book's terms. Table 5 illustrates that two of the terms in the book title were highly weighted, but the other three were only moderately weighted, suggesting the inclusion of the terms in the titles may not be skewing the term weights in an appreciable manner.

Additionally, each chapter's title in this book was on (almost) every other page in the header section of the chapter. Again, this could enable these terms to be weighted higher in the term set, but Table 6 suggests this is also not necessarily the case for all the chapter title terms.

Another factor to consider for strengthening the argument for no-to-low influence of title terms in headers/footers is the variations of terms previously mentioned. Table 6 provides the top 20-ranked terms in the aforementioned geosimulation book. Notice that even if 'geosimulation' was removed due to the term being present in the book's title, its plural form would become the highest term. However this is not the same for 'automata-based' (observed in the book title) and 'automata' (observed the titles of chapters two and four): removing those would eliminate terms deemed highly important to both the book and the overall corpus. Conversely, variants of 'land-use' and 'self-organize' are observed in the top terms, but are not in any of the titles. Variants of the term 'multiagent', observed in chapter five's title, are in the

Table 5. Term Row Location and Weights for Title Terms of *Geosimulation: Automata-based Modeling of Urban Phenomena*.

Row location in spreadsheet	Term	Weight
1	geosimulation	71.36536506
6	automata-based	61.66918901
105	phenomena	29.42341852
136	urban	26.55173763
180	modeling	25.49972284

Table 6. Term Row Location and Weights for Chapter Terms of *Geosimulation: Automata-based Modeling of Urban Phenomena*.

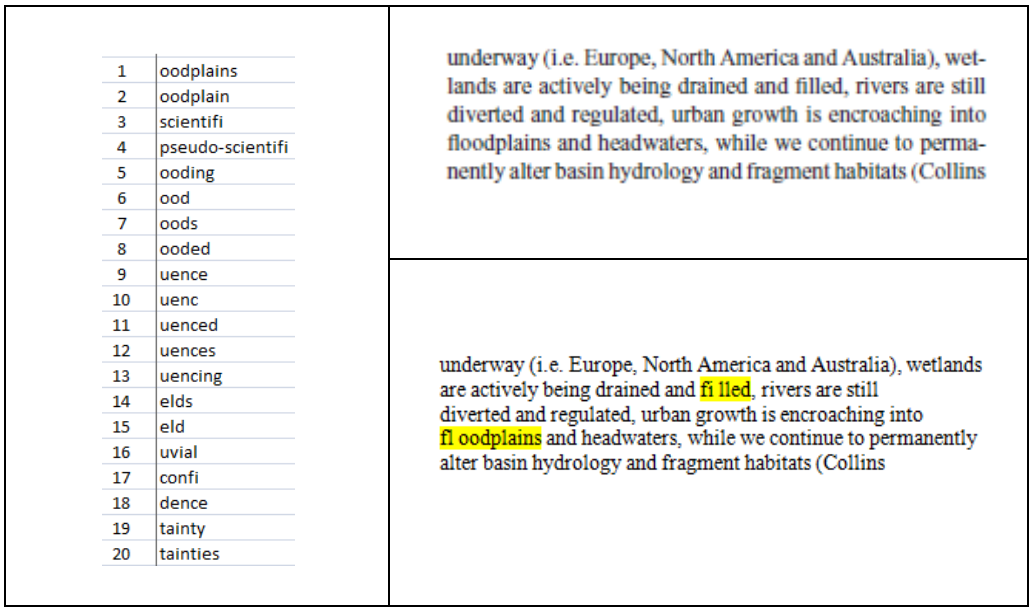
Row location in spreadsheet	Term	Weight
Chapter 1 title: *Introduction to Urban Geosimulation*		
1	geosimulation	71.36536506
136	urban	26.55173763
1173	introduction	13.57702696
Chapter 2 title: *Formalizing Geosimulation with Geographic Automata Systems (GAS)*		
1	geosimulation	71.36536506
5	automata	63.06311841
141	geographic	26.41368159
521	formalizing	17.39694076
618	gas	16.90798023
1254	systems	13.08629262
Chapter 3 title: *System Theory, Geography, and Urban Modeling*		
136	urban	26.55173763
158	geography	25.94346432
180	modeling	25.49972284
339	theory	20.48201591
1255	system	13.08629262
Chapter 4 title: *Modeling Urban Land-use with Cellular Automata*		
4	land-use	63.99997808
5	automata	63.06311841
135	cellular	26.63396569
136	urban	26.55173763
180	modeling	25.49972284
Chapter 5 title: *Modeling Urban Dynamics with Multiagent Systems*		
21	multiagent	42.89811282
136	urban	26.55173763
180	modeling	25.49972284
184	dynamics	25.3386457
1254	systems	13.08629262
Chapter 6 title: *Finale: Epistemology of Geosimulation*		
1	geosimulation	71.36536506
502	epistemology	17.46579093
4766	finale	5.553225027

list of top 20 terms, but the actual term is not (it was #21), so the removal of 'multiagent' could be offset by its remaining variants.

In conclusion, given the above cursory analysis that suggests the book and chapter title terms reprinted in headers/footers do not appreciably influence the top-weighted terms, I chose to leave them in the term sets. In other words, the inclusion of the top-weighted terms that also happened to be in the titles of books and chapters would not alter the overall objective of the dissertation project, which was to test a book classification method based on common origin.

Issues with quality of OCR processing. The first indication of issues stemming from the quality of the OCR process was in reviewing 9780470867082's terms. Figure 9 provides the first 20 terms. The title of this book is *River Restoration - Managing the Uncertainty in Restoring Physical Habitat*. As might be expected, 'oodplains' should be 'floodplains'. Conducting a search for 'oodplains' isolated the issue. The right side boxes in Figure 9 include the book's text (top box) and highlighted letters (bottom box). After either 'fl' or 'fi', a space was inserted, which explains several of the oddities in the top 20 terms, not only the 'oodplains' (floodplains) but also 'scientifi' (scientific), 'elds' (fields), and 'uence' (influence). Similar OCR issues were found with review of top terms of book 9781118665299 (see Figure 10). Even in books where the top terms seemed viable, OCR issues could be observed (see Figure 11).

Term set removals. After these discoveries, all 65 books were checked for OCR quality in at least two different areas of the book and searched using any suspicious terms observed in (roughly) the top 30 highly ranked terms. In summary, the book term sets removed had either numerous highly ranked terms that were clearly fragmented (3



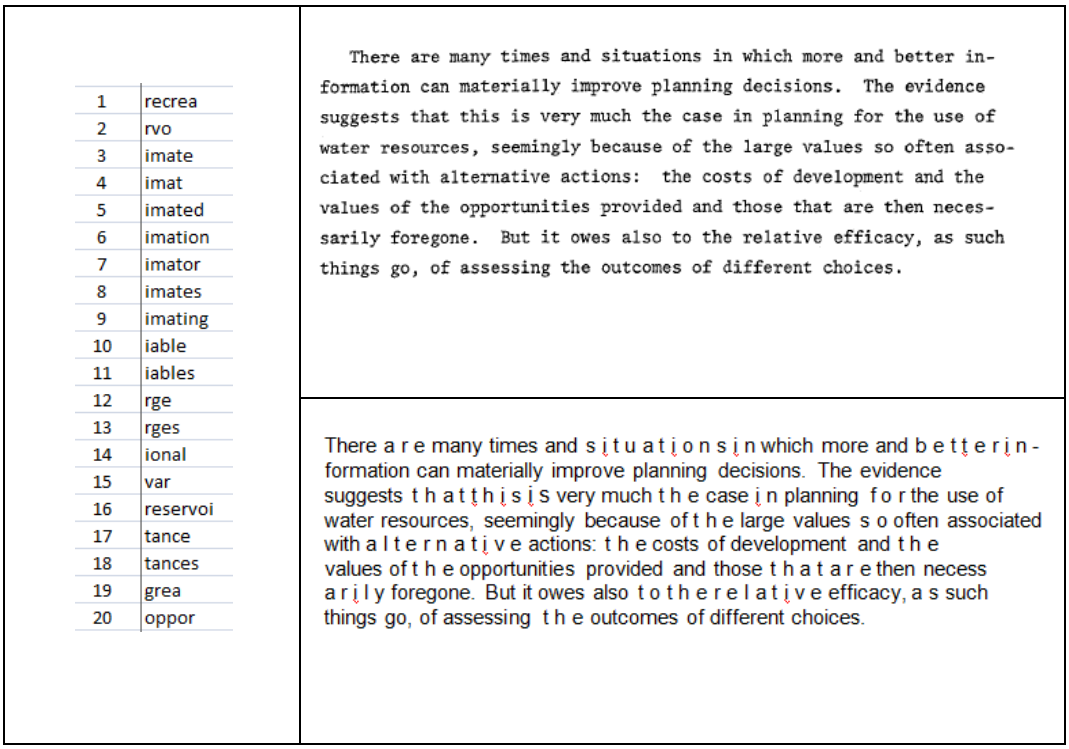
1	oodplains
2	oodplain
3	scientifi
4	pseudo-scientifi
5	ooding
6	ood
7	oods
8	ooded
9	uence
10	uenc
11	uenced
12	uences
13	uencing
14	elds
15	eld
16	uvial
17	confi
18	dence
19	tainty
20	tainties

underway (i.e. Europe, North America and Australia), wetlands are actively being drained and filled, rivers are still diverted and regulated, urban growth is encroaching into floodplains and headwaters, while we continue to permanently alter basin hydrology and fragment habitats (Collins

underway (i.e. Europe, North America and Australia), wetlands are actively being drained and fi lled, rivers are still diverted and regulated, urban growth is encroaching into fl oodplains and headwaters, while we continue to permanently alter basin hydrology and fragment habitats (Collins

Figure 9. Top 20 Terms and Examples of Human and Machine Readable Text for ISBN13 Book 9780470867082.

Left box provides the first 20 terms extracted by Piranha. The top right box is from the book's text and the bottom right box highlights spaces after 'fi' and 'fl' that the text-mining tool would have recognized due to an issue with the OCR process. Text source: *River Restoration - Managing the Uncertainty in Restoring Physical Habitat* (p. 22).



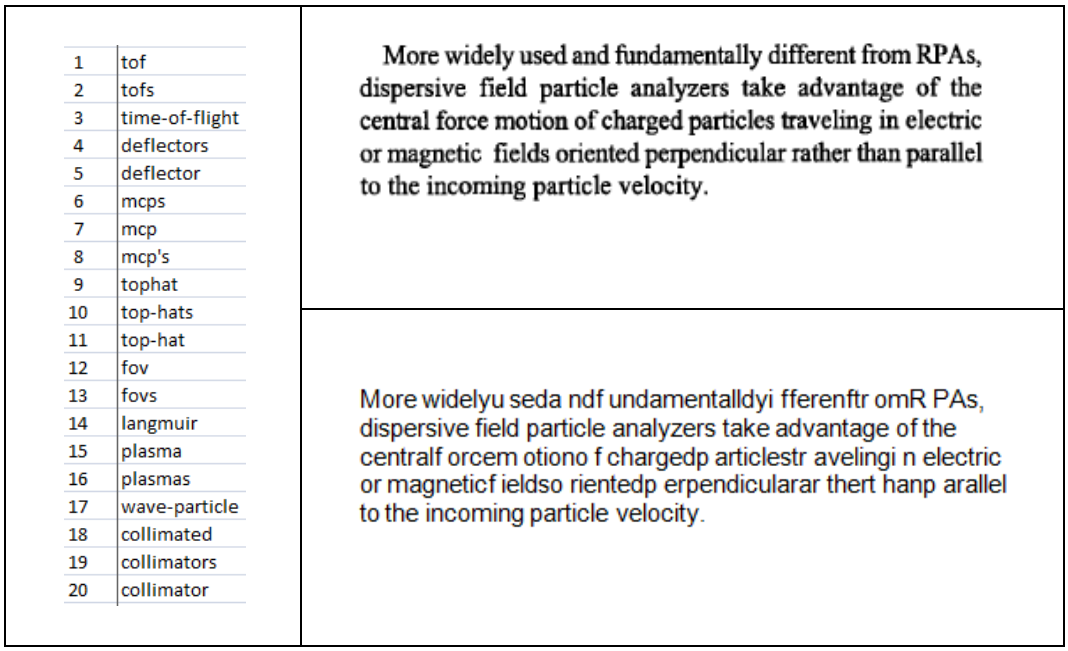
1	recrea
2	rvo
3	imate
4	imat
5	imated
6	imation
7	imator
8	imates
9	imating
10	iable
11	iables
12	rge
13	rges
14	ional
15	var
16	reservoi
17	tance
18	tances
19	grea
20	oppor

There are many times and situations in which more and better information can materially improve planning decisions. The evidence suggests that this is very much the case in planning for the use of water resources, seemingly because of the large values so often associated with alternative actions: the costs of development and the values of the opportunities provided and those that are then necessarily foregone. But it owes also to the relative efficacy, as such things go, of assessing the outcomes of different choices.

There a r e many times and s i t u a t i o n s i n which more and b e t t e r information can materially improve planning decisions. The evidence suggests t h a t t h i s i s very much t h e case i n planning f o r the use of water resources, seemingly because of t h e large values s o often associated with a l t e r n a t i v e actions: t h e costs of development and t h e values of t h e opportunities provided and those t h a t a r e then necess a r i l y foregone. But it owes also t o t h e r e l a t i v e efficacy, a s such things go, of assessing t h e outcomes of different choices.

Figure 10. Top 20 Terms and Examples of Human and Machine Readable Text for Book ISBN 13 Book 9781118665299.

Left box provides first 20 terms extracted by Piranha. The top right box is from the book's text and the bottom right box shows spaces between various words that would have prevented text-mining tool from accurately recognizing due to issue with the OCR process. Text source: *Outdoor Recreation and Water Resources Planning* (p. 1).



1	tof
2	tofs
3	time-of-flight
4	deflectors
5	deflector
6	mcps
7	mcp
8	mcp's
9	tophat
10	top-hats
11	top-hat
12	fov
13	fovs
14	langmuir
15	plasma
16	plasmas
17	wave-particle
18	collimated
19	collimators
20	collimator

More widely used and fundamentally different from RPAs, dispersive field particle analyzers take advantage of the central force motion of charged particles traveling in electric or magnetic fields oriented perpendicular rather than parallel to the incoming particle velocity.

More widelyu seda ndf undamentalldyi fferenftr omR PAs, dispersive field particle analyzers take advantage of the centralf orcem otiono f chargedp articlestr avelingi n electric or magneticf ieldso rientedp erpendicularar thert hanp arallel to the incoming particle velocity.

Figure 11. Top 20 Terms and Examples of Human and Machine Readable Text for Book ISBN 13 Book 9781118664384.

Left box provides the first 20 terms extracted by Piranha. The top right box is from the book's text and the bottom right box shows spaces between various words that would have prevented the text-mining tool from accurately recognizing due to an issue with the OCR process. Note that 'tof' is an acronym for *time-of-flight*, which incidentally would have been read by the text-mining tool as *electrostaticti,m e-of-flight* from the section in the book (p.5). 'Tof' is also found in 'cutoff' that is present in the book. Text source: *Measurement Techniques in Space Plasmas: Particles* (p. 3).

removed) or books that had poor OCR quality, even though their highly ranked terms were not fragmented (8 removed). Removing the latter was done because the terms may have been substantially different had the OCR quality been acceptable.

The term set data investigation resulted in the following 14 removals from the corpus (which includes the three provided as examples in the above tables; associated titles and print publication years are included in APPENDIX B: Term Set Removals): 9780470867082, 9781118663837, 9781118664384, 9781118664629, 9781118664698, 9781118664797, 9781118665190, 9781118665299, 9781118665442, 9781118665527, 9781118666050, 9781118669167, 9781118670354, 9781444323276.

Print publication years of the term sets removed included 1974, 1980, 1986, 1988, 1998 (2), 1999 (2), 2001, 2004, 2005, 2008, 2010, and 2011. The removal left 51 book term sets for phylogenetic analysis. Of these, 41 seemed to have good text quality 10 had acceptable text quality. The list of the 51 book titles and print publication year (1969 thru 2017) is included in Appendix A. Though this investigation of term sets took much more time than anticipated, a level of comfort with decisions for term sets to retain and term sets to remove needed to be attained, which required much interaction with both the data and the source of the data. This data interaction continued with the term-set boundaries process.

Setting term-set boundaries

Once the book term sets were finalized, the next step was to determine the terms to be used as characters in the phylogenetic analysis—i.e., setting the 'boundaries' of the overall term set. As mentioned previously, there are issues with the term sets, so not all terms are useable. The least number of terms extracted from a book was 1,702 and the most was 24,056,

with most book term sets ranging roughly between 5,000 – 8,000 terms. The highest weighted term was about 88.0 and the lowest weighted term was about 0.6. Though it should go without saying, the highest weighted terms are the terms that best differentiate one book from the other books in the corpus. And well over half of the terms in any given term set were weighted in the single digits (i.e., weights 1.0 – 9.9), suggesting at least half of the terms extracted from the books were not needed for this project. From another perspective, the highest weighted terms for each book are a relatively small percentage of the overall term set. Table 7 includes a list of eight (8) selected term sets.

