Journal of Open Humanities Data

Perseus and the Open-Source Ecosystem of Greco-Roman Studies

COLLECTION: LANGUAGE DATASETS REUSE: OPPORTUNITIES, CHALLENGES, AND BEST PRACTICES

DISCUSSION PAPER

GREGORY CRANE
ALISON BABEU
DAVID A. SMITH

*Author affiliations can be found in the back matter of this article

ubiquity press

ABSTRACT

This article provides a short overview of several key areas of digital classical scholarship where different types of open data from the Perseus Digital Library have been reused, enhanced and in some cases been reused by Perseus in turn. After providing some brief context regarding the history of and need for open data in the study of texts in classical languages it explores three areas of the reuse of Perseus data: TEI-XML texts, morphological and treebank data, and named entity data.

CORRESPONDING AUTHOR:
Gregory Crane
Department of Classical Studies, Tufts University, United States
gregory.crane@tufts.edu

KEYWORDS:
digital classics; data reuse; Greco-Roman studies; open data; treebanks; named entities

TO CITE THIS ARTICLE:
Crane, G., Babeu, A., & Smith, D. A. (2026). Perseus and the Open-Source Ecosystem of Greco-Roman Studies. *Journal of Open Humanities Data*, 12: 120. DOI: https://doi.org/10.5334/johd.595

1. CONTEXT AND MOTIVATION

Transparent, reproducible scholarship based on Greek and Latin sources requires open data. This paper describes an ecosystem that has emerged to provide that open data. We focus in particular on how texts from the Perseus Digital Library (PDL, 2026), formally released with a Creative Commons license in 2006, have provided a core upon which other projects have built, creating new open resources that enhanced the original Perseus content in a variety of ways.

While an enormous amount of Greek and Latin was converted into digital form in the 20th century, these collections circulated under license agreements that made it impossible to redistribute enhanced versions (e.g., texts to which TEI XML markup or linguistic annotations had been added). Producing openly licensed digital transcriptions of Greek and Latin texts had proven to be an enormous challenge. Few funders were willing to support digitizing Greek and Latin sources that were already available, however restrictively, in digital form. Much of Digital Classics involves the reuse, enhancement, and redistribution of digital texts and thus depends upon openly licensed textual materials (Crane, 2016; Berti, 2019; Middle, 2024; Farina et al., 2025). Proprietary search systems were easy for traditional classicists to understand – in fact, with core textual data locked up behind paywalls and restrictive agreements, there was little opportunity and less pressure for classicists to draw upon results from fields such as corpus and computational linguistics.

Researchers on Ancient Greek recognized at a very early stage that they needed digitized texts. This early recognition and bold work proved to be helpful for print culture research but deeply problematic for truly digital work. The initial model for the foundational digitized text collections for Greek and Latin led to a legal and technical framework that enhanced traditional print scholarship but imposed profound barriers to born-digital research.

Generations of scholars had labored to build a print culture infrastructure, one that included, for example, manually produced concordances for works such as the Homeric Epic. Such tools were so valuable that individuals labored for years to produce and publish each. In the preface to his 1875 *Concordance to the Iliad of Homer*, Guy Prendergast reports that work on the concordance extended from June 5, 1847, to October 18, 1863, – more than sixteen years (Prendergast, 1869–1875). Printing this work began almost five years later in July 1868 and was only completed in November of 1874, more than six years later. More than 27 years passed from the time work on the concordance began until its ultimate publication. A century later in 1975, when the lead author of this piece was a first-year student, that print concordance remained arguably the most important tool in Homeric studies – and not just for specialists. Gregory Nagy taught his students who did not know Greek how to pair that concordance with the line-by-line Lattimore translations (Lattimore, 1951, 1967) so that they could see how Greek words were actually used in context.

The importance of the concordance inspired the foundational project of Digital Classics: David Packard's *Concordance to Livy*, a multi-volume print concordance that Harvard University Press published in 1968. By today's standards, the production of this concordance was extraordinarily labor intensive: Packard and his friends keyed the entire text of Livy onto punch cards and then used the computing power available at the time to produce data for print publication. Where the Homeric *Iliad* contains c. 115,000 words, there are more than 500,000 – c. 517,000 – words in Livy's *History of Rome*. Difficult as it was to type the text of Livy, the collaborators only needed to enter the text once. The computer could then produce a keyword-in-context concordance that repeated the text of Livy many times over.

When Classicists gathered in 1972 to plan what would become the *Thesaurus Linguae Graecae* (TLG), they imagined a computerized concordance that could serve any Greek author. A decade later, in 1982, Harvard Classics Professor Albert Henrichs, who had participated in the 1972 meeting, told his students that they could send a letter to the TLG asking them to search for a word in a given author (he suggested the 225,000 – word *Description of Greece* by Pausanias as an example) and expected to receive a printout by mail in a few weeks.

The concordance model enabled a structure that would have consequences until the present time. The TLG became a monolithic, centralized resource that used computing technology to generalize concordance services to the entire surviving corpus of Ancient Greek (Brunner, 1991). Researchers could scan much more content than any print concordance or index could provide and could search for combinations of words but, in the end, researchers would work

with the results that the centralized resource provided. For a period of roughly 15 years, the TLG shared its data, first on magnetic tape and then on CD ROM, and made it possible for third parties to create their own search engines (e.g., Diogenes (Heslin, 2007)). The TLG, however, always insisted that access to its data be restricted to those who paid the TLG for a license. In the 1990s, the TLG stopped sharing its collection and insisted on being the sole source for researchers. The TLG began to track usage and block institutions that, in its view, were downloading too much material.

The Packard Humanities Institute published a digitized corpus of Classical Latin (PHI) in 1985, distributing it on CD ROM. A license agreement also formally prevented users from redistributing the PHI Latin texts, but the PHI did not charge licensing fees, and these restrictions were not rigorously enforced. When Perseus began adding Latin in the late 1990s, we went to great lengths to digitize Latin works directly – often re-digitizing editions that were already available in PHI. By the 2010s, however, PHI texts began to circulate as part of third-party Latin corpora (and they were also providing free online access (PHI, 2011) to the collection as well). The EU allows its member nations to assert copyright over reconstructed text of scholarly editions for up to 30 years after publication and so the European editions in PHI (which are the vast majority) are now in the public domain. All proprietary collections of Greek and Latin therefore contain, at the least, substantial amounts of public domain content. Contract law (i.e., the agreements that institutions and individuals sign) restricts use of these materials, whether or not the materials are in the public domain.

The lead author of this paper was confronted with the problem of restricted data when he began working with TLG texts in 1982. While no funders were eager to support re-digitizing Greek and Latin, Perseus would quietly include digitization within larger projects. By 2013, we had accumulated 8 million words of Greek and 5 million words of Latin. Support from the Alexander von Humboldt Foundation made it possible for the lead author to establish the collaborative Open Greek and Latin (OGL) project at Leipzig University (Köntges et al., 2019). The Harvard College Library, Harvard's Center for Hellenic Studies (under the leadership of Gregory Nagy), the University of Virginia Library and the Perseus team at Tufts University all contributed to building up these collections. Bruce Robertson of Mount Allison University in Canada spent a decade optimizing open-source Optical Character Recognition (OCR) systems and creating curation environments so that we could enter far more Greek than had been feasible before (Robertson and Boschetti, 2017).

