Title
The shifting sands of disciplinary development: Analyzing Library and Information Science (LIS) dissertations using Latent Dirichlet Allocation (LDA)

Authors
Cassidy R. Sugimoto* (corresponding author)
School of Library and Information Science
Indiana University Bloomington
1320 E. 10th St., Bloomington, IN 47405-3907
Phone: (812) 856-2323
Fax: (812) 855-6166
Email: sugimoto@indiana.edu

Daifeng Li
School of Library and Information Science
Indiana University Bloomington
1320 E. 10th St., Bloomington, IN 47405-3907
Phone: (812) 855-2018
Fax: (812) 855-6166
Email: ldf3824@yahoo.com.cn

Terrell G. Russell
School of Information and Library Science
University of North Carolina at Chapel Hill
216 Lenoir Drive, 100 Manning Hall, Chapel Hill, NC 27599-3360
Phone: (919) 962-8366
Fax: (919) 962-8071
Email: terrell@terrellrussell.com

Craig Finlay
School of Library and Information Science
Indiana University Bloomington
1320 E. 10th St., Bloomington, IN 47405-3907
Phone: (812) 855-2018
Fax: (812) 855-6166
Email: craigfinlay@gmail.com

Ying Ding
School of Library and Information Science
Indiana University Bloomington
1320 E. 10th St., Bloomington, IN 47405-3907
Phone: (812) 855-2018
Fax: (812) 855-6166
Email: dingying@indiana.edu
The shifting sands of disciplinary development: Analyzing North American Library and Information Science (LIS) dissertations using Latent Dirichlet Allocation (LDA)

Abstract
This work identifies changes in dominant topics in library and information science (LIS) over time, by analyzing the 3,121 doctoral dissertations completed between 1930 and 2009 at North American Library and Information Science programs. The paper utilizes Latent Dirichlet Allocation (LDA) to identify latent topics diachronically and to identify representative dissertations of those topics. The findings indicate that the main topics in LIS have changed substantially from those in the initial time period (1930-1969) to the present (2000-2009). However, some themes occurred in multiple time periods, representing core areas of the field: library history occurred in the first two time periods; citation analysis in the second and third time periods; and information-seeking behavior in the fourth and last time period. Two topics occurred in three of the five time periods: information retrieval and information use. One of the notable changes in the topics was the diminishing use of the word library (and related terms). This has implications for the provision of doctoral education in LIS. This work is compared to other earlier analyses and provides validation for the use of LDA in topic analysis of a discipline.

Introduction
Many evaluations of library and information science (LIS) have been conducted, primarily using the methods of content analysis and co-citation analysis on journal articles (e.g., Järvelin & Vakkari, 1993; White & McCain, 1998). The majority of these focus on a single time period in LIS (e.g., Järvelin & Vakkari, 1990; Kumpulainen, 1991). Subsequent analyses often use different source titles, techniques, or coding schemes, making comparison between the analyses difficult. While these studies constitute one lens on the field, there are some major limitations to
the current literature in the area. First, the focus on a single communicative genre (the journal article) provides a monocular view of the field. Research has shown that the writing and citing patterns of authors vary significantly by genre (Bazerman, 1988; Hyland, 2000). A different topic spectrum may be found by examining topics across multiple genres. Secondly, the focus has been on either a group of highly cited authors or a sample of journal articles. Previous analyses have been manually intensive, necessitating small sample sizes. This has the potential to skew the results in two ways: 1) highly cited works are not necessarily representative of the works produced and 2) a few articles/authors can heavily influence the results. Lastly, the analyses have been largely synchronic, rather than diachronic. Therefore, trend data rely on replication studies, which are not prevalent in the literature.

Understanding the development of the discipline and changes in topics over time is particularly necessary in LIS which has an extended literature of questioning its own disciplinary identity and the relationship of the library science and information science components of the name. Dissertations may serve a critical function in the exploration of disciplinary identity. During the doctoral process, students are acculturated in the ways of the discipline and are taught the central theories, methods and objects of scrutiny within the domain. The dissertation is seen as independent and original research that is meant to set the foundation for the rest of the scholar’s career. It should, therefore, accomplish two goals: situate the research within the domain, and explore new and original territory. Furthermore, although the primary focus of doctoral education is on the production of researchers (Sugimoto, 2010a), a secondary focus is the creation of new faculty. Therefore, the topics explored in doctoral dissertations may have an indirect effect on the education of the next generation of master’s students.
This paper provides a new lens on disciplinary identity, by identifying the main topics in LIS dissertations from 1930-2009. A topic modeling technique, Latent Dirichlet Allocation modeling, is used. This paper has both an historical and methodological objective: 1) to identify the main topics in LIS diachronically and 2) to examine the use of LDA in analyzing disciplinary development and change.

This work addresses gaps in the literature by providing a novel way of exploring disciplinarity, in terms of providing a new genre; an extended time period; and utilizing a new technique which allows for the examination of a large corpus of text data. This work has broad application to those interested in the history of LIS and to those providing and receiving LIS education. From a methodological standpoint, this work holds interest for scientometricians interested in new techniques for evaluating and describing disciplinary histories.

**Literature review**

This section reviews previous quantitative analyses of LIS, quantitative analyses of LIS dissertations, and studies utilizing Latent Dirichlet Allocation.

