Massive Digital Libraries (MDLs)

Andrew Philip Weiss
California State University – Northridge, USA

INTRODUCTION

To provide a clearer framework for analyzing the growth of digital libraries, Weiss and James have proposed the term Massive Digital Libraries (MDLs), which is based on the size, scope and increasing scalability of digitized book collections. Such MDLs rival the size, breadth, and depth of a physical library's print holdings, and often reach a scale seen among library consortia collections. (Weiss and James, 2013a, 2013b, 2014, 2015; Weiss, 2016)

The root of the concept begins in late 2004 when Google made its "resounding announcement" to digitize millions of the world's books—including works still under copyright protection—and to place them *all* online. (Jeanneney, 2005) 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.

Twelve years on, it is hard to imagine that Google's desire to create an online digital library on such a large scale should have come as such a shock. Yet at the time Google caused significant hand-wringing and soul-searching among institutions traditionally charged with producing or preserving cultural artifacts. (Jeanneney; Venkatraman, 2009) In retrospect, the controversy seems almost quaint in comparison to the current crop of issues – especially the current "disruptions" of established economic models by Uber/Lyft, Facebook, Twitter, Spotify, Snapchat, e-readers, et al. and the encroachments on civil rights via electronic digital surveillance and other intrusions of privacy.

DOI: 10.4018/978-1-5225-2255-3.ch454

A number of mass-digitization projects have grown in the wake of 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 have 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 will describe the characteristics of Massive Digital Libraries (MDLs) and outline their impact upon contemporary information science issues, especially with regard to digital collection metadata, copyright and the diversity of the source collections. Traditionally, libraries have been created to serve particular communities defined by geography, intellectual discipline, or specific end users. However, MDLs in their current trajectories promise–for better *and* for worse—to transcend such limits.

BACKGROUND: DEFINING MASSIVE DIGITAL LIBRARIES

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 dealing with 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 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 and rights ownership. The larger the institution or system the more unwieldy and slow-to-change it may become. It is therefore important to remain cognizant of how these MDLs approach the common characteristics of books in ways that stretch the boundaries of the original print medium. Print books, for better or worse, remain static during their lifetimes, changing only when new editions are created. Yet e-books could change into ongoing "works in progress"; they can be altered with relative ease. Identifying MDLs as unique entities would allow scholars and researchers to both utilize and safeguard these newly-created digital corpuses.

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 MDL include their starting dates, *modus operandi*, estimated sizes of the collections, subjects collected, the number of partner institutions, and the languages covered in their collections.

American-Based, Anglo-Centric MDLs

Google Books (25 Million Texts)

The *Google Books* project started in 2004. Envisioned by Google co-founder Sergei Brin as the replacement of a library's card catalog, *Google Books* currently holds approximately 25 million volumes within its collections, though exact figures have been difficult to verify.

Google Books is characterized in part by its extremely robust and flexible search engine. When searching in this MDL, a large amount of set results will often appear. Mistakes in the search are often auto-corrected. Their stated goal