2011-11-30

# Evolutionary Subject Tagging in the Humanities; Supporting Discovery and Examination in Digital Cultural Landscapes

# Evolutionary SUBJECT Tagging in the Humanities

**Supporting Discovery and Examination in Digital Cultural Landscapes**


**By**

**Jack Ammerman, Vika Zafrin, Daniel Benedetti, Garth Green**

**NOVEMBER 2011**

# Table of Contents

**ACKNOWLEDGEMENTS**

The authors want to thank the NEH Office of Digital Humanities for its encouragement of our research and the generous support of this project through the award of a Start-Up Grant (Award # HD-51166-10).

With this support, the project team was able to assemble a wonderfully diverse group of scholars to join us in exploring how to facilitate better discovery and examination of texts for humanities research. This group included:

- Andrew Ashton, Director, Center for Digital Scholarship, Brown University
- Rick Fitzgerald, Cataloger, Web Archives, Library of Congress
- Elli Mylonas, Associate Director, Center for Digital Scholarship, Brown University
- Stephen Paling, Assistant Professor, School of Library and Information Sciences, University of Wisconsin, Madison
- Eric Raimy, Assistant Professor, English, University of Wisconsin, Madison
- David A. Smith, Research Assistant Professor, Department of Computer Science, University of Massachusetts, Amherst
- Janis Young, Senior Cataloging Policy Specialist, Policy and Standards Division, Library of Congress

These scholars engaged the project team and each other around the topic, prompting the team at many points to more clearly articulate the issues identified and the desired outcomes for humanities research. The richness of this multidisciplinary conversation enabled the project team to deepen its understanding of the issues and to more clearly focus its efforts.

Support from the BU Libraries administrative staff made the weekend meeting with our consultants flow smoothly. We want to thank Cathy McLaughlin, Cathy Annunciata, and Gery Emory for attending to the many details that made the weekend a success. Robert E. Hudson, University Librarian, provided support and encouragement throughout the project. We continue to appreciate his interest in the research and support for this level of inquiry in the Boston University Libraries.

Members of the project team made presentations at the Digital Library Federation (http://www.diglib.org/) Annual Meeting in San Jose, California (November 2010) and to the Society for Digital Humanities (http://sdh-semi.org/) meeting in Fredericton, Canada (May 2011). Feedback from participants in the DLF session was especially helpful in early efforts to articulate issues and project possible solutions. Participants in the SDH-SEMI meeting were helpful in identifying ongoing research in Canada that parallels the efforts of this project. We are grateful to the participants of both meetings for their assistance.

The front cover image was created using Jonathan Feinberg's Wordle software. It is a word cloud created from this report combined with notes from the meeting with consultants.

# Introduction

Our research interest in "evolutionary subject tagging in humanities research" grew out of both our appreciation for the value of subject classification in organizing and discovering information, and our frustration with the limitations we have encountered in currently available systems. Even as subject terms highlight and focus attention on relationships between information objects, particularly within academic disciplines, they can hide and blur relationships when trying to bridge multiple disciplines in one's research. In early conversations that led to this project, the project team described the experience of multi-disciplinary research as trying to navigate through dark shadows.

Driven by desire to help humanities scholars more easily discover and examine information, our early articulations of the problem led us to explore how we might fix it. Would additional subject tags improve discoverability? Would layering subject terms from multiple disciplines help? Is there a way to merge them? Is translation between disciplinary thesauri required? If more subject terms are required, could we develop a scalable (sustainable) model for providing them? Would any of the efforts to improve the discoverability of humanities texts actually facilitate enhanced examination of the texts?

In this paper, the authors attempt to identify problematic issues for subject tagging, particularly those associated with information objects in digital formats. In the third major section, the authors identify a number of assumptions that lie behind the current practice of subject classification that we think should be challenged. We move then to propose features of classification systems that could increase their effectiveness. These emerged as recurrent themes in many of the conversations with scholars, consultants, and colleagues. Finally, we suggest next steps that we believe will help scholars and librarians develop better subject classification systems to support research in the humanities.

We should note that repeatedly throughout the project, team members found themselves challenging each other about very basic assumptions that underlie the subject classification and the use of subject terms for discovery and examination of information objects. Those conflicting opinions form a creative tension out of which the project and this paper have emerged. Undoubtedly readers will discover some of those points of tension in this paper. The authors hope readers will find these points as opportunities to engage the ideas presented.

## A note on terms used in this paper

We use the terms ***information*** and ***knowledge*** somewhat interchangeably, though we are influenced by the more precise formulation by Davenport and Prusak that "knowledge derives from information as information derives from data." (1998, 6) The transformation of data to information and information to knowledge is the product of human action. We recognize, however, that knowledge is cumulative, and the creation of new knowledge often treats older, more established knowledge as information or even data. Ultimately, it is this human action that we seek to better facilitate.

Subject classification is often associated with hierarchical classification systems. The term *subject heading* reinforces this notion as it is related to sub-headings. By using the term ***subject tagging***, we don't intend to diminish the importance of classification. We will argue that classification systems are snapshots of majority thinking. For humanities research, they may be most helpful when they are rhizomatic, emergent from multiple authorities, with connections that are as significant as the nodes.

When our use of the term ***discovery*** was not clear to our consultants, we realized we were using it in multiple ways. Modern search interfaces attempt to return a result set for virtually any query with easy mechanisms to narrow or broaden the query with the goal of identifying and retrieving the desired information object(s). Such mechanisms might include spelling correction, facets, suggestions of related terms, relevancy ranking, or simply starting a new search. This form of discovery is important and an essential part of research in the humanities. We think of this kind of discovery as being like traversing a landscape to find a landmark.

We do mean more, though, by our use of the term. Once an information object is identified, humanities researchers continue the discovery process by means of iterative queries to the information object that seek to understand it. What is the social context? The historical context? How has the information object been interpreted through the years? Has that changed? How does this information object fit into the intellectual universe? This examination is more like measurement or analyzing core samples from the information object. Discovery continues to be an appropriate term, but the focus shifts from identifying and locating an information object to continued evaluation of the object through a series of queries each informed by the results of prior queries. Throughout the paper, we have attempted to use the term ***examination*** to refer to this process, though we continue to see it as part of the discovery process.

# Genesis

Our interest in supporting scholarly research through enhanced subject tags is rooted in the research experience of humanities scholars. Early on in our conversations, Garth Green, the non-librarian of our project team, noted several of his research interests that were ultimately not only poorly served but also *hindered* by existing subject classification schemes. One example is the German philosopher and theologian Johann Gottlieb Fichte, who was widely understood to be one of the most important philosophers of his century but who, at least in Anglophone scholarship, is seldom studied. The reasons boil down to two things: Kant and Hegel eclipsed him; and, his work was too theological for philosophers and too philosophical for theologians. Fichte falls *both* within *and* between these disciplinary limits; but he is invisible to all of them, because of his uncomfortably close position to each of these communities' *others,* or opponents.

Not only does Fichte himself end up invisible; scholarship about him is also difficult to conduct. Subject headings attached to works about him are inconsistent, and that's for the works that *do* mention him. Much scholarship that would benefit from consideration of Fichte's thought does not, because of the self-perpetuating cycle of obscurity. Once he became interstitial, it is by definition more difficult to find him in the sea of the written word, and he falls further into obscurity. The project team believes this is largely an artifact of our knowledge classification systems, rather than Fichte's importance for the character and development of 19th and 20th century European thought. Researchers who study such "interstitial" subjects, or whose work falls into more than one field, are a core part of our target demographic. More about our target audience will follow. However, a second experience prompted us to consider alternatives to our current model for subject classification.

When we were setting up DSpace at Boston University to serve as our institutional repository software, colleagues at our medical campus demonstrated the MeSH Indexer Web Services application developed at Johns Hopkins University Medical School. When a medical article is uploaded using MeSH Indexer Web Services, the Indexer parses its text and automatically suggests multiple medical subject headings (or MeSH) to be assigned to it. This process is

semantic: *epidemiology* might be suggested for an article on AIDS even if the word "epidemiology" is found nowhere in it.

While automated indexing is not new,[1] early efforts have been exploratory. The MeSH Indexer is a wonderful advance over previous efforts. Recognizing the scalability problems of traditional modes of subject classification using human catalogers, it should not be surprising that many see great potential in such efforts for controlling the costs of cataloging and indexing. (Coyle, 2008) Not only would tools like the MeSH Indexer be a great boon to researchers in all fields, including within the humanities; they could mean big and positive changes for the usually time-consuming and far from granular library cataloging process. But building such a tool in the humanities is harder: word meanings are much more multivariate, more context-dependent.