*Quantitative analyses of LIS*

Many quantitative analyses have been conducted to analyze the domain of LIS: content analyses of journal articles (e.g., Fidel, 2008; Hider & Pimm, 2005; Enger, Quirk, & Stewart, 1989; Järvelin & Vakkari, 1990; Järvelin & Vakkari, 1993; Kumpulainen, 1991), bibliometric analyses of journal articles (e.g., Åström, 2007; Åström, 2010), and bibliometric analyses of authors (e.g., Åström, 2010; Bates, 1998; Budd, 2000; Pettigrew & Nicholls, 1994; Levitt & Thelwall, 2009a; Levitte & Thelwall, 2009b; White & McCain, 1998; White, 2001) to provide large-scale descriptions of the field. Some analyses have focused on particular journals (e.g., Lipetz, 1999; Liu, 2002; Park, 2010; Harter & Hooten, 1999), conference proceedings (e.g.,
Smeaton, et al., 2003), subject areas (e.g., Sugimoto & McCain, 2010), or countries (e.g., Uzun, 2002; Cano, 1999). Scholars have also performed bibliometric analyses to examine the relationship between LIS and other disciplines (e.g., Ellis, Allen, & Wilson, 1999; Borgman & Rice, 1992; Sugimoto, Pratt, & Hauser, 2008; Odell & Gabbard, 2008; Meyer & Spencer, 1996).

The majority of quantitative analyses of LIS share four things: 1) the journal article is the focal communicative genre; 2) they are synchronic, rather than diachronic; 3) they focus on relationships between journals and/or journal authors (rather than topic analysis); and 4) those focusing on topic analysis use methods of co-occurrence or content analysis. Some notable exceptions (on at least one of the four points) are the works by Smeaton et al. (2003) and Sugimoto and McCain (2010), which looked at changes in topics in information retrieval over time; Harter and Hooten’s (1992) bibliometric study of the Journal of the American Society for Information Science & Technology articles for three time slices; Åström’s (2007) co-citation analysis of LIS journal articles for three time periods; and Järvelin and Vakkari’s (1993) content analysis of journal articles for three time periods.

While synchronic studies provide a baseline for comparison, too few studies have replicated the methods, making comparison difficult. In addition, many of the studies assess topicality by means of secondary interpretative analysis, based on the dominant author or journal clusters identified. This can be problematic particularly for studies that only investigate the most-cited documents or authors in a field—production and impact are not necessarily equivalent. Lastly, there can be large differences in rankings of authors, sources, and topics based on the communicative genre under investigation (Sugimoto, 2010b). Therefore, further work is necessary that evaluates additional communicative genres, provides analyses across multiple
time periods for systematic comparison, and focuses specifically on changes in topics across those periods.

Some work has been done to examine the value of various methods of topic analysis, comparing the results found through co-citation with co-word analysis (e.g., Braam, Moed, & van Raan, 1991a; 1991b) and the difference between using titles, author-supplied, or indexer-supplied keywords (Whittaker, Courtial, & Law, 1989). Authors have noted limitations with each of these, including the “indexer effect” and the aging effect of keywords and the issue of non-standard titles (Whittaker, Courtial, & Law, 1989). Scholars have also criticized the use of co-occurrence analysis of terms, noting the large variance in meanings of individual terms based on the level of textual aggregation under investigation (Leydesdorff, 1997). Therefore, this work seeks to examine new methods that are not restricted to keywords and do not use strict co-occurrence analysis.

**Quantitative analyses of LIS dissertations**

Houser (1982) suggested that evaluations of dissertations may: “provide indications of an attempt to define the concepts and terms central to the discipline…articulate the range of problems believed to be important for an explanation of those concepts, and…suggest a move away from intellectual confusion and towards the consensus which any field requires for its identity and status” (p. 195). To this end, a number of evaluations of the doctorate and dissertations in LIS have been conducted. These have focused on the admissions criteria, demographics, and post-career trajectories of students (e.g., Bobinski, 1986; Whitbeck, 1991a; Whitbeck, 1991b (for a review of pre-1987 studies, see Abrera, 1987)). At the discipline level, Sugimoto, Russell and Grant (2009) studied the number of degrees awarded in LIS schools from 1930 to 2008. Bibliometric analyses of dissertation references have been conducted (Gao, Yu,
Luo, 2009; Sugimoto, 2010b; Buttlar, 1999; Keat & Kiran, 2008), in order to identify the top sources, authors, and subject areas cited.

Few studies have analyzed topics in LIS dissertations over time. Franklin and Jaeger (2007) classified the topics of all LIS dissertations written by African-American women between 1993 and 2003 into five areas: education, information issues, library/librarianship issues, literature, and technology. Danton (1959) classified the 129 dissertations completed between 1930 and 1959. The most frequent topics were background (including historical studies), resources, organization/administration, and technical processes/documentation. Schlater and Thomison (1974; 1982) conducted two surveys of dissertations (1925-1972 and 1973-1981) that evaluated the types of methods used in LIS dissertations. Their work showed that surveys, historical analysis, and citation/content analysis dominated the first time period while surveys, historical analysis and operations research dominated the second time period. This systematic study over time was one of the only studies to look at topic areas in LIS dissertations over time. However, the manual intensity of content analysis becomes difficult as the number of dissertations in the discipline has gone from a few hundred to a few thousand. While content analysis can provide high granularity for individual works, it becomes difficult to assess the entire body of work without automatic techniques. Therefore, this work applies Latent Dirichlet Allocation modeling to examine dominant topics in LIS diachronically.

Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) was proposed by Blei, Ng, and Jordan (2003) as a generative probabilistic model useful for discovering underlying topics in collections of data. Their initial experimentation focused on textual data, although they acknowledged the possibility of its application to other media. Applications to date have included detection in image data.
detection in video (e.g., Niebles, Wang, & Fei-Fei, 2008; Wang, Ma, & Grimson, 2009), automatic essay grading (Kakkonen, Myller, Sutinen, & Timonen, 2008), prediction of relationships between proteins (Aso & Eguchi, 2009), and fraud detection (Xing & Girolami, 2007). Expansions of LDA have also been used to understand correlations between topics (Blei & Lafferty, 2007), authors (Rosen-Zvi, Griffiths, Steyvers, & Smith, 2004), academic networks (Tang, Jin, & Zhang, 2008), social networks (McCallum, Wang, Corrada-Emmanuel, 2007), social tagging communities (Li, et al., 2010), and changes in topic over time (Pruteanu-Malinici, et al., 2010; Rzeszutek, Androutsos, & Kayn, 2010).

Testing and validation of the model and model expansions have utilized collections of scientific documents (e.g., Blei, Franks, Jordan, & Mian, 2006); however, there have been relatively few studies focused predominately on understanding the underlying topical structure of a field or domain (as opposed to focusing on developing or expanding the method). Two exceptions to this are Griffiths and Steyvers’ (2004) analysis of abstracts from the Proceedings of the National Academy of Science (PNAS) from 1991-2001 and Zheng, et al.’s (2006) analysis of the bioinformatics literature (from MEDLINE abstracts). These studies found that LDA performed well, in that the latent structure mimicked some characteristics of explicit structure (such as categorization schemes). Moreover, the studies displayed the ability of LDA to analyze the rich underlying structures of the domain—depicting emerging and sustained trends in a given discourse. Therefore, the present work utilizes LDA to evaluate the discourse of the discipline of LIS, as presented in dissertation abstracts, from 1930 to 2009.

Methods
This section describes the MPACT Project, reviews the process of identifying titles and abstracts, and describes the process of applying LDA.

**MPACT Project**

The MPACT database ([http://www.ibiblio.org/mpact/](http://www.ibiblio.org/mpact/)) was used as the sampling frame for identifying LIS dissertations for the study. The MPACT database is an academic genealogy project which creates a network of dissertations connected by means of disciplines, schools, and individuals (as authors, advisors, or committee members of dissertations). The MPACT database currently contains all dissertations received between 1930 and 2009 from schools with ALA-accredited programs (North American schools which contain a master’s degree program in library science accredited by the American Library Association). A complete entry in the MPACT database contains: author, school, year, discipline, title, abstract, advisor(s), and committee members for each dissertation. (For more information about the creation of the MPACT database and previous analyses of the data, see Marchionini et al. (2007); Sugimoto et al. (2008); Sugimoto, Russell, and Grant (2009); and Russell and Sugimoto (2009).) All 3,121 dissertations completed between 1930 and 2009 at schools with ALA-accredited program were included in this analysis.

**Titles and abstracts**

At the initiation of this project, MPACT only contained partial title and abstract information. Therefore, systematic collection of the remaining titles and abstracts was necessary. Titles for nearly all of the dissertations were found using either ProQuest’s *Dissertation and Theses* database or OCLC’s *WorldCat*. In some instances, contact with the author was necessary. Some of the most recent abstracts were available in ProQuest’s main entry for the item. Others were available using the Preview or full-text version of the documents. Most of the ProQuest versions
were image files, so the text had to either be retyped manually or converted using optical character recognition (OCR). For those dissertations not available online, a copy of the dissertation (either in monograph or microfiche form) had to be interlibrary loaned. The abstracts were then scanned, OCRed, and cleaned. (The large amount of dust and other imperfections in the microfiche made it particularly hard to OCR. In many cases, it was faster to retype the abstract by hand than to clean the OCRed copy.) In a few cases, the physical copy of the dissertation could not be obtained. In these cases, an attempt was made to contact the author of the dissertation in order to obtain the abstract. For a small number of cases, the dissertation could not be obtained (and therefore, no abstract could be examined) (Table 1). For many pre-1980 dissertations, the physical copy of the dissertation was obtained, but the dissertation did not contain an abstract. Table 1 describes the number and percent of dissertations with abstracts for each time period. The time slices are separated by decade with the exception of the first time period, which did not have a sufficient number of dissertations in each decade for the analysis. Therefore, the first time period is longer than subsequent time periods.

<table>
<thead>
<tr>
<th>Table 1. Number of abstracts located for each time period</th>
</tr>
</thead>
<tbody>
<tr>
<td>----------------------------------------</td>
</tr>
<tr>
<td># of dissertations with abstracts</td>
</tr>
<tr>
<td>(18.6%)</td>
</tr>
<tr>
<td># of dissertations without abstracts</td>
</tr>
<tr>
<td>(75.3%)</td>
</tr>
<tr>
<td># unknown</td>
</tr>
<tr>
<td>(6.1%)</td>
</tr>
<tr>
<td>Number of dissertations</td>
</tr>
</tbody>
</table>

All collected titles and abstracts were entered into the MPACT database. They were then exported and analyzed according to the process described below.

*Latent Dirichlet Allocation (LDA)*

LDA was developed as a topic modeling technique, particularly useful for “classification, novelty detection, summarization, and similarity and relevance judgments” of large text corpora
The main underlying assumption of this approach is the assumption of exchangeability for both words and documents. That is, the model assumes that there is no inherent function of the order of the words or the ordering of the documents in a given set. This leads to what is called “bag-of-words” modeling, under which a single word (unigram) is assumed to be conditionally independent and identically distributed (p. 995). In LDA, a topic is characterized by a distribution over words and then documents are represented as random mixtures over latent topics (p. 996). As a three-level hierarchical Bayesian model, each topic node is sampled repeatedly. The end result is that words may be repeated within topics and documents may be associated with more than one topic. The richness of this model is that it does not singularly classify the “aboutness” of a document, but recognizes that a document can contain many topics; and that a word does not necessarily imply a single topic, but can imply different topics based on combinations with other words.

