

# Conceptualizing Large-Scale Information Access Efforts: The Case for Historical Context

Elisabeth A. Jones  
The Information School  
University of Washington  
Seattle, WA 98195  
ej6@uw.edu

## ABSTRACT

Large-scale digitization initiatives (LSDIs) like Google Book Search and the Open Content Alliance have extraordinary potential to reshape the social world. However, the scholarly community currently lacks adequate conceptualizations to describe these phenomena and assess what they might portend. This paper examines the current state of the literature on LSDIs, finding hundreds of related articles and documents – only a fraction of which add to the conceptualization process. It then describes one way forward in developing a holistic concept of LSDIs upon which to ground further study: by placing them in historical context, as efforts to democratize access to information.

## Keywords

Digitization, digital libraries, information access, literature.

## INTRODUCTION

Since about 2004, books have been moving online at an unprecedented rate, thanks to the advent of “large-scale digitization initiatives” (LSDIs) such as Google Book Search (<http://books.google.com>) and the Open Content Alliance (<http://www.opencontentalliance.org/>). These projects have already changed the information ecosystem, by introducing millions of published works into a global, public digital space. Going forward, this expanded access to books has the potential to reshape myriad other social structures, from scholarship to citizenship to creativity (e.g., Kelly, 2006; Samuelson, 2010) – not to mention the economics of publishing and authorship. As of July 2010, Kindle e-books had already begun to outsell hardcover editions on Amazon.com (Miller, 2010) – and one imagines that this effect will only grow as the products of LSDIs’ scanning efforts become available.

Given the immense and varied social significance of LSDIs and their impacts, one might expect to find a robust body of

social scientific literature on the topic developed over the past six years, examining these projects’ existing, potential, or ideal forms and effects. However, this is thus far not the case. Though these initiatives have received copious coverage in the popular press, trade literature, and legal venues, the number of rigorous scholarly analyses of LSDIs as social and cultural phenomena remains quite small.

In this paper, I review the current state of the literature on LSDIs, and based on this review, argue that we presently lack a coherent holistic concept of what LSDIs actually *are*, as sociocultural (or sociotechnical) entities, outside of the legal issues they raise. Following from this analysis, I suggest that one route toward such conceptualization lies through analogy with past initiatives that similarly sought to democratize access to information – for example, the early American public library movement.

## THE OUTLINES OF AN ABSENCE

In order to develop a clear sense of the current state of scholarship on LSDIs, a very broad-scale review of the literature was conducted, which swept together as many materials about LSDIs as possible. That universe was then iteratively filtered down to discover the size and shape of the body of work making original, scholarly contributions to thinking about LSDIs as social and technical phenomena. This process is detailed below.

## Finding the Universe

As a first step in conducting a thorough literature search, nineteen library databases were selected<sup>1</sup> based on recommendations posted on library subject specialist pages at the author’s university.<sup>2</sup> Then, within those databases,

ASIST 2010, October 22–27, 2010, Pittsburgh, PA, USA.



The author has licensed this work under the Creative Commons Attribution-NonCommercial 3.0 Unported License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc/3.0/> or send a letter to Creative Commons, 171 Second Street, Suite 300, San Francisco, California, 94105, USA.

---

<sup>1</sup> Academic OneFile, Academic Search Complete, ACM Digital Library, America: History & Life, Business Source Complete, Communication & Mass Media Complete, Compendex, EconLit, Education Fulltext, ERIC, HeinOnline, Inspec, LISA, LLIS, LISTA, PAIS, PsycInfo, Sociological Abstracts, and Web of Knowledge.

<sup>2</sup> Subject pages used: Business, Communication, Computer Science, Economics, Education, Engineering, History of Science, Information Science, Political Science and Public Affairs, Psychology, and Sociology.

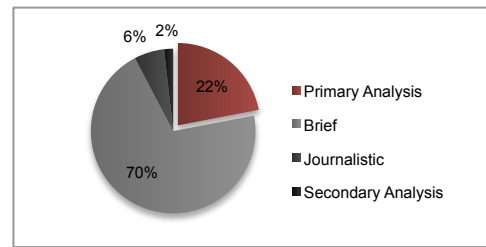
two very general keyword searches were executed (with some variances based on the structures of specific databases).<sup>3</sup> An initial pass through the results eliminated duplicates, clearly irrelevant articles, letters to the editor, and book reviews. Then, what remained was compared against Bailey’s “Google Books Bibliography” (2010), and all articles on that list that had not already been retrieved were added. This produced a starting value of 1306 articles. Though this process may not (and indeed, did not) retrieve a comprehensive set of writing about LSDIs, my prior experiences in this area indicate that the proportions of article types that it revealed would likely remain valid within that larger universe.

