February 10, 2011

HathiTrust Research Center

Fact Sheet

The HathiTrust Research Center (HTRC), recently formed, is dedicated to the provision of computational access to a comprehensive body of published works for scholarship and education. This great public good will provide sustained access to works in the public domain and, on more limited terms and at a later date, to publications in copyright for computational research purposes. The HTRC will provide a persistent and sustainable structure to enable original and cutting edge research. It will stimulate the development of new functionality and tools to enable new discoveries that would not be possible without the HTRC.

HTRC is founded as a joint venture between Indiana University and the University of Illinois Urbana-Champaign aimed at solving the difficult challenges of increasing access to the public domain and copyrighted material in HathiTrust.

Key architectural and organizational aspects of the center are as follows:

1) HTRC is architected to be modular and open. Its unique challenge is infrastructure that supports non-consumptive research. We anticipate accomplishing this through a novel grid and cloud based architecture, security, auditing and provenance collection. Non-consumptive research is defined as follows:
   a) *Non-consumptive research -- No action or set of actions on the part of HTRC users, either acting alone or in cooperation with other users over the duration of one or multiple sessions can result in sufficient information gathered from the HathiTrust collection to reassemble pages from the collection.*

2) HTRC will version the HathiTrust collection to enable researchers to tie research back to the version that was active when the research was carried out. Versions have unique identifiers (such as DOIs) that are long lasting and immutable.

3) HTRC will support interoperability through use of inCommon SAML identity for access by members of the academic federation. For non-academic federation members, we anticipate using one-time-passwords (OTP) via a hardware token that is provided to the user. The lighterweight OpenID is being considered for access to the public corpus.

4) HTRC may give preference to HathiTrust members.

5) HTRC is built on the sound and well-tested principles of a service-oriented architecture to enable interoperability. As part of this it will maintain a registry repository of text mining algorithms and retrieval tools available on-line for human and programmatic discovery. It may also register derived data sets, indexes, and versions in the registry repository.

6) HTRC is intended as a user driven resource, through an active advisory board, and community sharing model that allows users to share their algorithms and tools.

7) HTRC will provide access to familiar tools such as MONK and SEASR text mining and retrieval tools applied to the HTRC public domain works and later with safeguards in place, to the copyrighted collection.

It is important to delineate the structure of the HathiTrust Research Center with respect to HathiTrust itself. The HathiTrust repository offers long-term preservation and access services, including bibliographic and full-text search and reading capabilities for public domain volumes and some copyrighted volumes. The HathiTrust Research Center on the other hand, provisions for computational research access to the HathiTrust collection. Limited reading of materials will be possible in the Research Center to accommodate needs for reviewing results, etc., but the destination for reading-based research remains the HathiTrust repository.

To learn more or join the efforts, please contact project leads Dr. Beth Plale plale@indiana.edu (812) 855 4373 or Dr. Marshall Scott Poole mspoole@illinois.edu