Journal of the Chicago Colloquium on Digital Humanities and Computer Science
Volume 1 Number 3 (2011)

Finding the Canary for Text Mining: Analysis of the use and users of MONK text mining research software

Harriett E. Green, University Library, University of Illinois at Urbana-Champaign

Abstract

MONK is a text mining tool hosted by the University of Illinois Library that enables researchers to analyze digital texts from select databases and archives of digitized texts. Using web log statistical data generated by the MONK website over the twelve months of 2010, this study will present initial web log analysis on the use of MONK by researchers. This paper presents a preliminary analysis of web log statistical data from MONK to examine the ways in which MONK has been most commonly used by researchers, and analyze the possible needs of researchers in the future.

1. Introduction

Digital humanities tools are rapidly increasing in numbers, innovation, and power, but little is yet known about how researchers en masse actually use the tools. This study has taken the opportunity to study a digital humanities tool that has completed its first year as a public instance of a research tool: the web-based text mining software called MONK (Metadata Opens New Knowledge). MONK was a multi-institutional research project funded by a grant from the Andrew W. Mellon Foundation, and was transitioned to University of Illinois Library-hosted public research software in January 2010. As the first full year of its public release in 2010 drew to a close, enough data was gathered to begin a study of the use and users of MONK. The research questions explored in this study include: How are researchers accessing MONK? What does the data say about use patterns in MONK? How does this data reveal user needs and tool adjustments to improve MONK's functionality for researchers? This short paper presents preliminary analysis of web log statistics data in MONK for the twelve months of 2010, and examines what this early data reveals about usage of the tool and its users. Note: For larger, higher quality versions of the figures reproduced here, please refer to the *Supplementary Data* section accompanying this article online at http://jdhcs.uchicago.edu

2. MONK: The Background

MONK combines two previously developed text mining programs NORA (http://www.noraproject.org/) and WordHoard (http://wordhoard.northwestern.edu/), to create a robust data-mining environment on texts from publicly available and proprietary digital text collections.[1] The texts contained in MONK are from publicly available digital text collections such as Early American Fiction and Documenting the American South, as well as proprietary collections including Eighteenth Century Collections Online, Early English Books Online, and Chadwyck-Healey's Nineteenth-Century Fiction. TEI-A schema was created to implement a uniform mark-up of these texts, which were unevenly marked up per their various collections, and the corpora of texts were normalized into TEI-A using the Abbot software program.[2] Morphadorner then was applied the texts to mark them up for "tokenization, sentence boundaries, standard spellings, parts of speech

[1] MONK Documentation (2009).

[2] Brian L. Pytlik Zillig, "TEI Analytics, Converting Documents into a TEI Format for Cross-Collection Text Analysis," *Literary and Linguistic Computing* 24, no. 2 (2009): 190.

Source URL: http://jdhcs.uchicago.edu/
Published by: The Division of the Humanities at the University of Chicago

and lemmata."[3] The texts were then entered into a database that enabled extraction of the text for data mining, and the SEASR environment provides the tools for statistical analyses in MONK.

In its public instance, all users can run MONK on publicly available digital collections, and scholars affiliated with institutions in the Committee for Institutional Cooperation consortium are also able to use texts from their proprietary collections. Additionally, researchers can import texts into MONK with the use of Zotero and a MONK Firefox extension.

3. Literature Review

A number of studies have documented the research workflows of humanists with digital and online tools, from the early studies of Bates et al.'s log analysis of humanists using Dialog search system to recent ones including Duff and Cherry's exploration of humanities scholars' work with digitized materials.[4] Warwick, et al. examined the usage of digital humanities tools by scholars with the application of web log analysis.[5]

Among tools prominently used by literary scholars are text mining and statistical computations to conduct new types of textual analysis. John Burrows notes that "the real value of studying the common words rests on the fact that they constitute the underlying fabric of a text, a barely visible web that gives shape to whatever is being said."[6] A number of literary studies research works published in the past several decades have utilized computational tools to conduct stylometric analyses and data mining of texts on poems and prose, and studies such as Sinclair, Yu, and Sculley and Pasanek have explored the efficacy of various tools and analytic methods for literary text mining.[7] MONK as a tool itself has been briefly studied from various perspectives: Zillig documented the development of the Abbot software and its application of TEI-A mark-up to multiple collections of texts uploaded into MONK for text mining, and Tanya Clement analyzes the results from data mining of Gertrude Stein's *The Making of Americans* with MONK.[8] This study aims to explore the statistical usage of MONK as a tool and explore the extent of its usage during its first year as a public instance.