In addition to the substantial problem of building such a system in the humanities, subject classification is itself problematic. Before exploring possible solutions to the complexity of building an automated indexer for the humanities, gaining a better understanding of the problems associated with subject classification is essential. In the next section, we particularly explore the how those problems are affected by information objects in digital formats.

# Problematic Issues in Subject Classification

## 1. Knowledge is inherently difficult to classify.

Classifying knowledge has always been problematic. The digital age makes it more so, but not because we have a lot more to classify (though that is also a problem). Digital information is more difficult to classify because some of the constraints of physicality within which we'd worked in the past—the constraints of atoms—have fallen away. Efforts to attach keywords to digital information—social tagging, folksonomies, bookmarking, and reputation-based relevancy recommendations—are clearly not subject classification as we have conceived it thus far. They are, however, attempts to move beyond the constraints of information discovery in a physical information environment.

These emerging tagging systems are also problematic, from a librarian's perspective. Without the structure of authority control on the tagging, disambiguation becomes more complex, and precision searching more difficult. Until a certain scale is reached, social tagging seems more individualized, less helpful to others as a general information discovery mechanism. This leads us to consider defining new, conceptual constraints to guide classification processes.

David Weinberger described this change in constraint levels well in his 2007 *Everything Is Miscellaneous: The Power of the New Digital Disorder*. In it, he describes three orders, or levels, of ordering. First order is arranging objects directly, such as placing books on a shelf in relation to one another. The second order arrangement refers to arranging objects that represent other objects—like cards in a card catalog. Each representative object refers to one and only one represented object—but the reverse doesn't hold: a book, for example, might have

[1] John K. Vries, et. al. were exploring automated indexing of Medline articles in the early 1990s, and Wingert was exploring it in the late 1980s.

several cards in different parts of the catalog. Still, each card and each book can only exist in one place at a time.

"The point is that ... these are oriented around putting information *in its place.*" (Bailey and Gardiner, 86) Indeed, the traditional understanding of classification system is as a 'set of boxes (metaphorical or literal) into which things can be put,' and of categories as (ideally, at least) mutually exclusive, 'clearly demarcated bins, into which any object addressed by the system will neatly and uniquely fit.' (Bowker and Star, 10) Such a classification system works, or at least goes unchallenged, when arranging physical objects. With digital objects, it functions as an unnecessary constraint.

Weinberger's third order of order does not presume objects to exist in a single place: they are made no longer of atoms but of bits. With the physical constraint removed, ordering is no longer about geographical navigation or retrieval—but also gets more complicated, because there is no clear impetus to stop classifying. "As we arrange items in space, we're also determining the time it will take to reach them," writes Weinberger (4). Over-arranging items in cognitive space has the potential to create unnecessarily long cognitive paths to reach needed knowledge.

Classification does impose other constraints. An information object "is always to some extent contained and constrained by its label in the taxonomic order of the traditional archive in which, by virtue of its indexical function, the word specifies what the image is supposedly 'of', and not just where it is located." (Bailey and Gardiner, 87) This addition (or transformation) is easily observed in a physical manifestation of an object, when adjacency in a physical arrangement adds context or association to the object. The effect of association is not lost in a digital environment.

This becomes problematic when one recognizes that like all of knowledge, classification systems are socially constructed. All human knowledge is developed and transmitted in social situations. It both shapes and is shaped by a particular society. (Berger and Luckmann, 3) Sociologists have for decades been interested in questions of 'reality' and 'knowledge' because of the fact of their social relativity. The observable differences between societies in terms of what is taken for granted as 'knowledge' makes questionable any assertion that a single classification system might be universally accepted.

How have people dealt with knowledge classification in the past? Not very successfully, and not for lack of trying. We have attempted a universal alphabet (Weinberger 24-8), universal languages such as Esperanto, and multiple all-encompassing subject classification systems, all of which have proved to be inadequate even if they've gained widespread use, as did the Dewey Decimal System. It seems that knowledge workers have collectively abandoned the pursuit of a single, all-comprehensive subject classification system, moving instead toward interoperability among a multitude of systems, each of them again socially constructed and therefore inherently biased.

The failures of current and past knowledge classification systems are not total. Though they fail to achieve an impossible self-imposed goal of categorizing everything, the worldview each of reflects is both valuable and worth examining, especially if we can overlay other such worldviews on top, creating a customizable mosaic that's most useful to us in the moment. In other words: *these classification systems are a historical record of the evolution of our social*

*constructs of knowledge*. Much recent scholarship has shifted away from the notion of knowledge as a collection of facts, challenging us to question, or at least acknowledge, the socio-political origins of assertions.

So perhaps we do not need to discard current classifications; we need to be able to *create new ones*, and turn them on and off like facets. We need to be able to play with knowledge, which the *Encyclopedia Britannica*'s editors have acknowledged to be "atomistic" (Weinberger 31), as we play with building blocks, rearranging until something clicks and new knowledge is generated. Tools that enable us to convert classification systems into such overlapping facets would allow us to examine socially aggregated knowledge in much deeper context from the very beginning. This would in turn bring us closer to knowledge of "relationships that are in the world, not just in our heads." (Weinberger 42)

One logical reformulation of this thought experiment is that we can, in fact, have a single, all-encompassing subject classification system—provided that it is not static but emergent. We see no practical difference between this formulation and the one involving multiple static systems with crosswalks. We recognize the tension we have created by suggesting that all classifications systems are partial, limited by the worldviews of their creators and then suggesting that their connection might result in an all-encompassing system. Ultimately we conclude they are two perspectives on the same phenomenon, which we might summarize as follows:

- Most of human knowledge transmission takes place by way of language;
- Language is a fairly localized social construct;
- Knowledge is likewise a social construct;
- We are forever translating for each other our glimmers of understanding of the world around us by creating yet more social constructs—systems through which to describe what we "know" (or, see as true).

As the processes of translation among different natural languages are practically infinite, so too are our subject classification efforts never-ending. As we know from translating among natural languages, however, it is often impossible to achieve a complete carryover of meaning: nuances present in one language may be communicable through experience or by other means, but not in the words of another language. This is where we see multiple truths or realities existing at the same time: different worldviews produce and prevent different perspectives, and are sometimes mutually exclusive, yet they exist and cannot be discarded as false.

We have been drawn to Joseph B. Altepeter's discussion of quantum information because of the interesting parallels we see in our own research. We will attend to these in more detail later, but one is worth noting now. Altepeter notes, "Measuring a qubit changes its value to match the result of the measurement." (The quantum bit, or qubit, is the simplest unit of quantum information.) As noted above, the very act of classifying an information object changes it. It adds adjacency, relationship, context, or worldview. It validates, or invalidates, establishing (or disestablishing) the object as a part of knowledge or reality. We suspect this is helpful to those who share or at least understand those additions. We also suspect this obscures the information object for those who don't. This leads us to consider the limitations of current cataloging practices.

# 2. Limitations of Current Cataloging Practices

Existing systems for organizing, accessing, and even using knowledge are rooted in the historically physical artifacts containing them. Library buildings are segmented into shelves and drawers. Though the Dewey Decimal System is no longer the dominant classification system in academic libraries, Melvyl Dewey's notion that "the physical layout of libraries should reflect this basic structure of knowledge" (Weinberger, 46) continues. Prior to Dewey, libraries typically assigned books a fixed location on a shelf, often assigning an acquisition number that defined its shelving order, and relied on an alphabetical listing of books by author to locate a book on a shelf. This obviously works well for a known item, but presents problems when one knows little about who has written about a topic. Dewey advocated shelving books by subject, such that the floor plan of the library would become "a map of ideas." (Weinberger, 52)