This model was extended to what is called the author-topic model (Rosen-Zvi, Griffiths, Steyvers, & Smyth, 2004), which is used in the present analysis. In this model, not only is each document a mixture of probabilistic topics, but each author is also seen as a mixture of probabilistic topics. In the same way that a topic can be generated from multiple topics, this extended model recognizes that an author can be “about” multiple topics (within a single document or across documents). In the present study, each author is associated with a single dissertation; however, the surrogates for that dissertation are in two units: the title and the abstract. The author-topic model allows us to not only examine which topics were most salient across the various time periods, but also which authors are most associated with these topics. The author-topic model can be visualized in Figure 1.
The figure can be explained as follows:

1. $\Theta$ is the probability of a topic given an author $x$; $\alpha$ is a hyperparameter for $\Theta$.
2. $\phi$ is the probability of a word $w$ given a topic $z$; $\beta$ is a hyperparameter for $\phi$.
3. $a_d$ provides for the fact that multiple authors can write a single document; $x$ is a randomly selected author from $a_d$. [Note that there are only single authors in this selection; however, it is still necessary to identify author $x$.]
4. Given author $x$, we identify the topic $z$ most likely to be associated with the given author.
5. Given topic $z$, we identify the words $w$ most likely to be associated with the given topic.

Given the Bayesian network created from Figure 1, the joint probability of the author-topic pair was estimated using Gibbs sampling algorithm (Casella, 2001). This algorithm allows for an estimation of the unknown parameters of the model, namely, the probability of a topic given an author and the probability of a word given a topic. Gibbs sampling uses a successive Markov chain sampling to repeatedly draw $x$ and $z$ (as a pair), conditioned on all other variables. The process can be expressed as follows:

$$P(z_i, x_i | z_{-i}, x_{-i}, w, a_d, \alpha, \beta) \propto \frac{n w[m][j]}{n k w s u m[j] + V \times \beta} \times \frac{n a[x][j]}{n a k s u m[x] + T \times \alpha}$$

where $n w[m][j]$ is the number of times a single word is assigned to a topic; $n k w s u m[j]$ is the number of times any word is assigned to a topic; $n a[x][j]$ is the number of times a single author is assigned to a topic; $n a k s u m[x]$ is the number of times any topic is assigned to an author.

Once this process has been iterated 2,000 times, the results of these variables are used to calculate $\Theta$ and $\phi$. The formula appears below:
A simplified version of the code is presented in Appendix B. The actual code was a modified version of Tang, Jin, and Zhang’s (2008) ACT code, with the extraction of the conference portion. This formula is used to identify the most salient topics across time and the authors most likely to be associated with those topics.

However, before applying this model, one must first identify how many topics are appropriate for the data. This is done by selecting a sample test data set and running the program on an increasing number of topics. Perplexity analysis is used to estimate the performance of the model. This analysis is used when we have an unknown probability distribution in the data. A lower perplexity value indicates better performance. As shown in Figure 2, the performance for our data stabilized at 50 topics. This indicates that, for the data given, the identification of 50 topics for each time period will yield the most accurate results. That means, any less than 50 topics would be aggregating singular topics into super-topics; any more than 50 would be fragmenting topics into sub-topics. The identification of 50 topics will yield topics that are internally coherent and consistent.

\[
\theta_{xj} = \frac{na[x][j] + \alpha}{naksum[x] + T \times \alpha}
\]

\[
\phi_{mj} = \frac{nw[m][j] + \beta}{nkwsum[j] + V \times \beta}
\]
At this point, all data were analyzed according to the author-topic model, divided into the five time slices. For each time period, 50 topics were identified. Each topic contained a probability value—that is, the likelihood that the topic identified should be associated with the time period. These topics were ranked by probability values and the top five were selected as being most representative of the time period. Similarly, a probability for each word was calculated to represent the association between a word and the given topic. These were ranked by probability values and the top twenty were chosen as most representative of the topic. Lastly, the authors were assigned probability values for each topic and these too were ranked. The top five were chosen as highly representative authors for the given topic. These results are presented and discussed below.

To facilitate interpretation of the topic results, the representative authors were searched in the MPACT database and the year, title, abstract, and school for each was analyzed. This information was used to label and interpret the topics. The authors, years, titles, and schools are presented in the Appendices C-G. The abstracts can be found in the MPACT database.

**Results**

For each time period, a table is provided summarizing the LDA results. The topics, \( a \) through \( e \) for each time period, are organized in descending order according to their probability values (\( a \) having the highest probability value). Similarly, the top five authors and top twenty words are listed, ranked by probability value. A short description follows the table, interpreting and labeling the topics.

### 1930-1969:

Table 2 presents the LDA results for 1930-1969.

<table>
<thead>
<tr>
<th>Labels</th>
<th>Topic 1a</th>
<th>Topic 1b</th>
<th>Topic 1c</th>
<th>Topic 1d</th>
<th>Topic 1e</th>
</tr>
</thead>
<tbody>
<tr>
<td>Library</td>
<td>Library</td>
<td>Librarianship/education</td>
<td>Librarianship/recruitment</td>
<td>Subject</td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Extended LDA results for 1930-1969 (295 dissertations)
Evidence from high-loading keywords and most representative dissertations (see Appendix C) reveal that Topic 1a is about the history of American libraries and librarianship, with a focus on collections. Topic 1b is also about library history, but with an emphasis on regional public libraries and library service. High-loading dissertations in Topic 1c show a focus on training and education for librarians. Topic 1d also displays a focus on education and recruitment in librarianship, but displays a secondary emphasis on the history of academic libraries. The representative dissertations from Topic 1e share an emphasis on the analysis of subject headings.