The shades of the table communicate the following about the 30 example term weights in each book:

- Some terms start with high weights (80s, 70s) and drop to much lower weights (40s, 30s),
- Some start with high weights and remain relatively stable throughout (80s to 60s),
- Some start with medium weights (50s) and drop to much lower weights (20s),
- Some start with medium weights and remain relatively stable throughout (60s to 50s, 50s to 40s),
- Some start with lower weights (40s) and drop to much lower weights (20s),
- Some start with lower weights and remain relatively stable throughout (40s to 30s).

The point to the above is to illustrate there is no consistent weight given to the top ranked terms[30] and there is no consistent transition from higher ranked terms to lower ranked

[30] The reason for the relatively wide variation in the top weights in these examples is due to Piranha's term frequency calculations using both the number of times the term appears in both the document and across all the documents. For example, a term that is observed many times in a book would have a relatively high weight. But if

Table 7. Example from Various Term Sets of Top 30 Term Weights.

	9781118663615	9780470020999	9781118668665	9781118667262	9780470682104	9781118666012	9781118667842	9781118669693
1	83.69914115	71.36536506	84.25963642	54.45006657	65.77345794	54.76162140	43.87562118	48.43654025
2	83.69914115	71.36536506	84.25963642	46.33277943	65.77345794	54.76162140	41.96170507	48.43654025
3	71.48654796	63.99997808	84.25963642	46.33277943	65.45143631	51.62985560	39.62767346	48.24114164
4	64.14572880	63.99997808	73.99086854	46.33277943	64.14572880	51.62985560	39.21470488	46.27684162
5	59.45009625	63.06311841	73.99086854	41.07121605	64.14572880	50.21246278	35.30621268	45.32655323
6	59.43000051	61.66918901	71.57796554	37.76483754	63.02336705	50.21246278	35.30621268	44.99761237
7	57.38753080	55.93533952	71.57796554	37.76483754	63.02336705	50.21246278	34.65132625	44.38221532
8	57.38753080	47.71078314	71.57796554	37.76483754	63.02336705	49.35836428	34.65132625	44.38221532
9	55.92340251	47.71078314	71.57796554	37.76483754	63.02336705	49.35836428	34.55073629	44.38221532
10	55.92340251	47.71078314	71.57796554	37.76483754	62.73523306	49.35836428	34.55073629	44.30176714
11	55.92340251	47.71078314	70.85779190	37.33144947	62.73523306	49.35836428	34.01504174	44.30176714
12	55.92340251	47.71078314	70.85779190	37.02152840	61.93977758	49.35836428	34.01504174	44.30176714
13	51.11962581	46.57713432	70.85779190	37.02152840	60.94130073	49.13775697	34.01504174	43.25484583
14	51.11962581	45.53505324	68.66202411	34.59470048	60.94130073	49.13775697	33.74648667	43.18842037
15	51.11962581	44.99761237	68.66202411	34.59470048	59.76295479	49.12584578	33.45745898	40.80529180
16	48.74984672	44.65617885	68.66202411	34.01504174	56.79536941	49.12584578	32.88672897	40.80529180
17	48.74984672	44.61163919	68.66202411	34.01504174	56.79536941	48.75601459	32.88672897	40.80529180
18	45.32655323	44.30512622	67.87475729	31.50473625	53.55423446	48.75601459	32.61039667	40.80529180
19	45.23917865	43.57188274	67.87475729	30.96577350	53.55423446	48.74984672	32.55302214	39.60345757
20	45.23917865	42.89811282	67.87475729	30.85394108	53.46574901	48.74984672	31.96322922	39.60345757
21	45.23917865	42.89811282	67.31889879	28.69376540	52.94281030	48.74984672	30.96577350	39.38092031
22	45.23917865	41.33455256	67.31889879	28.69376540	52.92295564	48.53203880	30.96577350	39.38092031
23	44.53384517	40.13119473	67.31889879	28.38669352	52.76350441	47.15684554	30.77138944	39.38092031
24	44.30512622	40.11216265	67.15615436	28.38669352	52.55325014	46.27684162	30.74915289	39.38092031
25	44.26329539	39.16552813	67.15615436	28.38669352	52.55325014	45.96913354	29.99233195	39.05532305
26	43.95892745	39.16552813	67.15615436	28.21938305	52.55325014	45.96913354	29.99233195	38.81226045
27	43.95892745	37.93945025	67.15615436	28.21938305	52.32823424	45.96913354	29.99233195	37.93945025
28	43.95892745	37.93945025	66.51396551	27.80642466	50.49182914	45.96913354	29.88147740	36.69225583
29	43.95892745	37.93945025	66.51396551	27.56647199	50.49182914	45.70904351	29.24336746	36.69225583
30	43.65354949	37.93945025	65.66447461	27.56647199	50.49182914	45.70904351	28.69376540	36.69225583

Note. The number in the top row is the term set's online book ISBN13 number.

the term is observed in many documents of a corpus, it would be down-weighted somewhat. So a term in a single book might have a weight of 92.0 for that book, but if it appears in many books in the corpus, the final weight given to the term in that book might be 81.0.

terms, making the weight of a term difficult to use when considering the 'boundaries' of the overall term set. For example, if one book's term set starts with a weight of 83.7 and another book's term set starts with 48.4, those are both the highest weighted terms of the respective books. Therefore, in searching for the top terms for each book, the weight itself seems irrelevant; what does seem relevant is the *rank* of the term in a term set (i.e., the top 20, top 50, top 100, etc. ranked terms). In other words, looking again at Table 13, if only terms above the weight of 50.0 were considered for phylogenetic analysis, two books would not qualify to be in the classification, which is nonsense because all books in a collection need to be included in a classification.

Keeping with the principle of parsimony, the fewest number of terms that will differentiate a given book from another book will be the most parsimonious. However, some terms also need to be observed in other books to indicate (potential) similarity (which may suggest common origin). In other words, if every term set entered into the phylogenetic software were distinct (i.e., no words were ever repeated in other term sets), then there would be no way to create evidence for common origin. And one of the objectives of this study is to illustrate that terms extracted from the text of a book can be used to suggest common origin for a classification system.

In summary, the fewest terms with the highest rankings for each book are the most desired. But the higher the term is ranked, the less likely that same term will be observed in another book's set of highly ranked terms, which will likely not provide enough terms for phylogenetic analysis. Therefore, there will need to be a mix of highly ranked terms and lower ranked terms. But what is the mix needed?

Term weight characterization. To help with this decision (and in part, due to the evidence of OCR issues with the terms mentioned above), there was an intuitive need to conduct, at the least, a rough characterization of the terms to better understand the terms needed for data collection. Below is a very general characterization of the types of text terms for various weight classes after reviewing a few term sets.

Lower single weights (1.0 – 3.9). There are many general words and virtually no scientific/technical (S/T) words (e.g., club, broke, include, mayor, remarkable, safe, surpass, title, welcome, etc.). Also included were a few location and cultural proper nouns such as california, european, india, russian. (Note: the text mining tool displays terms in lowercase letters).

Middle single weights (4.0 – 6.9)*.* Emergence of general S/T single words (e.g., circulating, hemispherical, inflowing, neutrons, physiological, ultraviolet, etc.) and proper nouns of people (heider, kenworthy, mackin) were more visible. Still many general words and more location and cultural proper nouns. (Note: the text mining tool displays terms in lowercase letters).

Upper single weights (7.0 – 9.9)*.* Specific S/T single words become visible (e.g., hydroxyapitite, isobutyric, lamprophyre, neotrypaea, methanococcoides, n-cycling), multiple S/T variants of words emerging (hydrate-beating, hydrate-bearing, hydrate-derived, hydrogen-utilizing; paleolimnological, paleolimnology, etc.). More hyphenated terms emerging (e.g., diffusion-based, data-rich, decision-support, concept-oriented, county-level, ice-bedrock, open-ended, etc.). Still many general words (e.g., approximation, design, exclusion, famous, pile, quickly, separable, sort, sum, term, etc.). More names of people (brady, broecker, robles,

williamson, etc.) and more location and cultural proper nouns (e.g., derbyshire, german, louisiana, wichita, etc.). Abbreviations and fragments of words becoming much more visible (tmd, ces, dev, fro, usr, ver).

Lower- teens (10.0 – 12.9)*.* More variants with the same weights emerge (e.g., concentration, concentrate, concentrates, concentrated, concentrating, concentrations; connecting, connect, connected, connectivity, connects; determining, determines, determine, determined, determination, determinism; publisher's, publish, published, publishing; specific, specificity, specification, specifications, specifically, tolerate, tolerant, tolerance, tolerances, etc.). Less general single words but still many general words, though they were often variants (car, cars; likely, liking; paper, papers; road, roads; table, tables; tend, tends, etc.). More general S/T words and word variants: ecological, dichotomy; catalyzing, catalyzed; empirical, empirically; investigations, investigating, investigation, investigated, investigates, investigate; reactive, reactivation; statistically, statistical, statistics, etc.). More specific S/T single words (barnacles, cenozoic, iridium, valence, etc.). More abbreviations and fragments of words (nisms, respi, ticles, uring, etc.). Emergence of measurement abbreviations (n-min, gm-2, etc.).

Middle teens (13.0 – 16.9)*.* In general, more of everything mentioned above, but less of non-S/T words.

Upper teens (17.0 – 19.9). More specific S/T words, noticeable drop in proper names, virtually no non-S/T words.

20s and above. Mostly S/T words but noticeably more variants. Also noticeable were fragments of words and/or abbreviations and acronyms. Writing convention abbreviations (i.e.,

e.g., and al. from ‘et al.’) were observed from the weights from the 10s through the 50s, but primarily within the 30s and 40s.

In conclusion and in general, words weighted roughly 20.0 and above would be suitable for inclusion in a term set for science and technical books, though words weighted roughly between 7.0 and 20.0 may also contain relevant words. This latter statement is made with caution, but as stated previously, the text-mining tool’s creation and utilization of the corpus term set can assist in increasing *or decreasing* a single book’s terms.

Selection of term boundaries. Reiterating what is needed for data collection includes the fewest terms with the highest rankings for each book coupled with enough terms for phylogenetic analysis, thereby resulting in a mix of highly ranked terms and lower ranked terms. Given the above cursory term characterization, the highest ranked terms need to be weighted at 20.0 or above. Looking at the term sets, the number of terms to include as higher ranked terms would be 151 and above (i.e., the term ranked 152 in online ISBN13 9781118667262’s term set was 19.67). Of course, this is not necessarily the most parsimonious or optimized number. To reduce this further, an assumption of 10% of the 151 ranked terms would be sufficient, which means starting with first 15 terms for each book term. This number could be modified either higher or lower based on the results of the phylogenetic analysis (or could be replaced by a different approach)—but it provides for a method for the starting point for data collection. The top 15 terms from the 51 book term sets will be the characters used as data input into the phylogenetic software, giving an overall character set of 765 characters for analysis.

As for the lower ranked terms, and based on the above term characterization, no term in the term sets should be lower than 7.0. Searching for the ranked term closest to 7.0 placed the ending rank number for the all the term sets at 776 (i.e., online book ISBN13 9781118667842 was 7.01 at rank number 776). I chose not to reduce this because these will be used to find similarities among books, so I wanted to maximize those terms. In other words, I am minimizing the number of terms that differentiate the books and maximizing the number of terms that enable similarities to emerge to increase the likelihood that matches from the relatively few highly ranked terms will be found in other books. As with the selection method for identifying the boundaries of the highly ranked terms, the selection method for boundary identification of the lower ranked terms can be modified/replaced based on the results of the phylogenetic analysis.

In summary, the boundaries for the term set include the highly ranked terms 1 – 15 and the lowest ranked terms will end at the rank of 776 (see Table 8). This creates an initial term set of 39,576 terms, though this total was reduced during the term cleaning process (see *Term cleaning* section below). Furthermore, the boundary selection method developed above can easily be inserted into a fully automated classification system.

Conclusion. The term-set may have been stronger if 1) the text in headers/footers appearing on more pages than a title page were removed and 2) the text-extraction process after the initial results was repeated to remove/correct problematic terms, and possibly even repeat the extraction process again if more problematic terms emerged.

In future research projects of this type, the above steps should be performed, with the addition that the text extraction process would be used to identify text sources that needed

Table 8. Term Set Boundaries for Phylogenetic Analysis.

Book 1	**Book 2**	**Book 3**	**Book… *n***
Term 1	Term 1	Term 1	Term 1
Term 2	Term 2	Term 2	Term 2
Term 3	Term 3	Term 3	Term 3
…	…	…	…
Term 15	Term 15	Term 15	Term 15
…	…	…	…
Term 776	Term 776	Term 776	Term 776

Note. The highly ranked terms are from 1 – 15 and the lower boundary for the lower ranked terms is 776.

some other OCR processing. As for this dissertation's research project, a reasonable argument has been made for using text observed in headers/footers of multiple pages of a book, an analysis and subsequent removal of term-sets from potentially problematic text sources has been performed, and problematic terms were removed during the term-cleaning process discussed below.

Term cleaning

The 776 terms and weights from each of the 51 individual book term sets were placed into a separate spreadsheet for cleaning. The following summarizes the cleaning that was needed to create the top 15 terms for phylogenetic analysis (more details are provided in the next two subsections below):

- Removal of proper nouns that were not directly related to the intellectual content of the book (e.g., in-text citations of author names, author names appearing in headers/footers of multiple pages, locations used in a research project and repeated within one or two chapters);

- Removal of editorial conventions [for example, variants of e.g., i.e., and al. (fo*r et al.*)];
- Removal of numbers primarily observed in tables rather than the body of the text;
- Removal of terms in headers/footers that are not frequently observed in the text;
- Removal of fragments of words;
- Conversion of fragments if found in multiple variants (e.g., fragment *foli* and variants *folia*, *foliar*, *foliate*;
- Combined word variants that are the same or similarly weighted into one term using least amount of letters needed;
- Conducted search within spreadsheet of hyphenated words as potential non-hyphenated words and vice versa;
- Conversion of a non-US spelling of same word (e.g., organise would be written as 'organi').

Insight into the above process. Each of the top 15 terms were searched in the overall term corpus to both confirm the term and gain an understanding of the overall term corpus. If a term was observed in multiple-book term sets, that created confidence in the term. If a term seemed to be a fragment or was observed in only one book term set, that term was checked in the book and a decision was made to keep or remove.

Proper nouns (i.e. an individual person, place, organization) were removed unless they were used in referencing a scientific or technical concept, method, tool, landform feature, etc. and was observed in more than one chapter of the book or observed in more than one book (e.g., *Rossby* number, *Saltville* fault). The few proper noun terms removed were mostly author names, which appeared in the top 15 terms due to either multiple citations in-text or

combinations of in-text citations and author names included in chapter header sections. An argument could be made that highly cited authors are an important term. However, many chapter authors cite their own work (due to being an expert on the topic of the chapter) and those inclusions could inadvertently increase the term's weight. Furthermore, citation analysis is a well-known type of analytics in the information science domain, and this dissertation's project attempts to avoid any inclusion of citations in order to focus on the text itself.

For any questionable term, a search for the term (and its possible variants) was conducted in both the book from which the term set was extracted and the spreadsheet containing the other term sets before final decision was made. For example, if both 'multi-agent' and 'multiagent' was observed in the same term set, both would be searched in the spreadsheet. If 'multiagent' did not appear anywhere else in the spreadsheet, and if 'multi-agent' was the dominant of the two terms in the book, 'multi-agent' would be used and 'mutliagent' would be removed from the term set. However, if both 'multi-agent' and 'multiagent' are found in other term sets, both would be retained in the original term set.

During this term cleaning, there were other OCR issues discovered such as the terms 'lhe' and 'ihe' read by the text-mining tool, but a human would have read the word as *the*. When confirmed, terms such as these were removed.

Microsoft Excel was a useful application for this due to the search results providing all terms that had the connected letters. For example, in addition to 'two-dimensi' (the first letters in 'two-dimensional') the search results also provided 'two-dimensione', which would have been an unknown word to me.

Creating the top 15 term sets. The top 15 term cells for each book were checked for inclusion in other book term sets under two assumptions: 1) the top 15 terms are the most important terms that distinguishes a book from other books based on a relationship to the overall corpus via the Piranha algorithms and 2) if these top terms are observed in other books, there is the possibility the books are related, with a) the higher the weight of the term present in another book and/or b) the more terms shared with another book, the higher the probability the two will be related.

In creating the top 15 terms for each book, variants of terms were manually reduced to the same letters they all possessed (i.e., to their stems or roots). For example, *graptolitic*, *graptolite*, and *graptolites* were replaced by their stem 'graptoli'. The three terms appear with the same weights because the Piranha technology also contains a stemming feature, which reduces all stemmed words to the same weight. However, terms that contained stems were not always reduced to a single term. For example, *theca* was considered the same as its plural form *thecae* (which was removed), but not considered the same as *bithecae*, *epitheca*, or *intrathecal*, in part because there were different weights assigned to those, but also because these words have much different meanings than the stem. Reduction of terms with more complex variants (e.g., *folia*, *foliaceous*, *folial*, *folially*, *foliar, foliate*, *foliated*, *foliation*,) was limited to the terms with the same weights. For example, *foliate*, *foliated*, and *foliation* were reduced to 'foliat' rather than 'folia'.