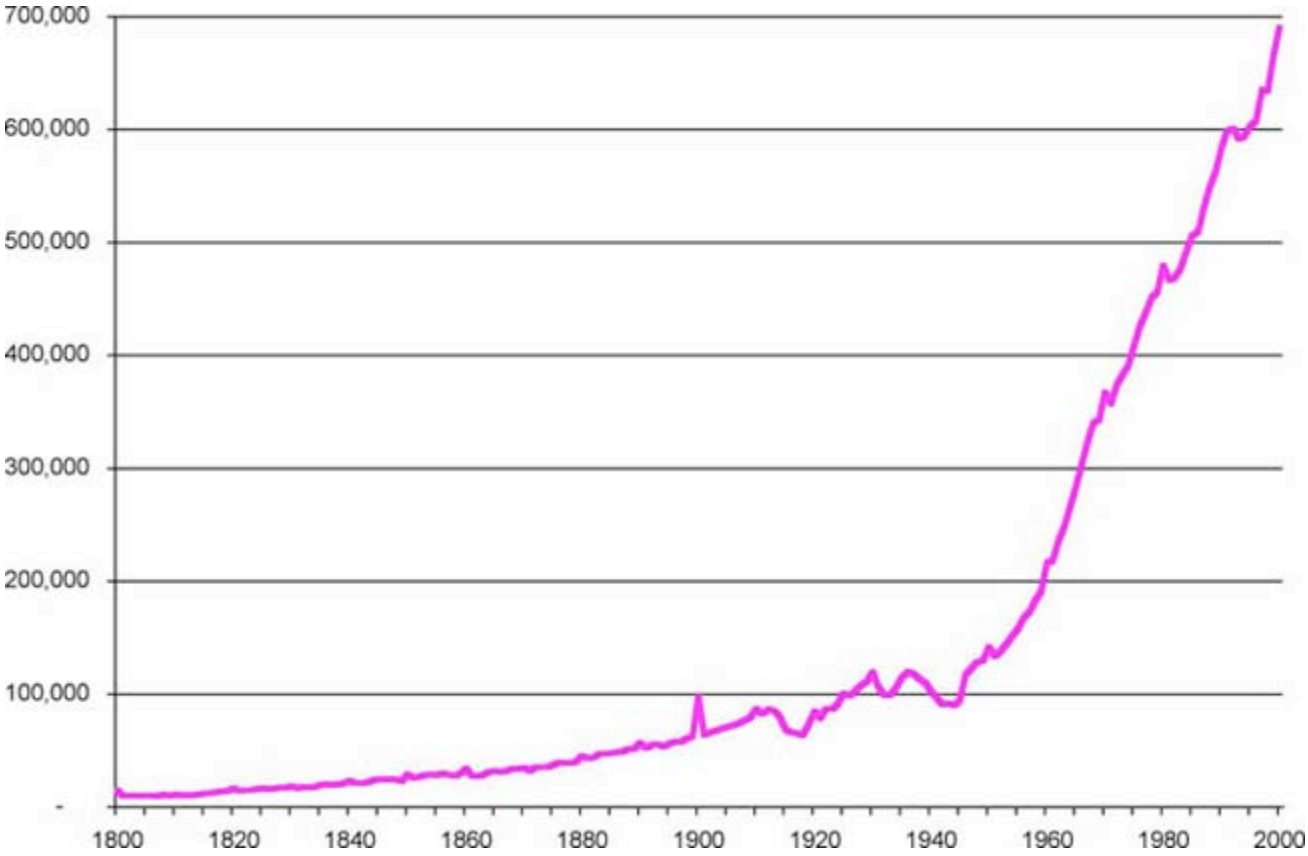
As long as cataloging tools and practices remain focused at the physical item level, shaped largely by the need of the user to locate the physical item on the shelf, libraries are not able to address the complex entanglement of ideas that might be contained within a single volume. And scholars notice the problem. Garth Green, a humanities scholar on the project team, noted:

> My book, for instance, is actually about what doesn't appear on the LC keyword list, and is not, per se, about what does appear there. But knowing that, what should I say the book is about? To depict this, should I choose disciplines, figures, concepts that appear in the book, concepts that appear the most, or the concepts that such frequently used concepts point to but which themselves never appear? (Green, 2011)

As noted earlier, knowledge (and the language used to communicate it) is a social construct. The analogy between language translation and subject classification made earlier helps us understand that the cataloging problem identified by Green is not simply an example of poor cataloging. In his classic work *Word and Object*, William Van Orman Quine identifies the problems associated with the "indeterminacy of translation." (Quine, 73-74) Without context, it is frequently possible to translate or interpret a word or phrase in multiple ways. One must rely on the available cues, the observable context, in order to accurately translate. Even then, one must recognize that one's interpretation of those cues is subjective, allowing for multiple translations or interpretations.

Green indicated an indeterminacy of significance in the case of his own, first book: "What is the significance of the book? Is it what was written in it? Or is it (as it is for me, the author) to be found in what the book wanted to say and didn't, that will be in book #2?" An indeterminacy of language points to an indeterminacy of significance. In this case, Green continued, "the book was about Kant; but it's significance will be shown in book #2 when book #1 becomes a book that introduces a completely new account and justification of Fichte." But it is at present, and perhaps even in the future, impossible to utilize this meaning in the determination of the book's significance. Further, he continues; "the book focused on a concept, inner sense, that is itself only significant in a wider context," say, of the theme of "self-consciousness" as such. However, "the book is not about that concept of self-consciousness as such," even though it has been categorized as such, but rather about Kant and about his specific usage of the specific concept of inner sense. In yet another acceptation, "the book is about rational theology, because Kant criticized the latter – once, for all and for ever, with a stunning success – with this doctrine of inner sense." However, "it never once mentions the latter, even though the latter is the book's terminus ad quem." This "suggests that what a book is 'really' about cannot be determined by a classificatory system until its significance has been determined in a wider context of use. The system has to codify the use, not the significance itself, because the latter 'is' only through the former." (Green, 2011)

Certainly, the situation Green describes might be better if each book could be cataloged by someone as expert as the author. Within academic libraries, cataloging practices would ideally have a subject area specialist who is able to place a particular work within the bibliographic universe of knowledge. However, with the rapid increase in the pace of knowledge creation and publication, this model is not scalable, if it ever was successful. Speaking of an analysis of bibliographic records for books in the OCLC WorldCat database, Lavoie and Schonfeld assert, "Approximately half of all books held in the system-wide collection were published after 1977." (Lavoie and Schonfeld 2006) Library staffing, particularly with subject area specialists has simply not been able to expand to meet the demand. Consequently, cataloging is accomplished within the very real constraints of a library's operational budget with the goal of providing an accurate bibliographic description and at least minimal subject classification so as to determine where the book should be placed on the shelf.




Print Manifestations by Year of Publication, 1800-2000
(Lavoie and Schonfeld 2006)

Yet scholars are often interested in different levels of discovery, access, and examination such as chapters, anthology sections, and digital collection items. What is the appropriate unit of scale for cataloged data not as a physical but as an information object? And how would libraries develop a scalable model for producing it? Theoretically, we could ask catalogers to codify these other levels of information. But economic constraints prevent us from spending a lot of time even on a book as a whole; often we do not have the resources to go deeper.

With the emergence of tools like Google Book Search and Google Books Ngram Viewer, new models for humanistic research have emerged. The research questions of scholars seem less bound by physical containers of knowledge. Both macro- and micro-analysis has become possible with digital information objects, prompting the need for alternative models for creating and maintaining metadata.

# 3. Knowledge Organization Is Power

Is it possible to assign subject tags to humanities objects automatically, as Johns Hopkins has done with medical articles? Maybe. But before we begin tackling this problem, we must answer a two-sided question: which subject tags? The question is both semantic and political. No matter what agency defines the categories into which we segment information, it has to be exclusive of someone: ontology this extensive can't be built by consensus. So there is power (and great responsibility) in being that entity.

Most academic libraries, and indeed librarians, take some pride in assuming a neutral stance to the information they collect and organize. Allan Sekula suggests that neutrality is illusory. Speaking more specifically of archives, he says, "Clearly archives are not neutral; they embody the power inherent in accumulation, collection, and hoarding as well as that power inherent in the command of the lexicon and rules of a language." (Sekula, 184) Later he states, "In short, photographic archives by their very structure maintain a hidden connection between knowledge and power. Any discourse that appeals without skepticism to archival standards of truth might well be viewed with suspicion." (186) One might say the same thing about the application of subject headings, even within widely accepted cataloging standards. Any assignment of subject headings or ordering of knowledge is done from a position of power.

In his *A Social History of Knowledge: From Gutenberg to Diderot*, Peter Burke reminds us that this history "is the story of the interaction between outsiders and establishments, between amateurs and professionals, intellectual entrepreneurs and intellectual rentiers. There is also interplay between innovation and routine, fluidity and fixity, 'thawing and freezing trends', official and unofficial knowledge." (51) Any cardinal reorganization of knowledge, to be conducted responsibly, needs both a channel for public feedback and an ultimate authority structure. Thus not only is there power involved, but also a tension of multiple interests, each of which must be addressed in order for the final product to achieve widespread acceptance.

But politics is only one side of the issue. Another aspect, at least as important, is the influence that all the opposites listed above have on shaping our implicit assumptions about knowledge organization. As the creators of major search engines have found, it is impossible to successfully guide users to knowledge discovery while working against their assumptions, regardless of whether the latter are conscious. So we must understand these expectations and take advantage of them in any effort to classify knowledge. Burke reminds us: "From Durkheim onwards anthropologists have developed a tradition of taking other people's categories or classifications seriously and of investigating their historical contexts." (81) It follows that any reclassification effort should include anthropological expertise. This approach would also acknowledge and make space for the inherently multi-institutionalized nature of knowledge, as well as for information that has not been institutionalized at all (and yet, as Burke states, is both valuable and inescapable if we are to understand what shapes our users' worldview.)