1970-1979

Table 3 presents the LDA results for 1970-1979.

<table>
<thead>
<tr>
<th>Authors</th>
<th>Authors</th>
<th>Words</th>
<th>Words</th>
<th>Words</th>
<th>Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Claude Sparks</td>
<td>Mary Anders</td>
<td>library</td>
<td>library</td>
<td>library</td>
<td>library</td>
</tr>
<tr>
<td>Irene Braden</td>
<td>George</td>
<td>public</td>
<td>study</td>
<td>students</td>
<td>cents</td>
</tr>
<tr>
<td>Winifred</td>
<td>Bobinski</td>
<td>libraries</td>
<td>analysis</td>
<td>undergraduate</td>
<td>list</td>
</tr>
<tr>
<td>Linderman</td>
<td>Violet</td>
<td>state</td>
<td>knowledge</td>
<td>reading</td>
<td>subject</td>
</tr>
<tr>
<td>Russell Bidlack</td>
<td>Coughlin</td>
<td>service</td>
<td>research</td>
<td>science</td>
<td>library</td>
</tr>
<tr>
<td>Lucy Maddox</td>
<td>John Larsen</td>
<td>development</td>
<td>libraries</td>
<td>programs</td>
<td>form</td>
</tr>
<tr>
<td></td>
<td>Irving</td>
<td>librarians</td>
<td>activities</td>
<td>study</td>
<td>congress</td>
</tr>
<tr>
<td></td>
<td>Verschoor</td>
<td>factors</td>
<td>data</td>
<td>patterns</td>
<td>subdivisions</td>
</tr>
<tr>
<td></td>
<td></td>
<td>government</td>
<td>positions</td>
<td>program</td>
<td>words</td>
</tr>
<tr>
<td></td>
<td></td>
<td>southeast</td>
<td>areas</td>
<td>school</td>
<td>references</td>
</tr>
<tr>
<td></td>
<td></td>
<td>population</td>
<td>reference</td>
<td>characteristics</td>
<td>study</td>
</tr>
<tr>
<td></td>
<td></td>
<td>growth</td>
<td>system</td>
<td>objectives</td>
<td>found</td>
</tr>
<tr>
<td></td>
<td></td>
<td>states</td>
<td>significant</td>
<td>student</td>
<td>word</td>
</tr>
<tr>
<td></td>
<td></td>
<td>region</td>
<td>performed</td>
<td>graduate</td>
<td>heading</td>
</tr>
<tr>
<td></td>
<td></td>
<td>corporation</td>
<td>core</td>
<td>academically</td>
<td>forms</td>
</tr>
<tr>
<td></td>
<td></td>
<td>social</td>
<td>librarians</td>
<td>amount</td>
<td>entry</td>
</tr>
<tr>
<td></td>
<td></td>
<td>interest</td>
<td>education</td>
<td>institution</td>
<td>number</td>
</tr>
<tr>
<td></td>
<td></td>
<td>larger</td>
<td>ballyhoo</td>
<td>free</td>
<td>punctuation</td>
</tr>
<tr>
<td></td>
<td></td>
<td>training</td>
<td>humanists</td>
<td>personal</td>
<td>based</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>substantially</td>
<td>academic</td>
<td>employed</td>
</tr>
</tbody>
</table>

Table 3. Extended LDA results for 1970-1979 (615 dissertations)
Information use in multiple contexts dominates Topic 2a. Issues of race are also prominent in the highest-loading dissertations (see Appendix D). Topic 2b indicates an interest in library history, with an emphasis on the social function of libraries. Topic 2c indicates a focus on citation analysis, with a focus on journal and document analysis. Dissertations and keywords for Topic 2d indicate a focus on information retrieval, with a secondary focus on indexing. Topic 2e is focused on education, particularly as relates to elementary and secondary school libraries.

1980-1989

Table 4 presents the LDA results for 1980-1989.

Table 4. Extended LDA results for 1980-1989 (707 dissertations)

<table>
<thead>
<tr>
<th>Topic 3a</th>
<th>Topic 3b</th>
<th>Topic 3c</th>
<th>Topic 3d</th>
<th>Topic 3e</th>
</tr>
</thead>
<tbody>
<tr>
<td>Labels</td>
<td>Regional libraries</td>
<td>Librarianship/permissions</td>
<td>Citation analysis</td>
<td>Library evaluation</td>
</tr>
<tr>
<td>Authors</td>
<td>Abdulla Saleh Isa</td>
<td>David Kim</td>
<td>Stephen Lawani</td>
<td>Ida Girvin Adams</td>
</tr>
<tr>
<td></td>
<td>Hisham A. Abbas</td>
<td>Blondell Strong</td>
<td>Shahrokh Afsharpanah</td>
<td>Suzy Queiroz</td>
</tr>
<tr>
<td></td>
<td>Robert Wagenknecht</td>
<td>Alixe Hambleton</td>
<td>Elizabeth Futas</td>
<td>Frances Casserly</td>
</tr>
<tr>
<td></td>
<td>Ruth C. Henry</td>
<td>Jo Anne Hall</td>
<td>Catalina Y. Diulvio</td>
<td>Robert H. Burger</td>
</tr>
<tr>
<td></td>
<td>Carole Taylor</td>
<td>Thein Swe</td>
<td>Michael Afolabi</td>
<td>Kathleen Haynes</td>
</tr>
</tbody>
</table>

1980-1989
Keywords for Topic 3a reveal an interest in both public and academic libraries. A review of the representative dissertations (see Appendix E) expose a shared concern for libraries in specific locations. Topic 3b is dominated by words reflecting quantitative and experimental research.