---

[3] MONK Documentation.

[4] Wendy M. Duff and Joan M. Cherry, "Use of Historical Documents in a Digital World: Comparisons with Original Materials and Microfiche," *Information Research* 6, no. 1 (2000).

[5] Claire Warwick, Melissa Terras, Paul Huntington, and Nikoleta Pappa, "If You Build It Will They Come? The LAIRAH Study: Quantifying the Use of Online Resources in the Arts and Humanities through Statistical Analysis of User Log Data," *Literary and Linguistic Computing* 23, no. 1 (2008).

[6] John Burrows, "Textual Analysis" in *A Companion to Digital Humanities*, ed. Susan Schreibman, Ray Siemens, John Unsworth (Oxford: Blackwell, 2004). http://www.digitalhumanities.org/companion.

[7] D. Sculley and Brad Pasanek, "Meaning and mining: the Impact of Implicit Assumptions in Data Mining for the Humanities," *Literary and Linguistic Computing* 23, no. 4 (2008); Bei Yu, "An Evaluation of Text Classification Methods for Literary Study," *Literary and Linguistic Computing* 23, no. 3 (2008); Stéfan Sinclair, "Computer-Assisted Reading: Reconceiving Text Analysis," *Literary and Linguistics Computing* 18, no. 2 (2003).

[8] Zillig, "TEI Analytics, Converting Documents into a TEI Format for Cross-Collection Text Analysis"; Tanya E. Clement, "'A thing not beginning and not ending': Using Digital Tools to Distant-Read Gertrude Stein's *The Making of Americans*", *Literary and Linguistic Computing* 23 , no. 3 (2008).

Source URL: http://jdhcs.uchicago.edu/
Published by: The Division of the Humanities at the University of Chicago

4. Data

Log data for MONK was gathered using the AWStats, a web statistics analyzer. Twelve months of web log statistics were analyzed for this initial study of MONK, from January 2010 through December 2010. The analyzed statistics included the number of visits on each URL recorded within MONK, the amount of data processed through each URL, and the number of entry and exit visits. The geographic locations of users based on recorded IP addresses, as well as qualitative survey and interview data are still being gathered and analyzed, and these will be incorporated into an advanced data analysis for a forthcoming article.

5. Analysis

5.1 Data Analysis

The monthly web log statistics were organized by the diverse types of URLs within MONK.[9] These various URLS were coded into three categories: Orientation, Workbench, and Functionality. Orientation URLs included the main menu, Shibboleth login functions, tutorial webpages, and webpages listing the terms and policies of the tool. Workbench URLs were functions that retrieved workflow tools in MONK, and began with the "/get" OR "/tool". These URLS were further broken down into "WorkbenchW" and "WorkbenchA" categories. "WorkbenchW" category contains the workset compilation functions whose URLs include terms such as "SearchManager," "ProjectManager," or "workset-manager." The "WorkbenchA" category contains web service analytic functions for the workset workflow that include terms such as "tools," "analytics," or "AnalyticsManager." The user builds a workflow in MONK as she creates worksets of texts and launches the tools for textual analysis, and the WorkbenchW and WorkbenchA URLs are either URLS called in as web services to launch and implement various functions on the user web interface or as tools in the workflow for compiling and analyzing worksets in MONK. Thus in the course of analyzing the data for these URLs, it was critical to distinguish not only the types of URLs, but the proportional frequency of their hits as reflective of the functions' interactions in the analytic workflow, which is detailed later in this paper.

The kilobytes of data processed through MONK were also analyzed for the average of each URL across the twelve months, the maximum and minimum for each page over the twelve months and among all of the pages each month; and the standard deviation of the average data processed per page. The entry and exit visits were analyzed for average number of visits per page, and the maximum and minimum per page and overall among all pages.

