









# The Sixth Generation of the Perseus Digital Library and a Workflow for Open Philology

Gregory Crane<sup>1</sup> , James Tauber<sup>2</sup> , Alison Babeu<sup>1</sup> , Lisa Cerrato<sup>1</sup> ,  
Charles Pletcher<sup>1</sup> , Clifford Wulfman<sup>3</sup> , Sergiusz Kazmierski<sup>4</sup> ,  
Farnoosh Shamsian<sup>5</sup> 

<sup>1</sup> Tufts University

<sup>2</sup> Signum University

<sup>3</sup> Princeton University

<sup>4</sup> Regensburg University

<sup>5</sup> Leipzig University

*Transformations, A DARIAH Journal*

Volume 1, 2025

<https://transformations.episciences.org>

## Dates

**Received:** 15/11/2024

**Accepted:** 03/04/2025

**Published:** 10/06/2025

**DOI:** [10.46298/transformations.14780](https://doi.org/10.46298/transformations.14780)

© The authors



Creative Commons Attribution 4.0  
International

## Abstract

This paper presents an overview of recent developments by the Perseus Digital Library in creating the Beyond Translation reading environment, a foundational component in the transition toward Perseus 6, built on the ATLAS (Aligned Text and Linguistic Annotation Server) architecture. It highlights the integration of diverse open data sources from multiple digital humanities projects, all brought together to support an innovative, richly layered digital reading experience. Following this, the paper details the key services offered by Beyond Translation, including text-translation alignments, advanced morpho-syntactic analysis, audio annotations and enhanced access to integrated reference resources such as commentaries and dictionaries. The paper concludes with a discussion of forthcoming enhancements and the future trajectory of the ATLAS architecture.

**Keywords:** multilingual corpora, text annotation, Linked Open Data, Named Entity Recognition (NER), cultural heritage

## **Zusammenfassung**

Dieser Beitrag gibt einen Überblick über die jüngsten Entwicklungen innerhalb der Perseus Digital Library, welche im Rahmen der Schaffung der Leseumgebung Beyond Translation erfolgt sind. Beyond Translation bildet im Kontext dieser Entwicklung eine grundlegende Komponente des Übergangs zu Perseus 6 und baut auf der ATLAS-Architektur (Aligned Text and Linguistic Annotation Server) auf. Der Beitrag beleuchtet insbesondere die Integration unterschiedlicher, offener Datenquellen, die aus zahlreichen Digital Humanities-Projekten zusammengeführt wurden, um eine innovative und vielschichtige, digitale Leseerfahrung möglich zu machen. Davon ausgehend werden die Schlüsselfunktionen von Beyond Translation erläutert und dabei das Alignment von Originaltexten und Übersetzungen, komplexe morpho-syntaktische Analysen, Audioannotationen und der erweiterte Zugriff auf integrierte Referenzressourcen wie Kommentare und Wörterbücher berücksichtigt. Abschließend werden in Arbeit befindliche und geplante Erweiterungen sowie Entwicklungsperspektiven der ATLAS-Architektur thematisiert.

**Schlüsselwörter:** Mehrsprachige Korpora, Textannotation, Verknüpfte Offene Daten, Erkennung benannter Entitäten (NER), Kulturelles Erbe

Gregory Crane was the lead author of this piece. James Tauber was primarily responsible for the new work on the ATLAS architecture performed in 2024. All images and figures in this paper are available freely under a Creative Commons - Attribution - CC-BY 4.0 licence.

## Introduction

We report here on the workflow that we needed to develop in order to integrate the growing range of openly licensed, born-digital and, increasingly, machine-actionable publications. Our developmental work focused upon textual data for Ancient Greek, Latin, Old English, Classical Arabic and Classical Persian, but the challenges that we have had to address are relevant to sources in a wide range of languages, both ancient and modern. During the course of the early twenty-first century, major projects have emerged that support forms of publication that either extend well beyond, or in fact, have no counterparts in print culture. Systems such as [INCEPTION](#) and [Perseids](#) allow us to add exhaustive linguistic annotations to our corpora and develop treebanks. The Ugarit platform allows us to align, both [manually](#) and [automatically](#), words and phrases in a source text with their equivalents in a translation. The [Recogito](#) system and the [ToposText](#) project have made it possible to align place names in source texts with gazetteers and then to generate customised maps. David Chamberlain individually developed [Hypotactic](#), a site that allows readers to visualise 250,000 lines of metrically analysed Greek and Latin poetry and to download the underlying data under a Creative Commons licence. Efforts such as these have provided new classes of data for automatic analysis and provided new ways for readers to visualise machine-actionable annotations of source texts.

The work presented here seeks to complement these and other efforts by addressing the challenge of bringing these streams of data together. This work lays the foundation for research into how human readers can best take advantage of the increasingly rich classes of annotation that are already—or soon will be—available online. How do we read in an environment where the answers to far more questions are available than was thinkable in print culture? How do we critically assess automated systems that allow us to interact directly with sources in languages and from cultural contexts of which we may have little or no knowledge? How do we not only translate but also localise knowledge about one cultural context (e.g. European scholarship on Greco-Roman culture) for readers from very different contexts (e.g. Persian-speakers encountering the *Iliad* and the *Odyssey*, for whom Ferdowsi's *Book of Kings* shapes their understanding of epic poetry)? And, of course, what are the roles that increasingly intelligent Large Language Models can—and should—play in helping us engage with a human record that is far too large and complex for any of us to master?

Perseus 6 has been designed to be a publishing workflow that organises complementary data into an integrated reading environment.<sup>1</sup> This paper focuses on the ways in which we have organised the data and describes the current state of

---

1. The work that we describe here was developed as part of the Perseus on the Web: Preparing for the Next Thirty Years project, with support from the US National Endowment for the Humanities, the Data Intensive Studies Center and the Faculty of Arts and Sciences at

the ATLAS (Aligned Text and Linguistic Annotation Server) architecture. While this is the sixth version of the Perseus Digital Library, Perseus 6 represents a major step beyond its predecessors. Whereas Perseus 5 (described below) can represent and integrate digital versions of print editions (e.g. critical editions with interactive textual notes, links to lexicon and commentary entries), Perseus 6 was designed to bring together an expandable range of born-digital classes of annotation. An online [ATLAS server](#) with some initial functionality is now available and public services will expand throughout 2025.<sup>2</sup> Most of the ATLAS data is, however, now available on GitHub and that data will be the focus of this paper in its current version.

The purpose of this publication is to introduce people to the problems that we have addressed. Others may build on what we have done or, having seen our work, choose to develop completely different solutions. The work that we have done, however, addresses the need to bring together complementary datasets that are currently available under an open licence, but split across multiple repositories and systems.<sup>3</sup> Our hope is that more people will see what can be done when we bring together the growing wealth of information that open scholarly projects are producing. The work presented here is a work in progress and represents development as of November 2024.

Our goal was to create a workflow to organise, rather than create, textual data that had been produced by, and was available in, platforms that were open but separate. In the quarter of a century since Creative Commons licences emerged in the early 2000s,<sup>4</sup> multiple projects in a range of countries have developed robust workflows to produce one or more classes of open data. Projects such as [Perseus](#), [Perseids](#), [PROIEL](#), [GLAUx Trees](#), [Daphne](#) and [Opera Graeca Adnotata](#) (Celano 2024) have all published treebanks of Greek and Latin (Hudspeth, O'Connor, and Thompson 2024; Gorman 2020; Keersmaekers et al. 2019; Keersmaekers 2021; Keersmaekers and Van Hal 2022). [Recogito](#) allows users to associate place names in source texts with gazetteers such as [Pleiades](#) and enables automatic mapping (Simon et al. 2017; Barker, Palladino, and Gordin 2024). [INCEpTION](#) (Eckart de Castilho et al. 2018; Berti 2019) enables linguistic and named entity annotation as well as links to authority lists, but we need to turn to separate workflows (such as the [Ugarit translation alignment editor](#) [Palladino et al. 2023], which supports

---

Tufts University. See also Crane (2019, 2023); Crane, Babeu, et al. (2023); Crane, Shamsian, et al. (2023); and Crane et al. (2024).