The semantic side of the question is the problem that spurred most of our thinking from the beginning. Existing ontologies have overwhelmingly hierarchical, tree-like structures. This emphasizes the segregation of knowledge into discrete disciplines, and—put together with the economic constraints we mentioned above—discourages interdisciplinary research. The real

structures of knowledge don't have vacuum-filled interstices. The tree branches of hierarchical ontologies create them, obscuring information. We need more of a rhizome (see Assumptions 1 and 2 below for more on this).

## 4. Subject Headings are Hard!

Another issue that we've known about for over ten years, yet have so far failed to address, is that nobody understands subject headings very well—not even librarians. In the late 1990s, Karen Drabenstott conducted a study (Drabenstott 1999, 10) whose primary objective was, "to determine the extent to which children, adults, reference, and technical services librarians understood subdivided subject headings." Her team chose twenty-four Library of Congress (LC) subject headings, had two librarians with decades of experience write definitions for them, and then asked study subjects to write their own definitions of the same subject headings. Here's what they found overall: 32% of the children and 40% of adult non-librarians gave correct definitions. Reference and technical services librarians came in at 53% and 56%, respectively. These are low numbers!

To help get perspective on the problem we brought in Dan Benedetti, bibliographer at BU's Mugar Library. He concurred that LC subject headings were difficult to understand, and added that many other such systems were, as well. Some such systems underlie the semantic structure of widely used library databases (Beale, 2010). As our conversation evolved, we began to investigate what current taxonomy practices (Dunsire, 2010) and discussions of them in the library literature might reveal.

Drabenstott's team observed, "About the only properties of subject headings that were likely to indicate subject headings to which respondents would have difficulty assigning correct meanings were subject headings that changed meaning across the various contexts and subdivision orders studied in this project." (1999, xviii) Drabenstott's study recommends that representatives from the various demographic groups studied be involved in the creation of subject classification systems, and that there may be benefit in providing "separate indexing systems for children, adults, librarians, and subject-matter experts," (1999, xviii) customized to the needs of each user group. Such a system would allow a person seeking information to choose the system best suited to her/his level of familiarity with the subject areas involved.

The problems Drabenstott identified with subject headings are easily extensible to scholars researching outside of their field of expertise. Nancy Fried Foster, in her 2010 "The Librarian-Student-Faculty Triangle: Conflicting Research Strategies?" frames learning as "movement from periphery to center in a community of practice." She contrasts students and researchers, pointing out that each of the two sees only the knowledge that they are enabled to see by what they already know. We can view anyone's level of experience in a field in the same light. Student or faculty, when entering a new field of inquiry, we start anew.

## 5. Diversity of Constituencies — Digital and Analogue Cultural Landscapes

Which brings us to the question of audience. Who are the practitioners we're addressing? The short answer: we don't fully understand.

We cannot lump all scholars together with regard to how much, or even how, they use networked technologies. Digital humanists are on Twitter in force, but there are fields not nearly

as well represented. There doesn't seem to be a large philosophers' community using it, for example, or Italian literature scholars'. Insofar as they use electronic communication technologies, the one philosophers of religion use almost exclusively for research is email. Then there are the disenfranchised, who may be so for a multitude of reasons. (Some scholars, for example, work with scripts for which there are no standard Unicode sets.) Issues like these will have a profound effect on how scholars engage with digital objects.

This is to say: digital scholarly communities may be even more varied in nature than analogue ones, in ways that we haven't fully explored. They potentially engage both digitally native cultural content, and all of our previously held analogue cultures.

We do know that researchers doing "interstitial"—interdisciplinary—work are a core part of our target demographic. We might separate them into two groups: those who use social media in their work and/or put their work online, and those who don't. The latter can only be found online if one already knows what one is looking for. They will appear in no more and no fewer places than they would in the analogue world, except that now they're string-searchable.

So as we think of digital cultural landscapes, we keep in mind that they encompass both digital and analogue cultures, to differing degrees, and that our target audience—the communities acting on and being affected by this project—is split between these two spheres. Individual members of this audience, and their work, are not represented equally by existing digital content. They come to digital content with very different attitudes to color their experience.

While we can't predict the full variety of our audiences, we recognize that much knowledge, even if born analog, is now encountered in digital formats. With this migration to digital formats, all audiences encounter rapid ongoing change in how they engage digital knowledge. Charlotte Hess and Elinor Ostrom assert

> The technologies that allow global, interoperable distribution of information have … dramatically changed the structure of knowledge as a resource. One of the critical factors of digital knowledge is the 'hyperchange' of technologies and social networks that affects every aspect of how knowledge is managed and governed, including how it is generated, stored, and preserved. (2007, 9)

Ultimately, it is hard to imagine an audience that does not, or at least will not, encounter digital knowledge and consequently discover and develop new models for scholarship based on the affordance of digital formats. Charles Henry describes this hyperchange[2] as 'rapid, exponential, discontinuous, and chaotic.' He goes on to say

> Aspects of hyperchange include increasingly permeable boundaries between knowledge creators, publishers, and readers; more flexible iterations of the processes and products of scholarly communication; the rise of new methodologies; greater collaboration within and among disciplines; and a more porous flow of original research among undergraduates, graduates, and faculty. (2010, 2)

Analogue knowledge by comparison is much more bounded. Models for organizing and analyzing analogue knowledge are much more established. For many scholars, analogue

[2] Derm Barrett introduced the term *hyperchange* in his *The Paradox Process: Creative Business Solution, Where you Least Expect to Find Them.* New York: AMACOM, 1998. He defined hyperchange change that is: "pervasive, disruptive, unpredictable, perplexing, transformative, explosive in its pace, and destined to remain permanent." (10)

research models seem rationale if not intuitive. By contrast, Henry describes today's digital commons as

> … a contested zone where bounded and unbounded impulses compete: intellectual property laws, copyright, and the commodification of information can struggle with open access, file sharing, social networks, and a much more free-form, nonhierarchical, even chaotic participation in the creation and distribution of knowledge. (2010, 2)

An obvious question for the project team is whether subject classification as it is practiced in the analogue information landscape can ever work in a digital information landscape. Even if computationally enhanced or automated, is it bound to an analogue knowledge landscape in ways that prevent it from functioning in a digital knowledge landscape? In the context of that question, there are a number of assumptions about the way knowledge has traditionally been organized that need to be challenged.

# Assumptions That Need To Be Challenged

The interdisciplinary conversation enabled by this grant highlighted several assumption of the project team that are not necessarily shared across knowledge-work fields. As these assumptions emerged in the course of our interaction during the work weekend with invited scholars, we concluded they should be challenged. That interaction, and the emergent properties of the knowledge we've gained, are representative of the logistical setup of the entire grant project: notably absent from it was a rigid framework for conversation. We let that emerge as well, and we believe that this allowed us flexibility for greater insight.

## 1. Knowledge is Analogue

In The Philosophy of Information (Oxford UP 2011), Luciano Floridi reminds us that we don't know whether reality is discrete (reducible to smallest elements with clearly defined boundaries) or analogue (continuous). (316ff) This applies to the physical and informational universes alike. Yet, one running assumption in library and information science literature — and practice — is that the sum of human knowledge is analogue, and we can only ever approximate it computationally.

Letting go of that assumption in our work on deriving subject headings from humanities texts may well bring us toward the theoretical work being done in quantum computing—research into computing systems based in something other than ones and zeroes. We believe that we will get further in conceiving automatically generatable models of knowledge by being flexible about the questions we are willing to ask of our qubits.

In "A tale of two qubits: how quantum computers work" (*Ars Technica* 2010), Joseph Altepeter describes quantum information theory as follows:

> Bits, either classical or quantum, are the simplest possible units of information. They are oracle-like objects that, when asked a question (i.e., when measured), can respond in one of only two ways. Measuring a bit, either classical or quantum, will result in one of two possible outcomes. At first glance, this makes it sound like there is no difference between bits and qubits. In fact, the difference is not in the possible answers, but in the possible questions. For normal bits, only a single measurement is permitted, meaning that only a single question can be asked: Is this bit a zero or a one? In contrast, a qubit is a system, which can be asked many, many different questions, but to each question, only one of two answers can be given.

It would seem that qubits, like bits, are discrete. While that may be true, the measurable values of qubits are a superimposed combination of two possible answers, and the answers depend on the question we are asking (the way in which we measure a qubit in a specific instance). For our purposes, then, qubits must be treated as neither discrete nor analogue.