The titles of the dissertations reflect a shared interest in the profession of librarianship, with a focus on role definition and the changing place in society’s perception of librarianship. Topic 3c is connected by a shared method: all studies in this topic are bibliometric studies.

Methodological words also dominate Topic 3d, with an emphasis on library evaluation. Topic 3e is focused on information use and communication, with a particular emphasis on science communication.

1990-1999

Table 5 presents the LDA results for 1990-1999.

<table>
<thead>
<tr>
<th>Table 5. Extended LDA results for 1990-1999 (741 dissertations)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Topic 4a</strong></td>
</tr>
<tr>
<td><strong>Labels</strong></td>
</tr>
<tr>
<td><strong>Authors</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>
Keywords and dissertations (see Appendix F) in Topic 4a is about model and framework development, with an emphasis on methods. Representative dissertations in Topic 4b suggest an interest in networking and outreach for libraries. The keywords imply a focus on academic libraries. Topic 4c reveals an interest in information-seeking behavior, with an emphasis on professionals. Keywords for Topic 4d indicate highly quantitative and experimental methods. Dissertation topics revolve around human resource and management issues. Topic 4e indicates an interesting split between the keywords and representative dissertations: the keywords would reveal a focused information retrieval topic. However, the dissertations all share a focus on telecommunication and media.

2000-2009

Table 6 presents the LDA results for 2000-2009.

<table>
<thead>
<tr>
<th>Labels</th>
<th>Information use</th>
<th>Internet</th>
<th>Information seeking behavior</th>
<th>Information retrieval</th>
<th>Information retrieval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Authors</td>
<td>Melissa H. Cragin, Minjuan Wang, Sarah Zach</td>
<td>Thomas King, Mohammed Allehaibi, Yunkeum Kim</td>
<td>Pamela McKenzie, Sarah Zach, James David Currier</td>
<td>Joel Uzi Eden, Wooseob Jeong, Boryung Ju</td>
<td>Abdullah Al-Shehri, Dana Sensuse, Brook Wu</td>
</tr>
</tbody>
</table>
Dissertations in Topic 5a indicate a focus on knowledge dissemination and use. The keywords imply methods such as interviews, case studies, and content analysis. Dissertation topics in Topic 5b focus primarily on the Internet, but also include some dissertations (see Appendix G) about the library profession. The keywords imply a focus on quantitative analysis. Topic 5c is firmly rooted around information-seeking behavior for both representative dissertations and topics. The focus of Topic 5d is on information retrieval, with a user-centered focus. Topic 5e is also on information retrieval, but from a system and classification/indexing focus.

**Discussion**

The results of the topic analyses are summarized in Figure 3, with the topics ranked from highest (top) to lowest for each time period.
Some topics occur across multiple time periods, including: library history, librarianship, information use, citation analysis, classification, information retrieval (abbreviated as IR in Figure 3), and information-seeking behavior. These are the core topics in LIS dissertations from 1930 to 2009. However, there are changes within these topics. Library history occurs only in the first two periods and is focused on collections and services in 1930-1969 and social functions in 1970-1979. Dissertations which are focused on the profession of librarianship tended to discuss education and recruitment in 1930-1969, education and school libraries in 1970-1979, evaluation and the perception of librarianship in 1980-1989, management in 1990-1999, and the impact of the Internet in 2000-2009. In 1970-1979, issues of race were central for the topic in information use; scientific use was the focus of 1980-1989, while dissemination was the focus of 2000-2009. Classification appears in the first time period with a focus on subject headings and reappears in the last time period as an information retrieval category. Information retrieval also
changes over time, from a focus on indexing in the 1970s, to media in the 1990s, and then as two distinct topics in the 2000s: one a user-centered and the other a classification topic.

Limitations

One limitation of the study was identifying a sufficient number of abstracts for the first time period. The issue was not only in obtaining the dissertations, but the actual presence of an abstract in the dissertation (these were relatively uncommon for pre-1970 dissertations (See Table 1). This was complicated by the fact that the inclusion of abstracts varied from school to school. Therefore, schools which encouraged/mandated abstracts (e.g., Columbia) showed more prominently in the data than those schools which did not (e.g., Chicago). Future analysis might attempt to examine the full-text of the dissertations, rather than just the titles and the abstracts for a more comprehensive picture of the field. In addition, the schools at which the dissertations were conferred were only within North America, producing an additional bias.

A limitation inherent with LDA analysis is in the manual interpretation and labeling of “topics”. While some topics were fairly straight-forward to label (e.g., Topic 5c, the top three loading words of which were (1) information, (2) seeking, (3) behavior), others proved more difficult to ascertain the content or methodological relationship that connected the words and dissertations.

Comparison with earlier works

The LDA analysis of the first time period corresponds well to Danton’s (1959) study of LIS dissertations from 1930 to 1959. Danton’s top five categories for the time period were (in order): background, organization/administration, resources, reader services, and technical processes/documentation. Danton considered background to include words on history, which was found in this study to be the most dominant topic for that time period. In addition, two topic
areas were concerned with professionalism, which Danton classified as organization/administration. The topic of resources and services equates with the collection and service components of the history topics. Lastly, the technical processes/documentation equates exactly with the 5th topic of subject headings. The similarity between the LDA results and Danton’s manual examination of each of the dissertations provides validation for the utilization of LDA for identifying topics across a discipline.

Kumpulainen (1975) evaluated LIS journal articles in 1975 according to topic and methods. The dominant topics were information storage and retrieval (with a primary focus on classification and indexing), library and information service activities, education in LIS, scientific and professional communication (which included citation analysis), and information-seeking (which included information use). These findings are consistent with the present study, although the focus on service is more historical in the present study.

