

DH Box: A Digital Humanities Laboratory in the Cloud

Level II NEH Office of Digital Humanities Start-Up Grant Proposal

Data Management Plan

The data generated by DH Box

Purpose and computing architecture

DHBox is a service which provides and configures [Docker](#) containers for digital humanities students, professors and researchers. This spreads the cost of configuration and installation of Free and Open Source (FOSS) digital humanities tools over a very large audience and only a single physical machine. Since students access DHBox via their browser, this obviates the need for each student to undergo what is known in the FOSS culture as the 'larval stage' of hacking (defined by [Jargon file](#)).

Describes a period of monomaniacal concentration on coding apparently passed through by all fledgling hackers. Common symptoms include the perpetration of more than one 36-hour hacking run in a given week; neglect of all other activities including usual basics like food, sleep, and personal hygiene; and a chronic case of advanced bleary-eye.

It is suspected that many potential users abandon free software because of the difficulties in installation and creation of the development environment. This is compounded by the fact that many critical FOSS projects have incorrect and incomplete documentation in three critical areas new features, installation and configuration¹. DHBox will generate information both about usage and abandonment so that this phenomena can be studied. Rustard (2011) also argues that this is likely a reason for lower participation on open source projects of both women and people of color.

Specifically, a docker container creates a specialized learning environment for each account or group of accounts with all of their data. Each docker container keeps libraries separate so that results can be replicated. However docker containers share one operating system kernel so that containers do not have to replicate multiple gigabytes worth of system libraries.

Data generated

The DHBox project generates two broad types of data, code and usage logs. The next section will address each of these in turn.

Code

¹See Rustard page 7.

DHBox is a platform. Creating and configuring the platform is automated through Ansible scripts. All of these scripts will be public except for security keys to maintain privacy. The DHBox website is written in Ruby on Rails and will be maintained in separate public repository on github. Version control is maintained via [git](#).

Usage data

Each of the client instances is a DHBox container rather than a virtual machine. This means that each resource it asks for file write lock, access to a program or library or read from an external data source is delivered from the linux kernel that resides on the DHBox server. By carefully logging these requests we will maintain a complete record of the client's usage. Obviously, this will be disclosed to users in the terms of service. These logs will be stored internally continuously and backed up to an external location in encrypted form nightly.

We anticipate analyzing the number of initial downloads, daily users, distribution of memory used by user and group, application usage, application usage time, application usage cycles, packages installed, data added by users, help request volume, help request author, help request topic, feature request volume, feature request author and feature request topic.

The data will allow us to answer the following research questions.

1. At what points do users have trouble accessing DHBox?
2. Which support topics are most often viewed during the installation process?
3. Which FOSS software packages are most popular?
4. When do users stop using DH Box?

Data plan

Data retention and exclusivity

There will be no period of exclusive use. Data will be public from the first month. Data will be retained for at least five years. If funds are available data will be retained in perpetuity.

Repository

Data will be updated monthly and stored in perpetuity at [Floss mole](#). This site is supported by NSF grants NSF Grants 07-08437 & 07-08767 for the study of open-source and collaborative software.

Metadata

Metadata will be stored using the [Text Encoding Initiative](#) (TEI) XML encoding. Metadata will be stored in English and in compliance ISO 639-2 in order to make these data more easily read by machines. It is anticipated that the program will join [Dublin core](#) which is the largest federation of digital humanities projects using metadata. The project will seek to comply with all emerging standards that are feasible given the size and budget of the DHBox project.

Compliance, departures and ownership

The team member in charge of data plan compliance is Evan Misshula (Data Manager). Backup will be provided by Micki Kaufman (Project Manager). The Project Director Mathew K. Gold (PD) will receive a report that data has been uploaded to the repository by the fifth day of the month. If the data is not uploaded and in compliance, the PD shall take appropriate corrective action including, if necessary, personnel changes. The project will remain at the Graduate Center of the City University of New York (CUNY-GC). Since the project is sponsored by the GC Digital Fellows numerous qualified graduate students are available to meet anticipated departures of staff. The principal investigator Mathew K. Gold is a tenured member of the faculty, as well as the director of the GC Digital Scholarship Lab where the project will be housed.

Dissemination

Data necessary to assess the success of the project (ie installs, usage patterns, and terminations) will be shared as tab-separated ASCII files (tsv) to facilitate analysis by FOSS software such as R and Python. Data will be captured from system log files and transformed. Publicly available data will be anonymized. Individual level data will be retained on site and encrypted but used for security and performance audits. It is not anticipated that this data will be used for published research from the project. However, we will facilitate its release to qualified research groups both internal to CUNY and external. However, the project requires that such groups remit funds in advance to cover the cost of all Institutional Review Board (IRB) compliance.