In quantum information, which Altepeter calls "the physics of knowledge," we gain new frameworks, and perhaps tools, to enable us to address the complexities and problems we have already identified in subject tagging in the humanities. A central concept in quantum theory is that an atom, or a subatomic particle, can exist in two or more places at once. Just as matter and energy can display the characteristics of both particles and waves, we might explore the possibility that units of information might display both digital and analogue characteristics—that is, the number of possible interpretations of an information unit may vary depending on the cultural context of the moment. Assuming this analog-digital duality, it seems worth questioning whether the assumptions for organizing information in an analogue landscape are applicable to organizing it in a digital landscape.

One of the investigative models for humanistic research that we continue to explore might be described by an iterative use of the query "I'm interested in ..." as an expression of the interest that animates the scholar's research. With the results received from the information object in response to the query, the scholar is able to refine the search. An analogue understanding of information might assume that this amounts to simply narrowing or broadening the search scope through the use of facets or additional search terms. Repeated queries are assumed to be helping one to discover the desired information object(s), hopefully with the least iteration. If we accept Charles Henry's notion of the boundedness of analogue information, the more expansive understanding of discovery as examination, is likely to be constrained. Much of the information of interest to researchers in the humanities lies beyond the information object itself. More permeable boundaries are required to discover other information objects with which this information object is related.

The project team developed a different understanding of this research model that seems more consistent with thinking about information objects as qubits. Thinking in more spatial terms, one might imagine the scholar's location in space being shifted around a sphere as a result of receiving the response from the qubit. (Or perhaps the location of the information object shifts.) From that new location (or perspective) the scholar asks again, "I'm interested in …." This time, however, the question to the same qubit is a different question, informed by the previous response. This research model is not focused on discovery of the identity or location of the information object (qubit), but on examination or measurement. The researcher seeks to develop a deeper understanding of the information object (qubit) by means of repeated queries, each informed by the response to the previous queries. Tools for examining a qubit might enable the researcher to discover the other qubits with which the qubit being examined is related (entangled).

It should be clear that both understandings of the research process are not only valid, but also required. Scholars need to be able to both locate and examine information objects. For examination, subject tags may be less useful, except as qubits themselves. For identifying and locating the information object, though, some form of subject term classification seems helpful, though as the team recognized, the interests that animate research are not always congruent with the disciplinary domains used to define a classification system. And as previously noted, subject terms are often misunderstood, representing a world view or systematizing of knowledge not shared by the individual searching for information.

If we understand subject terms to be socially constructed, we need to anticipate that different groups interpret the same information in different ways without any of them being incorrect. To

put it in different terms that we've explored in this paper: in the second and third orders of order, knowledge (being immaterial) can and often does exist in two or more different "locations"—two or more different contextual situations. It will be useful to make an effort to represent that computationally in future research on automatic subject indexing in the humanities.

It is also clear that scholars in the humanities are interested in more than discovery of information. Their interests prompt iterative interrogation of the same information objects. If we think of information objects as qubits rather than analogue or continuous objects, we might develop tools for measurement and examination designed to take advantage of their digital characteristics.

We previously identified the inability of standard human based cataloging and indexing models to scale to the level required by emerging models of research in the humanities. Here quantum computing may eventually provide tools to aid this process, but such computers are small and the data structures and algorithms do not exist yet. While quantum computing affords the possibility of developing theoretically efficient inference algorithms that could provide scalable tools for such a task, many of the issues raised in this paper could be tackled with probabilistic models. "Iterative interrogation" thus translates into conditioning the model on more information or refining the structure of the model. Probabilistic modeling is well understood and has current tool support. Such tools might prove to be far more efficient and more effective for research in the humanities than traditional models of subject classification.

# 2. Knowledge == Subject Classification == Tree Hierarchies

We've long conceived of knowledge as rhizomatic, but that is not reflected in extant subject classification schemes, which are hierarchical. By "rhizomatic" we are, of course, referring to a concept put forward by Deleuze and Guattari in 1987. Ian Buchanan's extract of key principles provides a helpful summary:

- The rhizome connects any point to any other point (connections do not have to be between same and same, or like and like).
- The rhizome cannot be reduced to either the one or the multiple because it is composed of dimensions (directions in motion), not units. Consequently no point in the rhizome can be altered without altering the whole.
- The rhizome operates by variation, expansion, conquest, capture and offshoots (not reproduction).
- The rhizome pertains to an infinitely modifiable map with multiple entrances and exits that must be produced.
- The rhizome is acentered, nonsignifying, and acephalous. (Buchanan, 2007)

Anthropologists Rachel Douglas-Jones and Salla Sariola describe the concept's helpfulness in creating a space in which to engage the complexities of anthropological research.

> The concept of rhizome has given us a theoretical space to work in, a tool to reflect upon the assemblages of our fieldwork when few conceptual tools existed. To us, these assemblages consist of the messy encounter of discursive, international and national forces where there is no centre to control the process of how ideas, practices and policies are adopted in any given context. These processes are rhizomatic in their ubiquitous, centre-less way; they materialize independent of each other in a number of locales, exist simultaneously in various fields. If we see the rhizome as a tool to understand slices of an 'ethnographic cake', we surrender to the fact that a rhizomatic fieldsite is temporal and constantly shifting, changing and alive. (2009)

A consistently recurring theme through the project was the need to re-contextualize information objects by reaching beyond of the usual two dimensions into more complex connective infrastructures. As we will discuss in the next section, humanities researchers are increasingly interested in the social networks within which an information object is exists. David Smith, a computer scientist and one of the participants in the weekend conference, suggested network modeling may be a helpful approach to discovery and examination of information objects. Those structural or generative models of networks—the most familiar being exponential random graph models—are widespread in fields such as computational social science.

Hierarchies by their nature assume that there is a center, a starting point and that an order can be imposed. We challenge that assumption as regards the sum total of human knowledge. Traditional subject classification systems tend toward reductionism, in that they attempt to classify information objects with a limited or constrained view of their ecosystems. This often obscures connections that exist when they aren't represented in the hierarchy. Additionally, such classification systems attempt to impose an order that masks the messiness of reality.

David Smith (p.c.) suggests exploring different network models that, in various ways, relax some of cataloging's too stringent assumptions. Exploring other models holds the potential for the development of more nuanced discovery and examination tools for humanities research. This leads us to consider the ecosystems, or social networks, within which information objects exist.

## 3. Social networks are vehicles for information discovery and analysis

At first glance, this assumption seems perfectly correct—and it is, but it is also incomplete. Social networks are also containers of knowledge work processes, inseparable from them.

The rhizomatic approach to classification detailed above mirrors the social context of knowledge production (which, we note, is different from the social context of an informational object). Ideas are generated in a less than orderly manner, usually aided by serendipity and regardless of whether they fit into an existing hierarchical knowledge structure. What they do always have are non-tiered cognitive connections to pre-existing knowledge. So also our subject classification schemes should consist of interconnected, non-tiered nodes—a rhizome.

Our insistence on mirroring social context (social networks) is a departure from most previous work in this field. As librarians, when organizing information, we have tended—of technological necessity—to ignore the social networks of discovery, as if the two could be separated. They never can be; and we've compensated for our technological limitations by setting up reference desks, or social networks of discovery—direct descendants of the oral cultures in which our knowledge was first transmitted through storytelling.

Though we speak of literacy as an age succeeding orality, the social essence of human knowledge transmission has not changed since we began. Reference points have spread out from physical desks to social media (notably telephone, and online instant messaging). But that still hasn't proved scalable enough: the minute social tagging became widely available, millions of people hungry to build informational scaffolding dove right in. We are looking to unite that drive with subject expertise, and with automated technologies that would aid us in seeing (and having the chance to accept or reject) more possibilities.

## 4. Classification is for Categorizing Information

Again, this is a truth: we classify information in order to find it later. Because classification has historically been time-consuming and difficult work, we have been understandably reluctant to classify the same information in more than one system. But there is immense potential in using classification for discovery of new knowledge. When we recognize that the product of classification is in fact a new information object, a qubit, it becomes possible to examine this new object just as we have the information object classified. With computational tools, we can make the results of that work more efficiently usable. With quantum computing tools, we might examine or measure the entangled qubits, the original information object and its classification, to discover how the classification affects the original object.

We arrived at the conclusion that the Earth is not the center of the universe by observation. Classification is, if freed from frameworks as much as possible, a powerful tool for recording observation. The records produced constitute data that can then be considered from a bird's-eye view. That, and not fitting an object into an existing knowledge framework in an attempt to make that framework all encompassing, should be the goal of classification.