After 40 years of development and in collaboration with partners at multiple institutions across a variety of countries, Perseus has published 41 million words of Greek and 17 million words of Latin in curated TEI XML, citable via the Canonical Text Services (CTS) protocol (Tiepmar and Heyer, 2017), and available under a Creative Commons (CC-BY-SA) license. These texts were developed specifically so that third parties could download, modify and reuse them. The next section describes some (though by no means all) examples of what others have done with this open data.

However much has been accomplished over preceding decades, we are poised for a rapid expansion in digital philology for both Greek and Latin. The quality of general-purpose OCR has increased dramatically. In June 2025, Harvard released OCR-generated text for c. 1 million books (IB1) published through the beginning of the 20th century (Cargnelutti et al., 2025). In summer 2025, Perseus collaborators David A. Smith, Alison Babeu and Cliff Wulfman were able to identify texts with 600 million words of classicizing Greek, from the Homeric epics up through katharevousa Greek in the 19th and 20th centuries. While large European digital libraries such as Gallica and the German Digital Library will surely produce far more Latin, the Harvard collection contains an estimated 12 billion words of Latin – twenty times larger than the already daunting Greek collection. We are now able to begin studying Greek and Latin both at scale, not only in the classical period but over thousands of years. Increasingly effective applications of machine learning – including not only Large Language Models (LLMs) but also traditional, task specific applications from Natural Language Processing (NLP) – will allow us to provide usable structure to these collections that are far too large for purely manual analysis.

But if vast amounts of data will transform our work, the carefully constructed efforts that will be described below will provide the data sets upon which future work will build (and, in fact, current LLMs have already built).

2. BACKGROUND: PRODUCING REUSABLE GREEK AND LATIN TEXTS

In response to the need for open textual data, researchers at Perseus and collaborators, both direct and indirect, developed a growing network of textual resources for Greco-Roman culture and, in particular, for the study of Ancient Greek. While Perseus was always designed to open up Greco-Roman culture to a wider audience, the emergence of Creative Commons (CC) licenses in the early 2000s provided a framework that allowed researchers to move on from sharing data on the basis of one-off, one-on-one agreements towards distributing content to a global community under well-defined conditions.

We at Perseus had supported a free open facing version of the PDL and were committed to opening up the Greco-Roman world to new audiences. But in an age of bilateral agreements about data sharing, we had had problematic experiences and were not among those who introduced the use of CC licenses and of GitHub in Digital Classics (Buchanan, 2015; Christensen, 2022; LaMonica and Foka, 2025). The papyrologists who built Papyri.info (Bagnall, 2010) and the collaborators who would create the Homer Multitext Project (Dué and Ebbott, 2019) led the way and, following their example, Perseus adopted a CC license in spring 2006. While some successful collaborations (such as Chicago's Perseus under Philologic) had already begun, most of this paper will discuss what happened when others were able to reuse our data without asking for permission and what is possible now for us because others have created entirely new data that we can ourselves reuse.

First, however, we explain the challenges that students of Greco-Roman culture faced in building an open intellectual ecosystem. Greco-Roman studies were, arguably, early-adopters of digital technology but that early adoption led many in the field to embed older assumptions about ownership and about the skills appropriate to philologists, assumptions that, in our view, have prevented this subject from flourishing in the intellectual life of society as fully as it should (Crane et al., 2006; Passarotti, 2026).

Second, we describe some of the most important ways in which third parties used openly licensed Perseus content as the basis for additional work of various kinds: improvements to machine actionable dictionaries, linguistic annotation, named entity linking, metrical analysis, rich metadata for speakers in drama and epic, and quotation detection (using the entire corpus of OGL to scour OCR-generated text from millions of books).

This new ecosystem not only enables new forms of research but also supports close reading by expert Classicists, by researchers from adjacent fields (such as Philosophy and the History of science), and by members of the general public. This new ecosystem has created opportunities for a much wider range of people to make tangible, credited contributions to the infrastructure for the study of Greco-Roman culture.

From its earliest development in the 1980s, Perseus viewed its text collections as the core around which additional annotations would cluster and has also long focused on the integration of textual data with machine actionable annotations. The Morpheus morphological analysis system, first composed in 1985, was designed to generate every possible morphological analysis for any given inflected Greek form (Crane, 1991). Because Ancient Greek is a highly inflected language, the resulting part of speech data was of immediate value to readers as they worked with unfamiliar forms. Equally important, Morpheus was able to generate the standard dictionary forms for inflected Greek words. Greek verbs have six principal parts and the inflected form that readers confront in a text may have little or no resemblance at all to the dictionary form (e.g., Latin *sustuli*, "I lifted/raised," is the perfect form of *tollo*, "I lift/raise"). Even if readers can themselves derive the dictionary form from the inflected form, tracking down the dictionary entry in large dictionaries was a tedious task. Furthermore, automatic lemmatization made smart word searches possible, as classicists could then use the dictionary form to retrieve all inflected forms of a given word.

The early work with morphological analysis thus predisposed us to explore other forms of annotation that could augment our core texts. In the 1990s, Perseus began adding named entity annotations (Smith and Crane, 2001). Treebanks as well as word and phrase level translation alignments followed in the 2000s (Bamman and Crane, 2007, 2011). The November 2022 release of GPT 3.5 set in motion, for the first time, a technological pathway that could justify

the term artificial intelligence. Transformative as LLMs already have been, we are still learning how to apply this technology to the production of reusable data at scale.

3. DATA SETS BASED IN WHOLE OR IN PART ON PERSEUS TEXTS

This section describes third party work, all of which is openly available. Unlike the datasets that typically feature in the *Journal of Open Humanities Data*, not all these sources have been placed in formal academic repositories, but all are available under open licenses for others to reuse (and preserve) and relevant details have been provided whenever available.

3.1 PROVIDING METADATA FOR THE CORE TEI XML TEXTS

While Perseus and its collaborators have spent decades slowly accumulating curated TEI XML texts, the core markup can always be enhanced. Here we mention four projects that have added crucial metadata that is based on, and can be applied to, Perseus texts to make them serve new research questions and provide new services for readers. Each of the projects below provides rich metadata that enables more precise forms of linguistic analysis than the encoding that we at Perseus have been able to provide.

Projects that identify speakers: Perseus contains an extensive collection of Greek and Latin epic poems. We did add TEI XML speaker tags (<sp>) to the *Iliad* and the *Odyssey*, allowing us to compare the vocabulary of speech vs. narrative but, even for these poems, we did not add information about who was speaking (e.g., so that we could compare the language of Achilles to that of Odysseus or the language of female vs. male characters).

DICES: Digital Initiative for Classics: Epic Speeches

Repository Name: Borealis: Mount Allison University Dataverse

Object Name -Citation: Forstall, Christopher, Berenice Verhelst, and Simone Finkmann. 2025. "Raw Data for DICES Epic Speeches Database." Borealis. https://doi.org/10.5683/SP3/N8LS2Y.