2. Preliminary versions for most ATLAS services are available on an internal Tufts.edu server and are ready to be published on the public-facing <https://atlas.perseus.tufts.edu/>.
3. Data use and reuse often flows in both directions. While one goal of Beyond Translation and Perseus 6 has been to aggregate and reuse data produced by relevant projects, some of the projects listed in this paper, such as [Perseids](#), [Opera Graeca Adnotata](#), [SEDES](#) and [Ugarit](#), have utilised data created by Perseus and its related partner project, Open Greek and Latin.
4. For more on the history of CC licences, see: <https://creativecommons.org/timeline/>.

word- and phrase-level alignments between source texts and translations). An entirely separate workflow emerged to produce and make available machine-actionable metrical analyses for more than 250,000 lines of Greek and Latin poetry (David Chamberlain’s [Hypotactic](#)). The [DICES project](#) publishes metadata identifying and classifying direct speech in Greek and Latin epic (Forstall, Finkmann, and Verhelst 2022). Individual scholars have published projects such as SEDES, which identifies words in Greek epic that are in statistically surprising metrical positions (Sansom 2021; Sansom and Fifield 2023). The [Ajax Multi-Commentary project](#) uses the rich tradition of scholarship on Sophocles to show how we can aggregate and organise multiple commentaries (Romanello and Najem-Meyer 2024). The [Homer Multitext project](#) publishes new high-resolution images and diplomatic editions of the *Iliad*, as well as *scholia*, with links between transcriptions and images (Dué and Ebbott 2019; Smith and Blackwell 2023).<sup>5</sup>

## Background

Before we outline the development of our own work over the past 40 years, we want to articulate some of the principles that have shaped this work and that are relevant to the most recent activities described in the rest of this paper. Sustainable integration of different categories of data has been a driving force behind the development of Perseus from the beginning. Planning for what is now the Perseus Digital Library began in the spring of 1985, with a substantial equipment grant from the Xerox Corporation and a planning grant from the Annenberg/CPB Project. Continuous development began in 1987 and has continued ever since.

The earliest versions of Perseus emphasised two classes of integration. **First**, one inspiration for Perseus was the realisation that a single system could display not only textual, but also visual information—a fundamentally radical idea in the early 1980s, when state-of-the-art computer terminals displayed monowidth (typewriter-style) ASCII characters without any formatting. As a graduate student, co-author Crane needed to move between two different Harvard libraries as he moved from philological to archaeological and art historical data, but even the best print publications had relatively few images—typically black and white and very low resolution in comparison with born-digital images. A major goal for Perseus was to combine, in a single digital space, textual data of various types with visual information such as images, maps, satellite photos and drawings (Crane 1996; Smith, Rydberg-Cox, and Crane 2000). To demonstrate such integration, we often used the description of Croesus’ dedications at Delphi, showing how we could bring together maps, images and drawings of the site and objects much more effectively than was feasible with print. In the late 1990s, as

---

5. For more on the data used from these projects, and on collaboration between such projects and Perseus, see Crane, Shamsian, et al. (2023) and Crane (2023).

many other projects (the German [Arachne project](#) in particular) began to make digital images and metadata about the art and archaeological record available online, we shifted our focus to the textual record. In 2024, however, we began (as will be discussed below) to exploit the [International Image Interoperability Framework](#) to begin integrating the textual and material records once again.

**Second**, as early as the 1980s, we harnessed automatic analysis to create new links between previously separate classes of textual data. The Morpheus system (Crane 1991) is a rule-based morphological analyser for Classical Greek and Latin that was first developed in 1985 in the C programming language and is still in use today (Keersmaekers et al. 2019; Keersmaekers 2021; Keersmaekers and Van Hal 2022). Morphologically, Greek and Latin are much more complex than languages such as English, French and even German. In print culture, readers have often struggled to match inflected forms on a printed page (e.g. *ênenkas*, “you (sg) carried”) with the relevant dictionary entry (e.g. *pherô*, “I carry”). Given an inflected Greek or Latin form, Morpheus provides every possible morphological analysis (e.g. *ênenkas* is second person singular aorist indicative active) and normalised dictionary form (e.g. *pherô*), matching its tables of stems, endings and combination rules. This system had two basic applications: (1) readers could click on a form and follow links to the morphological analysis and then to machine-readable Greek and (starting in the late 1990s) Latin dictionaries; (2) users could generate searches for those forms: for example, they could ask for *pherô*, “carry,” and retrieve *feréis*, “you (sg.) carry,” *eferon*, “I or they were carrying,” *oïsete*, “you (pl.) will carry,” *enêngektai*, “she/he/it has been carried” and so on.

**Third**, from the earliest stages of planning we designed Perseus to be as sustainable as possible. The software development and scholarly explorations that led to Perseus began with work on the *Thesaurus Linguae Graecae*, a corpus of Greek literature then available on magnetic tape for third party development (now locked behind a proprietary paywall). This experience made it clear that content could, and should, be separate from the software for publication and analysis. The life cycle of software is short; while data, especially in fields such as Greco-Roman studies, has value that persists not only over years, but over centuries and millennia.

Our work began before the Text Encoding Initiative (TEI) would document guidelines on the use of textual markup, but we were already familiar with the principles behind the TEI. On a snowy night in 1985, the authors of DeRose (1990) presented the case for a generalised model of text content (with a talk that bore the same title as their classic paper: “What is text, really?”) As a result, we adopted SGML (the predecessor to XML) and would later revise our markup to follow the TEI. The investment in TEI XML was onerous at first; tools for creating and validating SGML were still at an early stage of development and we had to throw out much of the laboriously added markup when we published our sources in



HyperCard. The investment would pay dividends over the years, however, as tools improved and the markup increasingly facilitated maintenance and updates. One of Perseus's first major digitisation projects, for example, the *Intermediate Liddell Scott Greek-English Lexicon* (based upon Liddell and Scott 1882), was completed in 1985 and is still in active use 40 years later. The current version is available on GitHub (<https://github.com/helmadik/MiddleLiddell>). The GitHub version has notably been enhanced by Helma Dik, a scholar who works with, but has never been part of, the Perseus project.<sup>6</sup> In our view, a sustainability strategy should—by choosing an open licence and making public statements—encourage others to take ownership of and enhance any digital scholarly product.

## Perseus before Perseus 6.0

The first five versions of the Perseus Digital Library augmented data extracted from print sources (e.g. TEI XML transcriptions of editions and reference works, and catalogue entries on art objects and archaeological sites converted into metadata) with automatically generated annotations (such as the Morpheus output described above and named entity annotations classifying people, places and organisations, which were then linked these to authority lists such as the [Pleiades Gazetteer](#)).

These five versions reflect two parallel threads of development. On the one hand, the evolution of Perseus reflects how rapidly software has changed and how agile we have needed to be. The publication platform for Perseus 1.0 and Perseus 2.0—Apple's [HyperCard](#)—had its final release in 1998. Perseus 3.0, written primarily in the [Perl programming language](#), appeared in its earliest form in 1995. It remained in use for almost a decade, giving way to Java-based Perseus 4 in 2004. Perseus 4, in turn, remains the most widely used version and has now been in use for more than two decades. Development of Perseus 4 ended in 2013, however, and the system now runs on virtual machines based primarily on a computing environment that is more than a decade old. We began work on a new platform, primarily written in Python, in 2018 and we continue to build upon this platform today.

On the other hand, whereas we have abandoned earlier code bases, we continue to use and enhance data that was collected decades ago—in one case (the so-called Middle Liddell Greek-English Lexicon based upon the *Intermediate Liddell Scott Greek-English Lexicon*) almost 40 years. Many of the older XML files have changelogs that go back to the late 1980s. While we hope that the longevity of software platforms will improve over time (and lessen the need to

---

6. For more than 15 years, Helma Dik has also led work at the University of Chicago that has made openly licensed materials from [Perseus and other sources available through its Philologic system](#), with its own browsing and advanced searching capabilities.

engineer new systems), we are pleased (and honestly relieved) with the extent to which our data has proven to be sustainable. We have of course had to spend substantial amounts of time updating our content—we still have not updated every Perseus text so that it is compatible with the Canonical Text Services data model—but each change has reflected an upgrade that made our data better structured and more sustainable.

- ▶ 1992: Perseus 1.0 *Interactive Sources and Studies on Ancient Greece*, published by Yale University Press. This included a videodisc, CD ROM and print documentation. Production language: Apple’s HyperCard.
- ▶ 1995: A web version of the Perseus Digital Library can be found at <http://www.perseus.tufts.edu/>. David A. Smith was the lead developer, and the primary programming language was Perl. As noted below, we were able to produce a web version of Perseus years before the CD ROM-based Perseus 2.0 could be published.
- ▶ 1997: Perseus 2.0: *Interactive Sources and Studies on Ancient Greece*, published by Yale University Press. This included five CD ROMs and print documentation. Production language: Apple’s HyperCard. Perseus 2.0 built upon the software backend workflow and frontend developed for Perseus 1.0. The five CD ROMs allowed us to move on from the analogue videodisc of Perseus 1.0 and to create the first fully digital version.
- ▶ 2000: Platform Independent Perseus, published by Yale University Press. This contained the same content as Perseus 2.0, but was accessible both on Macintosh and Windows systems.
  - ▷ 2000: The Perseus Digital Library on the Web became sufficiently well-developed for us to consider it a full version of Perseus and describe it as Perseus 3.0.
- ▶ 2004: Perseus 4.0 (known as [the Hopper](http://www.perseus.tufts.edu/hopper/)). This second web version of the Perseus Digital Library is available at <http://www.perseus.tufts.edu/hopper/>. David Mimno was the lead developer and the primary programming language was Java. Active development on Perseus 4.0 ended in 2013, but the system continues to run on a suite of virtual machines, supported by Tufts University. Two decades later, Perseus 4.0 remains (as of November 2024) the most commonly used version of the Perseus Digital Library.
- ▶ 2018: Perseus 5.0 (known as [the Scaife Viewer](#)).<sup>7</sup> James Tauber was the lead developer. The primary development language was Python. Scaife was based upon the Canonical Text Services (CTS) data model. Scaife built upon the [CapiTainS Software Suite and Guidelines for Citable Texts](#)<sup>8</sup> and

---

7. The Scaife Viewer is named after [Ross Scaife](#), a digital classics pioneer who embodied the spirit of collaboration and who set an early example in establishing open access and openly licensed data as the standards upon which digital classics now depends. The initial release of the Scaife Viewer was on March 15, 2018, the tenth anniversary of his premature passing on March 15, 2008.