**Drilling Down to Scholarly Conceptualization**

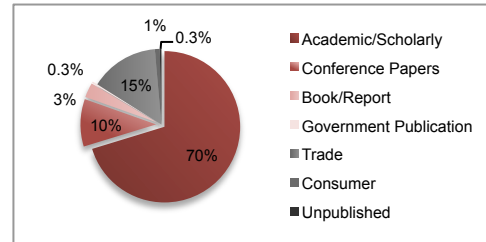
As previously noted, however, many of the references discovered in the above search process do not directly contribute to developing a generalizable conceptual framework for analysis of LSDIs. Thus, several filters were applied. The first of these had to do with article type. As shown in Figure 1, 70% of the articles found were very brief (less than 3 pages), 6% were journalistic (geared toward news, description, or opinion, rather than original research), and 2% were secondary analyses (discussions of previous articles, conferences, or reports). In fact, only 22% (286 references) could be called primary analyses (original, scholarly contributions).

Working from that 286, the second filter applied was the publication type. Publications appearing in traditional serials were classed according to categories assigned by the UlrichsWeb serials index: Academic/Scholarly, Trade, and Consumer. Non-serials-based publications were classed according to broader categories: Book/Report, Conference Paper, Government Publication, and Unpublished. Then, because the goal of this exercise is to gain perspective on *scholarly* conceptualization of LSDIs, these categories were used to filter out 16% of the remaining articles – those that were either unpublished or published in Trade and Consumer venues (Figure 2). Doing so leaves 240 articles.

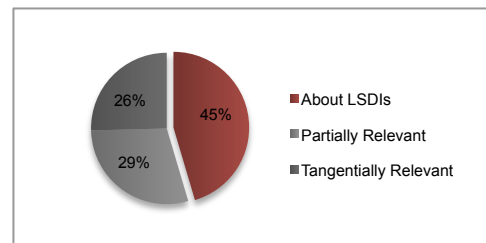
A third aspect of this body of literature that must be considered is the topicality of the works: that is, are LSDIs (a) the article’s primary focus, (b) used to illustrate some broader point, or (c) simply mentioned in passing, as a relevant but tangential issue? In fact, within the 240 remaining works, fewer than half were actually principally focused on the topic of LSDIs (Figure 3). Because LSDIs themselves are the primary focus here, this allows us to eliminate the combined 55% of articles that do not focus on LSDIs as phenomena, but rather as examples or illustrations of some other point. This leaves 108 articles.



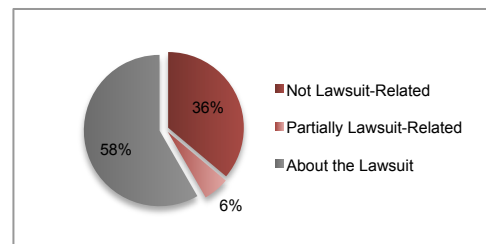
**Figure 1: All Articles Retrieved, by Article Type**



**Figure 2: Original Contributions, by Publication Type**



**Figure 3: Original, Scholarly Contributions, by Relevance**



**Figure 4: Original, Scholarly Contributions, About LSDIs, by Degree of Focus on Lawsuit/Settlement**

Finally, there is the issue of the lawsuit. The copyright lawsuit brought by the Association of American Publishers and the Authors Guild against the Google Books Library Project, and that lawsuit’s proposed settlement agreement, have been the subject of many analyses. However, these analyses tend to be less concerned with advancing conceptual understanding of LSDIs as independent phenomena than with using the lawsuit as an lens through which to analyze copyright and/or fair use laws. Thus, I also excluded those articles focusing exclusively on the lawsuit – more than half of those remaining (Figure 4). That done, we are left with just 45 references out of our original 1306 that qualify as original scholarly contributions principally focused on LSDIs, and not dealing exclusively

<sup>3</sup> (1) KW: “google book\*” or “google print” or “open content alliance,” YR: ≥2004; (2) SU: Library materials – digitization AND KW: google (or, in databases sans subject headings: KW: library AND digitization AND google).

with the lawsuit – a mere 3.4%.

All this citation-gerrymandering reveals many intriguing trends in the LSDI literature – more than can be covered here. Even preliminarily, however, it shows two things quite clearly. First, the academy (perhaps unsurprisingly) has fallen behind the journalistic and legal communities in thinking about LSDIs, to a problematic extent. And second, within the tiny set of relevant, scholarly literature we do have, little conceptual cohesiveness thus far exists. These insights are discussed below.

