The *Workset Creation for Scholarly Analysis (WCSA) Prototyping Project*: Background and Goals

J. Stephen Downie, Tim Cole, Beth Plale, Katrina Fenlon, Karen Wickett and Megan Senseney

The HathiTrust Digital Library (HTDL) houses over 10,000,000 volumes comprising some 3,800,000,000 pages. Of these volumes, approximately one third (3,000,000) are in the public domain and two thirds (7,000,000) under copyright restrictions, inhibiting open access by scholars and researchers. To help overcome the limitations imposed by copyright impediments, the HathiTrust Research Center (HTRC), has been developing a “non-consumptive” research model wherein analytic code developed by scholars from around the world is brought to the data collection rather than distributing copyrighted data to the scholars. In this way, analytic scholarship of large collections can take place under a restrictive copyright environment. The HTRC is a unique multi-institutional collaboration that brings together two interdisciplinary teams of digital humanities researchers from the University of Illinois at Urbana-Champaign and Indiana University at Bloomington.

Scholars rely on library collections like the HTDL to support their scholarship. Out of its collections, scholars strive to select, organize, and refine the worksets that will answer to their particular research objectives. The requirements for those worksets are becoming increasingly sophisticated and complex, both as humanities scholarship has become more interdisciplinary and as it has become more digital. It has also become quite evident how difficult it is for scholars to create and interact with worksets derived from collection at the scales associated with the HTDL. Neither current metatdata nor full-text searching techniques are up to the task.

To help overcome the problems associated with the creation and use of large-scale analytic worksets, the “Workset Creation for Scholarly Analysis: Prototyping Project” (WCSA) project began in July 2013. It will conclude in June 2015. The project will address three sets of tightly intertwined research questions regarding:

1) Enriching the metadata in the HathiTrust corpus;

2) Augmenting string-based metadata with URIs to leverage discovery and sharing through external service; and,

3) Formalizing the notion of collections and worksets in the context of the HTRC.

Also funded by the Andrew W. Mellon Foundation, WCSA builds upon the model of the Open Annotation Collaboration and it will similarly release an open, competitive Request for Proposals in November 2013 with the intent to fund four prototyping projects ($40,000 per project) that will build tools for enriching and augmenting metadata for the HathiTrust corpus. Concurrently, the HTRC will work closely with the Center for Informatics Research in Science and Scholarship (CIRSS) at the Graduate School of Library and Information Science, University of Illinois, to develop and instantiate a set of formal data models that will be used to capture and integrate the outputs of the funded prototyping projects with the larger HathiTrust corpus.