8. The CapiTainS suite of tools and guidelines were developed by Thibault Clérice, Bridget Almas and Matthew Munson. For an interesting discussion of their development and important lessons learned, see Almas and Clérice (2017).



allowed Perseus to publish new content. This currently includes 2,669 works in 3,776 editions and translations (1,941 in Greek and 631 in Latin), with 83.8 million words in all languages (40.6 million in Greek, 16.4 million in Latin). Brill adopted Scaife as the platform for all of its scholarly editions ([Brill's Scholarly Editions](#)). Users of Scaife who have access to these editions (most of which are available behind a paywall) will see the resemblance with Scaife in the page design. As of November 2024, two editions (*A Literary History of Medicine* and *The Pez Brothers' Correspondence*) are open access.<sup>9</sup> While Brill's content may be largely proprietary, the contributions its support team has made to the Scaife software platform are [available on GitHub](#). In effect, Scaife represents a collaboration between Perseus, an academic project committed to open scholarship, and Brill (now [De Gruyter Brill](#)), a traditional publisher that relies upon proprietary control (although it has begun to support open access as a publication option).

Readers scanning the above list of Perseus versions will notice a chronological anomaly: the version of the Perseus Digital Library, which we later came to refer to as Perseus 3.0, was first published in 1995—two years before Yale would be able to publish Perseus 2.0. Our earlier investment in structured data (such as TEI SGML/XML) allowed David A. Smith to create an initial version of the Perseus Digital Library as an unfunded side project. While we had worked with Yale University Press to publish the first editions of Perseus and had, in so doing, hoped to help Yale and other university presses develop the infrastructure to develop digital projects, such a development was premature in the 1990s, and partnership with a publisher no longer made sense when the World Wide Web emerged in the early 1990s.

The Scaife Viewer was able to support core features of digital editions, including interactive textual notes, automatic dictionary lookup and integration of commentary and reference works. Brill, in particular, applied Scaife not only to traditional editions, but also to editions of so-called fragmentary works (e.g. [Jacoby Online](#)), which document surviving evidence of texts that have disappeared. Editions of fragmentary works extract quotations from surviving texts that describe, paraphrase or explicitly quote works that are otherwise lost. Fragmentary editions are thus meta-editions, with one edition of a fragmentary work building on dozens or hundreds of other editions.<sup>10</sup>

---

9. The [Berlin-Brandenburg Academy of Sciences and Humanities](#) has also experimented with Scaife and has published sections of its Greek edition and German translation of the *De Locis Affectis* by Galen on Scaife: <https://scaife.perseus.org/library/urn:cts:greek-Lit:tlg0057.tlg057/>.

10. For more on the state of the art for digital editions of fragmentary texts, beyond the approach taken by Scaife and Brill, see for example Berti (2021).

## The Beyond Translation project

The Beyond Translation project<sup>11</sup> (2019–2023: Crane 2019, 2023; Crane, Babeu, et al. 2023; Crane, Shamsian, et al. 2023; Crane et al. 2024) allowed us to develop a prototype for Perseus 6. Our primary goal was to create a reading environment that could integrate and present a much wider range of born-digital annotations than had been feasible in Perseus 1 to 5. Treebanks, for example, contain not only part-of-speech tagging and regularised dictionary forms for each word in a corpus, but also syntactic role and dependency for each word in a sentence. We needed to be able to represent texts not only as texts with annotations, but also graphically as trees. We also needed to be able to represent multiple layers of annotations associated with a text. The following examples illustrate several categories of data that we wanted to represent and that required us to integrate data from standoff markup with one or more textual sources.

**First**, traditional commentaries constitute one well-known document class that requires standoff markup. Print commentaries can present text on the upper part of the page and commentary on the bottom, but the text and commentary are nevertheless separate—albeit parallel—documents.<sup>12</sup> The basic principle is that commentaries quote and comment upon phrases, words and (often) morphemes or other subsets of words. Perseus 4 simply used citations to link texts and commentary: for example, a reader looking at Thucydides book 1, chapter 33, section 2 would see any comments that were associated with that chunk of text. With Perseus 6, we now can link from spans in the source text to a commentary.

1.1.1 Θουκυδίδης Ἀθηναῖος **ξυνέγραψε** τὸν πόλεμον τῶν Πελοποννησίων καὶ Ἀθηναίων, **ὡς ἐπολέμησαν** πρὸς ἀλλήλους, **ἀρξάμενος** **εὐθὺς καθισταμένου** καὶ **ἐλπίσας** μέγαν τε ἔσσεσθαι καὶ **ἀξιολογώτατον τῶν προγεγενημένων**, **τεκμαίρομενος** ὅτι **ἀκμάζοντές τε ἦσαν** ἐς αὐτὸν **ἀμφοτέροι παρασκευῇ** τῇ πάσῃ καὶ **τὸ ἄλλο Ἑλληνικὸν ὅρων** **ξυνιστάμενον** πρὸς ἑκατέρους, τὸ μὲν εὐθύς, τὸ δὲ **καὶ διανοοῦμενον**.

1.1.2 κίνησις γὰρ αὕτη μεγίστη δὴ τοῖς Ἕλλησιν ἐγένετο

▼ COMMENTARY

Θουκυδίδης

**ξυνέγραψε**  
**ξυνέγραψε**—a characteristic word of Thuc., who is known to the ancient critics as ὁ συγγραφεύς, much as Homer is ὁ ποιητής. It denotes the bringing together in one work of many occurrences—composing in its etymological sense. (How some find a reference to the hunting up of materials is not clear.)

**Figure 1:** A commentary on the opening of Thucydides’ *History of the Peloponnesian War*

Beyond Translation. Thucydides. *History of the Peloponnesian War*. Commentary: <https://preview.scaife-viewer.eldarion.com/thucydides-commentary>

In figure 1, yellow highlighting identifies words that have comments. The reader has selected *xunegrapse*, “he composed/wrote,” and the system displays

11. <https://beyond-translation.perseus.org/>

12. Perseus team member Sarah Abowitz recently explored the content and structure of print commentaries and how they might be improved for digital library users, see Abowitz, Crane, and Babeu (2024).

the comment on this particular word. The ability to associate particular spans of a text with data can be applied to many classes of annotation, not just traditional commentary.

ILIAD (GREEK TEXT OF MUNRO & ALLEN)

1.1 μῆνιν ᾄειδε θεὰ Πηληϊάδεω Ἀχιλῆος  
1.2 οὐλομένην, ἣ μυρ' Ἀχαιοῖς ἄλγε' ἔθηκε,  
1.3 πολλὰς δ' ἰφθίμους ψυχὰς Ἄϊδι προΐαφεν  
1.4 ἡρώων, αὐτοὺς δὲ ἐλώρια τεῦχε κύνεσσιν  
1.5 οἴωνοίσι τε πᾶσι, Διὸς δ' ἐτελείετο βουλή,  
1.6 ἐξ οὗ δὴ τὰ πρῶτα διαστήτην ἐρίσαντε  
1.7 Ατρεΐδης τε ἄναξ ἀνδρῶν καὶ δῖος Ἀχιλλεύς.

From the *Didakta Modular Grammar* by Farnoosh Shamsian

→

GRAMMATICAL ENTRIES

Gen1. Possessor  
Gen7. Subjective  
Gen8. Objective

**Impf1. Imperfect of Continuance** x

The imperfect represents an action as ongoing in the past.

