

Final White Paper

Tesseract Intertext Service: Intertextual Search Access to Digital Collections in the Humanities

NEH Digital Humanities Advancement Grant HAA-258767-18

University of Notre Dame (Notre Dame, IN 46556-4635)

Walter J. Scheirer (Project Director: June 2017 to July 2020)

<https://www.wjscheirer.com/>

University at Buffalo (Buffalo, NY 14260-1660)

Neil Coffee (Co Project Director: August 2017 to July 2020)

<https://arts-sciences.buffalo.edu/classics/faculty/core-faculty/coffee-neil.html>

Project Activities and Accomplishments

For as long as there has been literature, there have been threads that link literary works to each other. Ideas from one work are referenced, built upon and sometimes simply taken by the writers that come after. Shakespeare drew from the work of the ancient Greeks; countless writers from Herman Melville to Margaret Atwood draw from Shakespeare's writing for their own creations. Pulling on these threads and following them from one work to another allows us to gain a greater depth of understanding of the original work and all of the subsequent works that link back to it.

The work of finding these links had long consisted of manually comparing two documents and finding common words, phrases or ideas. The Tesseract Intertext Service (TIS) uses automated detection of language similarity to allow researchers to find these links in a fraction of the time and in new and unexpected ways. Its ability to find instances of exact repetition of words or phrases, similar meaning and even the sound of different words or phrases allows researchers to explore how multiple texts are connected.

TIS is the culmination of over a decade of work. It initially began as the Tesseract search engine for locating allusions in Classical texts (<http://tesseract.caset.buffalo.edu/>). Researchers could visit the site to compare a limited number of pre-indexed Greek, Latin, and English texts. While this was considered a valuable tool for researchers in the humanities, it was limited in the number of texts that were available for search and could only be accessed through the original Tesseract website.

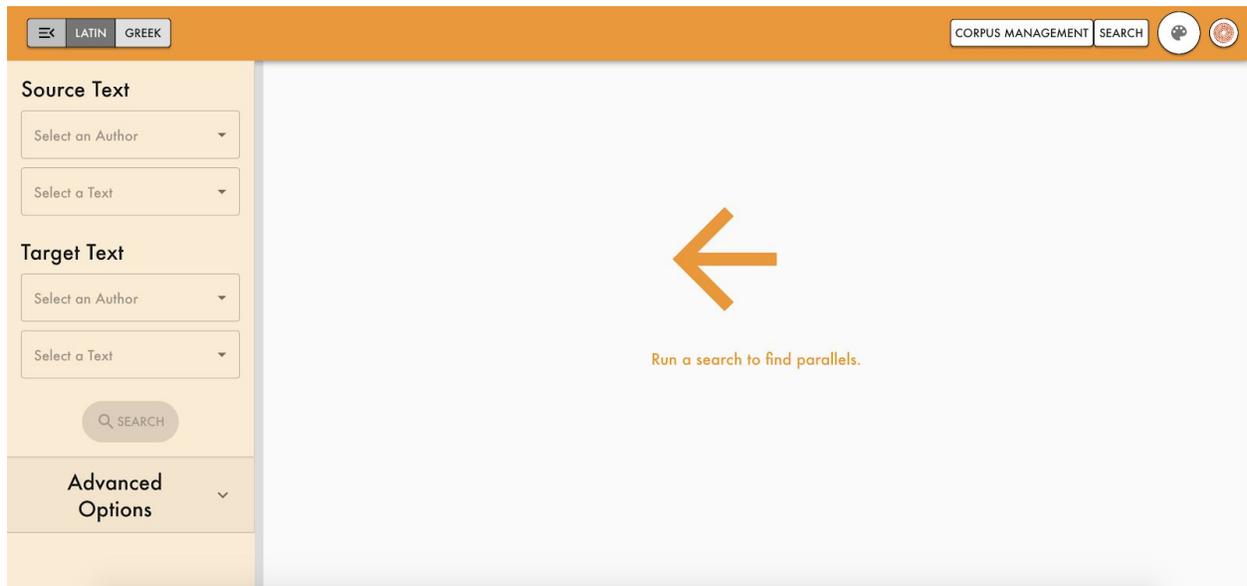
The work that has been supported by this National Endowment for the Humanities grant has allowed our team to transform the original website into a tool that can now be integrated for search into any digital humanities collection. With the new TIS downloadable version, individual users can even use the tool on texts they choose on their own machines.