5.2 Initial Findings

The distribution of pages saw the highest use for Orientation pages, followed by "WorkbenchW" workset URLs and "WorkbenchA" analytics URLs. The three most frequent URLs accessed on average were (fig. 1):

- **/secure/get/CorpusManager.getWork**: WorkbenchW web service used to compile worksets of texts
- **/secure/get/CorpusManager.getWorkList**: WorkbenchW URL possibly also connected to compilation of the worksets

[9] All URLs referred to in this paper begin with "https://monk.library.illinois.edu" unless otherwise noted.

Source URL: http://jdhcs.uchicago.edu/
Published by: The Division of the Humanities at the University of Chicago

- **/secure/get/ProjectManager.getToolSets**: WorkbenchA URL which enables users to select and utilize toolset

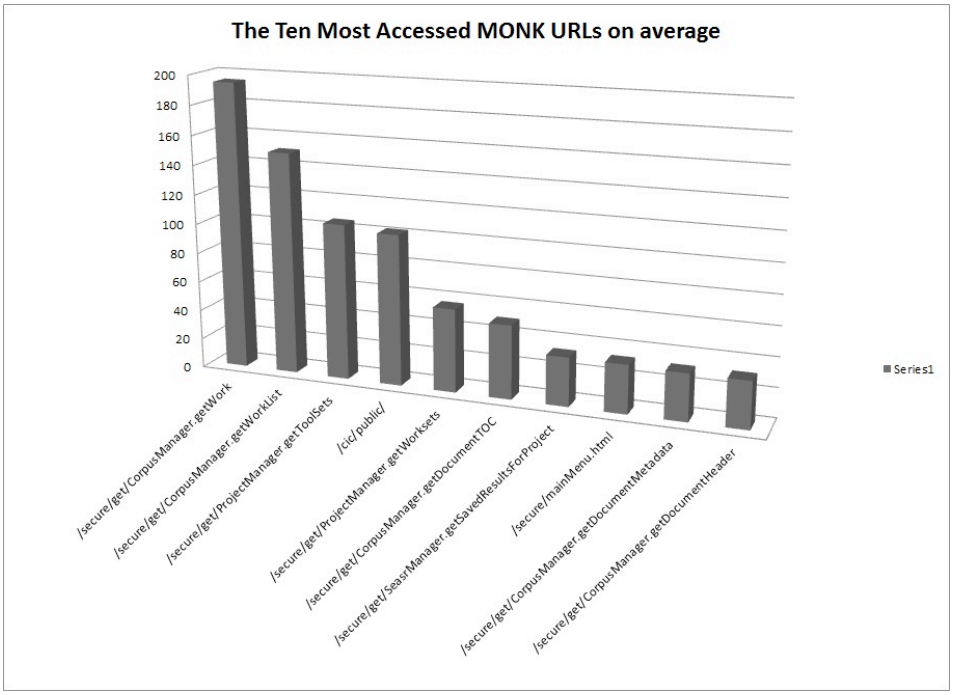

Figure 1. Ten most accessed MONK URLs.

One reason for the high numbers of hits of URLs with "/get/" is that when users begin work in MONK and click on the "Continue" button to conduct analysis with the selected toolset and workset, the **/secure/app/workflow** function is launched and calls in all "/get" URLs for MONK workflow functions. The dependency of all other URLs on **/secure/app/workflow** suggest implications for the number of hits on each URL: For every one **/secure/app/workflow** hit, there were 41 **/get/** hits and for every two **/secure/app/workflow** hits, there were 7 **/tool/** hits, pointing to a much higher call-in ratio for the **/get/** functions. And of the **/get/** functions, WorkbenchW constituted sixty-six percent of the **/get**/ URLs while WorkbenchA constituted thirty-three percent (Table 1).

