Massive Digital Libraries (MDLs) and the Impact of Mass-Digitized Book Collections

Andrew Philip Weiss
https://orcid.org/0000-0002-8900-2779
California State University, Northridge, USA

INTRODUCTION

To better analyze the growth of digital libraries, Weiss and James have proposed adopting the term Massive Digital Libraries (MDLs), a concept based on the increased size, scope and scalability of mass-digitized book collections. MDLs rival physical libraries' print holdings in size, breadth, and depth, often approaching a scale previously found only among library consortia or national libraries. (Weiss and James, 2013a, 2013b, 2014, 2015; Weiss, 2016) The concept further intersects the digital library with the wider development of 'big data,' which is driven to an ever-larger scale by the '5Vs' of Volume, Variety, Velocity, Variability, and Veracity. (Weiss, 2018)

The root of the concept begins in late 2004 when Google publicly announced its intention to digitize the world's books—including works still under copyright protection—and to place them *all* (roughly 130 million) online. Jean-Noel Jeanneney, head of Bibliothèque nationale de France at the time, interpreted Google's planned project as a wake-up call for European countries. Failure to catch up to the American company, he argued, would result in significant problems for non-American organizations, especially if they were unable to check the company's outsized influence. (Jeanneney, 2005)

Fifteen years on, it is hard to imagine that Google's desire to create an online digital should have come as such a shock. Yet Google spurred significant hand-wringing and soul-searching among institutions traditionally charged with producing or preserving cultural artifacts. (Jeanneney; Venkatraman, 2009) In retrospect, the controversy seems quaint in comparison to the current crop of issues – especially the current "disruptions" of established economic models by Uber/Lyft, Facebook, Twitter, Spotify, and the like; the looming unknowns circling around artificial intelligence; and the encroachments on civil rights via electronic digital surveillance and other intrusions of privacy.

A number of similar mass-digitization projects developed and matured since Google's announcement, including the *HathiTrust*, *Internet Archive*, *Digital Public Library of America* (*DPLA*), *California Digital Library, Texas Digital Library*, *Gallica*, and *Europeana*. These projects each transcend their roots as localized digital libraries and simultaneously adapted to and altered the digital landscape. These various MDLs have allowed for and contributed to the ascendancy of our current mass-digitization online culture.

This chapter describes the characteristics of Massive Digital Libraries (MDLs) and outlines their impact upon current information science issues, especially with regard to digital collection metadata, copyright, the diversity of source collections, and user privacy in an age of so-called "surveillance capitalism." (Zuboff, 2015) Traditionally, libraries have been created to serve particular communities *defined* as well as *bound* by geography, intellectual disciplines, or specific end users. MDLs in their current trajectories, however, promise to transcend such limits in ways that are simultaneously *con*structive and *de*structive.

DOI: 10.4018/978-1-7998-3479-3.ch123

BACKGROUND: DEFINING MASSIVE DIGITAL LIBRARIES

14

Defining Criteria

Massive Digital Libraries (MDLs) describes a specific class of digital libraries that correspond to the size of a traditional, large brick-and-mortar library. Although other disciplines have discussed digital libraries and archives in terms of computer science, such as in the Very Large Digital Library (VLDL) movement, none have framed the discussion in terms of the principles of library science or the services and content access provided by an actual, working library. (VLDL, 2011)

The following list of characteristics has been proposed to help define MDLs:

1. Collection size: surpasses 500,000 texts; prime MDLs comprise tens of millions of texts;
2. Acquisitions, collection development & copyright: numerous partnering members contribute content regardless of author or copyright holder permissions and regardless of end-user needs;
3. Content type: mass-digitized print books; the resulting searchable digital corpus of texts becomes as important as the individual titles;
4. Collection diversity: diversity is dependent upon self-chosen partner members, which can reflect distortions or biases inherent to the source collections;
5. Content access: varying degrees of open access exist within MDLs; content is searchable through single, uniform interfaces (search engines & portals) representing all the collections as members of a single entity regardless of source;
6. Metadata: Metadata is gathered and aggregated from multiple sources, with a reliance on new digital description schema;
7. Content / digital preservation: consortium members provide long-term digital preservation strategies at local levels as well as "in the cloud".

These criteria and their attendant issues, though not necessarily *unique* to digital libraries, require different approaches when dealt with as a Massive Digital Library. The issues involved with aggregating millions of previously published print materials into one uniform, yet decentralized, conceptual and online digital space become more complex as the size increases. It is important to differentiate MDLs from their smaller counterparts as they are more difficult to police and analyze, especially with metadata uniformity, copyright compliance, rights ownership, and user privacy. The larger an institution or system is, the more unwieldy and slow-to-change it may become. Furthermore, the lack of transparency among IT companies – especially Google – requires more robust oversight. As a result, it is important to remain cognizant of how MDLs approach the common characteristics of books in ways that stretch the boundaries of the original print medium. Print books remain static during their lifetimes, changing only when new editions are created. E-books promise the flexibility to change into ongoing "works in progress," without fixed texts. Unlike print books, their access can easily be tracked, eroding the user privacy that remains a central tenet of libraries. Identifying MDLs as unique entities would allow scholars and researchers to both utilize and safeguard these newly-created and widely-distributed digital collections.

Representative MDLs

The following is a look at several representative MDLs and their defining characteristics. This section will also help readers get a sense of how MDLs stack up against each other. Criteria described for each