→ "ἑξήθεσαν Ἀθηναίων πέντε καὶ εἴκοσι, οἱ ἐννεπολιορκούντο" T. 3.68; → they put to death twenty-five of the Athenians who were besieged (i.e. from the beginning to the end of the siege)

→ "ἔβασιλευεν Ἀντίοχος" T. 2.80; → Antiochus was reigning

**Figure 2:** Links to a language-specific (Ancient Greek) grammar

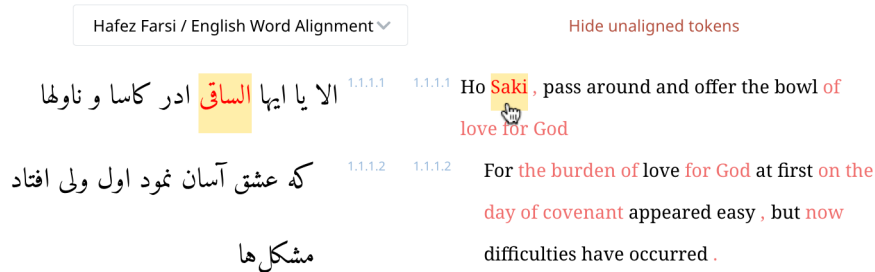
Beyond Translation, *Iliad*. Didakta Modular Grammar by Farnoosh Shamsian: <https://beyond-translation.perseus.org/reader/urn:cts:greekLit:tlg0012.tlg001.perseus-grc2:1.1-1.7?mode=grammatical-entries>

Figure 2 illustrates links from individual words to explanations in a machine-readable grammar. The reading environment uses highlighting to identify words with grammatical annotations and then displays just those grammatical features that appear in the selected passage. When readers select an annotated word, they see the grammatical explanation on the right, while all words with the same grammatical feature have an additional layer of highlighting: for example, *aeide*, “sing!”, *teuche*, “it was fashioning,” and *eteleieto*, “it was brought to completion,” have all been tagged as instances of the imperfect of continuance. Links from spans to external datasets allow us to support an open-ended range of annotation classes, of which the following provide some examples.

**Second**, we consider word- and phrase-level alignments between source texts and translations. We can generate these alignments between Greek or Latin source texts and English translations with reasonable accuracy (ca. 80%), but we can also manually create alignments between source texts and translations. Further, we can generate born-digital alignments designed to illustrate the literal meanings of a source text. And of course, in a digital reading environment, we can offer both literal and literary translations so that readers can choose which they wish to see.

In figure 3, Maryam Foradi manually aligned words and phrases taken from Persian poetry by Hafez with the 1891 [English translation](#) by Henry Wilberforce Clarke. In the figure above, the reader has moused over “Saki” and sees that name highlighted in the Persian original. More importantly, however, the reader can see that the words in light red have no equivalent in the original Persian. By showing what the translator has added, the annotator has revealed to readers with no knowledge of Persian that a layer of religious allegory has

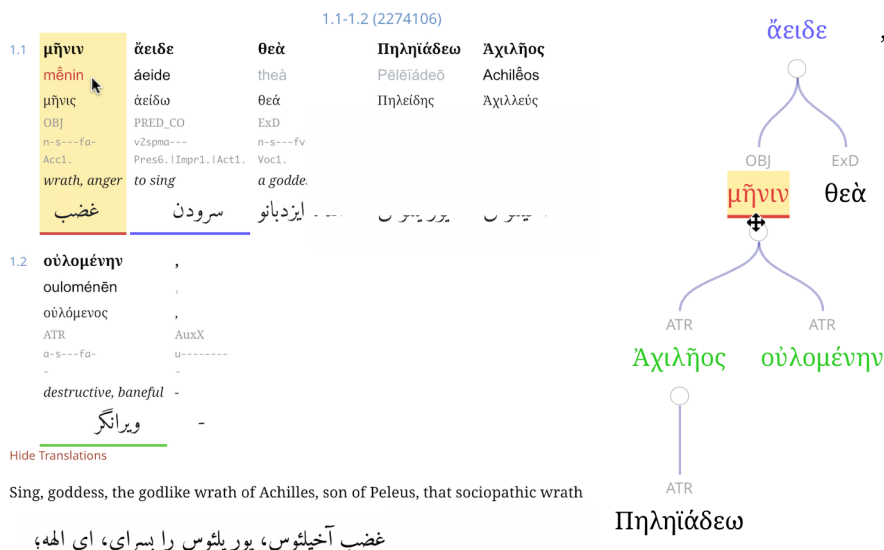
been imposed upon the original (Palladino, Foradi, and Yousef 2021; Foradi et al. 2019). Translation alignment can be revealing in itself, but becomes much more powerful when readers can go beyond the translation and explore the form and function of each word in a source text.



**Figure 3:** Manual alignment of a Persian poem by Hafez with a nineteenth-century English translation

Beyond Translation. *Divan*. Hafez Farsi/English word alignment: <https://beyond-translation.perseus.org/reader/urn:cts:farsiLit:hafz.divan.perseus-far1-hemis:1.1?mode=alignments&rs=urn%3Acite2%3Aascife-viewer%3Aalignment.v1%3Ahafz-farsi-english-word-alignment-temp>

This leads to the second class of data: linguistic annotations. The Perseus team had begun planning to develop rich linguistic annotation in 2001. More than one million words each of Greek and Latin are available in manually treebanked form, while machine learning allows us to produce automatically generated treebanks for any online Greek or Latin corpus.



**Figure 4:** The opening of the Homeric *Iliad* with linguistic annotations

Beyond Translation. *Iliad*. Greek text from Munro & Allen. Treebank by Gregory Crane from Ancient Greek Dependency Treebank: [https://beyond-translation.perseus.org/reader/urn:cts:greekLit:tlg0012.tlg001.perseus-grc2:1.1-1.7?mode=syntax-trees&collectionUrn=urn%3Acite2%3Abeyond-translation%3Atext\\_annotation\\_collection.atlas\\_v1%3Ail\\_gregorycrane\\_gAGDT](https://beyond-translation.perseus.org/reader/urn:cts:greekLit:tlg0012.tlg001.perseus-grc2:1.1-1.7?mode=syntax-trees&collectionUrn=urn%3Acite2%3Abeyond-translation%3Atext_annotation_collection.atlas_v1%3Ail_gregorycrane_gAGDT)

Figure 4 above (*left*) shows annotations for the *Iliad* that address the needs of readers who wish to push beyond translation alignments and explore the form and function of each word. There are seven layers of annotation for each word in the opening of the Homeric *Iliad*.

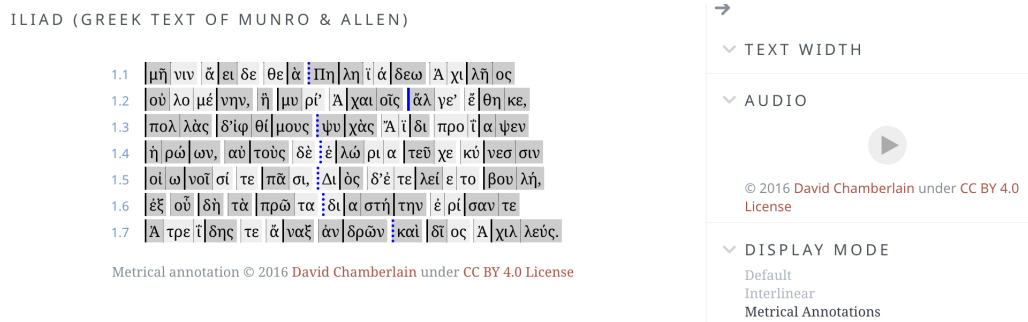
1. Transliteration
2. Regularised dictionary form
3. Syntactic role (using a dependency grammar for cross-lingual analysis)
4. Part-of-speech tagging
5. A (potentially language-specific) grammatical tag
6. An English gloss
7. A Persian gloss

These seven layers are drawn from different sources and, for all practical purposes, can only be managed as a series of linked datasets. Layer 1 was automatically generated—unlike languages such as Arabic and Persian, Greek regularly includes short vowels and can be automatically transliterated. Layers 2, 3 and 4 are derived from the Perseus Dependency Treebank. Layer 5 is based on a separate stream of annotations that link words in the text to a language-specific Greek grammar to shed light on grammatical features that cannot be inferred from a dependency grammar that was designed to represent shared features across multiple languages. Layers 6 and 7 are English and Persian glosses. Persian is included because collaborator Farnoosh Shamsian has completed a dissertation on localising the study of Ancient Greek in Persian.

Translations of the Greek sentence into English and Persian are visible below. The goal is to support an open-ended number of modern languages and to localise the platform as a whole (e.g. replace the English with Persian or some other language).

**Third**, the Greek quoted above is in poetic form and that poetic form is not easy for those studying Greek to decipher. Here we can draw upon yet another class of annotation: machine-actionable metrical analyses and recorded readings so that audiences can hear poetry as poetry and learn how to read poetry in metrical form.

The figure below presents a metrical analysis for the opening lines of the *Iliad*, with darker grey representing long syllables, lighter grey shorter syllables and dark vertical bars delineating breaks between the six metrical units of the hexameter. To the right, a button allows readers to listen to a recording of the lines being read. Together, the diagram and sound recording make it possible for readers with no Greek to learn how to read Greek poetry (this is now a regular assignment for students who do not know Greek). When readers combine the metre and recording with the aligned translation and linguistic annotation, they are able to engage directly with the Greek in ways that were not previously feasible.



**Figure 5:** Metrical analysis and recording of the opening of the *Iliad*

Beyond Translation. *Iliad*. Greek text from Munro & Allen. Metrical annotations by David Chamberlain, imported from Hypotactic: <https://beyond-translation.perseus.org/reader/urn:cts:greekLit:tlg0012.tlg001.perseus-grc2:1.1-1.7?mode=metrical>

Note that the metrical annotations do not correspond to word breaks: metrical units often begin and end in the middle of words. Thus, this class of annotation (along with other forms such as analysis of the morphemes within words) requires an ability to go beyond the token level and annotate chunks of a word.