#### *Leaving LSDIs to Journalists and Lawyers*

At the broadest level, the analysis above illustrates the extent to which the academic community has up to this point left investigation of the LSDI phenomenon up to the journalistic and legal communities. Of the 387 articles in the set longer than three pages, a combined total of 196 (51%) are either journalistic or legal. And though both communities have produced important and illuminating pieces on LSDIs (e.g., Grimmelmann, 2009; Roush, 2005; Samuelson, 2010; Toobin, 2007), neither has shown a strong interest in systematic conceptualization of what these projects are and what they mean. And there are good reasons for this. Journalism, which focuses on providing the public with timely and comprehensible information, tends not to be as interested in theoretical foundations or methodological rigor as academic research. And law reviews, for their part, quite reasonably privilege questions of jurisprudence and legal procedure over consideration of the broader social meaning of the phenomena that give rise to those questions. Still, exploring such broader questions of meaning is necessary for understanding LSDIs and their potential impacts going forward; it would thus behoove the non-legal academy to work toward closing this gap.

#### *Describing Different Parts of the Elephant*

The second point that emerges from the above analysis relates to the conceptual diversity of the 45 pieces of original scholarly literature about LSDIs that remained in the end. These articles, far from painting a cohesive, unified picture of LSDIs, seem to fall into roughly seven broad (and admittedly subjective) categories:

1. Library and preservation implications (12 works; e.g., Rieger, 2008; Sandler, 2005)
2. Metadata issues (8 works; e.g., Duguid, 2007; Mimno & McCallum, 2007)
3. Holistic and/or conceptual analyses (8 works; e.g., Jeanneney, 2006; Leetaru, 2008)
4. Descriptions or development of technological infrastructure (6 works; e.g., Choudhury *et al.*, 2006; Langley & Bloomberg, 2008)
5. Analyses of LSDIs' burgeoning collections (6 works; e.g., Kousha & Thelwall, 2009; Lavoie *et al.*, 2005)
6. Business/economic assessments (3 works: Ciriaci & Quaglione, 2009; Rosenblatt, 2005; van Dijk, 2005)
7. Postulations on LSDIs' potential role in scholarship (2 works; Courant, 2006; Martin, 2008)

From these works, we can learn a great deal about LSDIs: their implications for the use and custodianship of written culture; the challenges posed by attempts to organize their massive collections, enhance their searchability, and increase their cross-platform usability; the strengths and weaknesses of the current collections for particular domain applications and scholarship more broadly; and the projects' potential impacts on the publishing industry and others involved in information dissemination. Indeed, eight of these works (category no. 3, above) even deal fairly explicitly with exploration of what LSDIs *are* as complex, culturally-situated sociotechnical projects.

Still, to a great extent, these subsets of the primary literature on LSDIs do not operate in the same conceptual space. Instead, they resemble the tale of the blind men and the elephant, wherein each touches a different part – the trunk, the tail, the ear, etc. – and thus emerge with several radically divergent ideas of what the elephant as a whole must look like.<sup>4</sup> As it stands, the librarian looks at an LSDI and sees a massive digital extension of the library, with implications for preservation, document delivery, service provision, and so forth. The computer scientist, looking at the same object, sees an ideal laboratory for innovation in image standards, metadata generation, and cross-platform translation. The historian or the music theorist sees a third object still: a tool or resource with potential utility for work in her specific subject area. And on it goes, for university administrators, competing or complementary business leaders, and more.

Eliminating the diversity of these different perspectives would be both impossible and undesirable. However, in light of LSDIs' potential to influence and reshape social realities, some conceptual coherence needs to develop between or among these facets. Thus, I conclude this analysis by suggesting one path toward a more cohesive concept of LSDIs as an object of research: through their placement in historical context.

#### **CONCLUSION: THE VALUE OF HISTORICAL CONTEXT**

History, as Shera has noted, is essential to understanding how we have arrived where we are, and to avoiding the repetition of our collective missteps and failures; just as an individual relies on memories to guide present reasoning, so a society depends on history to shed light on present challenges (1952). For LSDIs, I would argue that this means looking to the history of the underlying urge they represent: the urge to provide open, egalitarian access to information, to as broad an audience as possible. And particularly when one considers existing LSDIs' top-down structure, their millionaire (or billionaire) benefactors, their democracy-building rhetoric, and their fundamental focus on *books*, a parallel historical exemplar of this urge comes into focus: the early American public library.

