Interim Project Report
Grant number: PW-50557-10
Title of Project: Pleiades: Content and community for ancient geography
Name of Project Director: Thomas R. Elliott
Name of Grantee Institution: New York University
Date Report is Submitted: 30 November 2013

This report covers the period 1 May 2013 - 31 October 2013

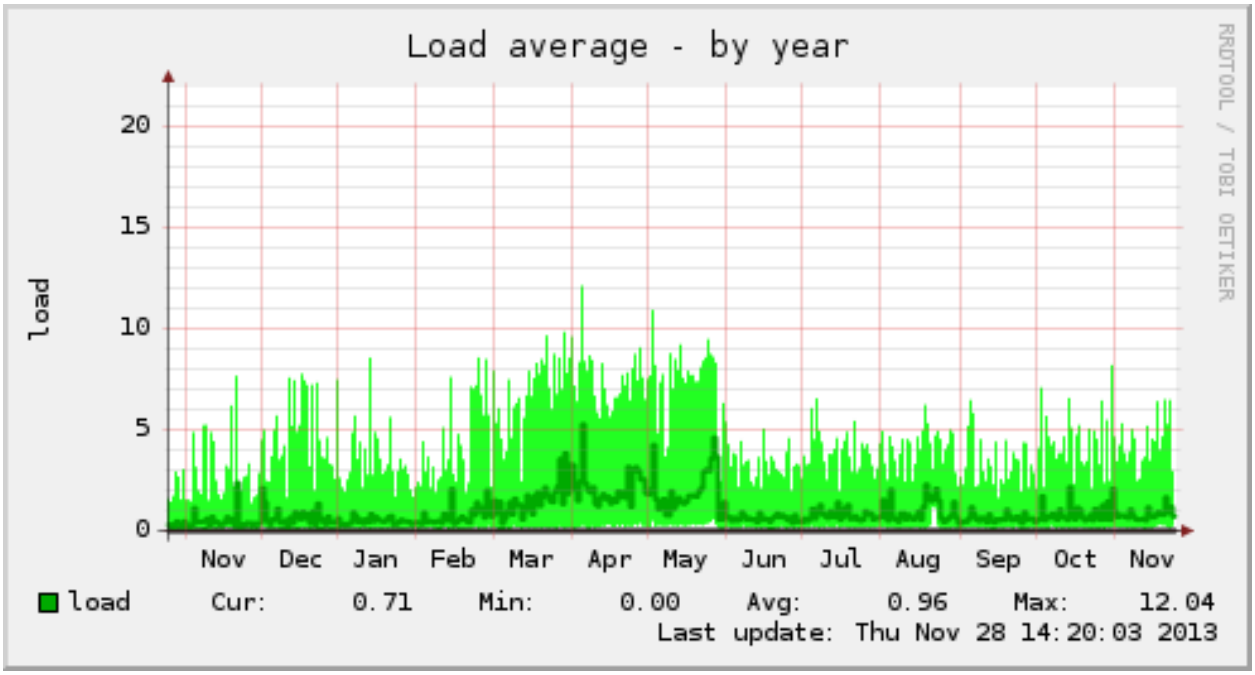
The Pleiades Project team is pleased to report steady growth in use of the Pleiades system and maturation of its content.

During the reporting period, 28,490 people visited the Pleiades site (a 21% increase over the preceding period). Nearly 33% of these were repeat visitors, such that the site logged almost 41,000 discrete visits (a 10% increase). On average, our visitors viewed 3.98 pages and spent an average 4:22 minutes on the site for each visit. Average visit duration is down by just over 12% from the preceding period; we believe this is directly attributable to site performance improvements that we began implementing in the previous period and to the tuning of which we have devoted significant effort this period (see below).

During the reporting period, 26 people requested and were given login credentials on the site, enabling them to contribute content (bringing the total number of Pleiades contributors to 265). We published 4,029 new Pleiades information resources (214 places, 1,192 locations, and 2,623 names) as well as 1,752 updates (1,185 places, 361 locations, and 206 names). Of the additions, Pleiades picked up 106 places, 178 locations and 2,544 names provided by the NEH-funded *Gazetteer of the Ancient Near East (GANE)* project (derived from the index to the *Tübinger Atlas des Vorderen Orients*). Because of irregularities in the data and the complexity of temporal periodization inherent in the TAVO data, ingest of this valuable dataset proved much more complicated than expected, and consumed fully half of our budgeted software developer hours during the period of performance. Of the 1,192 new locations published during the reporting period, 890 were created by individual Pleiades users via the OpenStreetMap integration added during the previous period and described in our last report.