WorkbenchW : secure/app/workflow	**workbenchA : secure/app/workflow**
2/3	1/3
/get/ URL : secure/app/workflow	**/tool/ URL : secure/app/workflow**
1/41	2/7

Table 1. Proportion of WorkbenchW and WorkbenchA URLs to "secure/app/workflow" function. But while WorkbenchW URLs had the highest frequency of users, the WorkbenchA URLs numbered the highest in the type of URLs called on during workflow processes, with 39 separate URLs or 44 percent of the URLs accessed (figs. 2 and 3).

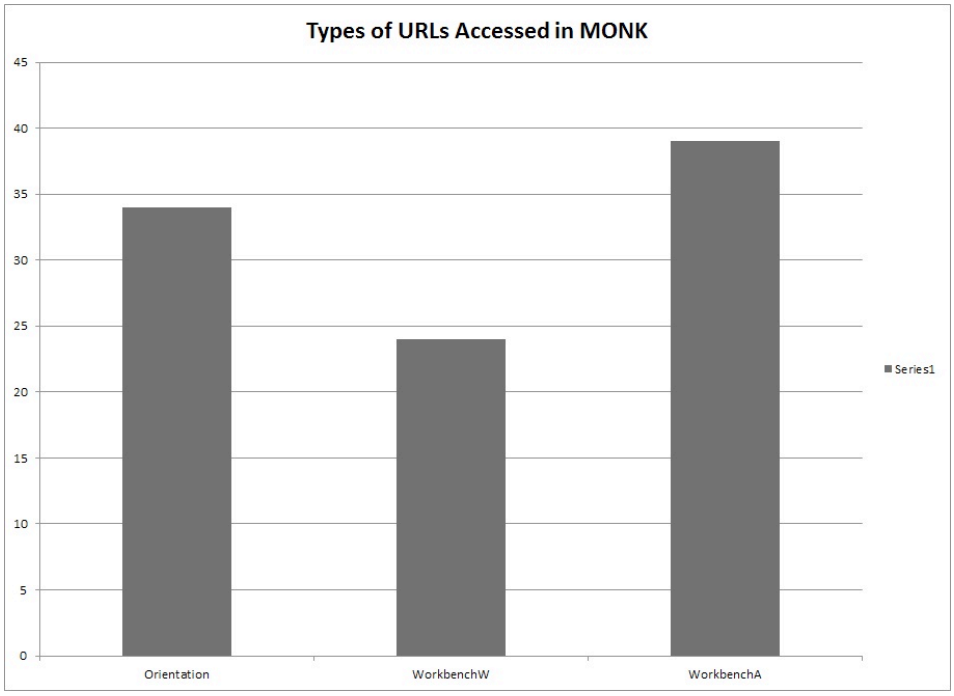

Figure 2. Types of URLs accessed.