---

<sup>4</sup> Several versions of this tale exist; one is available at: <http://www.jainworld.com/education/stories25.asp>

As a next step toward a fuller conceptualization of LSDIs, I plan to develop this parallel through a comparative case study analysis of current LSDIs and early American public libraries (1850-1920), using a theoretical lens informed by structuration theory and social construction of technologies (Giddens, 1984; Pinch & Bijker, 1984). Initially, I plan to explore the following questions:

- Through what processes have the architectures and policies of these initiatives been negotiated?
- What actors or phenomena have influenced their physical and political structures, and how? What actors or phenomena have failed to exert such an influence?
- Are there elements of these initiatives' architecture or policy that directly reflect the negotiation processes that created those structures? What elements, and what do they reflect?

This is clearly not an exhaustive set of questions; nor is it the only perspective from which the phenomenon of LSDIs might be interrogated. Indeed, I hope others in the ASIS&T community will add their own questions and viewpoints. Still, I suggest that such historical contextualization represents a feasible and useful path toward a stronger sense of LSDIs' position in the social world – and through that, I hope, toward a better understanding of how they might best be designed in order to maximize their enormous positive potential.

## REFERENCES

- Bailey, C. W., Jr. (2010, April 12). Google Books bibliography. Retrieved May 20, 2010, from <http://www.digital-scholarship.org/gbsb/gbsb.htm>
- Choudhury, G. S., DiLauro, T., Ferguson, R., Dorettboom, M., & Fujinaga, I. (2006). Document recognition for a million books. *D-Lib Magazine*, 12(3).
- Ciriaci, D., & Quaglione, D. (2009). Multilateral platforms and Google Book Search: Which competition after the settlement agreement? *Industria: Rivista di economia e politica industriale*, 30(4), 647-678.
- Courant, P. N. (2006). Scholarship and academic libraries (and their kin) in the world of Google. *First Monday*, 11(8).
- Duguid, P. (2007). Inheritance and loss? A brief survey of Google Books. *First Monday*, 12(8), 11 pp.
- Giddens, A. (1984). *The constitution of society: Outline of the theory of structuration*. Berkeley: University of California Press.
- Grimmelmann, J. (2009). *The Google Book Search settlement: Ends, means, and the future of books*. Washington, DC: American Constitution Society.
- Jeanneney, J.-N. (2006). *Google and the myth of universal knowledge: A view from Europe*. Chicago: University of Chicago Press.
- Kelly, K. (2006, May 14). Scan this book! *The New York Times*.
- Kousha, K., & Thelwall, M. (2009). Google Book Search: Citation analysis for social science and the humanities. *Journal of the American Society for Information Science & Technology*, 60(8), 1537-1549.
- Langley, A., & Bloomberg, D. S. (2008). Google Books: Making the public domain universally accessible. In *Proceedings of the SPIE - the international society for optical engineering* (10 pp.).
- Lavoie, B., Connaway, L. S., & Dempsey, L. (2005). Anatomy of aggregate collections: The example of Google Print for libraries. *D-Lib Magazine*, 11(9).
- Leetaru, K. (2008). Mass book digitization: The deeper story of Google Books and the Open Content Alliance. *First Monday*, 13(10).
- Martin, S. (2008). To Google or not to Google, that is the question: Supplementing Google Book Search to make it more useful for scholarship. *Journal of Library Administration*, 47(1/2), 141-150.
- Miller, C. C. (2010). E-books top hardcovers at Amazon. *New York Times*, p. B1.
- Mimno, D., & McCallum, A. (2007). Organizing the OCA: Learning faceted subjects from a library of digital books. In *Proceedings of the 7th ACM/IEEE-CS joint conference on digital libraries*. Vancouver, BC: ACM.
- Pinch, T., & Bijker, W. (1984). The social construction of facts and artefacts: Or how the sociology of science and the sociology of technology might benefit each other. *Social Studies of Science*, 14(3), 399-441.
- Rieger, O. Y. (2008). *Preservation in the age of large-scale digitization*. Washington, DC: Council on Library and Information Resources.
- Rosenblatt, B. (2005). Rights management and the revolution in e-publishing. *INDICARE Monitor*, 2(9).
- Roush, W. (2005). The infinite library. *Technology Review*, 108(5), 54-59.
- Samuelson, P. (2010). Google Book Search and the future of books in cyberspace. *Minnesota Law Review* (forthcoming).
- Sandler, M. (2005). Disruptive beneficence: The Google Print program and the future of libraries. *Internet Reference Services Quarterly*, 10(3-4), 5-22.
- Shera, J. H. (1952). On the value of library history. *The Library Quarterly*, 22(3), 240-251.
- Toobin, J. (2007, February 5). Google's moon shot: The quest for the universal library. *The New Yorker*.
- van Dijk, A. (2005). Primary source publishers in a Googling world. *Microform & Imaging Review*, 34(1), 31-34.