With the support of the NEH, our team has transformed Tesseract into an Application Programming Interface (API). The TIS-API is a set of rules that allow other programs to use TIS as a resource within the program's own environment. This greatly expands access to Tesseract and exponentially increases the number of texts that can be searched.

Tesseract works by comparing two documents based on the parameters set by the user. First, the text considered to be the source of an idea or phrase, or the **Source Text** is chosen. The text that is believed to be influenced by the Source Text, the **Target Text**, is then chosen. Users can then specify what language feature they want Tesseract to compare on. They can choose to compare two texts for exact word matches, matching lemmata (root words), words of similar meaning, or words with similar sounds. They can then choose a **Stoplist** to ignore results with common words that would likely be uninteresting.

Our work has focused on the humanities, but any field that can benefit from illuminating the relationships between texts and ideas over time would benefit greatly from having such a powerful tool at their disposal.

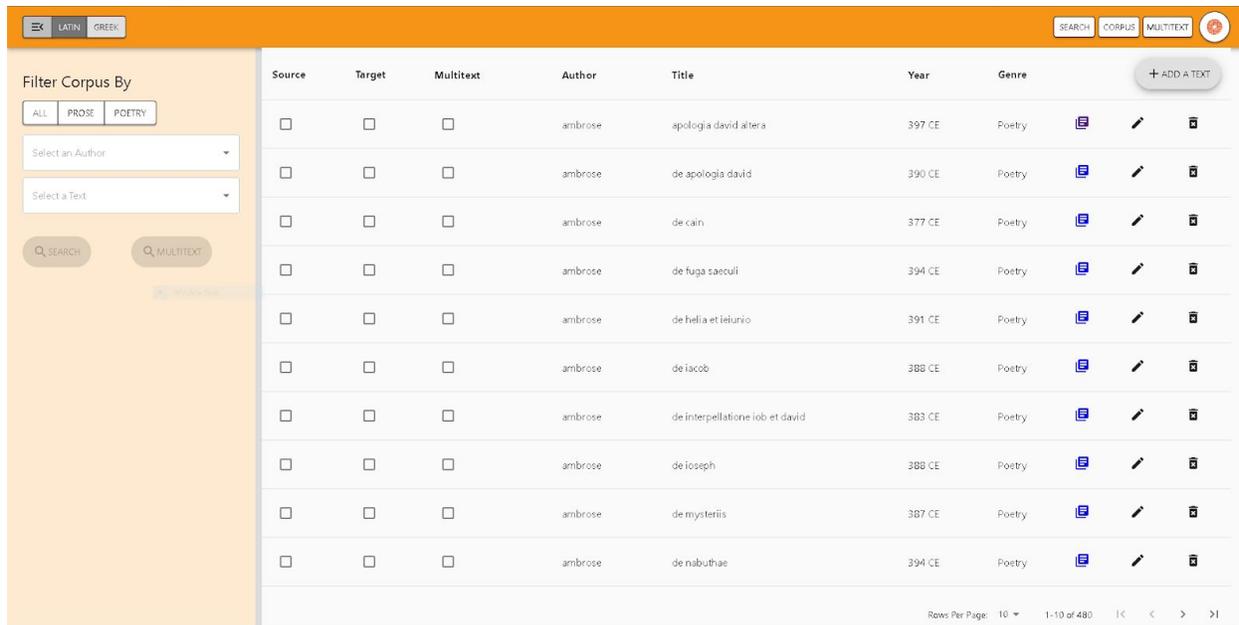
TIS provides the backend search functionality for the newly redesigned Tesseract website (Version 5.0), available at <https://tess-new.caset.buffalo.edu/front/>. This new website provides all of the features of the previous stable version, but with a more maintainable Python codebase for the backend, a database for text and result storage, a REST API for common access over the Internet, and a modern ReactJS frontend. This website offers a new stable version for Tesseract's web users, and also serves as a demonstration for the TIS API.



The newly redesigned Tesseract Version 5 website.