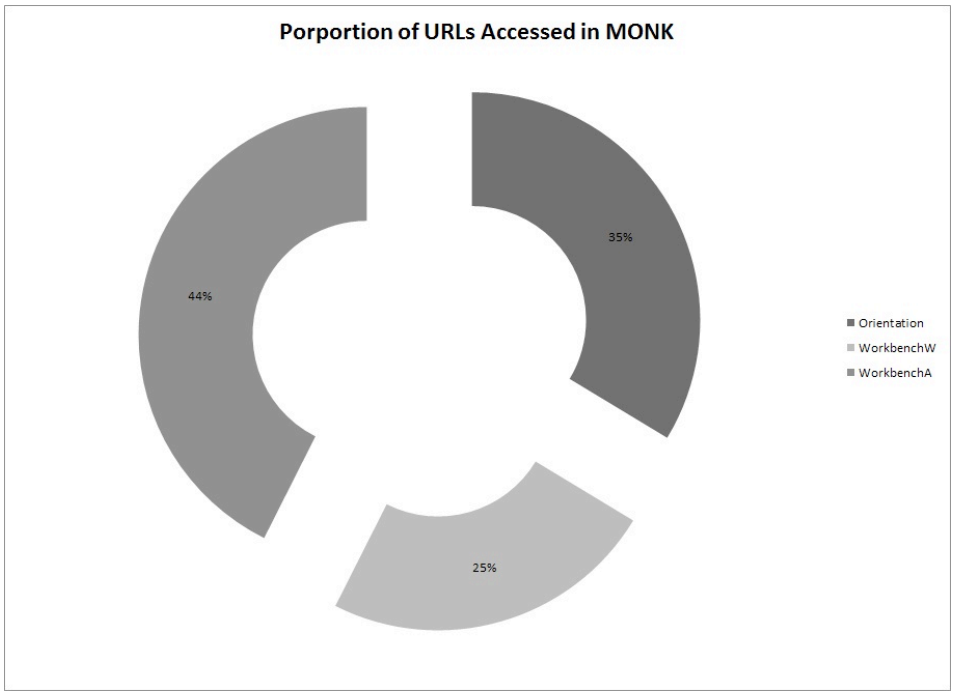

Figure 3. Percentage of URLs accessed.

Both categories also have prominent representation in the statistics for the kilobytes of data processed: The top ten URLs in amounts of processed data are predominantly from the WorkbenchW category, but **/secure/get/AnalyticsManager.compareFeaturesFrequencyDunning** in the WorkbenchA category

consumed by far the largest amounts of data at 293.55 KB on average per month (fig. 4).

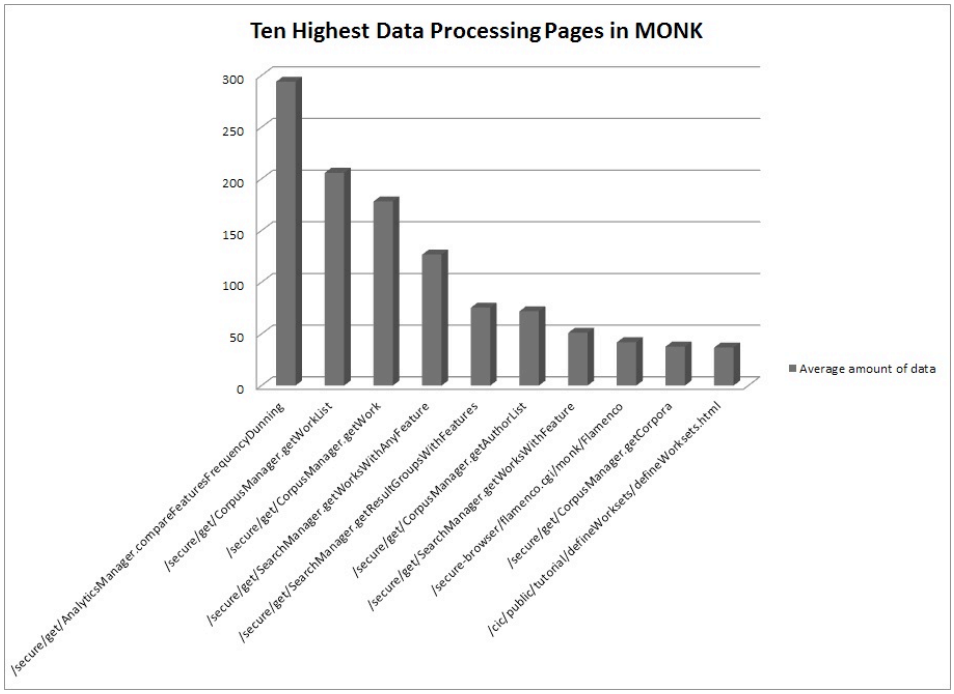

Figure 4. Data processing pages.

The data for the entry and exit visits revealed key points where users entered and left the tool: the most frequent entry point was the opening menu of MONK, **/cic/public**, with 35.8 entries on average over the twelve months. It was followed in frequency by **/ [https://monk.library.illinois.edu/]**, but this is a false positive due to the fact that it simply alternative URL for the opening menu of MONK. As such, the second most frequently entered page was actually **/cic/public/terms**, the webpage listing the terms and conditions policy of MONK, and was followed in frequency by **/cic/public/analytics/decisiontree.html**, a WorkbenchA category that may have been a starting point for users in the midst of MONK workflow (fig. 5).