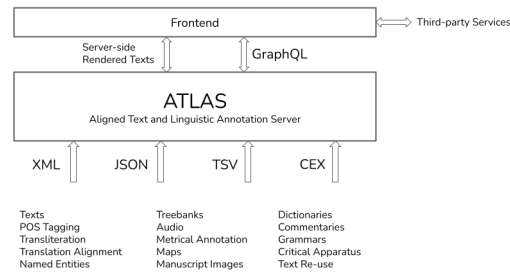
Other examples discussed elsewhere include the use of named entity annotation to generate maps and social networks and links at the word and character level between transcriptions and images of inscriptions, papyri or manuscripts.

The Beyond Translation project implemented initial frontends for born-digital annotations such as those listed above. In so doing, Beyond Translation also created an initial backend that imported different types of data from different projects into a coherent backend. In late 2023, a new National Endowment for the Humanities (NEH)-funded project, Perseus on the Web: Preparing for the Next Thirty Years, reviewed and reorganised the backend in general and the particular formats that we used to make data from multiple sources interoperable.

## Perseus 6, the ATLAS architecture and the CTS Data Model

While Scaife addressed a number of core needs, James Tauber and his collaborators (in particular Jacob Wegner) felt that they needed an approach to complement the Scaife architecture. With funding from Mellon and then from the NEH, they began developing ATLAS, the Aligned Text and Linguistic Annotation Server architecture.





**Figure 6:** The ATLAS architecture

The [CTS data model](#) allows us to integrate data between CTS-compliant TEI XML and a wider range of data in ATLAS. It also allows us to identify chunks of data with a URN (Uniform Resource Name)<sup>13</sup> such as the following:

`urn:cts:greekLit:tlg0012.tlg001.perseus-grc2:1.1-1.7`

- ▶ `cts`—This defines the CTS protocol. The data that follows in the URN is unique within this CTS protocol.
- ▶ `greekLit`—This is the name space that Perseus and the Open Greek and Latin project use to define Ancient Greek literature.
- ▶ `tlg0012`—This describes a text group. In most cases, a CTS text group will correspond to an author (e.g. Plato or Vergil), but the more general term text group allows us to also deal with cases (such as the New Testament) where we need to be able to address a collection of works by multiple authors as a single unit (tlg0031). Likewise, we use tlg0012 to describe the *Iliad* and the *Odyssey*, whether or not we view these to be the works of a single individual.
- ▶ `tlg001`—This describes the particular work and, in this case, tlg0012.tlg001 designates the *Iliad*.
- ▶ `perseus-grc2`—This designates the particular edition of the *Iliad* that we are citing (in this case, the 1908–1920 Munro & Allen *Oxford Classical Text of the Homeric Epics*).
- ▶ `1.1-1.7`—This defines a text range that extends from line 1 through to line 7 of book 1 of the *Iliad*.

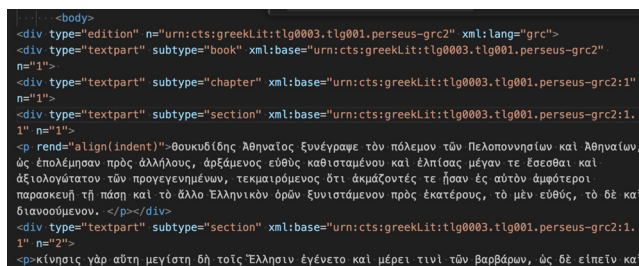
The following sections provide examples for some, but not all, of the annotation classes that we manage in ATLAS. Readers can follow the repositories available on GitHub for both the [tagging](#) pipeline and [data preparation](#) for more information.

13. [https://en.wikipedia.org/wiki/Uniform\\_Resource\\_Name](https://en.wikipedia.org/wiki/Uniform_Resource_Name)

## Scaife texts in ATLAS

While ATLAS helps us integrate data from many different sources, it also provides us with a simpler way to integrate basic textual data. We have completed an initial conversion into ATLAS of all of the texts that are now available in Scaife.

The CapiTainS CTS software library has explicit guidelines on how to structure XML markup, how to organise XML files in a directory and how to represent metadata. These requirements help develop more interoperable collections and can thus simplify search, analysis and visualisation. Nevertheless, the barrier to entry is substantial. ATLAS allows us to add texts that are not CapiTainS-compliant. ATLAS only requires a flat, tab-separated values (**TSV**) format that includes a unique reference and a chunk of text. We can convert TEI XML into this format, producing a TSV file that contains every book/chapter/section of a particular edition of Thucydides' *History of the Peloponnesian War* or every line from a particular edition of Sophocles' *Antigone*. Thus, the TEI XML version of Henry Stuart Jones' edition of Thucydides (available in Scaife) is in the following format:

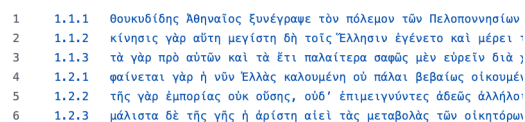


```
<body>
<div type="edition" n="urn:cts:greekLit:tlg0003.tlg001.perseus-grc2" xml:lang="grc">
<div type="textpart" subtype="book" xml:base="urn:cts:greekLit:tlg0003.tlg001.perseus-grc2"
n="1">
<div type="textpart" subtype="chapter" xml:base="urn:cts:greekLit:tlg0003.tlg001.perseus-grc2:1"
n="1">
<div type="textpart" subtype="section" xml:base="urn:cts:greekLit:tlg0003.tlg001.perseus-grc2:1.
1" n="1">
<p rend="align(indent)">Θουκυδίδης Ἀθηναῖος ξυνέγραψε τὸν πόλεμον τῶν Πελοποννησίων καὶ Ἀθηναίων,
ὡς ἐπολέμησαν πρὸς ἀλλήλους, ἀρξάμενος εὐθὺς καθισταμένου καὶ ἐλπίσας μέγαν τε ἔσεσθαι καὶ
ἀξιολογώτατον τῶν προγεγενημένων, τεκμαιρόμενος ὅτι ἀκμάζοντες τε ἦσαν ἐς αὐτὸν ἀμφοτέρω·
παρασκευὴ τῇ πάσῃ καὶ τὸ ἄλλο Ἑλληνικὸν ὄρων ξυνιστάμενον πρὸς ἑκατέρους, τὸ μὲν εὐθεὺς, τὸ δὲ καὶ
διανοοῦμενον. </p></div>
<div type="textpart" subtype="section" xml:base="urn:cts:greekLit:tlg0003.tlg001.perseus-grc2:1.
1" n="2">
<p>κίνησις γὰρ αὕτη μεγίστη δὴ τοῖς Ἕλλησιν ἐγένετο καὶ μέρει τινὶ τῶν βαρβάρων, ὡς δὲ εἰπεῖν καὶ
```

**Figure 7:** Opening of Thucydides in TEI XML

TEI XML file for Thucydides. GitHub. Perseus Digital Library: <https://github.com/PerseusDL/canonical-greekLit/blob/master/data/tlg0003/tlg001/tlg0003.tlg001.perseus-grc2.xml>

For Perseus 6, we create a TSV file with the name `tlg0003.tlg001.perseus-grc2.tsv` and then store the lines version of this file for annotation.



```
1 1.1.1 Θουκυδίδης Ἀθηναῖος ξυνέγραψε τὸν πόλεμον τῶν Πελοποννησίων
2 1.1.2 κίνησις γὰρ αὕτη μεγίστη δὴ τοῖς Ἕλλησιν ἐγένετο καὶ μέρει
3 1.1.3 τὰ γὰρ πρὸ αὐτῶν καὶ τὰ ἔτι παλαιότερα σαφῶς μὲν εὐρεῖν διὰ
4 1.2.1 φαίνεται γὰρ ἡ νῦν Ἑλλὰς καλουμένη οὐ πάλαι βεβαίως οἰκουμένη
5 1.2.2 τῆς γὰρ ἐμπορίας οὐκ οὐσης, οὐδ' ἐπιμεινόντες ἀδεῶς ἀλλήλο·
6 1.2.3 μάλιστα δὲ τῆς γῆς ἡ ἀρίστη αἰεὶ τὰς μεταβολὰς τῶν οἰκητόρων
```

**Figure 8:** ATLAS TSV version of the opening of Thucydides

TSV file for Thucydides. GitHub. Scaife Viewer: <https://github.com/scaife-viewer/tagging-shard-06/blob/main/tlg0003/tlg001/tlg0003.tlg001.perseus-grc2.tsv>

Because the TSV file becomes the basis for further computation, we can add new texts much more easily by converting them into this TSV identifier+text format. The more heavily structured format of the CapiTainS framework does