	Source	Target	Match Features	Score ↓
1	vergil aeneid 2.235: Accingunt omnes operi pedibusque rotarum subiciunt lapsus et stuppea vincula collo intendunt.	lucan bellum civile 10.493: Nec piger ignis erat per stuppea vincula perque Manantes cera tabulas	stuppeus, vinculo, vinculum	10
2	vergil aeneid 11.199: Tum litore toto ardentis spectant socios semustaque servant busta neque avelli possunt nox umida donec inverit caelum stellis ardentibus aptum	lucan bellum civile 9.3: Prodit busto semustaque membra relinquens Degeneremque rogam sequitur convesa Tonantis Qua niger astriferis connectitur axis aer Quisque patet terras inter lunaeque meatus (Semidei) manes habitant quos ignea virtus innocuus vita patientes aetheris limi Fecit) et aeternos animam collegit in orbes	semustus, bustum	10
3	vergil aeneid 7.516: audit et Triviae longe lacus audit amnis sulfurea Nar albus aqua fontesque Valini et trepidae matres pressere ad pectora natos	lucan bellum civile 1.473: Est qui tauriferis ubi se Mevania campis Explicat audaces ruere in certamina turmas Adferat et qua Nar Tiberino illabitur amnis Barbaricas saevi discurrere Caesaris alas	amnis, nar	10
4	vergil aeneid 9.715: tum sonitu Prochyta alta tremit durumque cubile Inarime lovis imperiis imposta Typhoeo Inarime lovis imperiis imposta Typhoeo	lucan bellum civile 5.100: Campana fremens ceu saxa vaporat Conditus Inarimes aeterna mole Typhoeus	typhoeus, inarime	10
5	vergil aeneid 3.291: Proflnus arias Phaeacum abscondimus arces litora que Epiri legimus portuque sublimis Chaorio et celsam Euthroti accedimus urbem	lucan bellum civile 5.418: Hic utinam summi curvet carthesia mali Incumbatque furens et Graia ad moenia perflet Ne Pompeiani Phaeacum e litori toto Languida lactatis comprehendant carbasa remis	phaeaces, phaeacus, litus	10
6	vergil aeneid 4.40: Hinc Osetulae urbes genus insuperabile bello et Numidae infreni cingunt et inhospita Syris	lucan bellum civile 1.367: Duc age per Scythiae populos per inhospita Syris Litora per calidas Libyae sitentis arenas	syris, inhospitus, inhospita	10
7	vergil aeneid 11.403: Nunc et Myrmidonum proceres Phrygia arma tremescunt nunc et Tydides et	lucan bellum civile 2.405: In laevum cecidere latus velocoque Metaurus Crustumiumque rapax et iunctus	undo, hadriacus, unda	9

View of a Tesseract source / target search.



The Corpus Management interface is a new addition to the Tesseract website. Using the filter on the sidebar, users can browse the corpus and even set up standard and multitext searches using the checkboxes.

Audiences

The primary audience for the Tesseract project has been Digital Classicists from around the world. The previous Tesseract website has been visited by over 12,000 users internationally.

As part of the TIS outreach effort, the Tesseract team delivered a presentation at *DH 2020*, a virtual digital humanities conference held in July 2020. In this virtual poster presentation, the team introduced the TIS web service that allows partner collections to integrate Tesseract search directly into their web applications. The abstract for this presentation is available on the DH 2020 website:

https://dh2020.adho.org/wp-content/uploads/2020/07/155_IntegratingIntertextualSearchintoYourWebApplicationTheTesseractIntertextServiceAPI.html

The team worked with two partners for testing the TIS functionality. Both Archimedes Digital and the Quantitative Criticism Lab at the University of Texas successfully integrated TIS into their text reuse detection workflows. Archimedes Digital is a start-up company helping digital humanities projects throughout the world with software and

data support. The Quantitative Criticism Lab is an academic research group working on the study of text reuse in literature. The Tesseract development team worked with both organizations to train them in the TIS and deploy the TIS within their software. Feedback from both organizations was incorporated into the first release of TIS. Word of the utility of TIS is already spreading. Subsequent to these tests, the Tesseract team received an inquiry from the Perseus Project regarding the possibility of integrating the TIS API into its Scaife Text Viewer, which will serve as the principal way of viewing and interacting with Perseus texts.

