9. Data Management Plan

A major drive of the proposed work is to establish a new platform for intertextual search over large collections of digitized documents. While our proposed work includes our own initial forays into this new space, we anticipate the greatest possible impact will be achieved by fully engaging and leveraging a much broader community of investigators in the spirit of “open science”. As such, a significant component of the proposed work is the creation and development of publicly-sharable material — both datasets and software.

This plan for the NEH is largely inspired by the National Science Foundation’s policy on the dissemination and sharing of research results within a reasonable time. In accordance with this policy, this plan does not include preliminary analyses (including raw data), drafts of scientific or humanistic papers, plans for future research, peer reviews, or communication with colleagues.

Furthermore, data to enable peer review and publication/dissemination and/or to protect intellectual property may be temporarily withheld from distribution and other proposed data management. This plan will make certain that the data produced during the period of this project is appropriately managed to ensure its usability, access and preservation.

Deliverables

Upon its completion, the project will make available the following six deliverables:

1. The core RESTful software service code package based primarily in Python, hosted on the Tesserae Github repository. The service will allow users to input two or more texts for comparison and a set of parameters, and return a list of parallel passages with the similar words or other language features marked. The languages serviced will be ancient Greek and Latin.

2. An operational version of the service hosted at UB, Notre Dame, and UCCS, to be used by the Tesserae project, as well as partner digital collections and interested maintainers of other digital collections.

3. A plugin to the Plokamos annotation framework for making the core Tesserae service compatible with its operation. This will be published on the Tesserae Github repository and presented to the Plokamos project for inclusion in its code base.

4. Interface code for making the core Tesserae service compatible with the Perseids collaborative editing platform via Plokamos. This will be published on the Tesserae Github repository and presented to the Perseids project for inclusion in its code base.

5. Interface code for making the core Tesserae service available to Open Philology, the Perseus Digital Library, and the Digital Latin Library via Perseids. This will be published

on the Tesserae Github repository and presented to each collection for inclusion in collection-specific code bases.

6. Documentation of the new software service, interface modules, and Tesserae front-end, both accompanying the code on Github and hosted on the Tesserae website.

Data and Code Sharing Timeline

We anticipate regular releases of data and code as they are completed. In keeping with the overall schedule presented in the timeline found in the proposal narrative, we anticipate the release of the TIS, TIS interface to Perseids, and the TIS interface to Plokamos to take place in the second year, ideally coinciding with the initial publication of results. We anticipate the initial releases of the interfaces to the Open Philology Project, Perseus Digital Library, and the Digital Latin Library at the end of the second year. Algorithm development will track the progress of each task, with stable code releases at the end of each project year, along with scientific papers describing our advances in intertextual search.