Access to the database: https://db.dices.mta.ca/app/

Format Names and Versions: CSV tabular data

Creation Dates: 2018–2025

Dataset Creators: Christopher Forstall, Berenice Verhelst, and Simone Finkmann.

Language: English

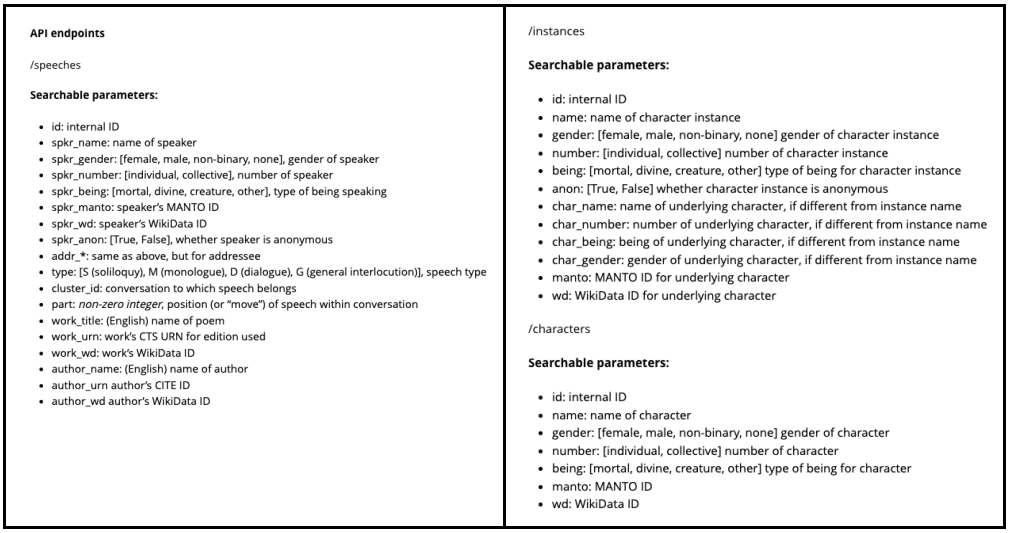
License: Open access but specific licensing terms unknown

Publication Date: 2025-11-12

The DICES (Digital Initiative for Classics: Epic Speeches) project is a research initiative launched in 2018–2019 to build a comprehensive, open-access database of direct speech in Greek and Latin epic poetry. DICES is a database of metadata on direct speech representation in Greek and Latin epic poetry. It contains information about who says what to whom in epics from Homer to Late Antiquity, including over 4500 speeches, 1000 characters, and 50 poems (Forstall et al., 2022).

DICES represents a best case in text reuse. To begin with, it provides elegant standoff markup that cleanly links to source texts without changing them. In part, this reflects the fortunate nature of the data: in most (if not all) Greek and Latin epics, speeches begin and end at line boundaries. Thus, in the first speech of the *Odyssey*, Zeus begins speaking at the beginning of Book 1, line 32 and continues through the end of line 43 in book 1. In Greek and Roman historians, by contrast, speeches occasionally begin or end in the middle of book/chapter/section chunks and precise annotation needs to identify the starting/ending word. Of course, annotations can address any word in a text by identifying the precise word in a given chunk of a given edition (e.g., the second instance of καί, 'and,' in a particular edition), but Greek and Roman epic poems provide an easier case.

Figure 1 provides some idea of the rich metadata that DICES metadata has published. Not only can we identify speeches by Achilles and Odysseus in Homer but also in post-homeric epics such as Ovid's *Metamorphoses* and Statius' *Achilleid*. We can distinguish among speeches by mortals, divinities and mythical creatures, among speakers who are male, female, or groups that combine both. In addition, the DICES metadata provided a starting point for publications by an international team, available now in open access (Forstall and Verhelst, 2026).



API endpoints

/speeches

Searchable parameters:

- id: internal ID
- spkr_name: name of speaker
- spkr_gender: [female, male, non-binary, none], gender of speaker
- spkr_number: [individual, collective], number of speaker
- spkr_being: [mortal, divine, creature, other], type of being speaking
- spkr_manto: speaker's MANTO ID
- spkr_wd: speaker's WikiData ID
- spkr_anon: [True, False], whether speaker is anonymous
- addr_*: same as above, but for addressee
- type: [S (soliloquy), M (monologue), D (dialogue), G (general interlocution)], speech type
- cluster_id: conversation to which speech belongs
- part: *non-zero integer*, position (or "move") of speech within conversation
- work_title: (English) name of poem
- work_urn: work's CTS URN for edition used
- work_wd: work's WikiData ID
- author_name: (English) name of author
- author_urn author's CITE ID
- author_wd author's WikiData ID

/instances

Searchable parameters:

- id: internal ID
- name: name of character instance
- gender: [female, male, non-binary, none] gender of character instance
- number: [individual, collective] number of character instance
- being: [mortal, divine, creature, other] type of being for character instance
- anon: [True, False] whether character instance is anonymous
- char_name: name of underlying character, if different from instance name
- char_number: number of underlying character, if different from instance name
- char_being: being of underlying character, if different from instance name
- char_gender: gender of underlying character, if different from instance name
- manto: MANTO ID for underlying character
- wd: WikiData ID for underlying character

/characters

Searchable parameters:

- id: internal ID
- name: name of character
- gender: [female, male, non-binary, none] gender of character
- number: [individual, collective] number of character
- being: [mortal, divine, creature, other] type of being for character
- manto: MANTO ID
- wd: WikiData ID

Figure 1 a selection of searchable fields in DICES (Forstall et al., 2022, 2022b).

DraCor: – Open Infrastructure for Drama Analysis

Repository location: https://github.com/dracor-org/greekdracor

Repository name: GitHub

Object name: Greek Drama Corpus

Format names and versions: XML

Creation dates: 2019-05-01, 2026-04-30

Dataset creators: Julia Jennifer Beine and Frank Fischer

Language: Ancient Greek, English, German

License: CC BY-SA 3.0

Publication date: 2019

The German DraCor project provides another example of reuse and major enhancement of Perseus TEI XML text. DraCor "is an open digital infrastructure developed for the computational study of (mostly) European drama from Greco-Roman antiquity to the 20th century". DraCor used TEI XML files from Perseus as the primary basis for 38 of 39 Greek and all 36 Roman plays in its collection (Fischer et al., 2019; Beine et al., 2024).

From our earliest work in the 1980s, we used markup to identify speeches and speaker names in drama, but we have not added unique identifiers for speakers. It is easy enough, when analyzing a single play, to extract and compare the language of Ὀδυσσεύς (Odysseus) and Μενέλαος (Menelaus) in the *Ajax* of Sophocles, but not nearly as easy to compare the language attributed to Odysseus or Achilles in Greek drama as a whole. The DraCor TEI XML files, originally drawn from Perseus, now include not only unique identifiers for each speaker (e.g., #odysseus and #menelaus) but also cast lists that identify gender and link characters with identities beyond a given play to entries in Wikidata (e.g., Q47231 for Odysseus and Q171839 for Menelaus) making it computationally much easier to compare the language and characterization of characters across plays. Likewise, the metadata would immediately support comparison of speech by gender. A researcher could also draw upon Wikidata to distinguish mortals from divinities and compare mortal and divine female language.