Six graduate students participated in the project, five at the University at Buffalo (Cari Haas, Laura Hambridge, Tessa Little, Nozomu Okuda, and Stephanie Richter) and one at Notre Dame (Jeffery Kinnison). All students received training in software development and literary criticism.

Evaluation

With respect to a quantitative evaluation of the TIS-powered backend of the new Version 5 website, benchmarking was performed assessing results returned, runtime and memory usage. As a continuation of the Tesseract project, it is expected that Version 5 will return results similar to Version 3 while maintaining or improving upon computational resource usage. Similar sets of top matches are returned by the legacy Version 3 system and the Version 5 system, thus preserving the core matching functionality, while enabling new features through the TIS, as well as improved recall in the total number of results. As we have done in previous evaluations, we used a source / target search of Vergil's *Aeneid* and Lucan's *Bellum Civile* as a baseline for benchmarking. Runtime is essentially the same compared to the legacy Tesseract codebase, and significant improvements were observed in memory usage. When results are cached into the database, the system runs even more efficiently.

Method	Matches Returned	Running Time (s)	Max Memory Usage (GB)
v3	66,609	19.14	2.90
v5	80,160	22.02	0.50
v5 from cache	80,160	3.51	0.19

Benchmark numbers for the new Tesseract Version 5 website compared to the legacy Version 3 site.

While the new Tesseract website demonstrates the advantages of the TIS API for members of the Tesseract Project, it was also relevant to consider how the TIS API could be used by people external to the project. For this purpose, we asked for feedback from groups outside of the Tesseract Project. Specifically, developers from Archimedes Digital and the Quantitative Criticism Lab at the University of Texas.

Feedback that we gathered following external assessment noted that “the API documentation was clear and easy enough to work through.” An external developer remarked that “the API produced results consistently in line with what I expected from the requests.” This suggests that people external to the Tesseract Project will have adequate documentation to work with the TIS API.

External evaluators expressed sentiment that the TIS API would have been useful in earlier work the projects performed: “This API would have been **extremely** useful in producing the data used for my 2017 JDMDH article, which relied on combining data from multiple Tesseract searches. If I were writing the paper now, I could, e.g., include a script using the API to better document the data collection process. I imagine this would also be the case for other Tesseract-data-heavy papers such as Bernstein, Gervais, and Lin 2015 in DHQ.”

External evaluators even went on to hold high hopes for future research with the TIS API: “I also imagine that the API, by reducing the difficulty of collecting and collating multiple searches will encourage more large-scale studies of this nature. I add that even for research making use of a single, simple search the ability to document the request in the form of an API request is valuable.” These comments were given in reflection on the current endeavors of the Quantitative Criticism Lab.

In addition to positive sentiments, criticism was provided for areas that could improve the TIS API with respect to reproducible research. Among the suggested changes are “including version information in json results” (of the Tesseract searches), “having clearer policies about cached data”, and presenting text snippets that “would make results more transparent and easier to work with in further data analysis.”

Nevertheless, one external evaluation concluded with the following statement: “I am looking forward to an official release so that I can incorporate the API into a Tesseract-based research workflow that currently relies on searches initiated through the web interface.”

Continuation of the Project

Tesseractae is a long-standing digital humanities project that continues to add resources to its staff, infrastructure, and funding. The Tesseractae team is currently pursuing new project support from a variety of federal agencies, private foundations, and internal resources. The first success in this new fund raising effort has been an award for the new Tesseractae Voyager project from the University at Buffalo. A \$10,000 OVPRED / HI Research Grant is enabling the initiation of the Tesseractae Voyager project by funding an RA to develop its first building block. The creation of Tesseractae Voyager will mark a new chapter in the digital study of text reuse by allowing for the standardized aggregation and exploration of connections between texts. Tesseractae Voyager will allow users to generate sets of parallel texts by selecting and saving them from Tesseractae searches, importing them from other text reuse search engines, collecting them from digitized published texts, or inputting them manually. Parallels chosen by a human being for ingestion will be considered “authorized” matches, as opposed to “found” matches suggested by search engines.

