Chapter 18

Big Data at Scale for Digital Humanities:
An Architecture for the HathiTrust Research Center

Stacy T. Kowalczyk
Dominican University, USA

Craig Willis
University of Illinois, USA

Yiming Sun
Indiana University, USA

Jiaan Zeng
Indiana University, USA

Zong Peng
Indiana University, USA

Milinda Pathirage
Indiana University, USA

Beth Plale
Indiana University, USA

Samitha Liyanage
Indiana University, USA

Aaron Todd
Indiana University, USA

Guangchen Ruan
Indiana University, USA

Loretta Auvil
University of Illinois, USA

J. Stephen Downie
University of Illinois, USA

ABSTRACT

Big Data in the humanities is a new phenomenon that is expected to revolutionize the process of humanities research. The HathiTrust Research Center (HTRC) is a cyberinfrastructure to support humanities research on big humanities data. The HathiTrust Research Center has been designed to make the technology serve the researcher to make the content easy to find, to make the research tools efficient and effective, to allow researchers to customize their environment, to allow researchers to combine their own data with that of the HTRC, and to allow researchers to contribute tools. The architecture has multiple layers of abstraction providing a secure, scalable, extendable, and generalizable interface for both human and computational users.

DOI: 10.4018/978-1-4666-9840-6.ch018

INTRODUCTION

Big Data is big news. The data deluge of scientific research data, social media data, and financial/commercial data is now mainstream. It is discussed in the public press, studied in academic situations, and exploited by entrepreneurs (Press, 2013). Until very recently, the humanities have not been included in the digital data wave (Anderson, 2007). However, the state of humanities research is undergoing a major transformation (Drucker, 2009; Dunn & Blanke, 2009). Digitally-based research methodologies are becoming more widely used and accepted in the humanities (Borgman, 2009). Big data is just becoming available for humanists (Manovich, 2012).

The initial efforts at doing *in silico* humanities research were personal projects of forward-looking researchers (Dunn & Blanke, 2009). These initial efforts entailed digital scholarly editions (Newman, 2013; Walsh, 2013), specialized applications for specific research projects (Walsh & Hooper, 2012), and specialized applications for a specific research domain (Almas et al., 2011). Some of these initiatives evolved into community-based efforts to develop a set of standards for text encoding that became the focus of much of the digital humanities work for the past 10 years (Drucker, 2009). Institutional support for digital humanities includes the trend towards Digital Humanities Centers in colleges and universities. Many of these centers have not been as successful as initially expected due to the resource intensive nature of digital humanities projects such as text digitization and encoding or specialty applications development (Borgman, 2009). Cross-institutional activities, including funding agency initiatives to develop shared technical infrastructures, have had mixed results (Friedlander, 2009). All of these efforts have not provided the desired results. Infrastructure for digital humanities remains unavailable to many researchers (Borgman, 2009).

This chapter describes the research efforts to create a cyberinfrastructure to support Big Data for digital humanities. In the sections that follow, the need for this research is placed in the context of digital humanities and cyberinfrastructure research, the nature of data in the humanities is discussed, the HathiTrust Research Center (HTRC) is introduced, the cyberinfrastructure research is described, and the architecture is explicated.

BACKGROUND

Digital Libraries and Digital Humanities

Digital Humanities encompasses many types of research domains, methodologies, and data types (Manovich, 2012). However, much of the work in digital humanities deals with textual data (Crane, 2006; Drucker, 2009; Unsworth, Rosenzweig, Courant, Frasier & Henry, 2006). Until very recently, the biggest barrier to digital humanities was getting textual data in digital form; digitizing materials or negotiating for access to digital materials was the first step in most digital humanities projects (Cunningham, 2011; Pitti, 2004). With the advent of massive digitizing efforts by many libraries and by Google, digital data for researchers in the humanities is now becoming available (Cunningham, 2011; Svensson, 2010; Wilford & Henry, 2012). These digital libraries offer services to researchers that include bibliographic and full text searching, results management, metadata, and text and image displays. In addition, some digital libraries can provide specialized or advanced functionality to specific communities of users within the scope and mission of the library (Candela et al., 2011).