The DraCor TEI XML texts (Figure 2) are based on earlier versions of Perseus files (including some that had not yet been updated to TEI XML P5). Aligning the new speaker metadata can be done efficiently – the only places where we cannot automatically import the new data would be instances where for whatever reason the speaker markup has gotten out of alignment.

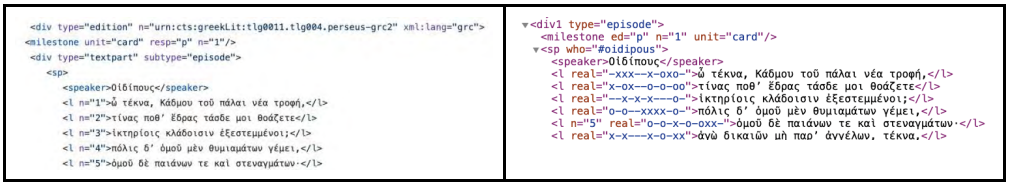


Figure 2 Comparison of the opening of *Oedipus Rex* in TEI XML from Perseus (left) and DraCor (right). The DraCor XML, based on an older Perseus version, adds two crucial data categories: a unique speaker identifier and metrical analysis for each line.

DraCor offers two other advantages. First, the DraCor team filled in a gap on the Greek side, creating a TEI XML version of the *Dyskolos* by Menander (which we had set aside because of the complexities of working with papyrological data). Julia Jennifer Beine and Frank Fischer worked with a team of students to convert the Wikisource version of this play into a standard and easily reusable TEI XML version.

Second, in June 2020, DraCor added metrical markup for Greek plays by Dmitry S. Krlov. While we can automatically generate possible metrical analyses for standard Greek meters, iambic trimeters can often be scanned in multiple ways, making a curated analysis extremely valuable. Even more importantly, Krlov provided the metrical analysis for the choral lyrics, a task that cannot be automated. This metrical analysis not only opens up new research questions but also provides materials for a new dimension of reading aid, opening up access to Greek meter to a wider audience.

Identifying quotations from external sources. Automatic quotation detection provides a foundational tool for those studying the influence of literary works over time (Ernst-Gerlach and Crane, 2008; Arnold et al., 2025). By seeing what passages from Homer were quoted we can gain insights into what topics interested audiences over time. At the same time, our ability to identify quotations at scale enhances our ability to trace language change: when we wish to analyze the language of Plato or Aristotle, we want to be able to recognize where these fourth century BCE prose authors are including quotations from Homer or Hesiod that are not only in poetry but also reflect an archaizing form of Greek.

More than a decade ago, researchers in the Proteus Project developed a robust system that could detect quotations from larger collections of uncorrected – and often noisy – OCR-generated text (Yalniz et al., 2011; Smith et al., 2015). The most technically challenging task was to identify quotations in OCR text for 19th century newspapers, a collection with a very high error rate. David A. Smith and his colleagues, however, also used Perseus corpora (Latin texts, translations of Latin texts and the plays of Shakespeare, c. 10 million words total) as an extended query with which they searched OCR-generated text from 3.6 million books in the Internet Archive (results here).[1]

IB1

Repository location: https://github.com/dasmiq/idi-cites; https://github.com/dasmiq/idi-latin/tree/main/docs/quotes; https://github.com/dasmiq/idi-latin/tree/main/docs/quotes

Repository name: GitHub

Object name: ID1 Cites, IDI Latin, IDI Greek

Format names and versions: Jupyter Notebooks, CSV

Creation dates: 2025–2026

Dataset creators: David A. Smith

Language: Ancient Greek, English,

License: MIT License

Publication date: 2026

1 http://books.cs.umass.edu/mellon/quotes.html.

A decade ago, general purpose OCR engines could not produce usable results for languages that were not printed in a Roman alphabet. That situation has greatly improved in recent years and the June 2025 million books of OCR-generated text from Harvard provide a collection in which we can, for the first time, explore ancient Greek as well as Classical Latin at scale. The amount of open Greek and Latin available has increased to 58 million words.

David A. Smith was able to extract one record for each of the 386 million pages in IB1 and one record for each of the 2,223 works in the API to OGL. We used the passim package (Smith, 2023) to find and align passages in IB1 pages with passages in OGL works. The initial indexing stage, which identified candidate document pairs, used character 15-grams (counting only Unicode alphanumeric characters) and a maximum gap of 100 characters between matching n-grams in the same passage. The full process produced 115 million aligned passages, with an average length of 83 characters.

Some of these passages are isolated quotations of an OGL work on a particular IB1 page. Others are parts of complete editions of OGL works. Using the information on citable chunks attached to each OGL work, we can identify, for instance, which books and lines of the Iliad or another work might be quoted on a given page. Several books in IB1 might contain quotations of all passages, or even most of the text, in an OGL work without being editions of those works. For example, the Index in Tragicos Graecos (Beatson, 1830) quotes every line in Aeschylus' Eumenides, and Proclus' and Olympiodorus' Greek commentaries quote every section of Plato's Alcibiades (Proclus and Olympiodorus, n.d.). These books do not, however, quote these entire works of, respectively, Aeschylus and Plato in full and in order. To distinguish commentaries, lexica, and other books from editions of OGL works, we compare the order of citable chunks in the OGL work with the order of the passim-aligned passages in IB1.

The result is a dataset that can be mapped back onto the Perseus corpus.

Figure 3 visualizes the frequency with which different passages of Aristotle's *Poetics* are quoted. The X axis follows the conventional 26 chapters. The flicker lines show the frequency as an aggregate but also as broken down by different languages (e.g., how often texts in English, French, German etc. quote particular passages). A heat map on the lower left side shows which particular words are most often quoted while the lower right shows a list of pages that cite a word on which the reader has clicked. The 115 million quotations provide a completely new dataset by which to track the influence of different sources over time.

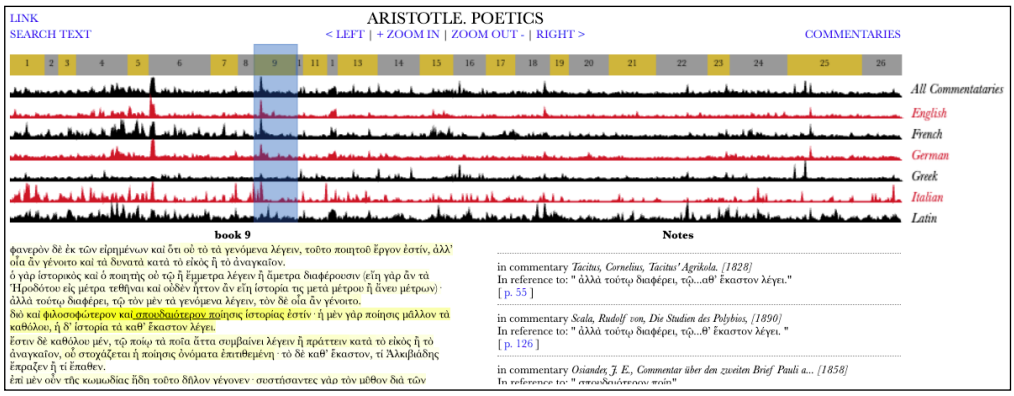


Figure 3 Visualization of the frequency and linguistic distribution of quotations across the conventional twenty-six chapters of Aristotle's *Poetics*.