And in the instance when a hyphen was used, the non-hyphenated term was also searched. For example, *self-organize*, *self-organizes*, *self-organized*, *self-organizing*, *self-organization,* and *self-organisation* would be reduced to 'self-organi' and the term also

searched as the non-hyphenated term ‘selforgani’. This also enabled other terms to be included in the top 15. Where there were multiple non-hyphenated terms but only one hyphenated word (e.g., *intracratonic* and *intra-cratonic*), the hyphenated word was replaced with the non-hyphenated word.

Terms that were not words or that did not directly characterize the book were removed. These included author names (e.g., *portugali*, *benenson*, *engelen*), years (*1960s*, *1970s*), measurement numbers (e.g., *68w*, *70w*, *72w*, etc., which were primarily limited to one book), customary abbreviations in writing (such as e.g., i.e., et al.) and in tables (including measurement units) were removed. However, a person’s name that had become a thing (e.g., *ruker* as the name of a mountain) and acronyms associated with a specific phrase in the book (e.g., *pcu* for Peru-Chile Undercurrent) and measurement units that were used extensively in the body of chapters were retained. Before removing these term deviations, they were all checked for 1) usefulness in characterizing the book and 2) presence in another book’s term set before deleting. The deletion enabled more useful term(s) to be included in the top 15 terms.

Terms that could easily be corrected (e.g., ‘lasma’ was clearly meant to be *plasma* from a search of the book) were considered for inclusion, with the understanding that the term weighting was based on ‘lasma’ not *plasma*. In this example, ‘lasma’ was ranked first in the term set and *plasma* was originally at row 41, before any changes to the term set. But after the changes to the terms set (e.g., removing term variants), it was ranked as 20. In contrast, after the same changes to the term set were made, ‘simulat’ ranked as 14, but *simulate*, *simulated*, *simulation*, and *simulator* ranked in the 230s rows, so ‘simulat’ was removed from the term set. Similarly, due to the OCR process, the term ‘fonnation’ was ranked first in a term set but the

term should have been *formation*, which would have been ranked 340 and thus 'fonnation' was removed from the term set.

Attention was given to distinguish between components of things and the things themselves. For example, the term *pyroxene* is a group of minerals/chemical compounds and the term *pyroxenite* is the rock comprised pyroxenes. Though the components and the things containing the components often appear in the same book term sets, in this example, more books (nine) contained the term for pyroxene than books (two) containing the term for pyroxenite.

Two term sets had top-tier terms that were removed due to too many observations of the term in header titles and not enough observations in the book chapters to warrant inclusion in the top-tier of terms: 'pi-interactions' in term set 9781119945888 and 'situ' in term set 9783527653218.

Character list creation

The top 15 terms for each term set were used to create the character list for input into the PAUP* phylogenetic software. The 15 terms from 51 book term sets generated 765 terms. A list was created of the 765 terms and 156 duplicates were initially removed, leaving 609 terms (characters) to begin character associations with the books (the taxa). In other words, each of the final terms would be checked for inclusion in each of the books' term sets and those books containing the term would be noted accordingly.

As discussed in chapter 3, the goal with creating the character list is to convert the taxa and characters into a row and column format with the taxa defining the rows and the characters defining the columns, and the cells in between illustrate whether or not the

characters are present in a given taxon. This structure ultimately becomes the matrix used as the input into the PAUP* software. Excel spreadsheet column identifiers (e.g., A, B, C… AA, AB, AC… etc.) were used as the surrogates for the book titles during this process because the cell's column is presented in the Excel search results. To determine which books contained the same characters, the characters were searched throughout the spreadsheet of terms discussed in the *Term cleaning* section above. Table 9 provides an example of this process.

Characters (terms) that were not shared with any other books represent distinguishing features of the book. And the more characters a book has with another book, the more likely they are to be 'related'. The characters shared by most books may be homologies (i.e., characters derived from a common ancestor) or they could be analogies (i.e., characters that look similar but are not derived from a common ancestor).

During this process, another 23 terms were removed due to duplication (i.e., same term

Table 9. Example of Characters (terms) Associated with Taxa (Books).

Characters	**Taxa Codes**						
deglaci	AF	BD					
delta	AH	BJ	AJ	CP	BD	AR	
denitrification	N	CD	BR				
denois	CH						
detrit	CV	CR	CN	BX	AR	BH	CP
deviatoric	BF						
devonian	BJ	AT	J	CN	AR	CT	CR
diabas	AP	AR	X	AJ			
diagene	CT	CN	CR	AL	AV	AJ	
diatom	BH	BD	CD	CF			
diffract	BL	CX	CN	AV	CR	CZ	
dimer	CJ	CX					

Note. The left column contains the characters and the other columns include the Excel column identifier for the taxa (books) where the character was observed. The character 'deglaci' was observed in both the AF and BD term sets.

or a variant of same term that already captured the sources under another term) leaving 586 unique terms as the characters for phylogenetic analysis. Once the above process was finished, the data matrix was developed.

Data matrix creation

To create the data matrix, the characters associated with each book were converted from the character table to the coded matrix form for the PAUP* analytical tool. To do this, the character list was copied into another spreadsheet using the transpose feature of Excel, which provided a vertical alignment for the 51 taxa (books) and horizontal alignment for the 586 characters (terms), enabling a transition step from the character table to the coded data matrix (see example in Table 10). During this process about 15 duplicate letters within columns (that were accidently included in the character list creation process) were identified and removed.

Each of the letters (representing a character present within the book) in the spreadsheet (D, F, H, AB, AH, etc.) were then replaced with the code [1], indicating the presence of the character, and where there were no letters in a cell (indicating the absence of that character in that book) the code [0] was inserted.

To make reading the trees easier, the spreadsheet column identifiers that represented the books were converted into quasi-subject names by modifying the book titles (see APPENDIX A: Corpus Dataset Book List for the list of books). The revised spreadsheet matrix was saved in a Tab-delimited text file format to enable the PAUP* software to import the matrix.

Outgroup: Hutton's *Theory of the Earth* (1788). A term set was also created from James Hutton's *Theory of the Earth*, which was used as the outgroup in the phylogenetic analysis. It was interesting to see how, in general, the important terms from 1788 seemed simpler in

Table 10. Example of Transition Step from Character Table to Coded Data Matrix.

Taxa (Books) Code	2-d (two-dimensi)	3-d (three-dimensi)	accelera	accret	adcp	adiabat	advect
D	D	D					
F	F	F					
H		H	H			H	H
J							
L		L	L			L	L
N		N			N		N
P							
R		R		R			
T							
V							
X				X			
Z		Z					Z
AB	AB	AB	AB			AB	
AD					AD		AD
AF					AF		AF
AH				AH			
AJ				AJ			

contrast with the terms observed from today's earth-science terms. Also noticed were simplification of terms: *schistus* is the older version of *schist* and *oeconomy* is an older version of what we consider today as *economy*. Not surprising, only six (6) of the top 15 terms characters matched terms in the 586 corpus set terms: *basalt*, *feldspar*, *observat*, *oolit*, *schist*, and *sediment*.[31] The selection of the outgroup is one that is either known or assumed to be the ancestor to all the other taxa and, therefore, should have fewer characters.

FINDINGS FROM PHYLOGENETIC ANALYSES

This section begins with some nomenclature for understanding a phylogenic tree. *Clades* are groupings of branches within the tree. *Leaves* are the terminal ends of the branches. *Nodes*

[31] The range of characters per books in the corpus was 21 (*Geosimulation - Automata-based Modeling of Urban Phenomena*) to 86 (*Tectonics of the Virginia Blue Ridge and Piedmont Culpeper to Richmond, Virginia, Field Trip Guidbook T363*) with an average of 50.8 characters per book (excluding Hutton's 1788 book).

are where the branches intersect and are considered the hypothetical common ancestor for the clade, thereby suggesting common origin, which was the primary theoretical impetus for this research project. The trees produced in this project are considered *unordered* because the books were not forced into a certain order—i.e., the book placements within the trees are the result of the phylogenetic analysis performed by PAUP*.

Data input

The final matrix file was imported into PAUP* by first selecting "Plain text (*.txt, *.dat)" file type, then selecting "Tab-delimited text" data format, and finally selecting the "Standard" data type. Once imported, the "Execute" command was selected, followed by the analyses and production of the trees. PAUP*'s default settings were used unless stated otherwise. Fifty-two (52) taxa were imported (51 ingroup, 1 outgroup) along with 586 characters. No weights were assigned to characters, and the taxa were rooted with terms from Hutton's *Theory of the Earth*, published 1788 and labeled OUTGROUP. Figure 12 provides a portion of the data matrix that was produced by PAUP*.

```
Begin data;
    Dimensions ntax=52 nchar=586;
    Format datatype=standard symbols="01" missing=? matchchar=.;
    Matrix
OUTGROUP                                00000000000000000000000000000000000000000000000000100
'Modeling -- Urban Phenomena'           00000011000000010000000000000000000000000010010000000
'Maps -- Self-Organising'               00000011000000000000000000000000000000000000000000000
'Meteorology -- Mesoscale'              00000001100011001000000000000001000000000000000011000
'Paleobiology -- Graptolites'           00000000000000000000000010000000000000000000000000000
'Atmosphere -- Fluid Dynamics'          00000001100011001000000000000001000000000000000011000
'Ocean -- Poleward Flows'               00000001000101000000001100000001010000000000000011001
'Cirripedia (Antarctic)'                00000000000000000000000000000010000010000000000000001
Subduction                              00000001010000000000000000010000001000001000001000100
'Crustacea Tanaidacea (Antarctic)'      00000000000000000000000000000010000000000000000000000
'Mineral Resources (Antarctica)'        00000000000000000000000000010110000000000001001000100
'Arcs - Trenches - Basins'              00000000010000000001000000010000000000001000001000111
Earth_Geophysics                        00000001000001100000000000001101000000000000001000100
Astrophysical_Plasmas                   00000011100010000010000000001000000000000001000000000
'Gulf of Mexico -- Circulation'         10000000000101000000011000000000100000000000000011001
```

Figure 12. Data Characteristics and Partial Data Matrix from PAUP* Output Display.

Analysis considerations

The PAUP* software enables characteristics of taxa data to be evaluated using different types of analyses to create phylogenetic trees, including *parsimony*, *distance*, and *likelihood*. In short, parsimony produces a tree based on the least number of character changes, distance produces a tree based on the amount of change between nodes, and likelihood produces a tree based on both character changes and branch distances. Each analytical method has various strengths and weaknesses, assumptions and limitations, and each has various parameters that can be changed. An evaluation of each method and their associated parameters was outside the scope of this project's objective of using a phylogenetic method to produce a sensible classification of books. Optimization—testing different analytical methods with different parameter settings in each method—could be conducted in future studies. Trees using each of the above methods were produced using the default settings in PAUP*.

Once the analysis method is chosen, the next decision is the type of search to be performed to find the optimal tree. One choice in PAUP* is an *exhaustive* search, which evaluates every possible tree. Unfortunately, my project has 52 taxa and PAUP* does not allow an exhaustive search for more than 12 taxa on a computer using a 32-bit operating system. According to Swofford and Bell (2017), "there are over 2 million trees for 10 taxa and 34 million trees for 11 taxa, so it is doubtful that exhaustive search strategies will be useful beyond 11 taxa" (p. 154). Therefore, a *heuristic* search was performed: "heuristic approaches… sacrifice the guarantee of optimality in favor of reduced computing time" (p. 160). The use of heuristic searching in PAUP* can be observed in related literature, for example Tehrani (2013) had 58 taxa and used the heuristic search in his phylogenetic study of similar folktales. To answer my

primary research question, a heuristic search will be sufficient, though high performance computers can be accessed from both universities and national laboratories for exhaustive searches of much larger datasets in future research.

Once the analysis has been completed, one of the next decisions is how to display a tree for evaluation. One of the aids for interpreting a tree is *rooting* the tree. Rooting helps indicate a direction of evolution (see Figure 13, (d) unrooted tree verses rooted tress of (a), (b), and (c)). Each node (junction of two branches) is considered a common ancestor. This means the root of the entire tree is considered the hypothetical common ancestor of all the other taxa being studied and, therefore, should have fewer characters. The outgroup is one of the taxa known to be outside all the other taxa. As previously mentioned, Hutton's (1788) *Theory of the Earth* was used as the outgroup to root the tree.

In addition to rooting with an outgroup, PAUP* also includes *midpoint* rooting, whereby the root is placed on the tree at the center point between the longest distance between two terminal taxa. This is conceptually intuitive: the first and last clades on a phylogenetic tree are the farthest apart from an evolutionary perspective. Using the midpoint method assumes that character changes happen across the tree at the same rate in every lineage. For a set of earth-science books, within a relatively small date range of 1969 - 2017, from English language sources, from the same publisher, this may be a fairly accurate assumption. However, a legitimate argument could be made that the more recent books represent more rapid change than the older books due to the increases in M-I-T-disciplinary sciences, more rapid dissemination via digitization and the Web, and no appreciable decrease in science funding that

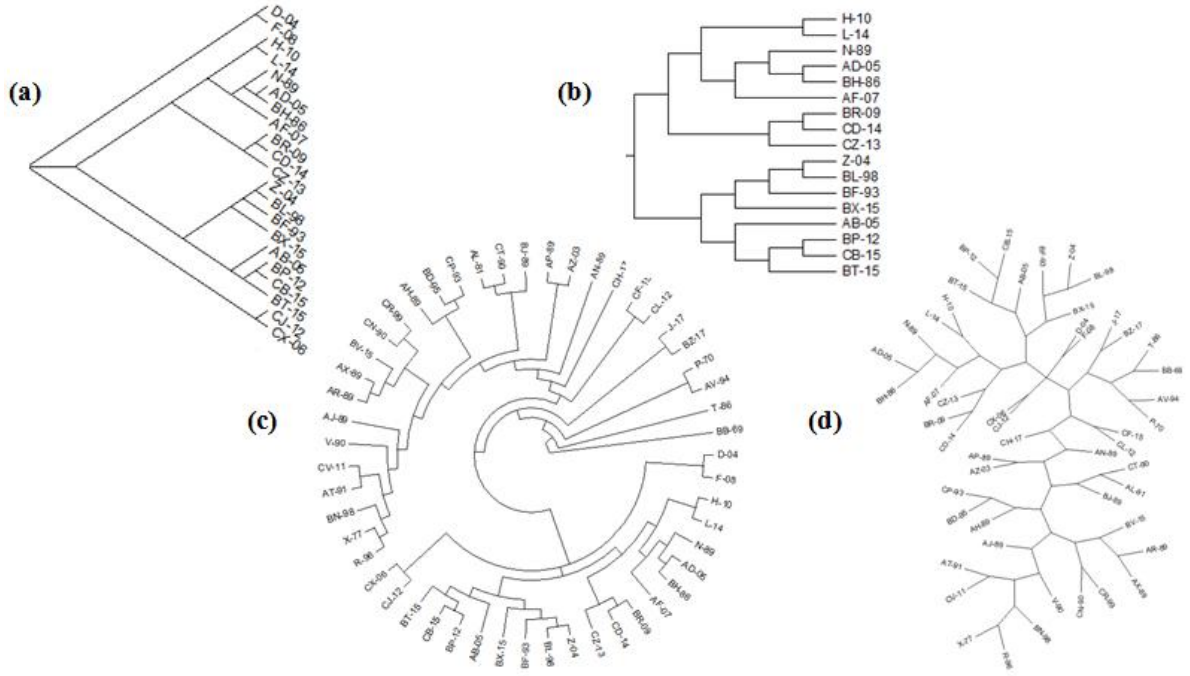


Figure 13. Different tree plot types available in PAUP*.
Slanted (a), rectangular (b), circle (c), and unrooted (d).

would have affected books published in 2017 and earlier.

Tree display considerations

Two types of trees can be displayed for evaluation: a *cladogram*, in which all the branches are considered equal (Figure 14, (a)) and a *phylogram*, in which the branch length represents the amount of character change in each taxon (Figure 14, (b)). These two types of trees can also be plotted in four variations in PAUP* (see Figure 13).

In summary, all of the above—the analytical criteria, the search method, the rooting method, the type of tree produced, and the plot type—are secondary to the primary research objective, which is creating a sensible classification of books. As stated previously, a heuristic search was used due to the limitations of my personal computer. I have chosen to root the tree

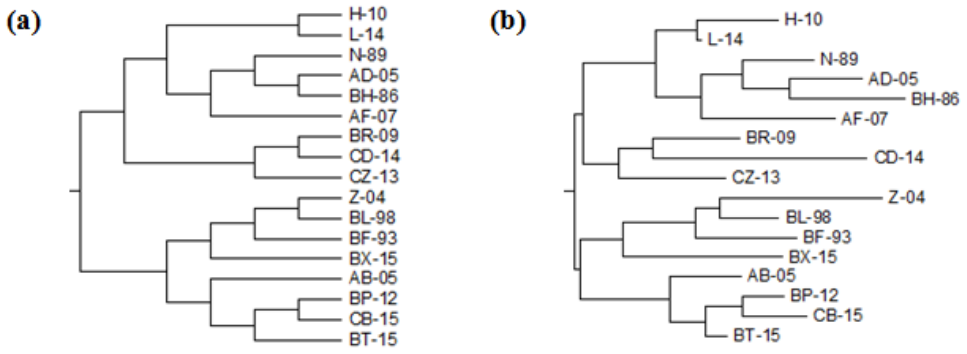


Figure 14. Two Tree Types of Displayed in PAUP*.

using a known outgroup related to the taxa. I do not believe a phylogram is superior to a cladogram for my particular project (though I use phylograms below for illustration purposes) and I prefer a rectangular view for the plot type. This leaves the decision regarding the type of analyses.