So we need more tools that would allow us to treat information like building blocks in order to create new knowledge. Allowing classification structures to emerge from the act of classifying, instead of the other way around, would be freeing. Structure ought to be the product, not the framework.

No system for automatically classifying humanities texts can operate without human intervention. We believe that at the point of intervention, we should be asking not whether a given piece of information is accurate, but whether it is *true*. Truths overlap and often contradict each other; multiple classification systems for human knowledge are inevitable.

# Seeking Solutions

Bearing that in mind, we wondered: what could everyone use? What information do we want to get at, that would be broadly applicable across user bases?

## Context

One overarching theme was context.  It has driven much of the Internet, online commerce, and social media development in the last decade. Librarians have always valued context, but we need it on more granular levels. What if we were able to automatically provide socio-historical context for individual works of literature, for people, for groups under study? The process of assigning subject headings would be greatly enriched if we could do it by mining semantic and social links among works.

For individual works of literature, we can imagine asking: who commented on this piece I'm looking at? Who commented on its author's other works?  Is there overlap between the two? What are the common subjects explored by author and commenters?

One fertile question to ask about individuals and groups under study is: did their social networks overlap, and if so, how? Such information is currently discoverable both in the classroom and in written prose, but can be difficult to trace in the latter; text mining can help us discover deeper social connections, and communicate them to the solitary researcher.

## Serendipity

We also want to make serendipitous discovery more likely to occur. In order to do that, we return to context as more than co-citation analysis. Analyzing citations automatically narrows our observation to institutions already in place to produce those citations, and participants in those institutions. Our field of inquiry becomes much more broad if we succeed in bringing in knowledge that's deducible, but not explicitly stated. Some respondents at the Digital Library Federation meeting suggested link analysis, similar to the approach used by Google, might be a productive approach. Network modeling tools hold potential for enabling such analysis. This is difficult both technologically (semantic text mining is a hot topic, but the needed technologies are not advanced enough yet), and because we are essentially setting out to find our own blind spots. We suggest, however, that serendipitous discovery is not confined to systems that enforce physical adjacency regardless of the classification system. In a digital environment, serendipitous discovery might take place by means of analysis of context defined in a variety of ways.

One way to do this is to look at what others are doing. Right now, in the grand scheme of the scale on which we need to operate, everyone is taking first steps. We simply don't have enough of this contextual knowledge codified to cross-reference anything yet. Digital humanists have been hand encoding word-based texts for a few decades, and it has proved educational and productive, but isn't generally regarded as a scalable solution. Still, we may need a broader base of hand-encoded knowledge to start off with.

## Interdisciplinarity

We want to enable researchers to identify and locate the information they need, regardless of whether they know where in a domain-based hierarchy this information is located. In order to do that, we need a more distributed classification system, with a larger number of subject tags per classified object. Crucially, we also need increased granularity of those headings to the object. For example, if a biography of a 19th-century German philosopher digresses into a discussion of the effect the Industrial Revolution had on German intelligentsia, a scholar researching the Industrial Revolution should be able to find this, though it makes up only a small part of the biography.

Such an approach to classification, as we discussed above, has historically been prohibitively expensive. For example, traditional library cataloging models for books provides subject classification only at the table of contents level. The increased granularity begins to look like a book index. We certainly have new knowledge to apply to the problem, and it is clear that we must combine expertise from historically diverse fields such as sociolinguistics, library and information science, various areas of the humanities, and software engineering. However, significant human effort of the traditional-humanities-research sort must still be applied.

One way such a scale of human effort might be harnessed, that has been explored in recent years, is crowdsourcing. This method has had limited success in knowledge work, but we might attempt to applying it by taking a cue from the reCAPTCHA experience.

reCAPTCHA, designed by Carnegie Melon University, is software used to determine that an online commenter is human and not a robot. When submitting a comment to a website (often a blog, a newspaper, or a social media site), the user is asked to type in two words that appear in a small image, usually distorted such that optical character recognition (OCR) programs cannot parse them. One of these words has a transcription known to the machine, and that is used to determine that the user is human. The other word is actually part of a book digitized during the course of a participating digitization project that OCR software has failed to recognize. The

typed character string serves as an act of proofreading on the part of the user. A tiny act, but the idea is so simple and effective that it has been adopted widely enough to have a significant effect.

To achieve true interdisciplinarity in subject classification, we would need to branch out beyond professional catalogers and get many more sets of eyes on the same content. Doing that in the first place, and after that getting people to participate, is a difficult and large-scale problem best approached institutionally. What if such contributions could be documented and certified, and graduate courses began to give credit for them? What if promotion and tenure committees viewed them as a service to the profession?

# Classification Process

Other themes have emerged in our conversations about the process of classification itself:

***Iteration***. We don't just need to find connections. We need to iterate the process of searching for them, and triangulate the results. This risks highlighting systems that are already prominent and further obscuring knowledge that's already difficult to find. A possible way to *partly* address that would be to iterate not only in the process of research but also in the process of classification *itself*, and that over time: since classification is closely tied to its creator communities of practice, and the communities are always evolving, it makes sense that classification must also evolve.

***Teaching, the User Interface, and Search Agents.*** Our investigation of the library literature led us to think about the researcher in the absence of a teacher who might help explicate classification. In a way, improvements to the usability of various databases, such as the way facets in modern discovery engines like Ex Libris Primo and ProQuest Summit organize hierarchies, and Google's suggestion of related terms under "something different" (Provine 2010), are really attempts to improve active learning of the user (Swartz 2008). We realized that novice researchers, but also more advanced ones coming from other disciplines may not readily understand whatever classifications might result from our work.

However, as long as we control our user interface to them, tracking characteristics of the researcher presents an interesting opportunity. A potential solution for many of these problems may be to explore a search interface that becomes more of a customized portal, in the way that Amazon.com's "recommendations" are personalized, or the way Pandora Radio suggests songs for listeners based on an individual's previous ratings. In offering "widgets" with advanced classifications to certain users that may be enabled - facets that activate as we recognize that they have more expertise in a particular area or are coming from a certain perspective - we have a "moment" in which to educate the user about a possibility.

***Flexibility*.** By definition, there's no way to predict what will be useful to future scholars, so we imagine subject heading lists being starting points that can be modified locally; a guided-social-tagging-like approach to classification. Being able to modify classification like this would more accurately *reflect* to the community the knowledge it has produced, and also be a more *self-reflective* exercise—not to mention a greater distribution of labor, which means more achievable sustainability and scalability.

***Weighted systems*.** A good subject classification system will incorporate a weighted approach, and not a binary one. In a book about John Locke, we need *some* way to indicate that a small portion might be about urban planning; but certainly that cannot have equal weight in discovery with other subject headings *more* relevant to the book.

And finally, ***layering***. Being able to layer different divisions of knowledge spheres on one another enables discovery in three ways.

1. First, of course, is the fuller perspective offered by a multidimensional view, much like depth perception offered to most of us by the two eyes we possess.

2. Second comes from the act of *choosing* the lenses through which you look at unfamiliar knowledge: such an act engenders a deeper and more constant awareness of our position within the noosphere than we now tend to practice.

3. Lastly, *any* classification system is going to have interstices. Layering multiple systems would help us *see* the interstices, and in some cases fill them in, in the process better understanding the systems themselves.

# Summary and Next Steps

The focus of our research is to enhance and expand humanities subject terms in order to help people find and use recorded knowledge in the course of their research. Our awareness of the difficulties of classification that we'd need to address in order to automate this process for the humanities has led us to assemble experts from several different fields to work on this issue.

We recognize the limitations of existing classification systems, and are at least in part interested in how we address issues of relevance and allow for an iterative tool that moves toward being capable of learning from the scholar. Ultimately, we would like to teach computers to make basic semantic connections in the humanities more quickly than scholars are able to do. But knowledge and systems for its classification are socially constructed; doing this work while addressing biases and power structures inherent in it involves a composite system in which worldviews expressed in different ways of classifying can be layered. We thus move from viewing subject classification as unambiguous declarative work to viewing it as a research process and a vehicle for advancing viewpoints. We position subject classification at the intersection between discovery and examination, and propose that it belongs in both categories.

We also propose that rigid hierarchies within subject classification systems impede research by creating the illusion of comprehensiveness while in fact obscuring important knowledge. We propose a less hierarchical, explicitly non-comprehensive approach justified in part by the flexibility of layering multiple sets of subject headings as described above. We believe that large classes of models, network modeling, and rhizomatic approaches would be better suited to a larger variety of users than are current subject classification systems, which have been shown to be difficult for users to understand.