SEDES Project

Repository location: https://github.com/sasansom/sedes

Repository name: GitHub

Object name: SEDES: Metrical Position in Greek Hexameter

Format names and versions: TEI-XML, CSV, JPG, and Python,

Creation dates: 2021–2026

Dataset creators: Stephen Sansom and David Fifield

Language: English

License: No license specified other than open access declaration

Publication date: 2023

We close this section by mentioning an individual project of particular interest. Stephen Sansom's SEDES Project is a computational initiative that analyzes the metrical position (the *sedes*) of words in ancient Greek hexameter poetry (Sansom and Fifield, 2023). Like the other projects mentioned, it heavily leverages the Perseus Digital Library by using its TEI XML texts as the base corpus for analysis. At least three features stand out from Sansom's work. First, Sansom did not simply publish conclusions as a PDF and/or as a dataset for specialists. He created a demonstration reading environment so that readers could see the impact of his work (see Figure 4). Second, Sansom had begun by working with Greek texts that were not openly licensed. After he had invested substantial work in his project, he decided that he needed to be able to publish the metrical analyses that he had generated and annotations to the source texts showing which words appeared in statistically unusual locations of the Greek hexameter. Third, he created a GitHub repository that included not only his data but also the source code he used to generate his data. That makes it possible for us at Perseus to integrate Sansom's work and provide it with another medium both for visibility and for preservation within the larger Perseus corpus.

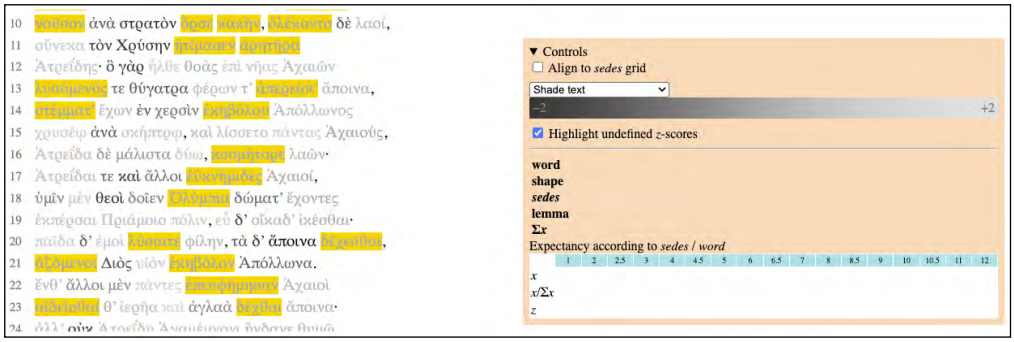


Figure 4 Interactive demonstration reading environment of Stephen Sansom's SEDES Project.

3.2 LINGUISTICALLY ANNOTATED TEXTS: TREEBANKS

In 2002, Perseus began planning to create treebanks – textual corpora which record the normalized dictionary form, the part of speech, syntactic role and syntactic dependency for each word (Crane, 2002). Developed for corpus linguistics, treebanks support far more sophisticated queries than traditional information retrieval. Researchers can identify all subjects of a given verb, the frequency with which a verb takes different cases, and many complex grammatical features (e.g., different varieties of Greek conditionals). Treebanks also represent an advance over a core feature of Perseus. The Morpheus morphological analyzer for Greek and Latin (mentioned above) was able to analyze most inflected Greek and Latin forms, providing part of speech data and the regularized dictionary entry, but Morpheus does not address the problem of ambiguity (Crane, 1991). If a form can have multiple morphological analyses or can be derived from more than one dictionary entry, Morpheus simply reports all possibilities and leaves it to readers to resolve the problem. Treebanks allow annotators to provide the precise part of speech tag and dictionary form, thus resolving structural ambiguities in Perseus services.

In 2007, Perseus, the Index Thomisticus and the Proiel Project, each of which had begun developing treebanks for Greek and/or Latin, collaborated to develop annotation guidelines that were as interoperable as possible. Perseus and the Index Thomisticus modelled their guidelines on the Prague Dependency Treebank (Bamman and Crane, 2007). Proiel adopted guidelines that differed in some measures but that have proven reasonably compatible. Other projects (such as Perseids) supported the production of manually curated treebanks, with roughly one million words of Greek and Latin each available in the early 2020s. As soon as there was any treebank data available, researchers began using it to support automatic analysis (Bamman and Crane, 2010), but as the amount grew larger, researchers began building models to publish analyses of much larger corpora.

Diorisis

Repository location: https://doi.org/10.6084/m9.figshare.6187256

Repository name: Figshare

Object name and citation: The Diorisis Ancient Greek Corpus. Vatri, Alessandro; McGillivray, Barbara (2018). The Diorisis Ancient Greek Corpus. figshare. Dataset. https://doi.org/10.6084/m9.figshare.6187256.v1

Format names and versions: XML

Creation dates: 2018–2026

Dataset creators: Alessandro Vatri and Barbara McGillivray

Language: English, Ancient Greek

License: CC BY 4.0

Publication date: 2018

GLAUx

Repository location: https://doi.org/10.5281/zenodo.10948374

Repository name: Zenodo

Object name: The GLAUx (the Greek Language Automated) corpus

Format names and versions: XML

Creation dates: 2021–2026

Dataset creators: Alex Keersmaekers

Language: English, Ancient Greek

License: CC BY-SA 4.0

Publication date: 2021

OGA

Repository location: https://doi.org/10.5281/zenodo.8158674

Repository name: Zenodo

Object name: Opera Graeca Adnotata

Format names and versions: XML

Creation dates: 2024–2026

Dataset creators: Giuseppe G.A. Celano

Language: English, Ancient Greek

License: CC BY-SA 4.0

Publication date: 2024

In 2018, Diorisis (Vatri and McGillivray, 2018) published 10 million words of lemmatized Greek. In 2021 Keersmaekers published versions of the GLAUx Treebank with 10 and then 20 million words (Keersmaekers et al., 2024). In 2024, Giuseppe Celano, building on his work with the Greek and Latin treebanks in PI Crane's chair for Digital Humanities at Leipzig, published the largest treebanks of Greek and Latin, with 45 million words of Greek (OGA) and 17 million words of Latin (OLA) (Celano, 2024). While GLAUx, OGA, and OLA primarily annotated texts from Perseus, they also digitized some new texts themselves and collected Greek texts that other projects had produced. Their treebanks thus not only contain automatically generated morpho-syntactic annotations that add value to texts already in Perseus but also allow us to add a number of texts that plug gaps in Perseus coverage (e.g., Sappho). The Celano Greek corpus contains 4 million words of Greek from a range of other smaller collections that are not yet in Perseus.