The Tesseractae project has also added new project staff for the current (2020-21) academic year. Graduate student Joseph Miller has joined the project at Buffalo, supported by a university fellowship. Stephanie Richter, a linguistics graduate student, is continuing on the project supported by the Tesseractae Voyager funding. Infrastructure support will continue to be provided by the University at Buffalo to keep the project servers running. Notre Dame is currently recruiting a new graduate student to focus on digital humanities projects, including Tesseractae.

Long Term Impact

The Tesseractae website is a well-established open source text reuse matching engine that provides an important resource to researchers in multiple disciplines attempting to answer questions about literary influence. Starting with two texts, Tesseractae discovers potential meaningful connections based on different features of the texts, and presents the user with a ranked list of parallels sorted by similarity score for further evaluation. The core matching engine is largely based on an understanding of intertextuality drawn from the field of Classics, but has been shown to apply to contemporary writing as well (see Prof. Scheirer’s book, listed below under award products, which was published during the project period). Tesseractae situates itself within a cluster of digital humanities projects considering the problem of detecting text reuse in literature. The TRACER project examines text reuse based on similarity, with an emphasis on crowdsourcing reuse detection. In a similar vein, *Musisque Deoque* incorporates an online tool for

lexical and metric similarity analysis in Latin poetry. Additionally, Filum finds similar but non-identical phrases in literary texts using a sequence alignment algorithm.

Tesseractae implements a distinctive set of tools designed to support accurate searches via intuitive web-based user interfaces. The core tools and data that exists in the current version of the software deployed on the live Tesseractae website include:

- Lexical Search via Lemma Matches
- Semantic Search via Dictionary Definition Matches
- Sound Search via Character-Level n-grams
- Multi-text Search
- Support for Multiple Languages (Latin, Greek, English)
- Tesseractae Corpus of Curated Open Access Texts

The core tools of Tesseractae have been used to produce a number of scholarly studies, and its user base numbers in the thousands. From students learning new languages to senior researchers looking for new parallels between texts, a diverse pool of users visits the Tesseractae website each and every day. An initial limiting factor was the scale to which the platform could be deployed. In particular, for much of Tesseractae's existence, text reuse matching was generally limited to matching on a single digital collection at a time. As the community continues to grow and collections interact more frequently, distributed studies that can operate over multiple collections will be increasingly in demand. Such studies require an approach that can both coordinate the match discovery process and transparently incorporate multiple collections through a common interface. Thus the current version of Tesseractae now supports standalone operation on a user's computer, as well as REST-ful API that allows other projects and collections to integrate Tesseractae search on their own platform.

Award Products

Award products consist of the new Tesseractae Version 5 website, API code and documentation, and project notes. All are available on the following websites:

Project website: <https://tess-new.caset.buffalo.edu/front/>

Project blog: <http://tesseractae.caset.buffalo.edu/blog/>

Source code repository: <https://github.com/tesseractae>

TIS API documentation: <https://tess-new.caset.buffalo.edu/docs/api/>

PI Scheirer published a book on quantitative methods for the study of intertextuality, which incorporates elements of this project:

Forstall, C. and W. Scheirer, "Quantitative Intertextuality," Springer Nature, 2019.
<https://link.springer.com/book/10.1007/978-3-030-23415-7>

Other articles that were published during the grant period:

Coffee, N., C. Gawley. 2020. "How Rare are the Words that Make up Intertexts? A Study in Latin and Greek Epic Poetry." In: Coffee, N., et al (eds.). *Intertextuality in Flavian Epic Poetry*. Berlin, Boston: De Gruyter.

Coffee, N., C. Forstall, Lavinia Galli Milić, and Damien Nelis (eds.). 2020. *Intertextuality in Flavian Epic Poetry*. Berlin, Boston: De Gruyter. [Volume intro]

Coffee, N. 2019. "Intertextuality as Viral Phrases: Roses and Lilies." In Monica Berti (ed.). *Digital Classical Philology: Ancient Greek and Latin in the Digital Revolution*. De Gruyter, pp. 177-200. <https://doi.org/10.1515/9783110599572>.

Coffee, N. 2018. "An Agenda for the Study of Intertextuality." *Transactions of the American Philological Association* 148.1. Pp. 205-223.
<https://doi.org/10.1353/apa.2018.0008>.

Drafts of two additional papers on the TIS and Version 5 software were produced by the end of the grant period, and will be submitted for publication by the end of the year.