Creating trees from parsimony, distance, and likelihood analyses

Trees were created using each of the three analytical methods and applying the heuristic search and outgroup rooting.

Parsimony analysis. Parsimony produces the simplest tree, which is based on the least number of character changes within the tree. The total number of rearrangements of possible trees calculated by PAUP* was 1,288,741 and eight (8) most parsimonious trees were produced. All had a retention index (RI) values of 0.47. As previously discussed in the *Non-vertical transmission criticism (Galton's problem)* section of chapter 3, the RI indicates how bifurcating the data in the tree is. The closer to 1.0 the RI is, the more likely the taxa can be explained by a branching model of evolution (i.e. vertical transmission) and the closer the RI is to 0.0, the more likely the taxa are explained by a network model of evolution (non-vertical

transmission). As a reminder, Collard, Shennan, & Tehrani (2006) observed average RIs of 0.61 for biological datasets and 0.59 for cultural datasets. Though RI for the taxa in my project's dataset was 0.47, which is lower than either of the aforementioned datasets, the RI does not suggest a truly horizontal model of evolution either.

When there is more than one most parsimonious tree, PAUP* can summarize the trees using a *consensus* feature. I computed a single consensus tree from the eight trees using a *strict* (default) consensus setting (other settings include *semistrict* consensus, *majority-rule* consensus, and *Adams* consensus). "Strict consensus trees contain only those groups appearing in all of the rival trees... This can be considered to be the most conservative estimate of consensus..." (p. 215). Though Swofford and Bell (2017) warn that a consensus tree is not an optimal tree and cannot be interpreted as a phylogenetic tree (p. 215), for my research objectives, a consensus tree is acceptable. Figure 15 illustrates the parsimony consensus tree.

Distance analysis. The focus of the underlying algorithms is on the nodes of a tree and not necessarily the individual branches. So clades that are shorter (i.e., less amount of change) tend to be grouped together and clades that are longer tend to be grouped together. The total number of rearrangements of possible trees was 16,722 and one (1) tree was produced. The RI is not an output metric in distance analysis. Figure 16 includes the distance tree presented as a phylogram to highlight the shorter and longer groupings.

Likelihood analysis. Likelihood produces a tree based on both character changes and branch distances. Total number of rearrangements of possible trees was 38,667 and one (1) tree was produced. The RI is not an output metric in likelihood analysis. Figure 17 includes the likelihood tree.

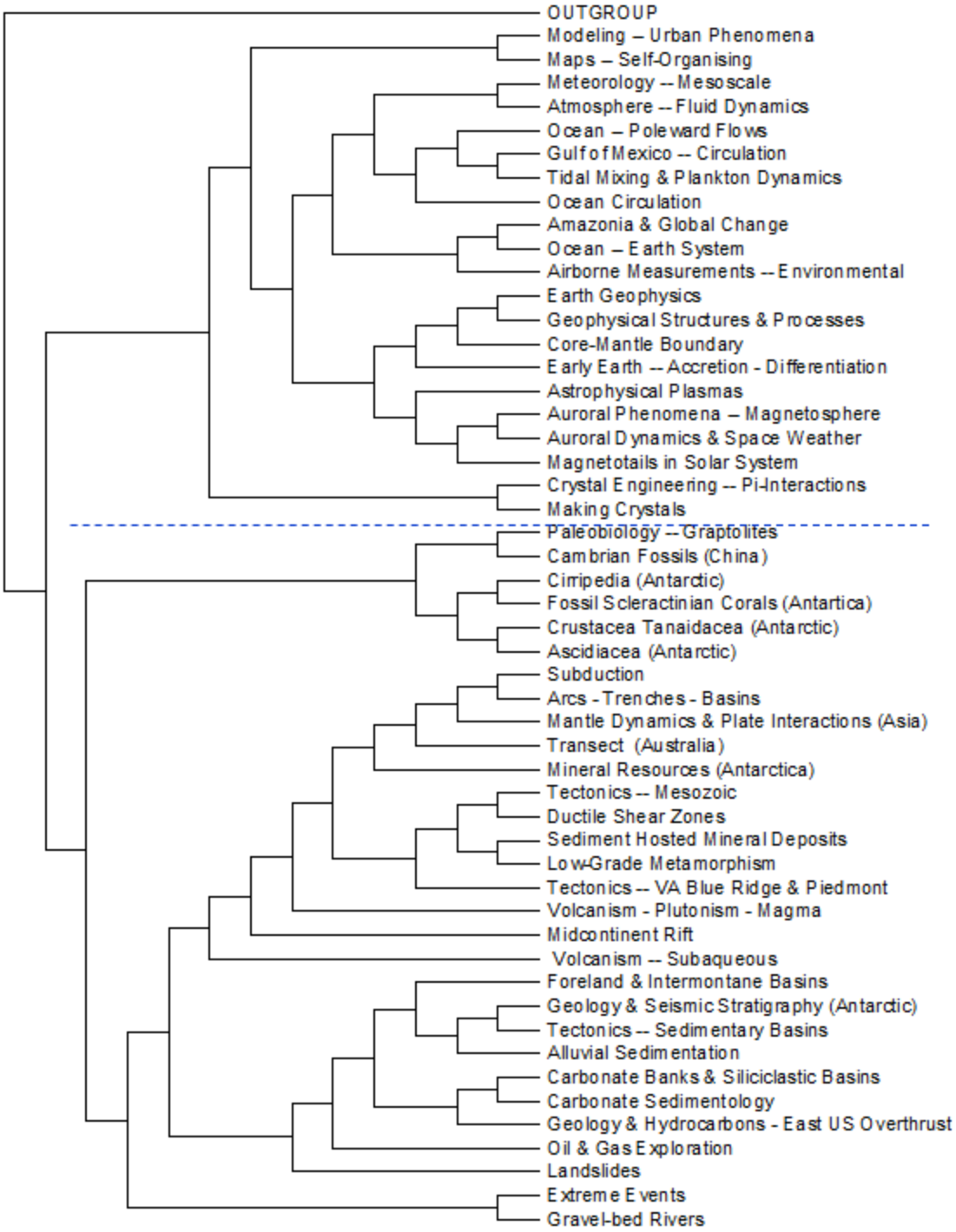


Figure 15. Cladogram Consensus Tree of Eight (8) Most Parsimonious Trees. The blue dashed line indicates the division between the two primary clades.

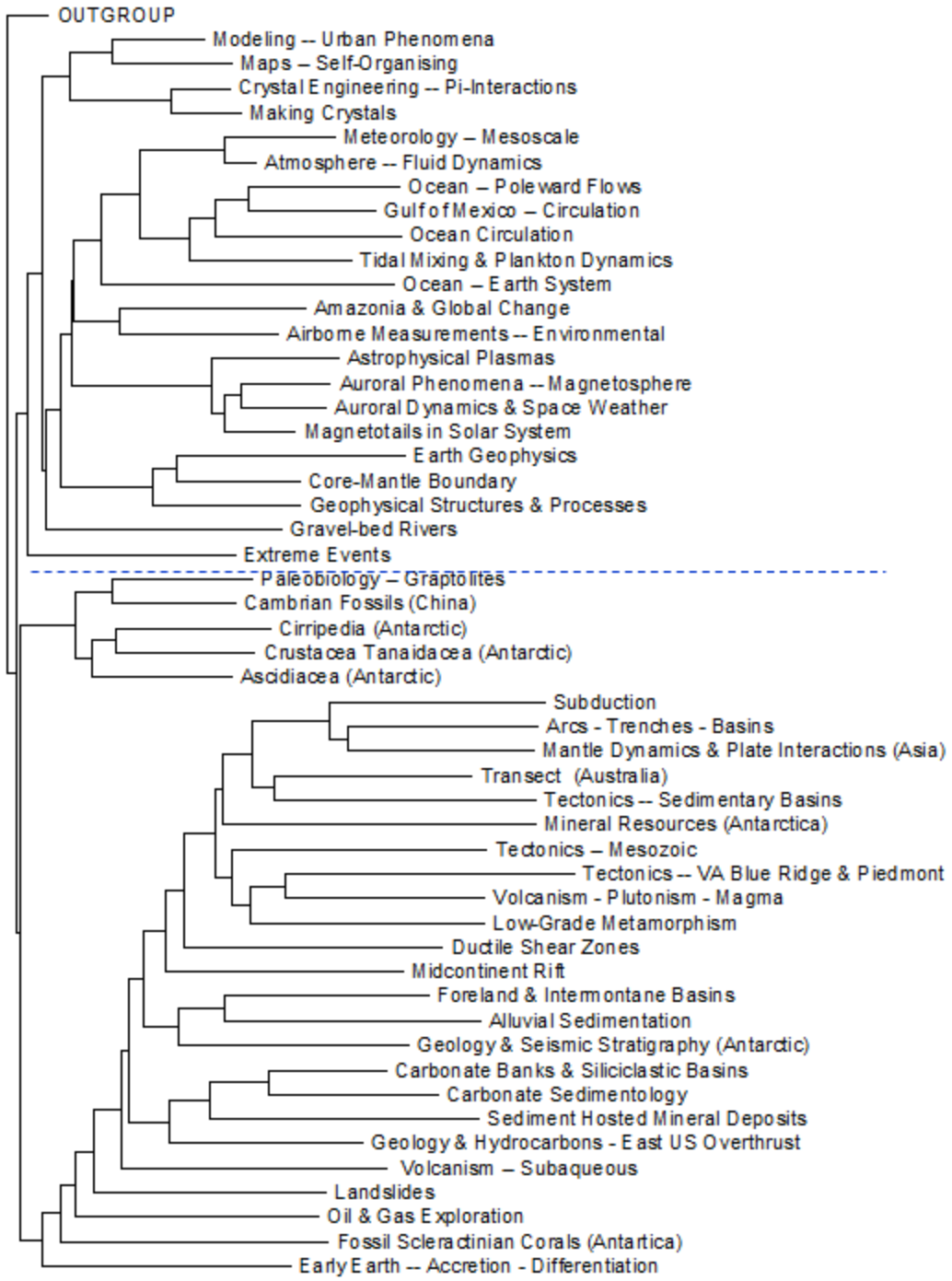


Figure 16. Phylogram Distance Tree Illustrating Groupings of Shorter and Longer Branch Lengths.

The blue dashed line indicates the division between the two primary clades.

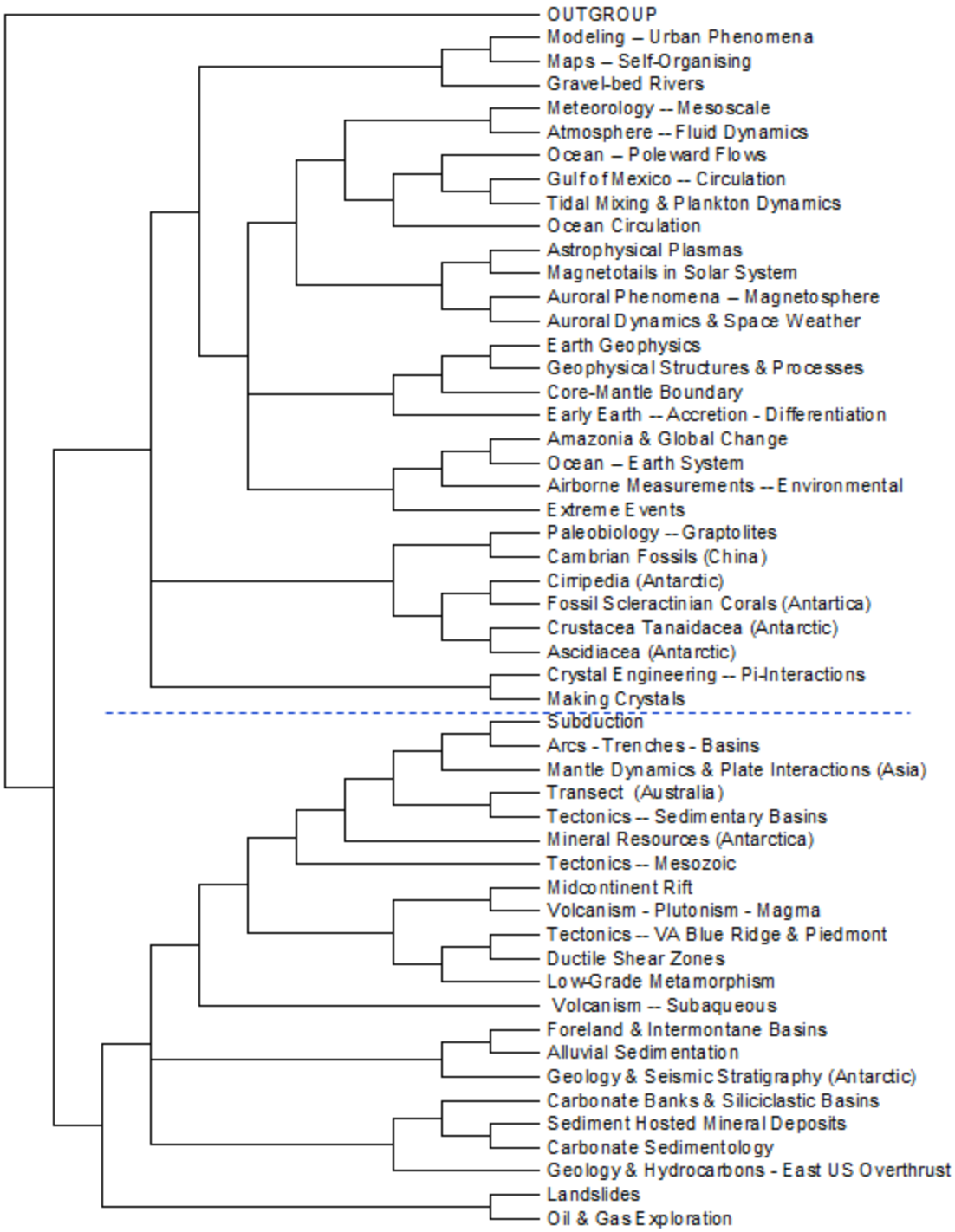


Figure 17. Cladogram of Likelihood Tree.
The blue dashed line indicates the division between the two primary clades.

Tree comparisons

There is more similarity than differences between the three phylogenetic trees. All three trees have two primary clades, what I term as *analytical* (computational, geophysics, earth systems) and *descriptive* sciences (e.g., fossils, biological, tectonics, geomorphology). And, for the most part, those sections mostly align. One glaring exception is a "biological" clade (which I believe uses a descriptive method based on the books in the clade) is observed inside the section of the likelihood tree I consider to be the analytical section of the tree. (More about this is included in the Likelihood tree section below.)

Parsimony tree. The *analytical clade* of the parsimony tree is divided into two clades. The smallest is a crystallography clade consisting of only two books. The larger clade can be divided into geographic and geophysical clades. The geographic clade is the smallest (two books) at the top of the parsimony tree. The other clade can be considered a geophysics clade, that is further divided into 1) lower atmosphere & ocean clades and 2) earth geophysics & upper atmosphere clades.

The two geographical books are computational-based (i.e., simulation and automation), which seems reasonable to include in an analytical grouping (this clade is also in the same position on each of the other trees). The books in the geophysical clade tend to be on the physics side of earth-sciences. And the crystallography (third clade) is also a mathematical-based discipline.

Within the *descriptive clade* of the parsimony tree, there are two primary clades. At top, there is a relatively small biological clade (six books including both extant and fossil life forms). The largest clade is a lithosphere clade and is divided into two clades. The smallest clade (two

books) is visible at the bottom of the parsimony tree and is best described as a miscellaneous phenomena clade. The extreme events book covers a wide range of events (e.g., space weather, asteroids, flooding, hurricanes, earthquakes, volcanoes, landslides, etc.), making it difficult to fit into the other clades. The gravel-bed rivers book is an anomaly due to being the only river-related book in the collection. It also appears in different places in each of the trees (bottom of the parsimony tree, near the middle of the distance tree, and at the top of the likelihood tree). It appears together with the extreme events book in the distance tree but not in the likelihood tree. It is notable that in the parsimony tree, the extreme events book is in the descriptive clade and next to a landslides book (but in an adjacent clade) and landslides are considered extreme events. In contrast, the extreme events book is in the analytical clade in the other two trees (i.e., nowhere close to the landslide book).

The largest lithosphere clade is divided into two distinct clades: 1) a lower lithosphere clade for processes at work deep within the earth (e.g., subduction, rifting, volcanism, and metamorphism) and 2) an upper lithosphere for deformations and activities near the upper part of the crust (e.g., alluvium, stratigraphy, sedimentation, field trips, and oil/gas exploration).

Distance tree. At the top of the distance tree's *analytical* section is a clade that includes the two geographic books and the two crystal books. The node that connects those is supposedly the hypothetical common ancestor. It is difficult to imagine what that common ancestor would be. In the *descriptive* section's first clade, one would think the taxon "Volcanism, Plutonism, and Magma" would be relatively close to the "Volcanism: Subaqueous" taxon. Similarly, the taxon "Fossil Scleractinian Corals (Antarctica)" should have been in the biological/fossil clade. Both of these are observed in the parsimony and likelihood trees (though

the proximity to each of the volcanic books is closer in parsimony tree than in the likelihood tree), but not in the distance tree.