Work on manually annotated treebanks has also continued. Francesco Mambrini, working at Perseus, originally created treebanks for the seven surviving plays of Aeschylus and for six of the surviving plays of Sophocles. Francesco, however, created new versions of these and other Perseus treebanks (including the *Iliad* and *Odyssey*) in the Universal Dependency Framework (see Figure 5 below) that emerged after Perseus and other projects had used the Prague dependency grammar as their initial framework (Mambrini, 2025).

```
# newdoc id = "tlg0011.tlg004.daphne_tb-grc1"
# sent_id = tlg0011.tlg004.1
# text = ὦ τέκνα, Κάδμου τοῦ πάλαι νέα τροφή, τίνας ποθ’ ἕδρας τάσδε μοι θοάζετε ἱκτηρίοις κλάδοισιν ἐξεστεμμένοι;
# Speaker = Oedipus
1   ὦ   ὦ   INTJ    i--------   _   2   discourse   _   Ref=1
2   τέκνα   τέκνον   NOUN   n-p---nv-   Case=Voc|Gender=Neut|Number=Plur   15   vocative   _   Ref=1|SpaceAfter=No
3   ,   ,   PUNCT   u--------   _   8   punct   _   Ref=1
4   Κάδμου   Κάδμος   PROPN   n-s---mg-   Case=Gen|Gender=Masc|Number=Sing   8   nmod   _   Ref=1
5   τοῦ   ὁ   DET   l-s---mg-   Case=Gen|Definite=Def|Gender=Masc|Number=Sing|PronType=Art   4   det   _   Ref=1
6   πάλαι   πάλαι   ADV   d--------   _   4   nmod   _   Ref=1
7   νέα   νέος   ADJ   a-s---fv-   Case=Voc|Gender=Fem|Number=Sing   8   amod   _   Ref=1
8   τροφή   τροφή   NOUN   n-s---fv-   Case=Voc|Gender=Fem|Number=Sing   2   appos   _   Ref=1|SpaceAfter=No
9   ,   ,   PUNCT   u--------   _   2   punct   _   Ref=1
10   τίνας   τις   DET   p-p---fa-   Case=Acc|Gender=Fem|Number=Plur|PronType=Ind   12   det   _   Ref=2
11   ποθ’   ποτέ   ADV   d--------   PartType=Ptcl   10   advmod   _   Ref=2
12   ἕδρας   ἕδρα   NOUN   n-p---fa-   Case=Acc|Gender=Fem|Number=Plur   15   obj   _   Ref=2
13   τάσδε   ὅδε   DET   p-p---fa-   Case=Acc|Gender=Fem|Number=Plur|PronType=Dem   12   det   _   Ref=2
14   μοι   ἐγώ   PRON   p-s---md-   Case=Dat|Gender=Masc|Number=Sing|Person=1|PronType=Prs   15   obl   _   Ref=2
15   θοάζετε   θοάζω   VERB   v2ppia---   Mood=Ind|Number=Plur|Person=2|Tense=Pres|Voice=Act   0   root   _   Ref=2
```

Figure 5 Universal Dependency Treebank for lines 1–2 of the *Oedipus Rex* on GitHub.

Treebanks may be the most important component for a digital philology infrastructure but treebanks become much more valuable when integrated with our resources. It is possible to add extra data (e.g., the current speaker in a play as in DraCor) but would not be feasible to add all the fields that DICES provides. When we combine data from sources such as DraCor and DICES with treebanks, we can develop increasingly sophisticated analysis (e.g., comparing the vocabulary and syntax of different characters by gender).

3.3 NAMED ENTITY DATA

Advanced linguistic analysis benefits from the ability to recognize proper nouns (e.g., distinguishing between Alexandria as a place name or as the name of a female individual) and to be able to link particular references to particular entities in an authority list (e.g., using the Pleiades gazetteer (2006) to distinguish a reference to the Alexandria in modern Afghanistan from the famous Alexandria in Egypt). The Recogito semantic annotation system allows individuals to add such annotations to sources in a variety of languages – often (though not always) using Perseus texts as a starting point (Simon et al., 2017). Perseus has already benefitted from such work in its Beyond Translation project. Figure 6 shows an automatically generated map plotting all place names mentioned in Book 3 of the *Odyssey*. The data was created using recogito.pelagios.org and published with named credit through the Beyond Translation project.

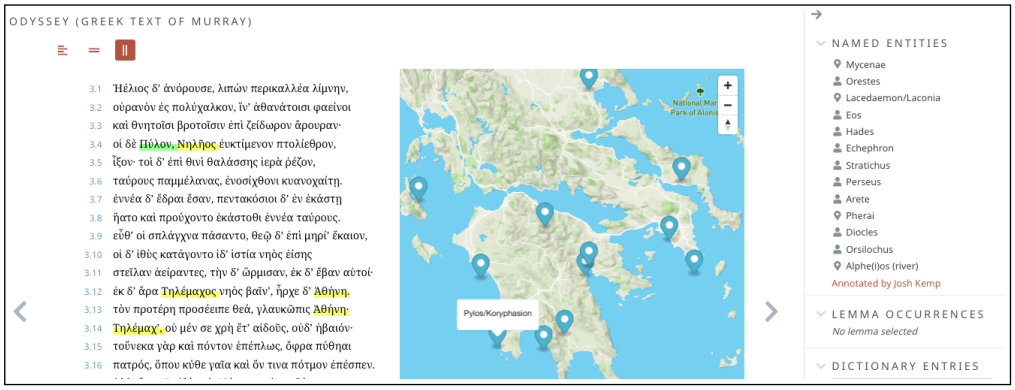


Figure 6 Greek named entities created by Josh Kemp and published in Perseus Beyond Translation.

Recogito/Pelagios

Repository location: https://github.com/pelagios/recogito2

Repository name: GitHub

Object name: recogito2

Format names and versions: Scala, JavaScript, HTML

Creation dates: 2016–2019

Dataset creators: Elton Barker, Leif Isaksen, Rebecca Kahn, Rainer Simon Valeria Vitale, Hugh Cayless

Language: English

License: Apache 2.0 license.

Publication date: 2019

ToposText

Repository location: https://topostext.org/TT-downloads

Repository name: N/A

Object name: ToposText places and text files

Format names and versions: kmz, geojson, json-ld

Creation dates: 2016–2026

Dataset creators: Bradly Kiessling

Language: English, Ancient Greek

License: Apache 2.0 License

Publication date: Ongoing

ToposText, however, has created the most imposing resource for named entity data in the Greco-Roman world. While ToposText used translations from Perseus as its largest starting point, it has added many new translations that are not yet in Perseus. ToposText contains 8,150 mapped historic places and archaeological sites from Spain to India, attested by 279,191 ancient references in 868 ancient texts; descriptions and links to Wikipedia/Wikidata, Pleiades, archaeological sites in Greece, and other sources. Originally focused on geotagging, ToposText began using Wikidata to disambiguate references to other proper names such as people, gods, festivals, animals, artworks, buildings and unmapped places. There are 430,860 references to 22,003 of these proper names, with links where possible to Wikipedia and/or Wikidata.

