JSTOR’s Big Data Challenge:
Mining Log Files to Improve Service to Users
CNI Fall 2011 Membership Meeting

December 12-13, 2011
Arlington, Virginia

In the 15 years that JSTOR has been in existence, a wealth of logging data has been generated and archived. This logging data represents many billions of user actions. Until recently this usage data has mainly been used for generating summary-level institution and publisher reports. The sheer volume and complexity of these data made multi-dimensioned longitudinal analysis impractical until just recently. Over the last year ITHAKA has made a significant investment in normalizing and organizing these data in the interest of better understanding user behaviors and trends in the consumption of academic materials.

In this presentation we will discuss the technological approach that we’ve taken in dealing with the data volume and complexity issues, including Big Data challenges such as storage, processing, and analysis. We’ll share some experiences from our original attempt to build this data warehouse using traditional relational database technologies and our decision to abandon this approach in favor of a solution based on the open source Hadoop infrastructure. Hadoop provides a robust, scalable and cost-effect solution to managing our Big Data. We’ve combined Hadoop with an open source indexing technology (Lucene/SOLR) and some custom-built software providing a web-based tool for the interactive exploration of this rich data set. In the presentation we’ll also share some top-level observations on user behaviors and content discovery and consumption trends that have been identified using these tools.

Contact:
Ron Snyder
Director of Advanced Technologies, ITHAKA
Ronald.Snyder@ithaka.org