Likelihood tree. At the top of the tree, a river-related book has been connected to the two geography books. If the river book was a map of the river, that might be somewhat possible, but it is not intuitive given the topics of modeling/simulating urban phenomena and self-organizing maps. The exact biological clade (i.e., same order, same nodes) is found in both the parsimony and likelihood trees, but the clade fails a "sensible" test in the likelihood tree due to the clade 1) being placed in what I perceive to be an analytical section of the tree and 2) it is located between the earth physics/systems clade and the crystal clade.

Conclusion. Even though there are more similarities among the three trees, the few differences mentioned above are enough to lose confidence in the ability of either a distance or likelihood tree to produce at sensible book classification. The parsimony tree has a clear, logical flow of books from the top of the tree to the bottom. The groupings from the clades seem reasonable as well. Any anomalies seem to be placed at the beginnings or endings of the tree of the two main sections of the tree. It seems distance—which is present in the distance analysis and, to a lesser degree, in the likelihood analysis—introduces more non-sensible clades than does the tree produced by parsimony analysis.

Library of Congress Classification comparison

In addition to comparing the three trees, another analytical approach to answering whether or not a phylogenetic tree can produce a sensible classification is to compare the trees with a known classification. The Library of Congress Classification (LCC) is appropriate for such

an analysis due to most academic libraries in the United States (and in other countries) using LCC.

Conversion of taxa into an LCC order. A search for each of the 51 books was conducted using the Library of Congress Online Catalog (https://catalog.loc.gov/vwebv/searchKeyword) and inputting each print (rather than online) book's thirteen-digit International Standard Book Number (ISBN13). Any that could not be identified with the ISBN13 were searched by the online book's ISBN13 and/or the titles. Of the 51 books, 39 were identified and their call numbers were retrieved. The other 12 books were searched using the Online Computer Library Center (OCLC) WorldCat (https://www.worldcat.org/). All but three (3) books were available through WorldCat and the LCC call numbers were obtained by member library holdings. The other remaining two book call numbers were obtained LCC search of books with the same or similar titles. The LC Subjects were obtained by browsing LCC call numbers for the closest matching numbers. For the book "Crustacea Tanaidacea of the Antarctic and the Subantarctic...," there were two LCC call numbers listed: QH95.58 .B56 vol. 18, etc. and QL444.M38. I used only the latter to be consistent with another similar book, "Antarctic Cirripedia," which did not include the LCC QH (Natural History) subclass.

The same taxa names used to produce the above trees were used in the LCC classification to represent the books for an easier comparison with the phylogenetic trees. Figure 18 is the classification tree created for the comparison. All taxa were ordered by the relevant LCC call numbers.

Discussion. Beginning with the top of the LCC tree in Figure 16, the first observation that catches attention is oceanography-related books are inserted in between two geographical-

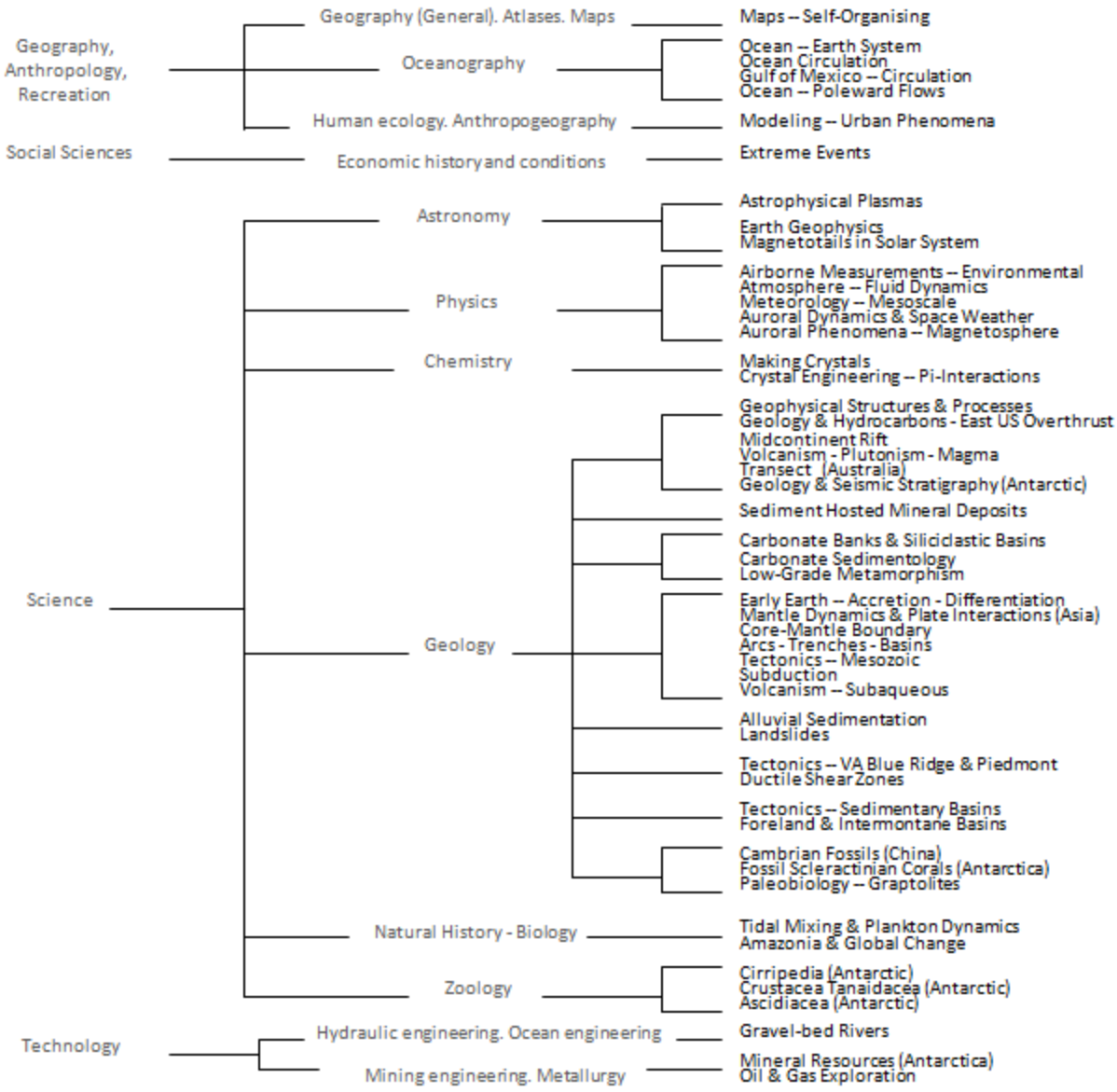


Figure 18. Taxa Ordered by Library of Congress Classification (LCC) Call Numbers. Relevant LCC classes and subclasses appear in a gray font.

related books. All three of the phylogenetic trees placed the two geographical-related books together, which I consider to be correct from a proximity perspective. “Astrophysical Plasmas and Magnetotails in Solar System” are placed in Astronomy in LCC and “Auroral Dynamics & Space Weather” and “Auroral Phenomena – Magnetosphere” are in Physics in LCC. The

likelihood tree does connect each pair as observed in LCC, but all phylogenetic trees group them together in one clade, which is not the case with LCC.

A separation between the extant lifeform books and the fossil books were made in LCC and they were mixed in the phylogenetic trees. The separation issue I raised with the two volcano-related books is also evident in LCC tree. The ocean books are closely located in the parsimony and distance trees, though “Tidal Mixing & Plankton Dynamics” was also included in the ocean clades, but is in the LCC Natural History – Biology subclass (though it is easily understood why a tidal/plankton book would be in the ocean clades).

It is evident from the LCC investigation that as science gets more complex and diverse it will become increasingly difficult to provide a book with a single call number. And as digital books continue to expand, providing multiple call numbers could be a solution due to not needing a single place on a shelf. But an evolutionary classification system is also promising when comparing the LCC tree with the parsimony tree.

CONCLUSION

Despite issues discussed in the findings from dataset development, a method for a sensible book classification using parsimony phylogenetic analysis with only words or phrases used as the characters for analysis, which was the primary research question, was supported by the overall findings. The nodes on the clades represent hypothetical common ancestor, therefore a successful proof-of-concept study was accomplished thereby proving a classification based on common origin can be used for a bibliographic classification. Due to the perceived logical flow of the ordering observed in the parsimony tree, the evolutionary classification does

seem to be intuitive even if a user has only a basic understanding of the corpus material, thereby eliminating the need for a subject matter expert to classify the books.

The close logical groupings found in the parsimony tree when compared to some of the distances between the same books in LCC suggest an evolutionary classification could be more efficient for classifying phenomena. And with the variety of different subjects of books observed in the LCC tree, there is evidence to suggest the evolutionary classification method presented in this dissertation would be sensible for any science and technical domain—whether it be for knowledge organization, information retrieval, or evolutionary research of a discipline / field.

CHAPTER 5 – CONCLUSIONS, DISCUSSION, AND SUGGESTIONS FOR FUTURE RESEARCH

INTRODUCTION

This chapter includes a summary of the findings from the dataset development and the phylogenetic analysis. Conclusions of the research questions are also provided, followed by suggestions for future research and final discussions of the how this project's results could contribute to society and scholarly domains.

SUMMARY OF FINDINGS

Findings from dataset development

In a phylogenetic study, the data is just as important as the outcome and data handling became a large part of overall project. In fact, the vast majority of my unexpected findings was from data handling. I had determined that 20 – 80 books would be needed for the proof-of-concept project. Eight-five (85) books were randomly sampled and cleaning was done reduce the books to the intellectual content of the book, meaning I did not want references to interfere with the text-mining selections. Nineteen (19) books needed to be removed due to not being able to remove the references sections. Another was removed because several chapters only had abstracts, which left 65 books for term extraction.

There were two notable observations during the initial review of the term sets: 1) the presence of fragmented words and variants of words and 2) the presence of highly weighted terms that were also observable in the titles of books and/or chapters printed in headers and footers of the books. Reviewing some of the books indicated the optical character recognition

(OCR) process was an issue, and was causing the fragments. This resulted in another 14 books removed, which left a total of 51 books for data set creation, which was still in the range of 20 – 80 books needed. A review was also conducted to make sure no header/footer terms were inflating the term sets, which led to two datasets having some of their top terms removed (and were replaced with the next available terms).

The next challenge was setting the term-set boundaries. The least number of terms extracted from a book was 1,702 and the most was 24,056, with most book term sets ranging roughly between 5,000 – 8,000 terms. This required an analysis of the types of terms and their changes according to their weights (the text-mining tool weighted the book terms with higher scores being more important than lower weights) to determine the range of terms to use. The fewest terms with the highest rankings for each book are the most desired. But the higher the term is ranked, the less likely that same term will be observed in another book. This resulted in selecting 15 terms from each book and limiting the total term set for each book at 776 terms (i.e., the terms that could be searched to determine if where the similarities were among the books, which were used as "characters" for phylogenetic analysis, in the way genetic material is used for such an analysis to determine which books might be "related").

The terms then needed to be cleaned such as removal of proper nouns that were not directly related to the intellectual content of the book, editorial conventions (variants of e.g., et al.), numbers primarily from tables, fragments of words, reducing words to their stems (the parts of the words that would enable the most variants to be found), etc. From that point forward, it was a matter of locating matching terms in other books, documenting those in a

spreadsheet, and converting the spreadsheet to a form recognizable by the phylogenetic software.

Findings from phylogenetic analysis

Challenges were encountered mainly from making a data conversion and learning new research software to create the classification tree. Findings included the creation of three (3) trees by three different types of analyses (parsimony, distance, and likelihood). A review of those trees indicated parsimony analysis would be able to support the primary research question and it seems that phylogenetic analyses containing distance measurements may not be suited for bibliographic classification. Distance analysis produces a tree based on the amount of change between nodes and likelihood produces a tree based on branch distances (but also considers character changes; somewhat of a combination of the other two analyses). Though distance-related analytical methods may be beneficial for biological species and maybe even other cultural datasets, this seemed to be detrimental for a book classification.

CONCLUSIONS

Research questions

Below are the answers to the five (5) research questions.

Can a phylogenetic-based method be developed to produce a sensible classification of books with only words or phrases used as the characters for analysis?

Yes. The parsimony tree produced a sensible classification of books (the trees produced by distance and likelihood analyses did not produce a sensible classification). More specifically, when looking at the grouping of books within the clades—whether a broad clade or a more

narrow clade—the classification tree produced by the parsimony classification were grouped in such a way that was logical: geographic with geographic, ocean with ocean, space with space, crystallography with crystallography, biological with biological, etc.

And if the same method was applied again to the same books and the same term extraction and phylogenetic tools were used, the results would be the same, making the method reliable. Additionally, the resulting parsimony tree is valid, in that the tree indicated what is was supposed to indicate—a sensible classification, with the caveat that 'sensible' is, of course, subjective. There is also the potential that a different term extraction tool may not produce the same or similar validity.

Does the classification have natural hierarchical groupings of books?

Yes. The clades in the parsimony tree produced relatively natural hierarchical groups: geographic, atmospheric, oceanic, biological, crystallographic, and lithospheric grouping were easily discernible. In other words, there was not just a collection of individual branches but rather logical groupings based on related phenomena. To better illustrate, below includes four of the highest hierarchical categories observed in the parsimony tree:

- Analytical clade
    - Geographic & geophysics clade
        - Geographic clade
        - Geophysics clade
            - Lower atmosphere & ocean clades
            - Earth geophysics & upper atmosphere clades
    - Crystallographic clade

- Descriptive clade
    - Biological clade
    - Lithospheric clade
        - Upper/lower lithospheric clade
            - Lower lithospheric clades
            - Upper lithospheric clades
        - Miscellaneous phenomena

Is this representation of books useful?

Yes. The visualization of the classification in a rooted tree structure provides a simple path to each book via clear hierarchical groupings (i.e., the clades). And the books shown at the tips of the branches enable browsing of similar books.

Consider the parsimony tree converted into an interactive classification whereby the two main clades would first be provided to a user, labeled accordingly. The user chooses (e.g., clicking with a mouse or touching the screen) one of the clades, which opens the next level of clades, with each clade labeled accordingly. This type of action continues until the user locates the most relevant book.

In another interactive use case, a user could enter keywords into a search box and the largest clades that contain the keywords would become visible to the user. The user then chooses one of the clades to 1) continue down the branch (i.e., selecting the next group of clades) and/or 2) reverse and select another of the largest clades to investigate.

Is there any insight from the project regarding the emergence of a new discipline or field similar to the emergence of a new biological species?

No. A larger number of books from different publishers would be needed to make that determination. The year of print publication for each book would likely also need to be incorporated into the classification.

Could this result in an automatic classification method?

Yes, this could be an unsupervised automatic classification, but only if the OCR processing consistently enables a text-mining tool to interpret a group of letters (e.g., a word) as a human would interpret those letters. Evidence from this dissertation project suggests several challenges to unsupervised automation of the entire method when the OCR processing is less optimum. For example, automating the term cleaning process would need to be resolved. There could also be a challenge with term review was with acronyms. For example, 'dle' was an acronym for 'discrete logistic equation' in one book, and the term appeared in several other term sets, but they were all fragments (e.g., of 'handle', 'middle'). When observed, these fragments were removed from the term sets. 'Two-dimensional' is also observed as '2-d' and 'multi-agent' could also be a variant of the acronym 'MAS' (multi-agent system). Another challenge is with stemming (the root of the word that will generate desired character matches). For example, *fusellum* can be reduced to 'fusell' or 'fusel'. If the former, a search of the terms accurately locates related words such as *fuselli*, *fusellar*, *fusellus*; if the latter, a search would also include *fuselage* in the results.

However, overcoming these challenges to automate the method seems doable, given that OCR technology has improved as the quality of text that is read by the technology has improved. And, of course, if a book publishing industry standard for term extraction emerged whereby publishers provided a table of ranked terms with each book, this would be a moot

point. The *Knowledge organization and information retrieval fusion* section below includes comments on automation of each step of the method.

SUGGESTIONS FOR FUTURE RESEARCH

Near-term research should be considered for optimization and enhancements. Longer term projects include expansion of the model.

Optimization

Higher quality term sets. The term-set could possibly be made stronger and lend itself to a more automated method if 1) the text in headers/footers appearing on more pages than a title page were removed and 2) the text-extraction process after the initial results was repeated to remove/correct problematic terms, and possibly even repeat the extraction process again if more problematic terms emerged.

In future research projects of this type, the above extra steps should be conducted with the understanding that the text extraction process could be used to identify text sources that need some other OCR processing. As for this dissertation's research project, a reasonable argument was made for using text observed in headers/footers of multiple pages of a book, an analysis and subsequent removal of term-sets from potentially problematic text sources was performed, and problematic terms were removed. However, more research in this area is needed to determine exactly what should and should not be included in the term sets. If natural intellectual content is the optimum, then removal of additional header/footer terms will likely need to be removed.