## Dictionaries

We have added most of the [dictionaries](#) available in Perseus to ATLAS and plan to add more (notably the *Intermediate Liddell Scott Greek-English Lexicon* and Slater’s *Lexicon to Pindar*). The format is JSON and captures the structure and inline formatting in the TEI XML. An entry from Cunliffe’s *Lexicon of the Homeric Dialect* follows below:

```
{
  "headword": "ἀγηνόπειη",
  "data": {
    "content": "<p>-ης, ῆ</p> <p>[ἀγῆνωρ.]</p>",
    "senses": [
      {
        "label": "1",
        "urn": "urn:cite2:exploreHomer:senses.atlas_v1:1.117",
        "definition": "Courage, spirit",
        "citations": [
          {
            "urn": "urn:cite2:scholarlyEditions:citations.v1:1.117_1",
            "data": {
              "ref": "Il. 12.46",
              "quote": null,
              "urn": "urn:cts:greekLit:tlg0012.tlg001.perseus-grc2:12.46"
            },
            "urn": "urn:cite2:scholarlyEditions:citations.v1:1.117_2",
            "data": {
              "ref": "Il. 22.457",
              "quote": null,
              "urn": "urn:cts:greekLit:tlg0012.tlg001.perseus-grc2:22.457"
            }
          ]
        },
        {
          "label": "2",
          "urn": "urn:cite2:exploreHomer:senses.atlas_v1:1.118",
          "definition": "The quality in excess or with arrogance.",
          "citations": [],
          "children": [
            {
              "label": "",
              "urn": "urn:cite2:exploreHomer:senses.atlas_v1:1.119",
              "definition": "In pl.",
              "citations": [
                {
                  "urn": "urn:cite2:scholarlyEditions:citations.v1:1.119_1",
                  "data": {
                    "ref": "Il. 9.700",
                    "quote": null,
                    "urn": "urn:cts:greekLit:tlg0012.tlg001.perseus-grc2:9.700"
                  }
                }
              ]
            }
          ]
        }
      ]
    },
    "urn": "urn:cite2:exploreHomer:entries.atlas_v1:1.60"
  }
}
```

**Figure 10:** An entry from Cunliffe’s *Lexicon of the Homeric Dialect*

Scaife Viewer. GitHub: <https://github.com/scaife-viewer/atlas-data-prep/tree/main/test-data/dictionaries/cunliffe-1-lex>

## Textual notes

[Textual notes](#) documenting different versions of a source constitute a core element for any critical edition. The unofficial but de facto development partnership with Brill (which uses Scaife for its scholarly editions) has provided us with a solution to that problem. We can and will include textual notes wherever these are available.

For now, we store textual notes as a special type of commentary. Figure 11 below modifies the first line of “The Passionate Shepherd to His Love” by Christopher Marlowe. The default line is: “Come live with mee and be my love.” The annotation above it indicates that there is a note associated with the phrase “live with me” and that this note, in this case, is a textual variant. The underlying data uses the TEI XML guidelines for textual editions to encode the variants. The Marlowe edition was actually produced in the late 1990s and we were able to import TEI annotations that were more than 25 years old, illustrating the initial stability and sustainability of that particular work and of this aspect of TEI XML.

```
{
  "references": [
    {
      "urn:cts:engLit:mds822-32.tpsth1-1599.pdl-eng:1.1"
    }
  ],
  "commentary": "<span>If thou wilt live</span>",
  "fragment": "live with mee",
  "ve_refs": [
    "1.1.t2",
    "1.1.t3",
    "1.1.t4"
  ],
  "idx": "1",
  "urn": "urn:cite2:scaife-viewer:commentary.v1:commentary2",
  "witnesses": [
    {
      "value": "Rs",
      "label": "MS Rodenbach"
    }
  ]
},
```

**Figure 11:** A textual note for a poem by Christopher Marlowe

Beyond Translation.

## Text alignments

Another category of data and services provided is [text alignments](#).<sup>14</sup>

```
{
  "urn": "urn:cite2:scaife-viewer:alignment-record.v1:iliad-word-alignment-parrish-998078bc3bab42978b47fa8e8b852cae_3",
  "relations": [
    [
      {
        "urn:cts:greekLit:tlg0012.tlg001.parrish-eng1:1.1.t4",
        "urn:cts:greekLit:tlg0012.tlg001.parrish-eng1:1.1.t5"
      }
    ],
    [
      {
        "urn:cts:greekLit:tlg0012.tlg001.perseus-grc2:1.1.t1"
      }
    ]
  ]
},
```

**Figure 12:** A textual alignment between two words in an English translation and one word in a Greek edition of the *Iliad*

In figure 12, each alignment between one text and another is a unique annotation with a unique, citable identifier. The format allows for many-to-many alignments. In the example above, two tokens in Amelia Parrish’s translation of the first book of the *Iliad* are aligned with one token of Greek. An individual could create alignments showing how different translations analysed the same word or the individual could align a pre-existing translation with the source text. They could also create a born-digital translation designed for alignment. We use the term text alignments because the same method could be used to align different editions of a Greek text as well as a Greek text with an English or Persian translation.

## Syntax trees (treebanks)

Figure 13 shows part of a treebank represented as JSON. The tagset for this treebank is based on the [Perseus Dependency Treebank](#). Our goal is to begin

14. While only a brief example is given here, further discussion of text alignment and how it has been implemented in Beyond Translation and the beginning of Perseus 6 can be found here: <https://pdldatajournal.pubpub.org/pub/knjho2r7/release/1>. See also Crane, Shamsian, et al. (2023).

using the tagset from the [Universal Dependencies \(UD\) framework](#). That transition can largely be done automatically but will require some curation. The UD data, however, can be represented by the JSON that we have chosen so that we can manage treebanks in multiple formats.

```
{
  {
    "urn": "urn:cite2:beyond-translation:syntaxTree.atlas_v1:tlg0008-tlg001-grc-1",
    "treebank_id": "1",
    "words": [
      {
        "id": 1,
        "value": "φύλαρχος",
        "head_id": 79,
        "relation": "SBJ",
        "lemma": "φύλαρχος",
        "tag": "n-s---mn-"
      },
      {
        "id": 2,
        "value": "6'",
        "head_id": 79,
        "relation": "AuxY",
        "lemma": "6'",
        "tag": "d-----"
      }
    ]
  }
}
```

**Figure 13:** A treebank represented as JSON

Syntax Tree. ATLAS. Available on GitHub: [https://github.com/scaife-viewer/atlas-data-prep/blob/main/test-data/annotations/syntax-trees/gorman\\_syntax\\_trees\\_017\\_tlg0008.tlg001.perseus-grc1.json](https://github.com/scaife-viewer/atlas-data-prep/blob/main/test-data/annotations/syntax-trees/gorman_syntax_trees_017_tlg0008.tlg001.perseus-grc1.json)

## Audio annotations

The GitHub repository with audio annotations is available [here](#).

urn:cts:greekLit:tlg0012.tlg001.perseus-grc2:1.1	<a href="https://storage.googleapis.com/explorehomer-prod-media/tlg0012.tlg001/audio/1/line_1.mp4">https://storage.googleapis.com/explorehomer-prod-media/tlg0012.tlg001/audio/1/line_1.mp4</a>
urn:cts:greekLit:tlg0012.tlg001.perseus-grc2:1.2	<a href="https://storage.googleapis.com/explorehomer-prod-media/tlg0012.tlg001/audio/1/line_2.mp4">https://storage.googleapis.com/explorehomer-prod-media/tlg0012.tlg001/audio/1/line_2.mp4</a>
urn:cts:greekLit:tlg0012.tlg001.perseus-grc2:1.3	<a href="https://storage.googleapis.com/explorehomer-prod-media/tlg0012.tlg001/audio/1/line_3.mp4">https://storage.googleapis.com/explorehomer-prod-media/tlg0012.tlg001/audio/1/line_3.mp4</a>
urn:cts:greekLit:tlg0012.tlg001.perseus-grc2:1.4	<a href="https://storage.googleapis.com/explorehomer-prod-media/tlg0012.tlg001/audio/1/line_4.mp4">https://storage.googleapis.com/explorehomer-prod-media/tlg0012.tlg001/audio/1/line_4.mp4</a>
urn:cts:greekLit:tlg0012.tlg001.perseus-grc2:1.5	<a href="https://storage.googleapis.com/explorehomer-prod-media/tlg0012.tlg001/audio/1/line_5.mp4">https://storage.googleapis.com/explorehomer-prod-media/tlg0012.tlg001/audio/1/line_5.mp4</a>

**Figure 14:** Lines of the *Iliad* signed to recorded performances of each line

At present, we support alignments of text chunks to particular MP4 files. Each line in the TSV file points to a line of the *Iliad*, with the first field pointing to the text of a particular edition and the second to a recorded reading stored in a server.