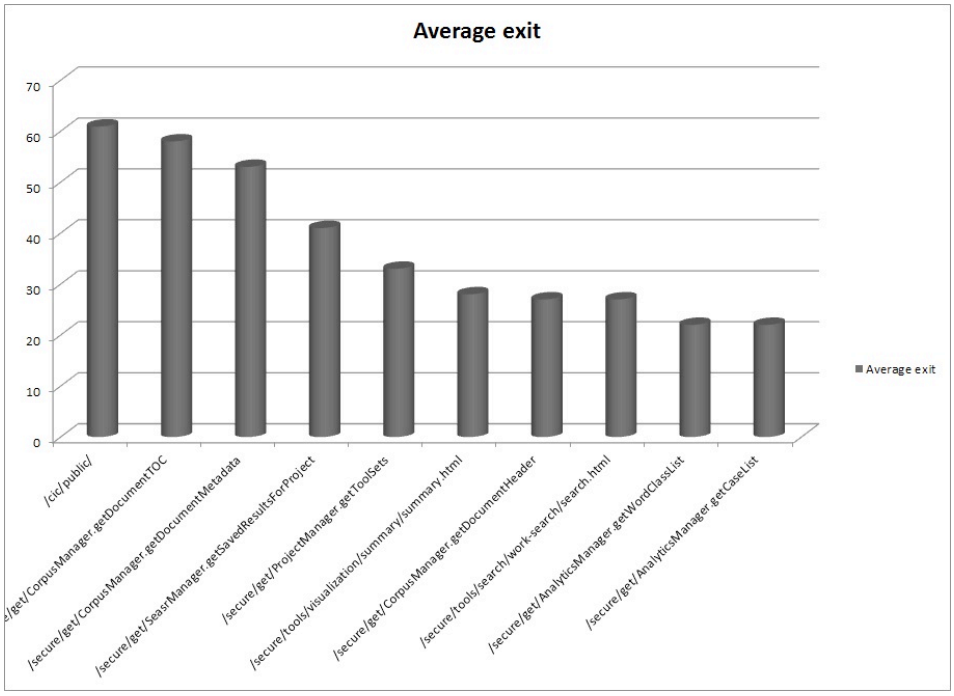

Figure 5. Entry frequencies.

The opening menu was also the most frequent exit point on average with 60.9 exits averaged over the twelve months, followed by **https://monk.library.illinois.edu/cic/public/terms** and **https://monk.library.illinois.edu/secure/public/analytics/clusterclassification.html**, a workbench analytic function for launching cluster analysis in MONK (fig. 6). These entry and exit data point to a high number of users entering the tool at the opening menu, and most frequently exiting early on in the analytic workflow process.