And as research continues, a list of problematic terms would emerge that can ultimately be used to eliminate acquisition of non-terms. For example, Figure 19 contains the list of top 100 terms from this project's corpus term-set to illustrate the fragmented terms (e.g., tions, ments) and writing conventions (e.g., e.g., i.e.) the text-mining tool recovered as terms. The initial corpus term-set could be used to program the text-mining tool to disregard these types of terms (a total of 176,993 unique terms were extracted from the corpus).

Parameter sweep. For future studies, a parameter sweep should be conducted, which was discussed with Dr. Robert Patton during this project[32]. For example, starting with 10 books, the text-mining tool would be used to extract the five (5) most important terms from each of the 10 books. Those terms would be used to create the book/term data matrix, then the matrix would be entered into the phylogenetic software to produce the tree. These steps are repeated

1	tion's	21	i-e	41	als	61	ing	81	fig
2	tions	22	i.e	42	aling	62	inging	82	figs
3	tion	23	i-e.	43	microal	63	tives	83	.sediments
4	.tion	24	ies	44	ales	64	tively	84	sedim
5	tioning	25	al'	45	al.'s	65	tiveness	85	sedime'nts
6	tione	26	al'.	46	littl	66	tive	86	sediment
7	tioned	27	ale	47	little	67	longs	87	sediments
8	eges	28	al	48	ment's	68	ultra-long	88	sedime
9	e.g	29	al_.	49	ments..	69	longed	89	sediment's
10	e.g	30	a_l	50	ments	70	longe	90	sedimented
11	eg	31	al's	51	ment.	71	long	91	sediment.
12	e-.g	32	al	52	ments	72	micro-gradients	92	sedimenting
13	e-g	33	al.	53	mented	73	gradient.	93	.sediment
14	e-g	34	a.l	54	menting	74	gradients	94	variabl
15	ege	35	aled	55	ment	75	gradient	95	variability
16	e.g.	36	al.	56	me-nt	76	fig	96	variabilities
17	e.-g	37	al	57	ingful	77	fi_g	97	variable
18	ie	38	a-l	58	inge	78	f'ig	98	variably
19	i.e.	39	al	59	ingness	79	f.ig	99	variables
20	.i.e.	40	al'e	60	ings	80	fig	100	turing

Figure 19. Top 100 Terms from Corpus Term-set.

[32] R. Patton, personal communication, September 2017.

with 10 top terms from each book, increasing the number of terms by five (5) until there is some indication of optimal quality (e.g., the most sensible classification tree). Once optimized, add books by doubling the number of books and determine if changes need to be made to the number of top terms. It should be noted that as the number of terms increases, there is a diminishing return phenomenon that occurs due to the increase in 'noise', which is a known issue in data/text-mining. Thus, the principle of parsimony should be considered: the least number of terms required to characterize a book for a sensible classification will be the preferred number of terms to use.

Other optimization. Once the system is optimized for the original set of books, the addition of more books is a natural next step to explore. For example, including other similar earth science books from other publishers, followed by books from other scientific domains. Other text-mining tools should also be tested to learn whether or not the same classification results always emerge. This could result in a decision for using only one text mining tool or possibly several tools in different ways. Ultimately, fully automating the various steps in the method would be needed for society to get the most benefit from the technology.

Expansion

Additional data. Adding additional data could also be explored to enhance the system. For example, adding citation and/or sales data to the classification to enable a user to visualize the major works in the evolutionary histories. Adding publication dates would enable modification of tree branches for accurate placement of the book on the evolutionary tree.

Beyond science and technology. The data was collected from scientific and technical books, so there are very specific words/phrases used that likely made this classification project

easier to demonstrate. Another research track is to modify this methodology to create a classification of less technical books—for example, fictional works such as romance novels. For this, exploration into other machine-based text analysis tools such as genre analysis, content analysis, discourse analysis, and sentiment analysis would need to be considered.

Beyond books. This project relied on earth science books, many of which were written by multiple authors in a single book. This is somewhat like a journal article. Therefore, another expansion would be into journal articles and other text-based information packages. Another interesting possibility would be an evolutionary classification based on analysis of music notation (i.e., sheet music).

Speciation. An assumption is that if enough books within a certain domain are present on an evolutionary tree, we should be able to see when a new species has emerged (i.e., speciation). In nature, speciation often occurs when dramatic changes take place in the environment (termed a 'speciation event'). One approach to understanding this would be to first start with a fully populated evolutionary tree and study a known speciation to gain insight need to define other speciations. If this is possible, then it may also be possible to predict speciations in advance or as near to the time of speciaion as possible. And, of course, prediction is one of the primary goals of scientific research.

DISCUSSION

This final section covers three topics to indicate the importance of this type of project. The subsections include fusions between knowledge organization and information retrieval, information needs beyond science, and the connection with a unified science of cultural evolution.

Knowledge organization and information retrieval fusion

Classification schemes can be categorized within the broader category of knowledge organization systems (along with subject headings, thesauri, ontologies, etc.). Ultimately, all knowledge organization (KO) systems are designed, developed, and used to make management and retrieval of knowledge and information easier (Mazzocchi, 2018, p.54). But this project is an example of how information retrieval (IR) technology (in this project, text-mining term-extraction and weighting) can be used to enhance KO systems in at least two ways: first by providing a "first cut" at a classification and second by suggesting possible alternative placements within a traditional classification scheme. This method could possibly be a replacement for an knowledge organization (in this project, with a new type of classification). In other words, the line between KO systems and IR continues to blur as more advancements are made in machine learning.

This dissertation project aids in visualizing how this is possible: as long as text is in a form that enables a machine to interpret a group of letters (e.g., a word) the same way a human would interpret that group of letters, the method presented herein could be automated. Specifically:

- A collection of digitized books is scanned with a text-mining tool, and rank-order term sets for each book are automatically created (acronyms can be identified/included with current or additional computer programming);
- Non-words can be eliminated from the term sets with dictionaries; if needed, synonyms can be added to the term sets with thesauri);
- Words having the same stem can be collapsed into a single term (i.e. the stem);

- Parameters for the terms to be included in the character matrix (i.e., number of top terms, lowest ranked terms to be included) can be automated;
- A character matrix can be automatically created from the term sets;
- Automatically inputting into phylogenetic software, automatically running a parsimony analysis, and automatically producing a classification tree can be accomplished, provided the software owner permits modifications made directly to the source code.

As of this writing, there is much information being generated regarding artificial intelligence, in particular, deep learning via neural networks. Considering the speed at which new technology emerges today, intuition suggests that integrated KO/IR examples, such as the one provided in this dissertation, will continue to be produced and the lines between the two will continue to blur into fully automatic, integrated systems.

Information needs beyond science

At the end of chapter 4 there was the suggestion that the evolutionary classification could be used for classifying phenomena, which is more amenable with the projection of increases in the volume, diversity, and complexity of scientific and technical information. But a phenomena-based classification may also be more amenable to *any* information user. Kumbhar (2012) suggested that even though a formal classification (e.g., LCC) "is a satisfactory method of linking a catalogue record to an item on the shelf, it does not facilitate browsing in the areas of most interest to public library users" (p. 87). The author discussed research suggesting that when public libraries organized collections primarily according to topics rather than following a traditional library classification scheme, patrons either benefited from the topic classification or did not notice any difference.

Additionally, in the era of "big data," there is the recognition of the need for graphical displays to help us quickly obtain relevant information. Therefore, a graphical representation of a library's holdings should also be incorporated into a classification system. Returning to Julien et al., 2012 (mentioned in chapter 2), after creating the Library of Congress Subject Headings (LCSH) tree and conducting the subsequent analysis of the tree, one of the conclusions reached by the researchers is relevant here:

> The tree was characterized in terms of its size, children per node, and depth. This revealed that the structure was large, highly redundant due to multiple inheritances, very deep, and unbalanced. ***The complexity of the LCSH tree is a likely usability barrier for subject browsing and navigation of the information collection*** (p. 2417, emphasis mine).

As evidence from this project suggests, an evolutionary classification based on common origin would likely provide an intuitive browsing experience. It may also provide a more affective searching experience if a user can actually see the evolutionary relatedness of a book or group of books. In other words, for those of us who like a broader understanding of 1) the universe in which a group of books (or documents) resides and/or 2) the genealogy that lead to a book or group of books, but do not want to spend hours browsing just to satisfy that affective need, an evolutionary classification could provide quick satisfaction and/or knowledge.

Unified science of cultural evolution and cultural systematics

Finally, this classification research project continues investigations that apply biological phylogenetics to cultural phenomena. As mentioned in the literature review chapter, Mesoudi, Whiten, & Laland (2006) suggested a science of cultural evolution could be conceptually

structured according to a framework similar to that of evolutionary biology (refer back to Figure 7), which could ultimately result in a unified science of cultural evolution.

However, I have found virtually no formal discussion of the cultural analog of biological *systematics* discipline of which phylogenetics is a sub-discipline. The tasks of systematics include:

1. Identifying, naming, describing, and cataloging organisms;
2. Conducting comparative studies of all aspects of organisms;
3. Classifying organisms into groups of higher taxonomic rank;
4. interpreting the contributions of lower and higher taxa to the operations of nature and to evolutionary history;
5. Discovering the phylogenetic (genealogical) relationships among organisms, (Abbas, 2010, pp. 83-84; Futuyma, 1998, p. 12; Mayr & Ashlock, 1991, p.5).

Given the above tasks, information science classification researchers are well-suited to play a vital role in cultural systematics. Additionally, a classification system developed for this unified science could provide:

1. A process to normalize disparate cultural data for research;
2. A data source for discovery of evolutionary relationships between cultural phenomena;
3. A guide for scholars to more easily identify research gaps and areas of interest for potential research;
4. An information source to provide an answer to the basic question “How did we get here?” when contemplating a cultural phenomenon;

5. A visualization platform that illustrates both the diversity and similarities of all human cultures;
6. The symbol of a unified science, which, by default, would also communicate to the world the existence of such a science.

The development of a classification system that would unify major areas of cultural evolution science (e.g., evolutionary psychology, archaeology, linguistics, cultural anthropology, and sociology) can easily be understood as an information science problem to be solved.

Closing remarks

From a biological perspective, the *tree of life* is one way of communicating the story of life on Earth. Thus, a classification is one way of telling a story. From a library classification perspective, it is the story of information that has been encapsulated in books to become the preservation of written human knowledge over time. From a cultural systematist perspective, the written record is a treasure trove of human cultural past.

Imagine if humans could one day explain cultural diversity throughout the world and over time in the way biological sciences can explain past and present biological diversity today. Imagine if cultural evolution could be communicated as simply as the biological tree of life. Imagine if social science could explain culture at the "genetic" level. More broadly, imagine if a meaningful discovery or insight in one area of social science was transmitted across all social sciences or if the social sciences communicated in one voice.

I see classification of cultural phenomena as one way to communicate the story of culture's emergence and evolution, which is also the story of humankind. For me, the grand questions include *how and why did one biological species on Earth become capable of creating*

its own ecosystems full of interacting living systems? A unified science of cultural evolution could ultimately answer these questions and information science classification researchers (via cultural systematics) could play a major role in this unification by looking at culture through the lens of a biologist.

REFERENCES

Abbas, J. (2010). *Structures for organizing knowledge: Exploring taxonomies, ontologies, and other schema*. New York, NY: Neal-Schuman Publishers.

Bailey, K. D. (1994). *Typologies and taxonomies: An introduction to classification techniques* (Vol. 102). Thousand Oaks, CA: Sage Publications, Inc.

Balter, M. (2009). On the origin of art and symbolism. *Science*, *323*(5915), 709-711.

Bergel, G., Howe, C. J., & Windram, H. F. (2015). Lines of succession in an English ballad tradition: The publishing history and textual descent of The Wandering Jew's Chronicle. *Digital Scholarship in the Humanities*, *31*(3), 540-562.

Bisio, F., Meda, C., Gastaldo, P., Zunino, R., & Cambria, E. (2016). Sentiment-oriented information retrieval: Affective analysis of documents based on the senticnet framework. In: W. Pedrycz & S. M. Chen (Eds.), *Sentiment analysis and ontology engineering*, (pp. 175-197). Cham, CH: Springer.

Boc, A., Diallo, A. B., & Makarenkov, V. (2012). TREX: a web server for inferring, validating and visualizing phylogenetic trees and networks. *Nucleic Acids Research, 40*(W1), W573–9.

Bornmann, L. (2011). Scientific peer review. *Annual review of information science and technology*, *45*(1), 197-245.

Bornmann, L., & Mutz, R. (2015). Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology*, *66*(11), 2215-2222.

Brown, T. A. (2002). Molecular phylogenetics. In T. A. Brown (Ed.), *Genomes* (2nd ed.). Oxford, UK: Wiley-Liss. Available from: https://www.ncbi.nlm.nih.gov/books/NBK21122/

Bryant, D., & Moulton, V. (2004). Neighbor-net: an agglomerative method for the construction of phylogenetic networks. *Molecular Biology and Evolution*, *21*(2), 255-265.

Buchanan, B., & Collard, M. (2007). Investigating the peopling of North America through cladistic analyses of Early Paleoindian projectile points. *Journal of Anthropological Archaeology*, *26*(3), 366-393.

Busagala, L. S., Ohyama, W., Wakabayashi, T., & Kimura, F. (2012). Multiple feature-classifier combination in automated text classification. In *10th IAPR international workshop on document analysis systems*, (pp. 43-47). Los Alamitos, CA: IEEE Computer Society.

Cabrera, F. (2017). Cladistic Parsimony, Historical Linguistics and Cultural Phylogenetics. *Mind & Language*, *32*(1), 65-100.

Carneiro, R. L. (Ed.). (1967). *The evolution of society*. Chicago, IL: University of Chicago Press.

Caron, F., d'Errico, F., Del Moral, P., Santos, F., & Zilhão, J. (2011). The reality of Neandertal symbolic behavior at the Grotte du Renne, Arcy-sur-Cure, France. *PloS one*, *6*(6), e21545.

Carstensen, K. U., Diekmann, B., & Möller, G. (2000). GERHARD (German Harvest Automated Retrieval and Directory). In R. Decker & W. Gaul (Eds.) *Classification and information processing at the turn of the millennium*, (pp. 441-450). Berlin, DE: Springer.

Cavalli-Sforza, L. L., & Feldman, M. W. (1981). *Cultural transmission and evolution: A quantitative approach*. Monographs in population biology (No. 16). Princeton, NJ: Princeton University Press.

Clarivate Analytics. Web of Science. Philadelphia, PA.

Collard, M., Shennan, S. J., & Tehrani, J. J. (2006). Branching, blending, and the evolution of cultural similarities and differences among human populations. *Evolution and Human Behavior*, *27*(3), 169-184.

Copleston, F. (1994). *A history of philosophy, vol. VIII. Modern philosophy: Empiricism, idealism, and pragmatism in Britain and America*. New York, NY: Doubleday.

Cowlishaw, G., & Mace, R. (1996). Cross-cultural patterns of marriage and inheritance: A phylogenetic approach. *Ethology and Sociobiology*, *17*(2), 87-97.

d'Errico, F. (1998). Palaeolithic origins of artificial memory systems: an evolutionary perspective. In C. Renfrew & C. Scarre (Eds.), *Cognition and material culture: The archaeology of symbolic storage*, (pp. 19–50). Cambridge, UK: McDonald Institute for Archaeological Research.

d'Errico, F., Henshilwood, C., Lawson, G., Vanhaeren, M., Tillier, A. M., Soressi, M., Bresson, F., Maureille, B., Nowell, A., Lakarra, J., Backwell, L. & Julien, M. (2003). Archaeological evidence for the emergence of language, symbolism, and music–an alternative multidisciplinary perspective. *Journal of World Prehistory*, *17*(1), 1-70.

Darwin, F. (Ed.). (1890). *The expression of the emotions in man and animals*. London, UK: John Murray.

Dawkins, R. (1982). *The extended phenotype: The long reach of the gene*. Oxford, UK: Oxford University Press.

Dawkins, R. (1989). *The selfish gene*. Oxford, UK: Oxford University Press.

de Schryver, G. M., Grollemund, R., Branford, S., & Bostoen, K. (2015). Introducing a state-of-the-art phylogenetic classification of the Kikongo Language Cluster. *Africana Linguistica*, *21*, 87-162.

Demarest, B., & Sugimoto, C. R. (2015). Argue, observe, assess: Measuring disciplinary identities and differences through socio-epistemic discourse. *Journal of the Association for Information Science and Technology*, *66*(7), 1374-1387.

Dennett, D. C. (1995). *Darwin's dangerous idea: Evolution and the meanings of life*. New York, NY: Simon & Schuster.

Desale, S. K., & Kumbhar, R. M. (2013). Research on automatic classification of documents in library environment: A literature review. *Knowledge Organization*, *40*(5), 295-304.

Dipper, S., & Schrader, B. (2008). Computing distance and relatedness of medieval text variants from German. In: A. Storrer, A. Geyken, A. Siebert, & K. M. Würzner (Eds.), *Text resources and lexical knowledge. Selected Papers from the 9th Conference on Natural Language Processing KONVENS 2008,* (pp. 39-51). Berlin, DE: De Gruyter

Donald, M. (1991). *Origins of the modern mind: Three stages in the evolution of culture and cognition*. Cambridge, MA: Harvard University Press.

Dousa, T. (2011). Evolutionary order in the classification theories of C. A. Cutter and E. C. Richardson: Its nature and limits. *NASKO, 2*(1), 76-90.