We are conscious of the dangers of information overload. But the potential benefits of producing more subject metadata, which we believe to be an essential process in current knowledge work, are too great to allow ourselves to be dissuaded from doing the work. If we can automate subject metadata generation, we will be able to address information overload practically.

We also recognize that in the digital cultural landscape, quantum information and quantum computing approaches that utilize probabilistic modeling and theoretically efficient inference algorithms may ultimately be better positioned to address the needs of humanities researchers

for discovery and examination than the kind of subject tagging that humanities scholars have long valued. To this end, we propose some directions for further work.

## Specific Directions/Next Steps

**We need to gather more data.** It is clear that we need more data to guide our decisions about what might be done to improve humanities scholars' ability to discover and analyze texts. Several studies, for example, point to problems with subject headings. Many of these were done prior to major digitization efforts. Very few, if any, of these have been replicated. It would be very helpful to replicate these studies to verify their findings. Previous studies suggest problems, but we don't have enough data to be statistically significant or to identify causal relationships.

It is also important to construct controlled studies that measure differences for discovery between analogue and digital resources. In single search box, Google-like search engines, are subject headings helpful?

In addition to replicating previous studies about the use and understanding of subject terms, new controlled studies should be developed that explore more deeply the use of subject tagging:

- Introduce prototypical ways of using subject tags for searching that are a superset of those previously explored.
- Measure the effectiveness of user modifiable subject tags versus subject tags that are authority controlled.
- Measure whether additional subject tags helpful, or create confusion in the discovery process.
- Measure the level granularity at which subject tags are effective. Identify the cost benefit ratio that would help libraries decide what level of tagging is most effective.

**Explore the relationship between discovery and examination in humanities research**.

- Do the two processes require different kinds of subject tagging?
- Are user modifiable subject tags more appropriate to examination than discovery?
- Are tools and approaches such as network modeling more appropriate to examination than discovery?

**Explore different models for application and use of subject tags.**

- Is it possible to use computer analysis such as the MESH Indexer Web Service developed by Johns Hopkins University? Could this be done across disciplines, or would disciplinary thesauri be required?
- What are effective ways to implement social tagging to enhance both discoverability and examination of humanities texts? What is the scale required for social tagging in order to make such tagging useful for a broad community of users?
- If multiple layers of classification systems are developed, what is the most effective way of displaying them to the end user?

**Explore the potential of probabilistic and network modeling and eventually quantum information to provide effective tools for humanities research.**

- How might probabilistic or network modeling help us to develop more nuanced discovery and examination tools?
- Are such tools better suited to aiding humanities research in a digital cultural landscape then traditional subject tagging?
- And eventually, how might quantum computing enable the development of contextualized knowledge in the humanities?

## Cited Works


Altepeter, Joseph B. “A tale of two qubits: how quantum computers work.” *Ars Technica* (2010): Web. 3 Nov. 2011.

Bailey, Chris, and Hazel Gardiner. *Revisualizing Visual Culture*. Farnham, Surrey, England: Ashgate, 2010. Print.

Barrett, Derm. *The Paradox Process: Creative Business Solutions, Where You Least Expect to Find Them*. New York: AMACOM, 1998.

Berger, Peter. *The social construction of reality a treatise in the sociology of knowledge,*. [1st ed.]. Garden City N.Y.: Doubleday, 1966. Print.

Bowker, Geoffrey C, and Susan L. Star. *Sorting Things Out: Classification and Its Consequences*. Cambridge, Mass: MIT Press, 1999. Print.

Buchanan, I. (2007). Deleuze and the Internet. *Australian Humanities Review*. 43 December. Online. «http://www.australianhumanitiesreview.org/archive/Issue-December-2007/Buchanan.html» (accessed 14 November 2011).

Coyle, K. “Machine Indexing.” *The Journal Of Academic Librarianship* 34.6 (2008) : 530-531. Web. 31 Oct 2011.

Deleuze, Gilles, and Félix Guattari. *A Thousand Plateaus: Capitalism and Schizophrenia*. London [u.a.: Continuum, 2010. Print.

Davenport, Thomas H, and Laurence Prusak. *Working Knowledge: How Organizations Manage What They Know*. Boston, Mass: Harvard Business School Press, 1998. Print.

Green, Garth W. E-mail correspondence, October 30, 2011.

Henry, Charles, et. al. "The Idea of Order: Transforming Research Collections for 21st Century Scholarship. Washington, DC: Council on Library and Information Resources, 2010. Web. 11 Nov 2011.

Hess, Charlotte, and Elinor Ostrom. *Understanding Knowledge As a Commons: From Theory to Practice*. Cambridge, Mass: MIT Press, 2007. Print.

Schonfeld, Roger C.; Lavoie, Brian F. “Books without Boundaries: A Brief Tour of the System-wide Print Book Collection.” *Journal of Electronic Publishing* 9.2 (2006) . DOI: http://dx.doi.org/10.3998/3336451.0009.208

Sekula, Allan. “Reading an archive: photography between labour and capital.” *Visual culture : the reader*. Ed. Jessica Evans & Stuart Hall. London ;;Thousand Oaks: SAGE Publications in association with the Open University, 1999. Print.

Smith, David. E-mail correspondence, November 23, 2011.

Vries, J. “An automated Indexing System Utilizing Semantic Net Expansion.” *Computers and Biomedical Research* 25.2 (1992) : 153-167. Web. 31 Oct 2011.

Wingert, F. Medical linguistics: Automated indexing into SNOMED. Crit. Rev. Med. Znf. 1, 333 (1988).

**Evolutionary Subject Tagging in the Humanities**
**Select Bibliography**

Abbott, A. "Library Research and its Infrastructure in the Twentieth Century." (2008)Web.

---. "Publication and the Future of Knowledge." *Keynote address to the Association of American University Presses (electronic version).Retrieved* 4 (2008)Web.

---. "The Traditional Future: A Computational Theory of Library Research." *College & Research Libraries* 69.6 (2008): 524. Web.

Allison, S. D., et al. *Quantitative Formalism: An Experiment*. Stanford Literary Lab, 2011. Web.

Altepeter, Joseph B. “A tale of two qubits: how quantum computers work.” *Ars Technica* (2010). Web. 3 Nov. 2011.

Bailey, Chris, and Hazel Gardiner. *Revisualizing Visual Culture*. Farnham, Surrey, England: Ashgate, 2010. Print.

Beall, J. "Academic Library Databases and the Problem of Word-Sense Ambiguity." *The Journal of Academic Librarianship* (2010)Web.

Berger, Peter. *The social construction of reality  a treatise in the sociology of knowledge,*. [1st ed.]. Garden City  N.Y.: Doubleday, 1966. Print.

Bogers, Toine, Willem Thoonen, and den Bosch van. *Expertise Classification: Collaborative Classification Vs. Automatic Extraction*. Ed. Joseph T. Tennis. Web.

Bohannon, J. "Google Opens Books to New Cultural Studies." *Science* 330.6011 (2010): 1600. Web.

Bollen, J., M. A. Rodriguez, and H. Van de Sompel. "MESUR: Usage-Based Metrics of Scholarly Impact". *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries.* Web.

Bowker, Geoffrey C, and Susan L. Star. *Sorting Things Out: Classification and Its Consequences*. Cambridge, Mass: MIT Press, 1999. Print.

Broughton, Vanda, and Aida Slavic. *Building a Faceted Classification for the Humanities: Principles and Procedures*. Web.

Chaudhry, Abdus Sattar. "Assessment of Taxonomy Building Tools." *Electronic Library* 28.6 (2010): 769-88. Web.

Coleman, Anita Sundaram, and Paul Bracke. "Controlled Vocabularies as a Sphere of Influence." Ed. K. N. Prasad. Web.

Coyle, K. “Machine Indexing.” *The Journal Of Academic Librarianship* 34.6 (2008) : 530-531. Web. 31 Oct 2011.

Douglas, D. "The Multi-Dimensional Approach to Linguistic Analyses of Genre Variation: An Overview of Methodology and Findings." *Computers and the Humanities* 26.5 (1992): 331-45. Web.

Douglas-Jomes, Rachel, and Salla Sariola. “Rhizome Yourself: Experiencing Deleuze and Guattari from Theory to Practice.” *Rhizomes: Cultural Studies in Emerging Knowledge* 19 (2009). Web. 14 Nov. 2011.

Drabenstott, Karen M., Schelle Simcox, and Marie Williams. “Do Librarians Understand the Subject Headings in Library Catalogs?(Report on the Extensive Study Underway to Determine if Users and Librarians Understand the Subject Headings in Catalogs).” *Reference and User Services Quarterly* 38.4 (1999): 369. Web.