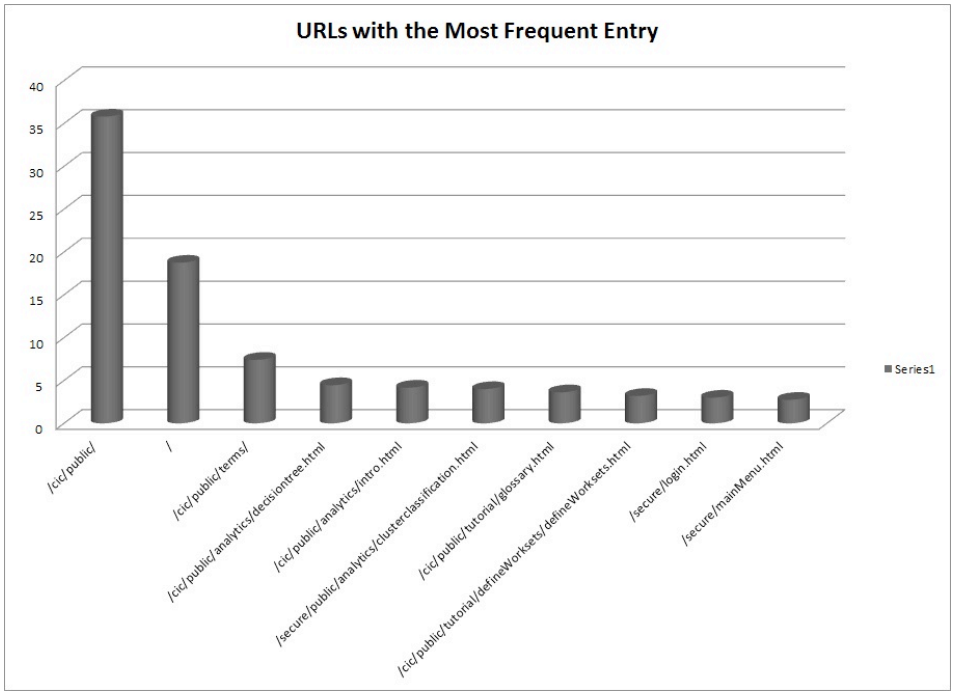

Figure 6. Exit frequencies.

6. Conclusion

This preliminary data analysis reveals that researchers are exploring the use of MONK as a text mining tool and working through the various levels of data analysis. Users explored and used MONK at varying levels and frequencies: they most often entered at the log-ins, browsed tutorials, and frequently began their workflows with the tools for compiling texts into worksets. And despite the contrasting lower numbers in usage of analytics tools, the most significant amounts of data were processed through these functions. This may point to a need to ensure that MONK will be able to handle data loads of text mining processes with increased numbers of users.

It should be noted there are several data points that AWStats did not record, such as deeper metrics for the proportions in URL hits, which would have been valuable for analysis of MONK usage. Another challenge was posed by the design of MONK itself, as its applications, functions and servlet URLs did not fit well with AWStats' method of recording statistics, producing a not insignificant amount of noise in the data.

Yet as the MONK usage data grows and deepens in complexity, our analysis should begin to critically reveal ways that the tool can be improved and revised to fit the research needs of users. And critical analysis of digital humanities tools such as MONK will enable us to examine the research processes of humanities scholars as they increasingly integrate digital tools and resources with traditional modes of scholarship.

7. Acknowledgements

Special thanks to Kirk Hess at the University of Illinois of Urbana-Champaign for his assistance with the data analysis.

Bibliography

Burrows, John. "Textual Analysis." In *A Companion to Digital Humanities*, edited by Susan Schreibman, Ray Siemens, and John Unsworth, chapter 23. Oxford: Blackwell, 2004. Accessed June 13, 2011. http://www.digitalhumanities.org/companion.

Clement, Tanya E. "'A thing not beginning and not ending': Using Digital Tools to Distant-Read Gertrude Stein's *The Making of Americans*." *Literary and Linguistic Computing* 23, no. 3 (2008): 361-381.

Duff, Wendy M. and Joan M. Cherry, "Use of Historical Documents in a Digital World: Comparisons with Original Materials and Microfiche." *Information Research* 6, no. 1 (2000). Accessed June 1, 2011. http://informationr.net/ir/6-1/paper86.html.

MONK documentation. Accessed March 8, 2011. http://monkpublic.library.illinois.edu/monkmiddleware/public/index.html.

Sculley, D. and Brad Pasanek. "Meaning and mining: the Impact of Implicit Assumptions in Data Mining for the Humanities." *Literary and Linguistic Computing* 23, no. 4 (2008): 409-424.

Sinclair, Stéfan. "Computer-Assisted Reading: Reconceiving Text Analysis." *Literary and Linguistics Computing* 18, no. 2 (2003): 175-184.

Warwick, Claire, Melissa Terras, Paul Huntington and Nikoleta Pappa. "If You Build It Will They Come? The LAIRAH Study: Quantifying the Use of Online Resources in the Arts and Humanities through Statistical Analysis of User Log Data." *Literary and Linguistic Computing* 23, no. 1 (2008): 85-102.

Yu, Bei. "An Evaluation of Text Classification Methods for Literary Study." *Literary and Linguistic Computing* 23, no. 3 (2008): 327-343.

Zillig, Brian L. Pytlik. 2009. TEI Analytics: converting documents into TEI format for cross-collection text analysis. *Literary and Linguistic Computing*. 24: 187-192.

Source URL: http://jdhcs.uchicago.edu/
Published by: The Division of the Humanities at the University of Chicago