Dousa, T.M. (2010). The simple and the complex in E. C. Richardson's theory of classification: Observations on an early KO model of the relationship between ontology and epistemology. In C. Gnoli & F. Mazzocchi (Eds.), *Paradigms and conceptual systems for*

knowledge organization: Proceedings of the Eleventh International ISKO Conference, (pp. 15-22). Würzburg, DE: Ergon.

Dunnell , R. (2000). Evolution, scientific. In L. Ellis (Ed.), *Archaeological method and theory: An encyclopedia*, (pp. 190-193). New York, NY: Garland Publishing, Inc.

Dunnell, R. C. (1980). Evolutionary theory and archaeology. *Advances in archaeological method and theory*, *3*, 35-99.

Eagleton, C., & Spencer, M. (2006). Copying and conflation in Geoffrey Chaucer's treatise on the astrolabe: A stemmatic analysis using phylogenetic software. *Studies in History and Philosophy of Science Part A*, *37*(2), 237-268.

Edwards, W. H. (2005). *An introduction to Aboriginal societies*. Southbank, Victoria: Social Science Press.

Farber, P. L. (2000). *Finding order in nature: The naturalist tradition from Linnaeus to E.O. Wilson*. Baltimore, MD: Johns Hopkins University Press.

Fiorini, N., Harispe, S., Ranwez, S., Montmain, J., & Ranwez, V. (2016). Fast and reliable inference of semantic clusters. *Knowledge-Based Systems*, *111*, 133-143.

Fortunato, L. (2011). Reconstructing the history of marriage strategies in Indo-European—speaking societies: Monogamy and polygyny. *Human Biology*, *83*(1), 87-105.

Fortunato, L. (2011). Reconstructing the history of residence strategies in Indo-European—speaking societies: Neo-, uxori-, and virilocality. *Human Biology*, *83*(1), 107-128.

Fortunato, L., Holden, C., & Mace, R. (2006). From bridewealth to dowry?. *Human Nature: An Interdisciplinary Biosocial Perspective*, *17*(4), 355.

Frank, E., & Paynter, G. W. (2004). Predicting library of congress classifications from library of congress subject headings. *Journal of the Association for Information Science and Technology*, *55*(3), 214-227.

Freedman, D. J., Riesenhuber, M., Poggio, T., & Miller, E. K. (2002). Visual categorization and the primate prefrontal cortex: neurophysiology and behavior. *Journal of Neurophysiology*, *88*(2), 929-941.

Futuyma, D. (1998). *Evolutionary biology*, 3rd ed. Sunderland, MA: Sinauer Associates.

Georges, P. (2017). Western classical music development: a statistical analysis of composers similarity, differentiation and evolution. *Scientometrics*, *112*(1), 21-53.

Giddens, A. (1984). *The constitution of society: Outline of the theory of structuration*. Berkeley, CA: University of California Press.

Gnoli, C. (2006). Phylogenetic classification. *Knowledge Organization*, *33*(3), 138-152.

Gnoli, C. (2017). Classifying Phenomena Part 2: Types and Levels. *Knowledge Organization*, *44*(1).

Gnoli, C., & Ridi, R. (2014). Unified theory of information, hypertextuality and levels of reality. *Journal of Documentation*, *70*(3), 443-460.

Grube, G. M. A. (1992). *Republic (Grube Edition)*. Indianapolis, IN: Hackett Publishing Company, Inc.

Guo, J. L., Wang, H. C., & Lai, M. W. (2015). A feature selection approach for automatic e-book classification based on discourse segmentation. *Program*, *49*(1), 2-22.

Hedges, A. (2018). *Random number generator / picker* https://andrew.hedges.name/experiments/random/

Hennig, W. (1966). *Phylogenetic systematics*. Urbana, IL: University of Illinois Press.

Henshilwood, C. S., d'Errico, F., Yates, R., Jacobs, Z., Tribolo, C., Duller, G. A., Mercier, N., Sealy, J. C., Valladas, H., Watts, I., & Wintle, A. G. (2002). Emergence of modern human behavior: Middle Stone Age engravings from South Africa. *Science, 295*(5558), 1278-1280.

Henshilwood, C., d'Errico, F., Vanhaeren, M., Van Niekerk, K., & Jacobs, Z. (2004). Middle stone age shell beads from South Africa. *Science*, *304*(5669), 404-404.

Hernando, A., Moya, R., Ortega, F., & Bobadilla, J. (2014). Hierarchical graph maps for visualization of collaborative recommender systems. *Journal of Information Science*, *40*(1), 97-106.

Hoffmann, D. L., Angelucci, D. E., Villaverde, V., Zapata, J., & Zilhão, J. (2018). Symbolic use of marine shells and mineral pigments by Iberian Neandertals 115,000 years ago. *Science Advances, 4*(2), eaar5255. http://advances.sciencemag.org/content/4/2/eaar5255.full

Hoffmann, D. L., Standish, C. D., García-Diez, M., Pettitt, P. B., Milton, J. A., Zilhão, J., ... & Pike, A. W. G. (2018). U-Th dating of carbonate crusts reveals Neandertal origin of Iberian cave art. *Science, 359*(6378), 912-915.

Horn, R. E. (2002). Beginning to conceptualize the Human Cognome project. Retrieved from https://web.stanford.edu/~rhorn/a/topic/cognom/artclCncptlzHumnCognome.pdf

Huson, D. H., & Bryant, D. (2005). Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution*, *23*(2), 254-267.

Huson, D. H., Dezulian, T., Klopper, T., & Steel, M. A. (2004). Phylogenetic super-networks from partial trees. *IEEE/ACM Transactions on Computational Biology and Bioinformatics, 1*(4), 151–158.

Hutton, J. (1788). Theory of the Earth; or an investigation of the laws observable in the composition, dissolution, and restoration of land upon the Globe.*Transactions of the Royal Society of Edinburgh, 1*, 209-304. Retrieved from http://pages.uwc.edu/keith.montgomery/Hutton/Hutton.htm

Ibekwe-Sanjuan, F. & Bowker, G. C. (2017). Implications of big data for knowledge organization. *Knowledge Organization*, *44*(3), 187-198.

Janmaat, K. R., Ban, S. D., & Boesch, C. (2013). Taï chimpanzees use botanical skills to discover fruit: what we can learn from their mistakes. *Animal Cognition*, *16*(6), 851-860.

Jmila, H., Khedher, M. I., & El Yacoubi, M. A. (2017). Estimating vnf resource requirements using machine learning techniques. In D. Liu, S. Xie, Y. Li, D. Zhao, & E. El-Alfy (Eds.), *International conference on neural information processing*, (pp. 883-892). Cham, CH: Springer.

Joorabchi, A., & Mahdi, A. E. (2011). An unsupervised approach to automatic classification of scientific literature utilizing bibliographic metadata. *Journal of Information Science*, *37*(5), 499-514.

Jordan, P., & Shennan, S. (2009). Diversity in hunter–gatherer technological traditions: mapping trajectories of cultural 'descent with modification' in northeast California. *Journal of Anthropological Archaeology*, *28*(3), 342-365.

Julien, C. A., Tirilly, P., Leide, J. E., & Guastavino, C. (2012). Constructing a true LCSH tree of a science and engineering collection. *Journal of the Association for Information Science and Technology*, *63*(12), 2405-2418.

Kim, J. & Lee, K. (2002). Designing a knowledge base for automatic book classification. *Electronic Library, 20*(6), 488-495.

Klump, B., Patton, R., Potok, T., Reed, J., Treadwell, J., Cunic, C., Martin, P. (2010). PIRANHA: A Knowledge Discovery Engine [Software]. Oak Ridge, TN: UT-Battelle, LLC.

Kuhn, T. (1996). *The structure of scientific revolutions*, 3rd ed. Chicago, IL: University of Chicago Press.

Kumbhar, R. (2012). *Library classification trends in the 21st century*. Oxford, UK: Chandos Publishing.

Lakoff, G. (1987). *Women, fire, and dangerous things: What categories reveal about the mind*. Chicago, IL: University of Chicago Press

Larsen, A. W. (2011). Evolution of Polynesian bark cloth and factors influencing cultural change. *Journal of Anthropological Archaeology*, *30*(2), 116-134.

Larsson, S. (2016). Conceptions, categories, and embodiment: Why metaphors are of fundamental importance for understanding norms. In M. Baier (Ed.), *Social and legal norms: Towards a socio-legal understanding of normativity* (pp. 121-139). New York, NY: Routledge.

Lee, M. S., & Palci, A. (2015). Morphological phylogenetics in the genomic age. *Current Biology*, *25*(19), R922-R929.

Library of Congress. (1900). *Tree of Library Classifications*. 1900. [?] [Photograph] Retrieved from the Library of Congress, https://www.loc.gov/item/2016650285/

Lipo, C. P. (2001). *Science, style and the study of community structure: An example from the central Mississippi River valley*. BAR International (No. 918). Oxford: British Archaeological Reports.

Lycett, S. J. (2007). Why is there a lack of Mode 3 Levallois technologies in East Asia? A phylogenetic test of the Movius–Schick hypothesis. *Journal of Anthropological Archaeology*, *26*(4), 541-575.

Mace, R., Holden, C. J., & Shennan, S. (Eds.). (2005). *The evolution of cultural diversity: A phylogenetic approach*. Walnut Creek, CA: Left Coast Press.

Mayr, E. and Ashlock, P.E. (1991). *Principles of systematic zoology*. New York, NY: McGraw-Hill.

Mazzocchi, F. (2018). Knowledge Organization System (KOS): An Introductory Critical Account. *Knowledge Organization*, *45*(1), 54-78.

Mesoudi, A. (2011). *Cultural evolution: How Darwinian theory can explain human culture and synthesize the social sciences*. Chicago, IL: University of Chicago Press.

Mesoudi, A., Whiten, A., & Laland, K. (2004). Perspective: is human cultural evolution Darwinian? Evidence reviewed from the perspective of "The Origin of Species." *Evolution, 58*(1), 1-11.

Mesoudi, A., Whiten, A., & Laland, K. (2006). Towards a unified science of cultural evolution. *Behavioral and Brain Sciences, 29*(4), 329-383.

Miller, .J. G., & Miller, L. J. (1990). The nature of living systems. *Behavioral Science, 35*(3), 157-163.

Miller, J. G. (1978). *Living systems*. New York, NY: McGraw-Hill.

Morrison, J. (2014). China becomes world's third-largest producer of research articles. *Nature News*, Nature Publishing Group. Retrieved from https://www.nature.com/news/china-becomes-world-s-third-largest-producer-of-research-articles-1.14684

Mulder, M. B., George-Cramer, M., Eshleman, J., & Ortolani, A. (2001). A study of East African kinship and marriage using a phylogenetically based comparative method. *American Anthropologist*, *103*(4), 1059-1082.

Mullins, D. A., Whitehouse, H., & Atkinson, Q. D. (2013). The role of writing and recordkeeping in the cultural evolution of human cooperation. *Journal of Economic Behavior & Organization*, *90*, S141-S151.

Mushegian, A. R. (2007). *Foundations of comparative genomics*. Burlington, MA: Academic Press.

Noman, N., & Iba, H. (2016). A brief introduction to evolutionary and other nature-inspired algorithms. In N. Norman & H. Iba (Eds.), *Evolutionary computation in gene regulatory network research*, (pp. 1-29). Hoboken, NJ: Wiley.

O'Brien, M. J., & Dunnell, R. C. (1996). *Evolutionary archaeology: Theory and application*. Salt Lake City, UT: University of Utah Press.

O'Brien, M. J., Darwent, J., & Lyman, R. L. (2001). Cladistics is useful for reconstructing archaeological phylogenies: Palaeoindian points from the southeastern United States. *Journal of Archaeological Science*, *28*(10), 1115-1136.

Online Computer Library Center [OCLC]. (n.d.). Automatic classification research at OCLC. Retrieved from http://www.oclc.org/research/activities/auto_class.html

Online Computer Library Center [OCLC]. (n.d.). Classify: An experimental classification web service. Retrieved from http://classify.oclc.org/classify2/

Online Computer Library Center [OCLC]. (n.d.). Scorpion. Retrieved from http://www.oclc.org/research/activities/scorpion.html

Osborne, F., Salatino, A., Birukou, A., & Motta, E. (2016). Automatic classification of Springer Nature proceedings with Smart Topic Miner. In *International semantic web conference*, (pp. 383-399). Cham, CH: Springer.

Panigrahi, P., & Prasad, A. R. D. (2007). Facet sequence in analytico synthetic scheme: A study for developing an AI based automatic classification system. *Annals of Library and Information Studies, 54*(1), 37-43.

Parrochia, D., & Neuville, P. (2013). *Towards a general theory of classifications*. Basel, CH: Birkhäuser.

Perszyk, D. R., & Waxman, S. R. (2016). Listening to the calls of the wild: The role of experience in linking language and cognition in young infants. *Cognition*, *153*, 175-181.

Pike, A. W., Hoffmann, D. L., Garcia-Diez, M., Pettitt, P. B., Alcolea, J., De Balbin, R., González-Sainz, C., de las Heras, C., Lasheras, J. A., Montes, R., Zilhao, J. (2012). U-series dating of Paleolithic art in 11 caves in Spain. *Science, 336*(6087), 1409-1413.

Platnick, N. I., & Cameron, H. D. (1977). Cladistic methods in textual, linguistic, and phylogenetic analysis. *Systematic Biology*, *26*(4), 380-385.

Prentiss, A. M., Walsh, M. J., Foor, T. A., & Barnett, K. D. (2015). Cultural macroevolution among high latitude hunter–gatherers: a phylogenetic study of the Arctic Small Tool tradition. *Journal of Archaeological Science*, *59*, 64-79.

Radcliffe-Brown, A. R. (1940). On social structure. *The Journal of the Royal Anthropological Institute of Great Britain and Ireland, 70*(1), 1-12.

Reed, J. W., Jiao, Y., Potok, T. E., Klump, B. A., Elmore, M. T., & Hurson, A. R. (2006). TF-ICF: A new term weighting scheme for clustering dynamic data streams. In *Proceedings of the 5th international conference on machine learning and applications*, (pp. 258-263). Washington, DC: IEEE Computer Society.

Richardson, E. C. (1901). *Classification: Theoretical and practical*. New York, NY: Charles Scribner's Sons.

Rivero, D. G. (2016). Darwinian archaeology and cultural phylogenetics. In L. M. Straffon (Ed.) *Cultural phylogenetics: Concepts and applications in archaeology*, (pp. 43-72). Cham, CH: Springer.

Rokach, L., & Maimon, O. (2009). Classification trees. In O. Maimon & L. Rokach (Eds.), *Data mining and knowledge discovery handbook*, 2nd edition (pp. 149-174). Boston, MA: Springer.

Roos, T., & Heikkilä, T. (2009). Evaluating methods for computer-assisted stemmatology using artificial benchmark data sets. *Literary and Linguistic Computing*, *24*(4), 417-433.

Runciman, W. G. (2009). *The theory of cultural and social selection*. Cambridge, UK: Cambridge University Press.

Sahlins, M., & Service, E. (Eds.). (1960). *Evolution and culture*. Ann Arbor, MI: University of Michigan Press.

Saitou, N., & Nei, M. (1987). The Neighbor-Joining Method: a New Method for Reconstructing Phylogenetic Trees. *Molecular Biology and Evolution 4*(4), 406–425.

Schillinger, K., Mesoudi, A., & Lycett, S. J. (2016). Copying error, evolution, and phylogenetic signal in artifactual traditions: An experimental approach using "model artifacts". *Journal of Archaeological Science*, *70*, 23-34.

Shera, J. H. (1965). *Libraries and the organization of knowledge*. Hamden, CT: Archon Books.

Smiraglia, R. P., & Cai, X. (2017). Tracking the evolution of clustering, machine learning, automatic indexing and automatic classification in knowledge organization. *Knowledge Organization*, *44*(3), 215-233.

Smith, J. D., Zakrzewski, A. C., Johnson, J. M., Valleau, J. C., & Church, B. A. (2016). Categorization: The view from animal cognition. *Behavioral Sciences*, *6*(2), 12.

Smith, W. J. (1980). *The behavior of communicating*. Cambridge, MA: Harvard University Press.

Steward, J. H. (1963). *Theory of culture change: The methodology of multilinear evolution*. Urbana, IL: University of Illinois Press.

Stubbersfield, J., & Tehrani, J. (2013). Expect the unexpected? Testing for minimally counterintuitive (MCI) bias in the transmission of contemporary legends: A computational phylogenetic approach. *Social Science Computer Review*, *31*(1), 90-102.

Suominen, A., & Toivanen, H. (2015). Map of science with topic modeling: Comparison of unsupervised learning and human-assigned subject classification. *Journal of the Association for Information Science and Technology*, *67*(10), 2464-2476.

Swofford, D. L. (2003). PAUP*. Phylogenetic analysis using parsimony (*and other methods) (Version 4) [Software]. Sunderland, MA: Sinauer Associates. Available from http://paup.phylosolutions.com/get-paup/.

Swofford, D.L., & Bell, C. D. (2017), *PAUP* manual*. Available from http://paup.phylosolutions.com/.

Taylor, A. G. (2004). *The organization of information* (2nd ed.). Westport, CT: Libraries Unlimited.