The editorial board continues its work to improve the Pleiades help documentation (http://pleiades.stoa.org/help). Guidance for citation of several common ancient texts (e.g., Ptolemy) has been added to the Citation Guide. Entirely new documents covering the creation and titling of resources, as well as the editing and review processes, have been drafted and are now in review.

We have continued to spend a significant amount of time trying to manage and improve site performance and reliability. Although we believe we are reaching the scale limits of the Plone content management system platform that underpins Pleiades, we nonetheless were successful during this period in mitigating some of the most disruptive performance problems on the site. At the end of May 2013, we restructured load balancing between Plone web application processes on our server, resulting in significant performance improvements. The following graph of average CPU loading on our server over the past year further illustrates the fundamental improvement. Load is calculated as a rolling average of the number of processes that are in the server's run queue (i.e., they are scheduled to run immediately, but are waiting for access to the CPU because other processes are currently running). Note the reduction in average load significantly visible at the beginning of June.



The following table illustrates the resulting gains during the two-month period after 10 June 2013 as compared to the preceding two-month period.

measure	4/10 - 6/9	6/10 - 8/10	% improvement
average page load time	9.76 sec	5.78 sec	40.79%
average server response time	4.17 sec	1.53 sec	63.42%
average page download time	0.41 sec	0.17 sec	57.83%

Viewed on a longer timescale, however, these improvements are less impressive (e.g., improvement in page load average during the period 1 June - 31 October as compared to the preceding 5 months is only 0.38 seconds (a 5% improvement). Our average site performance has been skewed by another problem: complete saturation of the site by web crawlers and other automated systems. We have accordingly implemented a series of changes down through the end of the reporting period as follows:

Our robots.txt file (cf. http://www.robotstxt.org/) now has a crawl-delay of 27 (i.e., it asks polite bots to wait a minimum of 27 seconds between each request for content on the site). It also tells them where we do and don't want them to go.

We have "connected" Pleiades to "webmaster" management accounts at Google and Bing. This allowed us to ask Google to also observe the 27-second rule (they ignore the crawl-delay directive). They are complying. Bing doesn't have a similar option, but they do allow one to

specify periods of the day for greater or lesser indexing. We have taken advantage of this to try to reduce their engagement with our site during daytime hours in Europe and North American, whence most of our traffic.

Several bots, spiders, and crawlerss are banned from access to our search interface, user publication lists, and json resources because all of these are expensive, in terms of processing time, to produce. This includes anything with the substrings "bot", "crawler", or "spider" in its user-agent string. Some other bots, spiders, and crawlers have now been banned outright from access to any content on Pleiades because they have a history of pummeling our site at higher rates than our specified crawl delay during periods of heavy site loading or of engaging in attempts to probe for vulnerabilities in various web frameworks. These include, unfortunately, the big Chinese search engine, Baidu.

We have also tweaked instructions for server- and client-side caching of some of our site resources to reduce the frequency with which "friendly" web browsers and bots ask for content that rarely changes (e.g., site themes via CSS).

Monitoring software installed on the site now watches all aspects of the server infrastructure for the unresponsiveness characteristic of high-load periods and selectively restarts server components in an effort to clear such "log jams." Repeated restarts of this kind trigger an automatic notice to users, via http://pleiades-site-status.tumblr.com/, as well as emails and SMS messages to the managing editors so that they can intervene manually as necessary.

It is our impression that these combined steps will have improved site performance and the user experience for most users; however, statistical confirmation will require a few additional months of data. We remain convinced that migration out of the Plone content management system is an essential next step if Pleiades is to continue to expand and to improve to meet users' expectations. Unfortunately, the proposal to the NEH Office of Digital Humanities Phase 2 Startup Grants program on which we reported previously was not funded. We continue to seek opportunities to expand and improve Pleiades content and services.

In September 2013, Sean Gillies (our Chief Engineer and the architect of Pleiades' technical underpinnings) declined to apply for a new, expanded position at NYU and was laid off because the Pleiades 2 grant was sufficiently spent down that it could no longer support his dedicated work on the project. Gillies continues as an uncompensated, volunteer editor on the project, but has relinquished all technical responsibilities. Elliott now has sole technical responsibility for Pleiades.

We were unable to complete our promised review of Pleiades editor Adam Rabinowitz's experience using Pleiades with his undergraduate classical archaeology class. This assessment has been deferred to the end-of-project report.