Järvelin and Vakkari (1990) analyzed journal articles in 1985 with the same coding scheme as Kumpulainen (1975) (the latter based on the former). The top five topics identified were: information storage and retrieval, library and information service activities, scientific and professional communication, information-seeking, and the profession. These topics, broadly construed, match with the top five topics in this analysis.

The top three topics by time period identified in Järvelin and Vakkari’s (1993) content analysis of LIS journal articles is displayed in Table 7.

<table>
<thead>
<tr>
<th></th>
<th>1965</th>
<th>1975</th>
<th>1985</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Classification and indexing</td>
<td>Classification and indexing</td>
<td>Information retrieval</td>
</tr>
<tr>
<td></td>
<td>Automation</td>
<td>Information retrieval</td>
<td>Collections</td>
</tr>
<tr>
<td></td>
<td>Methodology</td>
<td>Education in LIS</td>
<td>Professions</td>
</tr>
</tbody>
</table>
These topics share some similarity with the present findings: classification is certainly a top theme in the first time period, but there is an overwhelming emphasis on history in this time period that is not represented by Järvelin and Vakkari’s results. The top three topics from 1975 explain the 4th and 5th topics in the 1970s in this analysis. However, Järvelin and Vakkari’s findings do not show dominant themes in information use, history, or citation analysis. The 1980s are the only time period in the present analysis which do not have IR as a main topic. However, it is the main topic in Järvelin and Vakkari’s analysis. In both analyses there is an emphasis on the library profession, collections and evaluations. The discrepancies between these findings lead to questions relative to genre and the possible impact of genre upon the topics. Further analysis should explore whether certain topics appear first in particular communicative genres and whether genres consistently emphasize different topics within a single field.

Additional differences can be seen when using citation rather than content analysis. White & McCain’s (1998) study of the 120 most highly cited information science researchers from 1972 to 1995 found the following specialties, in order of decreasing variance explained: experimental retrieval, citation analysis, online retrieval, bibliometrics, general library systems, science communication, user theory, OPACs, imported ideas, indexing theory, citation theory, and communication theory. These factors combined explained 84% of the total variance. As can be seen, theory is one of the lowest loading of the factors. In terms of the highest loading specialties, this work confirmed an interest in information retrieval, bibliometrics, and information use consistent with the findings of White and McCain. However, the current analysis places a stronger emphasis on library evaluation, management, and education than was the case with White and McCain’s study. While White and McCain did not undertake a detailed
analysis on the changes in topics over time, they provided evidence for a shift toward the cognitive and user side of information retrieval. This is also reflected in the present work.

Åström (2007) similarly evaluated LIS using co-citation analysis. He identified the following main clusters by time period (Table 8):

<table>
<thead>
<tr>
<th>Table 8. Summary of the Åström’s (2007) findings by time period</th>
</tr>
</thead>
<tbody>
<tr>
<td>------------------</td>
</tr>
<tr>
<td>Information-seeking and retrieval</td>
</tr>
<tr>
<td>Informetrics/Research Collaboration</td>
</tr>
<tr>
<td>Information retrieval/Information search</td>
</tr>
<tr>
<td>Bibliometric mapping</td>
</tr>
<tr>
<td>Information-seeking and retrieval</td>
</tr>
<tr>
<td>Informetrics</td>
</tr>
<tr>
<td>Information Technology</td>
</tr>
<tr>
<td>Experimental information retrieval</td>
</tr>
<tr>
<td>Bibliometrics/Research performance</td>
</tr>
<tr>
<td>Webometrics</td>
</tr>
<tr>
<td>Information-seeking and retrieval</td>
</tr>
<tr>
<td>Children’s information behavior</td>
</tr>
<tr>
<td>Health informatics</td>
</tr>
</tbody>
</table>

Åström’s (2007) clusters vary considerably from those found in the present study, particularly in the focus on bibliometrics over time. This may demonstrate the effects of genre, source titles chosen, and the use of citations rather than content analysis in identifying dominant topics in the field. The current study may provide a complementary lens to previous studies on LIS by exploring a new form of analysis (LDA), a longer time period than previously explored, and a new genre (dissertations).

Dissertation as a communicative genre

One difficulty in comparing the results of this study with previous studies is the difference between communicative genres. Citing behaviors between communicative genres vary considerably (e.g., Cronin, Snyder, & Atkins, 1997; Sugimoto, 2010) and one may expect that the topics covered by disciplines may vary as well. For example, numerous studies summarized above noted the lack of theoretical work published in LIS literature. It may be the case the journal articles are more likely to cover experimental work, while dissertations tend toward the theoretical. In addition, not all those who complete a dissertation continue on to become active researchers in the field—therefore, large differences may be seen in the topic distribution among
dissertations, in comparison to journal articles. Furthermore, analyses of journal articles often take into consideration the most cited or most productive authors. This inevitably skews the results towards those subdomains that are both productive and highly cited and may not accurately display the range of interests presented in the literature.

An additional factor that differentiates analyses of dissertations is the filtering of eligible authors—authors must have received a doctoral degree within the field (and, for the purpose of this study, within North America). However, publishing in an LIS journal does not carry such a limitation—individuals may contribute to the literature of the field without necessarily being “trained” in the discipline. This factor could be seen from a positive or negative vantage point: dissertation may more accurately portray the interests of the discipline, as they represent those who are being trained in disciplinary norms and behaviors; alternatively, one may conceive of the field more broadly, allowing contributions from any member to constitute the values and interests of the discipline. Future work should be done to examine this, by applying the same methods to multiple communicative genres in a single discipline.