Tehrani, J. J. (2013). The phylogeny of little red riding hood. *PloS One*, *8*(11), e78871.

Tehrani, J. J., & Collard, M. (2009). On the relationship between interindividual cultural transmission and population-level cultural diversity: a case study of weaving in Iranian tribal populations. *Evolution and Human Behavior*, *30*(4), 286-300.

Tehrani, J., & Collard, M. (2002). Investigating cultural evolution through biological phylogenetic analyses of Turkmen textiles. *Journal of Anthropological Archaeology*, *21*(4), 443-463.

Tehrani, J., Nguyen, Q., & Roos, T. (2016). Oral fairy tale or literary fake? Investigating the origins of Little Red Riding Hood using phylogenetic network analysis. *Digital Scholarship in the Humanities*, *31*(3), 611-636.

Tëmkin, I., & Eldredge, N. (2007). Phylogenetics and material cultural evolution. *Current Anthropology*, *48*(1), 146-154.

Tenopir, C., & King, D.W. (2009). The growth of journals publishing. In A. Phillips (Ed.), *The future of the academic journal*, (pp. 159-178). Oxford, UK: Chandos Publishing.

The Center for Human Emergence (n.d.). Human memome project. Retrieved from http://www.humanemergence.org/humanMemome.html

Thompson, R., Shafer, K., & Vizine-Goetz, D. (1997). Evaluating Dewey concepts as a knowledge base for automatic subject assignment. In *Proceedings of the second ACM international*

conference on digital libraries, (37-46). New York, NY: Association for Computing Machinery.

Toivanen, H., & Suominen, A. (2014). Epistemic integration of the European Research Area: The shifting geography of the knowledge base of Finnish research, 1995–2010. *Science and Public Policy*, *42*(4), 549-566.

Trigger, B. (2004). Writing systems: a case study in cultural evolution. In S. Houston (Ed.), *The first writing: Script invention as history and process*, (pp. 39–70). Cambridge, UK: Cambridge University Press.

United Nations Department of Economic and Social Affairs. (2017). *World's population increasingly urban with more than half living in urban areas*. Retrieved from https://www.un.org/development/desa/en/news/population/world-urbanization-prospects.html

Vinicius, L. (2010). *Modular evolution: How natural selection produces biological complexity*. Cambridge, UK: Cambridge University Press.

Vogt, L. (2008). The unfalsifiability of cladograms and its consequences. *Cladistics*, *24*(1), 62-73.

von Lieven, A. F., & Humar, M. (2008). A cladistic analysis of Aristotle's animal groups in the "Historia Animalium". *History and Philosophy of the Life Sciences*, *30*(2), 227-262.

von Waldenfels, R. (2017). The expansion of the preposition do+genitive in North Slavic. *Russian Linguistics, 41*(1), 79-108.

Wang, J. (2009). An extensive study on automated Dewey Decimal Classification. *Journal of the Association for Information Science and Technology*, *60*(11), 2269-2286.

Ware, M., & Mabe, M. (2015). *The STM report: An overview of scientific and scholarly journal publishing*. The Hague, The Netherlands: International Association of Scientific, Technical and Medical Publishers.

White, L. A. (1959). *The evolution of culture: The development of civilization to the fall of Rome*. New York, NY: McGraw-Hill.

Wiley, E. O., & Lieberman, B. S. (2011). *Phylogenetics: Theory and practice of phylogenetics systematics*. Hoboken, NJ: Wiley-Blackwell.

Wilkins, J.S., & Ebach, M.C. (2014). *The nature of classification: Relationships and kinds in the natural sciences*. New York, NY: Palgrave Macmillan.

Willey, G. R., & Sabloff, J. A. (1974). *A history of American archaeology*. London, UK: Thames and Hudson.

Windram, H. F., Charlston, T., & Howe, C. J. (2014). A phylogenetic analysis of Orlando Gibbons's Prelude in G. *Early Music*, *42*(4), 515-528.

Windram, H. F., Shaw, P., Robinson, P., & Howe, C. J. (2008). Dante's Monarchia as a test case for the use of phylogenetic methods in stemmatic analysis. *Literary and Linguistic Computing*, *23*(4), 443-463.

Witze, A. (2016). Research gets increasingly international. *Nature News*, Nature Publishing Group. Retrieved from http://www.nature.com/news/research-gets-increasingly-international-1.19198

Woodrow Wilson International Center for Scholars (n.d.). Mapping China's cultural genome. Retrieved from https://www.wilsoncenter.org/mapping-chinas-cultural-genome

APPENDICES

APPENDIX A: Corpus Dataset Book List

Below includes the list of 51 books used as the corpus dataset for the classification project, presented alphabetically by title.

Book Name in Classification Tree	Title	Print Publication Year	Online Book ISBN13
Airborne Measurements -- Environmental	Airborne Measurements for Environmental Research - Methods and Instruments	2013	9783527653218
Alluvial Sedimentation	Alluvial Sedimentation	1993	9781444303995
Amazonia & Global Change	Amazonia and Global Change	2009	9781118670347
Ascidiacea (Antarctic)	Antarctic Ascidiacea	1969	9781118668801
Cirripedia (Antarctic)	Antarctic Cirripedia	1970	9781118664407
Auroral Dynamics & Space Weather	Auroral Dynamics and Space Weather	2015	9781118978719
Auroral Phenomena -- Magnetosphere	Auroral Phenomenology and Magnetospheric Processes: Earth and Other Planets, Geophysical Monograph 197	2012	9781118670286
Transect (Australia)	Broken Hill-Sydney Tasman-Sea Transect: New South Wales, Eastern Australia	1991	9781118667842
Cambrian Fossils (China)	Cambrian Fossils of Chengjiang, China 2e, The	2017	9781118896372
Carbonate Banks & Siliciclastic Basins	Cambro-Ordovician Carbonate Banks and Siliciclastic Basins of the United States Appalachians: Knoxville, Tennessee to Hagerstown, Maryland, June 30–J	1989	9781118667217
Carbonate Sedimentology	Carbonate Sedimentology	1990	9781444314175
Gulf of Mexico -- Circulation	Circulation in the Gulf of Mexico: Observations and Models, Geophysical Mongraph 161	2005	9781118666166
Volcanism - Plutonism - Magma	Cordilleran Volcanism, Plutonism, and Magma Generation at Various Crustal Levels, Montana and Idaho Western Montana and Central Idaho, Field Trip T33	1989	9781118668306

Book Name in Classification Tree	Title	Print Publication Year	Online Book ISBN13
Core-Mantle Boundary	Core-Mantle Boundary Region, The	1998	9781118669747
Crustacea Tanaidacea (Antarctic)	Crustacea Tanaidacea of the Antarctic and the Subantarctic: 1. On Material Collected at Tierra del Fuego, Isla de los Estados, and the West Coast of the Antarctic Peninsula	1986	9781118664827
Ductile Shear Zones	Ductile Shear Zones - from micro- to macro-scales	2015	9781118844953
Midcontinent Rift	Duluth Complex and Associated Rocks of the Midcontinent Rift System: Minneapolis to Duluth, Minnesota	1989	9781118667422
Early Earth -- Accretion - Differentiation	Early Earth: Accretion and Differentiation, The	2015	9781118860359
Tectonics -- Mesozoic	Early Mesozoic Tectonics of the Western Great Basin, Nevada: Battle Mountain to Yerington District, Nevada, July 1-7, 1989, Field Trip Guidebook T122	1989	9781118667071
Foreland & Intermontane Basins	Evolution of Resource-Rich Foreland and Intermontane Basins in Eastern Utah and Western Colorado: Salt Lake City, Utah to Grand Junction, Colorado, July 20–24, 1989, 1	1989	9781118667033
Volcanism -- Subaqueous	Explosive Subaqueous Volcanism, Geophysical Monograph 140	2003	9781118668665
Extreme Events	Extreme Events: Observations, Modeling, and Economics	2015	9781119157052
Atmosphere -- Fluid Dynamics	Fluid Dynamics of the Mid-Latitude Atmosphere	2014	9781118526002
Fossil Scleractinian Corals (Antartica)	Fossil Scleractinian Corals from James Ross Basin, Antarctica	1994	9781118668009
Geology & Hydrocarbons - East US Overthrust	Geology and Hydrocarbon Potential of the Eastern Overthrust: Knoxville, Tennessee to Washington, D.C., July 20–23, 1989	1989	9781118669693
Geology & Seismic Stratigraphy (Antarctic)	Geology and Seismic Stratigraphy of the Antarctic Margin	1995	9781118669013
Modeling -- Urban Phenomena	Geosimulation - Automata-based Modeling of Urban Phenomena	2004	9780470020999

Book Name in Classification Tree	Title	Print Publication Year	Online Book ISBN13
Paleobiology -- Graptolites	Graptolite Paleobiology	2017	9781118515624
Gravel-bed Rivers	Gravel-bed Rivers - Processes, Tools, Environments	2012	9781119952497
Crystal Engineering -- Pi-Interactions	Importance of Pi-Interactions in Crystal Engineering - Frontiers in Crystal Engineering, The	2012	9781119945888
Arcs - Trenches - Basins	Island Arcs Deep Sea Trenches and Back-Arc Basins	1977	9781118665756
Landslides	Landslides in Central California: San Francisco and Central California, Field Trip Guidebook T381	1989	9781118667262
Low-Grade Metamorphism	Low-Grade Metamorphism	1999	9781444313345
Magnetotails in Solar System	Magnetotails in the Solar System	2015	9781118842324
Making Crystals	Making Crystals by Design - Methods, Techniquesand Applications	2006	9783527610112
Mantle Dynamics & Plate Interactions (Asia)	Mantle Dynamics and Plate Interactions in East Asia, Geodynamics Series Volume 27	1998	9781118670132
Meteorology -- Mesoscale	Mesoscale Meteorology in Midlatitudes	2010	9780470682104
Mineral Resources (Antarctica)	Mineral Resources Potential of Antarctica	1990	9781118664926
Ocean Circulation	Ocean Circulation: Mechanisms and Impacts	2007	9781118666241
Ocean -- Earth System	Ocean in the Earth System	2014	9781119007678
Oil & Gas Exploration	Oil and Gas Exploration: Methods and Application	2017	9781119227519
Astrophysical Plasmas	Particle Acceleration in Astrophysical Plasmas: Geospace and Beyond	2005	9781118666104
Ocean -- Poleward Flows	Poleward Flows Along Eastern Ocean Boundaries	1989	9781118663615
Geophysical Structures & Processes	Relating Geophysical Structures and Processes: The Jeffreys Volume	1993	9781118669129

Book Name in Classification Tree	Title	Print Publication Year	Online Book ISBN13
Sediment Hosted Mineral Deposits	Sediment Hosted Mineral Deposits IAS 11	1990	9781444303872
Maps -- Self-Organising	Self-Organising Maps - Applications in Geographic Information Science	2008	9780470021699
Earth Geophysics	State of the Planet: Frontiers and Challengesin Geophysics, The	2004	9781118666012
Subduction	Subduction Top to Bottom, Geophysical Monograph 96	1996	9781118664575
Tectonics -- Sedimentary Basins	Tectonics of Sedimentary Basins - Recent Advances	2011	9781444347166
Tectonics -- VA Blue Ridge & Piedmont	Tectonics of the Virginia Blue Ridge and Piedmont Culpeper to Richmond, Virginia, Field Trip Guidbook T363	1989	9781118667446
Tidal Mixing & Plankton Dynamics	Tidal Mixing and Plankton Dynamics	1986	9781118669457

APPENDIX B: Term Set Removals

Below includes the list of 14 books removed from the term sets, presented alphabetically.

Title	Print Publication Year	Online Book ISBN13
Antarctic Subglacial Aquatic Environments, Geophysical Monograph 192	2011	9781118670354
Assessment of Non-Point Source Pollution in the Vadose Zone	1999	9781118664698
Coastal Ocean Prediction	1999	9781118665527
Derivation, Meaning, and Use of Geomagnetic Indices, Geophysical Monograph 22	1980	9781118663837
Geology of the Central Transantarctic Mountains	1986	9781118664797
Interactions Between Macro- and Microorganisms in Marine Sediments, Coastal and Estuarine Studies Volume 60	2005	9781118665442
Mathematical Modelling of Tides and Estuarine Circulation: The Coastal Seas of Southern British Columbia and Washington State	1988	9781118669167
Measurement Techniques in Space Plasmas: Particles	1998	9781118664384
New Perspectives on the Earth's Magnetotail, Geophysical Monograph 105	1998	9781118664629
Nitrogen Loading in Coastal Water Bodies: An Atmospheric Perspective	2001	9781118665190
Outdoor Recreation and Water Resources Planning	1974	9781118665299
River Restoration - Managing the Uncertainty in Restoring Physical Habitat	2008	9780470867082
Sea Salt Aerosol Production: Mechanisms, Methods, Measurements, and Models - A Critical Review, Geophysical Monograph 152	2004	9781118666050
Understanding Sea-level Rise and Variability	2010	9781444323276

APPENDIX C: Copyright Permissions

Regarding this dissertation's use of image on page 331 in Mesoudi, A., Whiten, A., & Laland, K. (2006). Towards a unified science of cultural evolution. *Behavioral and Brain Sciences, 29*(4), 329-347.

Dear Customer,

1 figure from: Mesoudi, A., Whiten, A., & Laland, K. (2006). Towards a unified science of cultural evolution.*Behavioral and Brain Sciences, 29*(4), 329-347 © Cambridge University Press

Thank you for your request to reproduce the above material in your forthcoming PhD thesis, for non-commercial publication. Cambridge University Press are pleased to grant non-exclusive permission, free of charge, for this specific one time use, on the understanding you have checked that we do not acknowledge any other source for the material. This permission does not include the use of copyright material owned by any party other than the authors. Consent to use any such material must be sought by you from the copyright owner concerned.

Please ensure full acknowledgement appears in your work.

Should you wish to publish your work commercially in the future, please reapply to the appropriate Cambridge University Press office, depending on where your forthcoming work will be published. Further information can be found on our website at the following link:

http://www.cambridge.org/about-us/rights-permissions/permissions/

Kind regards

Permissions Sales Team
Cambridge University Press
University Printing House
Shaftesbury Road
Cambridge CB2 8BS, UK

http://www.cambridge.org/about-us/rights-permissions/permissions/

Regarding this dissertation's use of image on page 158 in Miller, .J.G. & Miller, L.J. (1990). The nature of living systems. *Behavioral Science, 35*(3), 157-163.

Dear David,

Thank you for your email.

Permission is granted for you to use the material requested for your thesis/dissertation subject to the usual acknowledgements (author, title of material, title of book/journal, ourselves as publisher) and on the understanding that you will reapply for permission if you wish to distribute or publish your thesis/dissertation commercially.

You should also duplicate the copyright notice that appears in the Wiley publication in your use of the Material. Permission is granted solely for use in conjunction with the thesis, and the material may not be posted online separately.

Any third-party material is expressly excluded from this permission. If any material appears within the article with credit to another source, authorisation from that source must be obtained.

Many thanks,

Orla Davies
Rights Assistant
John Wiley & Sons Ltd

Regarding the classification tree figure at the end of this dissertation's chapter 2.

Dear David,

You can also find the record for this item in our Prints and Photographs Online Catalog here:http://www.loc.gov/pictures/item/2016650285/
You will notice in the rights advisory the term "No known restrictions on publication." Generally, "No known restrictions on publication" is the best that we can say, meaning there are no donor restrictions and as far as we can tell, no copyright restriction is in effect. For further information about the use of this phrase, see this explanation in our rights overview document: http://www.loc.gov/rr/print/195_copr.html#noknown.

The Library of Congress, as a publicly supported institution, does not own rights to material in its collections. Therefore, it does not charge permission fees for use of such material and cannot give or deny permission to publish or otherwise distribute material in its collections.

The Library does request the courtesy of a credit when publishing items from the collections in order to assist researchers in locating the materials. For information about supplying credits, please see the suggested credit line at the end of the relevant rights statement. Additional information on crediting is available at: http://www.loc.gov/rr/print/195_copr.html#question4.

Please let me know if I may be of any further assistance.

Lara Szypszak

~*~

Reference Section
Prints and Photographs Division
Library of Congress
telephone: 202-707-6394
fax: 202-707-6647
URL: http://www.loc.gov/rr/print/
email: http://www.loc.gov/rr/askalib/ask-print.html

Visit our:
online catalog: http://www.loc.gov/pictures
blog - "Picture This": http://blogs.loc.gov/picturethis/

VITA

David Sims earned his master's degree in information science from the School of Information Science at The University of Tennessee in 2008. He earned a bachelor's degree in dramatic arts from the College of Speech Communication at Louisiana Tech University in 1986 and a bachelor's degree in record industry management from the College of Mass Communication at Middle Tennessee State University in 1993. He completed his education with a Ph.D. degree in communication and information from the College of Communication and Information at The University of Tennessee.

He has worked in various jobs, including project manager and treasurer among others. Since obtaining his master's degree, he has worked primarily as a commercialization manager (i.e., intellectual property management, marketing, and licensing) in Technology Transfer at the Department of Energy's Oak Ridge National Laboratory in Oak Ridge, TN. He has made many presentations related to technology commercialization.

David's research interests include identification, classification, and evolutionary history/relationships of cultural taxa. His dissertation is a method of producing an evolutionary classification of science books based on biological phylogenetics.