## Attributions/credits

Arguably the most important challenge we face is to preserve and aggregate fine-grained credits for born-digital annotations.<sup>15</sup> Credits are easily represented for articles, editions, and even individual commentary notes. In a true digital library, however, a researcher may contribute one or more machine-actionable annotations: for example, they may provide the syntactic structure of a

15. The repository for annotation and credit information is available here: <https://github.com/scaife-viewer/atlas-data-prep/tree/main/test-data/annotations/attributions>.



sentence, identify which Antonius is which in a given context, create alignments between a Greek word and its translations in half a dozen passages or may publish metrical analyses for 250,000 lines of Greek and Latin poetry.

In the original workflow for the Perseus Dependency Treebank, for example, two independent annotators analysed each sentence while a senior editor reviewed and adjudicated places where the annotators differed. A final corrected version was published, with three credits for each sentence. When another project, run by a former Perseus project member and long-time collaborator, adapted the treebank to a new format, the credits were initially lost. There was no ill intention. There was simply no existing framework to preserve credits in the new format and the project needed to change format. Likewise, another major treebank project includes both automatically generated and manually curated treebank data. This second project indicates whether a sentence was automatically or manually produced, but the treebank does not identify the annotator. Because the manually produced treebanks are all available on GitHub, it is possible to identify who did what, but this second treebank project did not do so. Its focus was to aggregate existing treebanks and produce new ones as quickly as possible. Both treebank projects were working with limited time and probably intended to go back to resolve the credits issue, but did not find that practical.

In *Beyond Translation*, we preserved all the credit information that we received. We now have an initial framework by which to represent credits from multiple projects and to make it possible for contributors to develop portfolios showing their contributions across projects.

```
[
  {
    "role": "Annotator",
    "person": {
      "name": "Alex Lessie"
    },
    "organization": {
      "name": "University of Pennsylvania, Philadelphia, PA, USA"
    },
    "data": {
      "references": [
        "urn:cite2:exploreHomer:syntaxTree.v1:syntaxTree-tlg0012-tlg001-perseus-grc2-2277120",
        "urn:cite2:exploreHomer:syntaxTree.v1:syntaxTree-tlg0012-tlg001-perseus-grc2-2277121",
        "urn:cite2:exploreHomer:syntaxTree.v1:syntaxTree-tlg0012-tlg001-perseus-grc2-2277122",
        "urn:cite2:exploreHomer:syntaxTree.v1:syntaxTree-tlg0012-tlg001-perseus-grc2-2277123",
        "urn:cite2:exploreHomer:syntaxTree.v1:syntaxTree-tlg0012-tlg001-perseus-grc2-2277124",
        "urn:cite2:exploreHomer:syntaxTree.v1:syntaxTree-tlg0012-tlg001-perseus-grc2-2277125",
        "urn:cite2:exploreHomer:syntaxTree.v1:syntaxTree-tlg0012-tlg001-perseus-grc2-2277126",
        "urn:cite2:exploreHomer:syntaxTree.v1:syntaxTree-tlg0012-tlg001-perseus-grc2-2277127"
      ]
    }
  },
  {
    "person": {
      "name": "Farnoosh Shamsian"
    },
    "role": "Annotator",
    "organization": {
      "name": "Universität Leipzig: Leipzig, Sachsen, DE"
    },
    "data": {
      "references": [
        "urn:cite2:scaife-viewer:alignment.v1:crito-greek-english-word-alignment-7b34509f15734bd7a20b873aeb08eaa1",
        "urn:cite2:scaife-viewer:alignment.v1:crito-greek-farsi-word-alignment-tr1-7b34509f15734bd7a20b873aeb08eaa1",
        "urn:cite2:scaife-viewer:alignment.v1:crito-greek-farsi-word-alignment-tr2-7b34509f15734bd7a20b873aeb08eaa1"
      ]
    }
  }
]
```

**Figure 15:** Sample attribution data for treebanks and translation alignments

Treebank alignment: Alex Lessie. Text alignment: Farnoosh Shamsian.

Figure 15 shows credits for treebanking individual sentences in the *Iliad* (above) and for aligning particular words and/or phrases between the Greek text of the *Crito* and an English translation, and between a Greek text and a Persian translation. We can now begin to aggregate credits as seen in figure 16.

Translator	Farshid Rahimi	3
Translator	Mahdi Shojaian	3
Annotator	Alex Lessie, University of Pennsylvania, Philadelphia, PA, USA	2,081
Annotator	Daniel Lim Libatique, College of the Holy Cross, Worcester, MA, USA	293
Annotator	Florin Leonte, Central European University of Budapest, Hungary	89
Annotator	Jack Mitchell, Tufts University, Medford, MA, USA	2,410

**Figure 16:** An initial report showing what contributions individuals have made to the ATLAS server

For now the data is relatively simple. In figure 16 we can see that Farshid Rahimi and Mahdi Shojaian each contributed three translated sentences. Alex Lessie helped treebank 2,081 sentences. However, the goal is to provide richer information (e.g. allow viewers to see the translated or treebanked sentences) and also to show where annotators have contributed to more than one project (e.g. treebanks and translations), but the above is a first step in that direction.

## Next steps

As the ATLAS backend takes shape, our focus will shift to the next stage of work:

1. We need to build out the services available on the evolving ATLAS server: <https://atlas.perseus.tufts.edu/>.
2. We need to refine the ATLAS data already available on GitHub.
3. We need to add frontend support, developed in Beyond Translation (and discussed above) into the Scaife architecture. We will implement Perseus 6 by adding the ATLAS backend and Beyond Translation UI widgets to the earlier Scaife architecture.

## Conclusions and ongoing work

This paper provides information about the first release of the ATLAS architecture and representative data.

## References

- Abowitz, Sarah, Gregory Crane, and Alison Babeu. 2024. “Bridging the Understanding Gap: Helping Readers Engage Directly with Foreign-Language Sources More Easily.” Paper presented at the 2024 ACM/IEEE Joint Conference on Digital Libraries (JCDL ’24) (Hong Kong, China, 16–20 December).
- Zenodo. Version v1. <https://zenodo.org/records/14278721>.
- Almas, Bridget, and Thibault Clérice. 2017. “Continuous Integration and Unit Testing of Digital Editions.” *Digital Humanities Quarterly* 11 (4). <https://hal.science/hal-01709868/>.

- Barker, Elton, Chiara Palladino, and Shai Gordin. 2024. "Digital Approaches to Investigating Space and Place in Classical Studies." *The Classical Review* 74 (1): 1–19. <https://doi.org/10.1017/S0009840X23002858>.
- Berti, Monica. 2019. "Named Entity Annotation for Ancient Greek with INCEpTION." In *Proceedings of the CLARIN Annual Conference 2019*, edited by Kiril Simov and Maria Eskevich, 1–4. Leipzig, Germany: CLARIN.
- Berti, Monica. 2021. *Digital Editions of Historical Fragmentary Texts*. Digital Classics Books 5. Heidelberg: Propylaeum. <https://doi.org/10.11588/propylaeum.898>.
- Celano, Giuseppe G. A. 2024. "Opera Graeca Adnotata: Building a 34M+ Token Multilayer Corpus for Ancient Greek." ArXiv. <https://arxiv.org/abs/2404.00739>.
- Crane, Gregory. 1991. "Generating and Parsing Classical Greek." *Literary and Linguistic Computing* 6 (4): 243–45. <https://doi.org/10.1093/lilc/6.4.243>.
- Crane, Gregory. 1996. "Building a Digital Library: The Perseus Project as a Case Study in the Humanities." In *DL '96: Proceedings of the First ACM International Conference on Digital Libraries (Bethesda, USA, 20–23 March)*, 3–10. New York: Association for Computing Machinery. <https://dl.acm.org/doi/pdf/10.1145/226931.226932>.
- Crane, Gregory. 2019. "Beyond Translation: Language Hacking and Philology." *Harvard Data Science Review* 1 (2). <https://doi.org/10.1162/99608f92.282ad764>.
- Crane, Gregory. 2023. "Beyond Translation: A Reading Environment for the Next Generation Perseus Digital Library." *Perseus Journal of Data Preservation and Sustainability*. <https://pdldatajournal.pub-pub.org/pub/el65xygp/release/1>.
- Crane, Gregory, Alison Babau, Lisa Cerrato, Farnoosh Shamsian, James Tauber, and Jacob Wegner. 2023. "Perseus 6.0: Towards a Next Generation Reading Environment for Born-Digital Editions and Corpora." In *Proceedings of the 2023 ACM/IEEE Joint Conference on Digital Libraries (JCDL) (Santa Fe, USA, 26–30 June)*, 297–98. IEEE. <https://doi.org/10.1109/JCDL57899.2023.00066>.
- Crane, Gregory, Alison Babau, Lisa Cerrato, Farnoosh Shamsian, Jacob Wegner, James Tauber, and Charles Pletcher. 2024. "Beyond Translation: Translation as Gateway Rather Than Endpoint." White paper on a concluded project. Tufts University.
- Crane, Gregory, Farnoosh Shamsian, Lisa Cerrato, Alison Babau, Amelia Parrish, Carolina Penagos, James Tauber, and Jake Wegner. 2023. "Beyond Translation: Engaging with Foreign Languages in a Digital Library." *International Journal of Digital Libraries* 14 (March): 163–176. <https://doi.org/10.1007/s00799-023-00349-2>.
- DeRose, Steven J., David G. Durand, Elli Mylonas, and Allen H. Renear. 1990. "What Is Text, Really?" *Journal of Computing in Higher Education* 1 (2): 3–26. <https://doi.org/10.1007/BF02941632>.
- Du  , Casey, and Mary Ebbott. 2019. "The Homer Multitext within the History of Access to Homeric Epic." In *Digital Classical Philology: Ancient Greek and Latin in the Digital Revolution*, edited by Monica Berti, 239–56. Berlin, Boston: De Gruyter Saur. <https://doi.org/10.1515/9783110599572-014>.
- Eckart de Castilho, Richard, Jan-Christoph Klie, Naveen Kumar, Beto Boullosa, and Iryna Gurevych. 2018. "Linking Text and Knowledge Using the INCEpTION Annotation Platform." In *2018 IEEE 14th International Conference on E-Science (e-Science) (Amsterdam, Netherlands, 29 October to 01 November)*, 327–28. IEEE. <https://doi.org/10.1109/eScience.2018.00077>.
- Foradi, Maryam, Jan Ka  el, Johannes Pein, and Gregory R. Crane. 2019. "Multi-Modal Citizen Science: From Disambiguation to Transcription of Classical Literature." In *HT '19: Proceedings of the 30th ACM Conference on Hypertext and Social Media (Hof, Germany, 17–20 September)*, 49–53. New York: Association for Computing Machinery. <https://doi.org/10.1145/3342220.3343667>.
- Forstall, Christopher W., Simone Finkmann, and Berenice Verhelst. 2022. "Towards a Linked Open Data Resource for Direct Speech Acts in Greek and Latin Epic." *Digital Scholarship in the Humanities* 37 (4): 972–81. <https://doi.org/10.1093/lilc/fqac006>.
- Gorman, Vanessa B. 2020. "Dependency Treebanks of Ancient Greek Prose." *Journal of Open Humanities Data* 6 (1): 1. <https://doi.org/10.5334/johd.13>.
- Hudspeth, Marisa, Brendan O'Connor, and Laure Thompson. 2024. "Latin Treebanks in Review: An Evaluation of Morphological Tagging Across Time." In *Proceedings of the*