As Figure 7 shows, ToposText encodes named entity data in its translations, linking placenames to the ToposText gazetteer and Pleiades. Personal names link to WikiData, and these annotations can be aligned to the Perseus XML.



Figure 7 Named entity data in the HTML from ToposText for Thucydides book 1, chapter 114.

The annotated English translations are valuable by themselves. The annotations can also be transferred with reasonable accuracy back to the Greek and Latin sources upon which the translations are based.

CONCLUSION

The sections above have described only a subset of the openly licensed data that was built to a substantial degree upon content from Perseus (for a discussion of parallel text alignment and training data derived from Perseus texts, example, see Palladino et al., 2022; Yousef et al., 2022 and Crane et al., 2025). Releasing data under an open license twenty years ago in 2006 has led to a parade of efforts, most of which were initially unfamiliar to us, that has done new, often surprising and always gratifying new work more effectively because of the data that they could reuse. A great deal of work remains to structure open content so that it can flow with less data wrangling from project to project, but a vigorous open ecosystem has emerged for the study of Greek and Latin language and literature.

FUNDING STATEMENT

The work discussed in this paper has been made possible through funding from the National Endowment for the Humanities, the Institute of Museum and Library Services, the National Science Foundation, the Andrew W. Mellon Foundation, the Schmidt Sciences foundation, private funding and Tufts University.

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

Gregory Crane: Conceptualization, Data Curation, Funding Acquisition, Investigation, Methodology, Project Administration, Resources, Software, Supervision, Writing – original draft.

Alison Babeu: Conceptualization, Data Curation, Funding Acquisition, Investigation, Methodology, Project Administration, Resources, Supervision, Writing – original draft.

David Smith: Conceptualization, Investigation, Software.

AUTHOR AFFILIATIONS

Gregory Crane orcid.org/0000-0001-6847-2811
Department of Classical Studies, Tufts University, United States

Alison Babeu orcid.org/0009-0000-6353-5588
Department of Classical Studies, Tufts University, United States

David A. Smith orcid.org/0000-0002-6636-6940
Khoury College of Computer Sciences, Northeastern University, United States

REFERENCES

Arnold, F., Jäschke, R., & **Kraut, P.** (2025). Advances and challenges in the automatic identification of indirect quotations in scholarly texts and literary works. In M. Hämäläinen, E. Öhman, Y. Bizzoni, S. Miyagawa, & K. Alnajjar (Eds.), *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities* (pp. 179–190). Association for Computational Linguistics. https://aclanthology.org/2025.nlp4dh-1.15/

Bagnall, R. (2010). Integrating digital papyrology. *Online Humanities Scholarship: The Shape of Things to Come*. http://hdl.handle.net/2451/29592

Bamman, D., & **Crane, G.** (2007). The Latin Dependency Treebank in a cultural heritage digital library. *Proceedings of the Workshop on Language Technology for Cultural Heritage Data, LaTeCH@ACL 2007*. Prague, Czech Republic, June 28, 2007 (pp. 33–40). Association for Computational Linguistics. https://aclanthology.org/W07-0905/

Bamman, D., & **Crane, G.** (2010). Corpus Linguistics, Treebanks and the Reinvention of Philology. *Informatik*, *2010*, 542–551.

Bamman, D., & **Crane, G.** (2011). The Ancient Greek and Latin Dependency Treebanks. In C. Sporleder, A. van den Bosch, & K. Zervanou (Eds.), *Language Technology for Cultural Heritage* (pp. 79–98). Springer Berlin Heidelberg. http://hdl.handle.net/10427/75560

Beatson, B. W. (1830). *Index in tragicos graecos v. 3*. W.P. Grant. https://babel.hathitrust.org/cgi/pt?id=hvd.32044025675752

Beine, J. J., Fischer, F., & **Illmer, V. J.** (2024, August 6). Just the type: Analysing character typology in Roman comedy with RomDraCor. *Proceedings of DH2024: "Reinvention & Responsibility."* https://doi.org/10.5281/zenodo.13801481

Berti, M. (2019). *Digital Classical Philology: Ancient Greek and Latin in the digital revolution*. Walter de Gruyter GmbH & Co KG. https://doi.org/10.1515/9783110599572

Brunner, T. F. (1991). The Thesaurus Linguae Graecae: Classics and the computer. *Library Hi Tech*, *9*(1), 61–67. https://dl.acm.org/doi/10.5555/123828.123843

Buchanan, S. A. (2015). The emerging tradition of digital classics. *Annual Review of Cultural Heritage Informatics: 2014*, 149–164. https://doi.org/10.5040/9798216428367.ch-0010

Cargnelutti, R., et al. (2025). *Institutional Books 1.0 (IB1)* [Dataset]. Harvard Law School Library. https://huggingface.co/datasets/institutional/institutional-books-hl

Celano, G. G. A. (2024). Opera Graeca Adnotata: Building a 34M+ token multilayer corpus for Ancient Greek. https://arxiv.org/abs/2404.00739

Christensen, J. P. (2022). Digital classics. *TAPA*, *152*(1), 43–54. https://doi.org/10.1353/apa.2022.0005

Crane, G. (1991). Generating and parsing Classical Greek. *Literary and Linguistic Computing*, *6*(4), 243–245. https://doi.org/10.1093/llc/6.4.243

Crane, G. (2002). Don't miss the lexicographers for the treebanks: Philology in an electronic age. *Conference on the Cambridge Greek Lexicon*. https://www.academia.edu/82054421/Dont_miss_the_lexicographers_for_the_treebanks_Philology_in_an_electronic_age.

Crane, G. (2016). Greco-Roman Studies in a digital age. *Daedalus*, *145*(2), 127–133. https://doi.org/10.1162/DAED_a_00383

Crane, G., Babeu, A., & **Shamsian, F.** (2025). Greek, Latin and augmented intelligence: The other AI. *The Classical Review*, 1–16. https://doi.org/10.1017/S0009840X25000356

Crane, G., Bamman, D., Cerrato, L., Jones, A., Mimno, D., Packel, A., Sculley, D., & **Weaver, G.** (2006). Beyond digital incunabula: modeling the next generation of digital libraries. In J. Gonzalo, C. Thanos, M. F. Verdejo, & R. C. Carrasco (Eds.), *Research and Advanced Technology for Digital Libraries* (Vol. 4172, pp. 353–366). Springer Berlin Heidelberg. https://doi.org/10.1007/11863878_30

Dué, C., & **Ebbott, M.** (2019). The Homer Multitext within the history of access to Homeric Epic. In M. Berti (Ed.), *Digital Classical Philology: Ancient Greek and Latin in the Digital Revolution* (pp. 239–256). De Gruyter Saur. https://doi.org/10.1515/9783110599572-014

Ernst-Gerlach, A., & **Crane, G.** (2008). Identifying quotations in reference works and primary materials. *Research and Advanced Technology for Digital Libraries*, 78–87. https://doi.org/10.1007/978-3-540-87599-4_9