Library Science in LIS

There has been continuing concern in the literature over the place of librarianship and libraries in LIS doctoral dissertations. Danton (1959) noted that “many, and very likely most, dissertations, highly specialized and often theoretical in nature, are of a kind which hold no interest for the librarian ‘in the field’ and have no direct impact upon the work-a-day library world” (p. 449). This concern has been exacerbated in recent years by the dropping of the word library from the school labels and the emergence of the iSchool caucus (this development is chronicled in Olson and Grudin (2009)). Some argue that the nomenclature is not important—
that information science is an umbrella term that includes library science and library-related
topics will still be valued and taught within the iSchools. However, the implications for doctoral
programs are rarely discussed. As shown by the latest Association for Library and Information
Science Education (ALISE) statistical report, the majority of faculty members in LIS hold
doctorates in LIS. Therefore, it may be necessary to identify the relationship between what is
being studied at the doctoral level and what is being taught at the master’s level.

While the data produced cannot directly address the degree to which dissertations are or are
not dealing with issues of library science, an examination of the top-loading words is surely
illustrative. Table 9 displays the top-loading word for each of the five topics in each time period.
For those words that do not begin with librar*\(^1\), the rank of the next top ranked word (within the
top 20) that begins with librar* is noted in brackets. For example, the highest-loading word for
the third topic in 1930-1969 is information, but the sixth highest-loading word is libraries. For
those topics where no librar* word appears in the top 20 words, [no] is written.

| Table 9. Top word for each of the five topics for each time period and rank of librar* words |
|-----------------------------------|-----------------------------------|-----------------------------------|-----------------------------------|-----------------------------------|-----------------------------------|
| Topic a | Library | Library | Library | Research [no] | Study [no] |
| Topic b | Library | Library | Library | Librarians | Library |

As can be seen, there is a steadily diminishing use of librar* words in the top topics over
time. In the first time period, librar* appears in all of the five topics. In 2000-2009, librar* does
not appear in any of the top 20 words for the 5 topics (although librar* appears in some of the
titles). This is depicted in Figure 4, which displays the total number of occurrences of librar* in
all 100 words of the highest-loading topics for each year.

\(^1\) The asterisk denotes that it includes any derivative word that begins with librar, such as library, libraries,
librarian, librarians, etc.
Figure 4. Number of occurrences of librar* in the 100 words of the highest-loading topics by year

The sampling frame for this study included all ALA-accredited schools in the U.S.: schools at which the MLS program is often the bedrock. These programs must conform to ALA standards and guidelines in order to receive accreditation. However, the findings suggest that the bulk of dissertations completed at these schools do not have explicit connection to library practice.

As doctoral programs create the faculty of the future, the apparent disjoint between the professional programs and the research programs within LIS schools may need further examination. Upon closer examination, we may find that the context is still libraries, but the vocabulary has changed. It may also be the case that the dissertations may not directly address librarianship, but nonetheless have implications for it. For example, the theoretical or experimental studies may have direct application for practicing librarians. Alternatively, we may find that there is no explicit relationship between the content of the two programs. Further research needs to examine the doctorate and how it reflects efforts to prepare future faculty of MLS programs.

Conclusion
A number of scholars have explored the degree to which LIS meets the criteria of disciplinarity (idiosyncratically applied), whether Library Science and Information Science should be conjoined, and how these areas should be defined (e.g., Houser 1982; Cronin, 1995; Cronin, 2002; Zins, 2007a, 2007b; Bawden, 2008; Robinson, 2009). Houser (1982a) states that a “discipline is formed to solve a range of problems about some natural or social phenomena…[t]hese problems have a genealogy, that is, a continuity which forms the domain of the discipline” (p. 97). Criteria for quantitatively defining and operationalizing disciplinarity often resolve around institutional structures (Palmer & Cragin, 2008), with many scholars explicitly noting the presence of doctoral programs to be a cornerstone of disciplinarity (e.g., Abbott, 1999; Klein, 1990; Abbott, 2001; Price, 1963; Golinski, 2005). Turner’s (2000) definition of a discipline focused on the identity of a shared name for a specialization and the exchange of scholars within that discipline (thereby propagating the discipline). In this way, dissertations may be more emblematic of a discipline, while journal articles and other forms of communicative genres may be more representative of the domain or field. In examining dissertations, this paper was able to identify some dominant themes in LIS: these can be broadly defined as information-seeking, use, access, organization, and retrieval; and the education and training of the professionals providing these services.

The data presented here provide a foundation for examining changes over time in the field. LDA have proved itself to be a useful tool for identifying the latent themes in the corpus. Some topics were organized around similar objects of study; other topics were organized around methodological approach. This analysis served as an accurate way to provide an overview of the field using all dissertations from all years. The results of this work can serve as a platform for
future, more granular, examinations of the field. Additionally, the technique can be used across other fields to provide large-scale overview of topical changes over time.

Acknowledgements

The authors would like to thank Siqi Xiao, Ariela Bernstein, Brenna Bychowski, and all other members of the MPACT team for their contributions to data collection. We would like to thank Blaise Cronin for edits and discussions on the work. We would like to thank the anonymous reviewers for their thoughtful suggestions for improving the piece. We would also like to thank Jie Tang (Tsinghua University, China) for sharing his ACT code with us. Funding for Daifeng Li was provided by the China National Natural Science Foundation (70971083), Graduate Innovation Fund of Shanghai University of Finance and Economics (CXJJ-2008-330), China 211 project fund (211-5-1), and the 2009 Doctoral Education Fund of Ministry of Education in China (20090007811001).
Cited References


*Library & Information Science Research, 30*, 265-272.


*Journal of the American Society for Information Science, 43*(9), 583-593.