- 1st Workshop on Machine Learning for Ancient Languages (ML4AL 2024) (Bangkok, Thailand and online, August), 203–18. Kerrville, Texas: Association for Computational Linguistics. <https://aclanthology.org/2024.ml4al-1.21>.
- Keersmaekers, Alek. 2021. “The GLAUx Corpus: Methodological Issues in Designing a Long-Term, Diverse, Multi-Layered Corpus of Ancient Greek.” In *Proceedings of the 2nd International Workshop on Computational Approaches to Historical Language Change 2021 (online, August)*, edited by Nina Tahmasebi, Adam Jatowt, Yang Xu, Simon Hengchen, Syrielle Montariol, and Haim Dubossarsky, 39–50. Kerrville, Texas: Association for Computational Linguistics. <https://doi.org/10.18653/v1/2021.lchange-1.6>.
- Keersmaekers, Alek, Wouter Mercelis, Colin Swaelens, and Toon Van Hal. 2019. “Creating, Enriching and Valorizing Treebanks of Ancient Greek.” In *Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019)*, 109–17. Kerrville, Texas: Association for Computational Linguistics. <https://doi.org/10.18653/v1/W19-7812>.
- Keersmaekers, Alek, and Toon Van Hal. 2022. “Creating a Large-Scale Diachronic Corpus Resource: Automated Parsing in the Greek Papyri (and Beyond).” *Natural Language Engineering* 30 (5): 1035–1064. <https://doi.org/10.1017/S1351324923000384>.
- Liddell, Henry George, and Robert Scott. 1882. *An Intermediate Greek-English lexicon: Founded Upon the Seventh Edition of Liddell and Scott’s Greek-English Lexicon*. Oxford: Oxford University Press.
- Palladino, Chiara, Maryam Foradi, and Tariq Yousef. 2021. “Translation Alignment for Historical Language Learning: A Case Study.” *Digital Humanities Quarterly* 15 (3). <http://www.digitalhumanities.org/dhq/vol/15/3/000563/000563.html>.
- Palladino, Chiara, Farnoosh Shamsian, Tariq Yousef, David J. Wright, Anise d’Orange Ferreira, and Michel Ferreira dos Reis. 2023. “Translation Alignment for Ancient Greek: Annotation Guidelines and Gold Standards.” *Journal of Open Humanities Data* 9 (1): 22. <https://doi.org/10.5334/johd.131>.
- Romanello, Matteo, and Sven Najem-Meyer. 2024. “A Named Entity-Annotated Corpus of 19th Century Classical Commentaries.” *Journal of Open Humanities Data* 10 (1): 1. <https://doi.org/10.5334/johd.150>.
- Sansom, Stephen A. 2021. “Sedes as Style in Greek Hexameter: A Computational Approach.” *TAPA (Society for Classical Studies)* 151 (2): 439–67. <https://doi.org/10.1353/apa.2021.0017>.
- Sansom, Stephen A., and David Fifield. 2023. “SEDES: Metrical Position in Greek Hexameter.” *Digital Humanities Quarterly* 17 (2). <https://digitalhumanities.org/dhq/vol/17/2/000675/000675.html>.
- Simon, Rainer, Elton Barker, Leif Isaksen, and Pau De Soto Cañamares. 2017. “Linked Data Annotation Without the Pointy Brackets: Introducing Recogito 2.” *Journal of Map & Geography Libraries* 13 (1): 111–32. <https://doi.org/10.1080/15420353.2017.1307303>.
- Smith, Neel, and Christopher Blackwell. 2023. “Analytical Developments for the Homer *Multitext*: Palaeography, Orthography, Morphology, Prosody, Semantics.” *International Journal on Digital Libraries* 24 (3): 179–84. <https://doi.org/10.1007/s00799-023-00380-3>.
- Smith, David A., Jeffrey A. Rydberg-Cox, and Gregory R. Crane. 2000. “The Perseus Project: A Digital Library for the Humanities.” *Literary and Linguistic Computing* 15: 15–26. <https://doi.org/10.1093/lc/15.1.15>.

## Websites cited

- Ajax Multi-commentary Project: <https://github.com/mromanello/ajax-multi-commentary>.
- Ancient Greek and Latin Perseus Dependency Treebanks: [https://perseusdl.github.io/treebank\\_data/](https://perseusdl.github.io/treebank_data/).
- Arachne: <https://arachne.dainst.org/>.
- Brill’s Scholarly Editions: <https://scholarlyeditions.brill.com/>.
- CapiTainS Software Suite and Guidelines for Citable Texts: <http://capitains.org/>.

Daphne: <https://github.com/francescomambrini/Daphne>.  
De Gruyter Brill: <https://degruyterbrill.com/>.  
DICES (Digital Initiative for Classics: Epic Speeches): <https://www.dices.uni-rostock.de/en/about-dices/>.  
GLAUx Trees: <https://glaux.be/>.  
Homer Multitext: <https://www.homermultitext.org/>.  
Hypotactic: <https://hypotactic.com/>.  
INCEpTION: <https://inception-project.github.io/releases/32.1/docs/user-guide.html>.  
International Image Interoperability Framework: <https://iiif.io/>.  
Jacoby Online: <https://scholarlyeditions.brill.com/bnjo/>.  
*A Literary History of Medicine*: <https://scholarlyeditions.brill.com/lhom/>.  
Opera Graeca Adnotata: <https://github.com/OperaGraecaAdnotata/OGA>.  
Pedalion Trees. GitHub: <https://github.com/perseids-publications/pedalion-trees>.  
Perseids Treebanking Environment: <https://github.com/perseids-publications>.  
Perseus 4.0 (the Hopper): <https://www.perseus.tufts.edu/hopper/>.  
Perseus 5.0 (the Scaife Viewer): <https://scaife.perseus.org/>; source code: <https://github.com/scaife-viewer>.  
Perseus—Beyond Translation: <https://beyond-translation.perseus.org/>.  
Perseus 6.0—Scaife ATLAS Server: <https://atlas.perseus.tufts.edu/>.  
Perseus under Philologic: <https://perseus.uchicago.edu/>.  
Pleiades: <https://pleiades.stoa.org/>.  
PROIEL: <https://www.hf.uio.no/ifikk/english/research/projects/proiel/>.  
Recogito: <https://recogito.pelagios.org/>.  
SEDES: <https://github.com/sasansom/sedes>.  
Text Encoding Initiative (TEI): <https://tei-c.org/>.  
*The Pez Brothers' Correspondence*: <https://scholarlyeditions.brill.com/pez/>.  
Ugarit translation alignment editor: <https://ugarit.ialigner.com/>.  
Universal Dependencies framework: <https://universaldependencies.org/>.