Farina, A., Marongiu, P., Bru, M., & **Borkowski, D.** (2025). When data meets the past: data collection, sharing, and reuse in Ancient World Studies. *Open Information Science*, *9*(1), 20250014. https://doi.org/10.1515/opis-2025-0014

Fischer, F., Börner, I., Göbel, M., Hechtl, A., Kittel, C., Milling, C., & **Trilcke, P.** (2019, July 9). Programmable Corpora: Introducing DraCor, an Infrastructure for the Digital Study of Drama. *Proceedings of the Digital Humanities 2019 (DH2019)*. https://doi.org/10.5281/zenodo.4284002

Forstall, C., Finkmann, S., & **Verhelst, B.** (2022b). *Open Access Database—DICES: Digital Initiative for Classics: Epic Speeches* [Dataset]. University of Rostock. https://www.dices.uni-rostock.de/en/open-access-database/

Forstall, C., & **Verhelst, B.** (Eds.) (2026). *Direct speech in Greek and Latin epic: Expanding the methods and canon*. Brill. https://doi.org/10.1163/9789004750227

Forstall, C. W., Finkmann, S., & **Verhelst, B.** (2022). Towards a linked open data resource for direct speech acts in Greek and Latin epic. *Digital Scholarship in the Humanities*, *37*(4), 972–981. https://doi.org/10.1093/llc/fqac006

Heslin, P. (2007). *Diogenes, a tool for searching and browsing the databases of ancient texts*. Version 4.7.2 https://d.iogen.es/d/

Homer. (1951). *The Iliad of Homer* (R. Lattimore, Trans.). University of Chicago Press.

Homer. (1967). *The Odyssey of Homer* (R. Lattimore, Trans.). Harper & Row.

Keersmaekers, A., Pietowski, F., Van Hal, T., & **Depauw, M.** (2024). The Browser-based GLAUx Treebank infrastructure: Framework, functionality, and future. *Cybernetics and Information Technologies*, *24*(4), 164–174. https://doi.org/10.2478/cait-2024-0041

Köntges, T., Lesage, R., Robertson, B., Sellick, J., & **Stylianopoulos, L. W.** (2019). Open Greek and Latin: Digital humanities in an open collaboration with pedagogy. *IFLA WLIC 2019*- Athens, Greece – Libraries: Dialogue for Change. https://repository.ifla.org/handle/20.500.14598/6602

LaMonica, C. R., & **Foka, A.** (Eds.). (2025). *Evolving perspectives on Digital Classics*. Routledge. https://doi.org/10.4324/9781003584155

Mambrini, F. (2025). *Daphne: The Daphne treebank repository of Ancient Greek poetry* (Version [insert version if known]) [Computer software]. GitHub. https://github.com/francescomambrini/Daphne

Middle, S. (2024). Linked Ancient World Data: Implementation, advantages, and barriers. *Digital Classics Online, Bd. 10*, 16–49 Seiten. https://doi.org/10.11588/DCO.2024.10.104105

Packard, D. W. (1968). *A concordance to Livy*. Cambridge, Harvard University Press.

Palladino, C., Shamsian, F., & **Yousef, T.** (2022). Using parallel corpora to evaluate translations of ancient Greek literary texts: An application of text alignment for digital philology research. *Journal of Computational Literary Studies*, *1*(1). https://doi.org/10.48694/jcls.100

Passarotti, M. (2026). From data to knowledge (towards wisdom) in linguistic computing for ancient languages. In P. d'Hoine, D. Kohler, & W. Decock (Eds.), *Lectio* (pp. 77–98). Brepols Publishers. https://doi.org/10.1484/M.LECTIO-EB.5.151803

Perseus Digital Library. (2026). Retrieved from https://www.perseus.tufts.edu/hopper/ (last accessed: 2 June 2026).

PHI Latin Texts. (2011). Classical Latin Texts: A Resource Prepared by The Packard Humanities Institute. Retrieved from https://latin.packhum.org/ (last accessed: 2 June 2026).

Pleiades. (2006). *Pleiades hosted by the STOA Consortium*. Retrieved from https://pleiades.stoa.org/ (last accessed 3 June 2026).

Prendergast, G. (1869–1875). *A concordance to the Iliad of Homer*. Longman.

Proclus, & **Olympiodorus.** (n.d.). *Commentary on Plato's Alcibiades*. Retrieved May 14, 2026, from https://babel.hathitrust.org/cgi/pt?id=hvd.hxjfe7

Robertson, B., & **Boschetti, F.** (2017). Large-scale Optical Character Recognition of Ancient Greek. *Mouseion*, *14*(3). https://doi.org/10.3138/mous.14.3-3

Sansom, S. A., & **Fifield, D.** (2023). SEDES: Metrical position in Greek hexameter. *Digital Humanities Quarterly*, *17*(2). https://dhq.digitalhumanities.org/vol/17/2/000675/000675.html

Simon, R., Barker, E., Isaksen, L., & **De Soto Cañamares, P.** (2017). Linked data annotation without the pointy brackets: Introducing Recogito 2. *Journal of Map & Geography Libraries*, *13*(1), 111–132. https://doi.org/10.1080/15420353.2017.1307303

Smith, D. A. (2023). *Passim: A tool for detecting and analyzing text reuse*. Computational Humanities Research (CHR).

Smith, D. A., Cordell, R., & **Mullen, E. B.** (2015). Computational methods for uncovering reprinted texts in antebellum newspapers. *American Literary History*, *27*(3), E1–E15. https://doi.org/10.1093/alh/ajv029

Smith, D. A., & **Crane, G.** (2001). Disambiguating geographic names in a historical digital library. *ECDL 2001: Proceedings of the 5th European Conference on Research and Advanced Technology for Digital Libraries*, 127–136. https://doi.org/doi.org/10.1007/3-540-44796-2_12

Tiepmar, J., & **Heyer, G.** (2017). An Overview of Canonical Text Services. *Linguistics and Literature Studies*, *5*(2), 132–148. https://doi.org/10.13189/lls.2017.050209

Vatri, A., & **McGillivray, B.** (2018). The Diorisis Ancient Greek corpus: Linguistics and literature. *Research Data Journal for the Humanities and Social Sciences*, *3*(1), 55–65. https://doi.org/10.1163/24523666-01000013

Yalniz, I. Z., Can, E. F., & **Manmatha, R.** (2011). Partial duplicate detection for large book collections. *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, 469–474. https://doi.org/10.1145/2063576.2063647

Yousef, T., Palladino, C., Wright, D., & **Berti, M.** (2022). Automatic translation alignment for Ancient Greek and Latin. *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, 101–107. https://doi.org/10.31219/osf.io/8epsy

TO CITE THIS ARTICLE:
Crane, G., Babeu, A., & Smith, D. A. (2026). Perseus and the Open-Source Ecosystem of Greco-Roman Studies. *Journal of Open Humanities Data*, 12: 120. DOI: https://doi.org/10.5334/johd.595

Submitted: 16 May 2026
Accepted: 21 August 2026
Published: 16 September 2026

COPYRIGHT:
© 2026 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See https://creativecommons.org/licenses/by/4.0/.

Journal of Open Humanities Data is a peer-reviewed open access journal published by Ubiquity Press.