Drabenstott, Karen M. "Why I Still Teach Online Searching." *Journal of Education for Library and Information Science* 45.1 (2004): 75-80. Web.

Drabenstott, Karen M., Schelle Simcox, and Eileen G. Fenton. "End-User Understanding of Subject Headings in Library Catalogs." *Library Resources & Technical Services* 43.3 (1999): 140-60. Web.

Dunsire, Gordon. "Signposting the Crossroads: Terminology Web Services and Classification-Based Interoperability." *Knowledge Organization* 37.4 (2010): 280-6. Web.

Fadaie Araghi, Gholamreza. "Users Satisfaction through Better Indexing." *Cataloging & Classification Quarterly* 40.2 (2005): 5-12. Web.

Foster, Nancy Fried. "Librarian□Student□Faculty Triangle: Conflicting Research Strategies? Paper Presented at the Association of Research Library’s Library Assessment Conference, Baltimore, Maryland, October 26, 2010".2010. Print.

Gnoli, Claudio. "Is there a Role for Traditional Knowledge Organization Systems in the Digital Age?" *The Barrington Report on Advanced Knowledge Organization and Retrieval (BRAKOR)* 1.1 (2004)Web.

Kipp, Margaret E. I., and D. G. Campbell. *Patterns and Inconsistencies in Collaborative Tagging Systems : An Examination of Tagging Practices*., 2006. Web.

Kipp, Margaret E. I. "Exploring Inter Tagger Consistency Measures".-11, Web.

Kipp, Margaret E. I. "Searching with Tags: Do Tags Help Users Find Things? “Web.

Knott Malone, Cheryl. *When More is Better: A Counter-Narrative regarding Keyword and Subject Retrieval in Digitized Diaries*. Ed. Joan Lussky. Web.

Lepori, Benedetto, Lukas Baschung, and Carole Probst. "Patterns of Subject Mix in Higher Education Institutions: A First Empirical Analysis using the AQUAMETH Database." *Minerva: A Review of Science, Learning and Policy* 48.1 (2010): 73-99. Web.

Lieberman, E., et al. "Quantifying the Evolutionary Dynamics of Language." *Nature* 449.7163 (2007): 713. Web.

Lin, Xia, et al. *Exploring Characteristics of Social Classification*. Ed. Joseph T. Tennis. Web.

Louwerse, M. M. "Semantic Variation in Idiolect and Sociolect: Corpus Linguistic Evidence from Literary Texts." *Computers and the Humanities* 38.2 (2004): 207-21. Web.

Lu, Caimei. "User Tags Versus Expert-Assigned Subject Terms: A Comparison of LibraryThing Tags and Library of Congress Subject Headings." *Journal of Information Science* 36.6 (2010): 763-79. Web.

Matthews, Brian. "An Evaluation of Enhancing Social Tagging with a Knowledge Organization System." *Aslib Proceedings* 62.4 (2010): 447-65. Web.

McIntyre, D., and D. Archer. "A Corpus-Based Approach to Mind Style." *Journal of Literary Semantics* 39.2 (2010): 167-82. Web.

Michel, J. B., et al. "Quantitative Analysis of Culture using Millions of Digitized Books." *Science* 331.6014 (2011): 176. Web.

Obaseki, T. I. "Automated Indexing: The Key to Information Retrieval in the 21st Century." *Library Philosophy and Practice (e-journal)* (2010): 338. Web.

Osinska, Veslava. "Visual Analysis of Classification Scheme." *Knowledge Organization* 37.4 (2010): 299-306. Web.

Potter, R. G. "Statistical Analysis of Literature: A Retrospective on Computers and the Humanities, 1966–1990." *Computers and the Humanities* 25.6 (1991): 401-29. Web.

Provine, John. "Understanding the Web to Find Short Answers and “something Different”." *Official Google Blog*. Google, 12 May 2010. Web. 10 Nov. 2011. <http://googleblog.blogspot.com/2010/05/understanding-web-to-find-short-answers.html?m=0>.

Qiang, Jin. "Is FAST the Right Direction for a New System of Subject Cataloging and Metadata?" 45.3 (2008): 91. Web.

Rodriguez, M. A., J. Bollen, and H. Van de Sompel. "A Practical Ontology for the Large-Scale Modeling of Scholarly Artifacts and their Usage". *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*. Web.

Romero, Lisa. "An Evaluation of Classification and Subject Cataloging in Entry-Level Cataloging Copy: Implications for Access and Instruction." *Journal of Education for Library and Information Science,* 36.3 (1995): 217-29. Web.

Schallier, Wouter. *Subject Retrieval in OPAC's: A Study of Three Interfaces*. Ed. Amadeu Pons. Web.

Schonfeld, Roger C. and Brian F. Lavoie, “Books without Boundaries: A Brief Tour of the System-wide Print Book Collection.” *Journal of Electronic Publishing* 9.2 (2006) . DOI: http://dx.doi.org/10.3998/3336451.0009.208

Schwartz, Candy. "Thesauri and Facets and Tags, Oh My! A Look at Three Decades in Subject Analysis." 56.4 (2008): 830. Web.

Sekula, Allan. “Reading an archive: photography between labour and capital.” *Visual culture : the reader*. Ed. Jessica Evans & Stuart Hall. London ;Thousand Oaks: SAGE Publications in association with the Open University, 1999. Print.

Smith, Tiffany. *Cataloging and You: Measuring the Efficacy of a Folksonomy for Subject Analysis*. Ed. Joan Lussky. Web.

Strober, Myra H. "Communicating Across the Academic Divide." *Chronicle of Higher Education* Print.

Vries, J. “An automated Indexing System Utilizing Semantic Net Expansion.” *Computers and Biomedical Research* 25.2 (1992) : 153-167. Web. 31 Oct 2011.

Winget, Megan. *User-Defined Classification on the Online Photo Sharing Site Flickr ... Or, how I Learned to Stop Worrying and Love the Million Typing Monkeys*. Ed. Megan Winget. Web.

Yi, Kwan. "A Semantic Similarity Approach to Predicting Library of Congress Subject Headings for Social Tags." *Journal of the American Society for Information Science & Technology* 61.8 (2010): 1658-72. Web.

# Biographies of the Principle Investigators

**Jack Ammerman** is Associate University Librarian for Digital Initiatives and Open Access at Boston University. As Head Librarian at the Boston University Theology Library and as Director of the Library and Information Technology at Hartford Seminary Ammerman developed several digital library projects. He held several positions in libraries at Emory University where he was heavily involved in library automation and systems support. Ammerman received a D.Min. from Princeton Theological Seminary, an Masters of Librarianship from Emory University, and an M.Div. from the Southern Baptist Theological Seminary.

**Dan Benedetti** received his Master's of Library and Information Science from Simmons College in 2000. From 2004-2010 he was a Bibliographer/Librarian, Selector for Philosophy and Religion at Boston University's Mugar Memorial Library. In December, 2010, Dan was appointed to be Head Librarian for the Pickering Educational Resources Library.

**Garth Green** (M.A., Religious Studies [Boston University]; M.A., Philosophy, [University of Leuven (Belgium)]; Ph.D., Philosophy of Religion [Boston University]) was Assistant Professor of Philosophy of Religion in the School of Theology at Boston University until 2011. He teaches Medieval Theology, Modern Philosophy of Religion, and Contemporary Phenomenology. He has held fellowship and research positions at the Institut für die Wissenschaften vom Menschen (Austria), the University of Leuven (Belgium), the Institut Catholique de Paris (France), and the Istituto Italiano per gli Studi Filosofici (Italy). His first book is *The Aporia of Inner Sense: The Self-Knowledge of Reason and the Critique of Metaphysics in Kant* (Brill's Critical Studies in German Idealism, Leiden, 2010). Garth has lectured and taught in both Europe and the United States, and is the author of several articles, in each of his areas of concentration; in medieval neo-Platonic theology, in 19th-century philosophy and philosophy of religion, and in 20th-century phenomenology. In 2011, Garth was appointed Associate Professor of Philosophy of Religion in the Faculty of Religious Studies at McGill University.

**Vika Zafrin** is formerly the Digital Collections Librarian at the School of Theology, Boston University. In 2010, she was appointed as the Institutional Repository Librarian for Boston University Libraries. Vika holds a PhD in Humanities Computing from Brown University. Her research interests include semantic text encoding, use of social media in pedagogy and scholarship, the library as a locus for scholarly collaboration, issues in open access, digitization technologies, and social tagging of scholarly resources. Zafrin worked on two Brown University projects previously funded by the NEH -- The Decameron Web and Virtual Humanities Lab (the latter as Project Director).