HUMAN REVIEW | Número Especial, 2021 | ISSN 2695-9623
International Humanities Review / Revista Internacional de Humanidades
DOI: https://doi.org/ 10.37467/gkarevhuman.v10.3012

BUILDING A CORPUS OF OPEN ANCIENT GREEK: NEW WAYS TO LEARN, TEACH, AND PUBLISH

LEONARD MUELLNER

Brandeis University, United States

KEY WORDS

Open data
Corpus-building
Ancient Greek
Catalogue
Verification
Crowdsourcing
Wordnet
Text reuse
Open commentaries

ABSTRACT

Documents the history and rationale as well as the technology for the creation of an open online corpus of Ancient Greek texts, the Free First Thousand Years of Greek. Reveals the energy on a worldwide basis for building open data along with the boundless creativity around developing software tools for teaching, research, and publishing that exploit open data in ways that were impossible before.

Received: 19/ 05 / 2021
Accepted: 13/ 09 / 2021

The Free First Thousand Years of Greek (=FF1K) is an attempt to provide a free, open corpus of Ancient Greek texts attested in manuscript — in other words, not including inscriptions on stone and not yet including fragmentary texts on papyrus — from their earliest appearance up to around the 3rd Century CE. The corpus also includes later texts that reference and illuminate earlier work like the *Suda*, an encyclopedic compilation that dates to the 10th Century CE but incorporates earlier work and that is indispensable for research on the work of the first thousand years of Greek literature.

The FF1K began in 2008-2009 as a brainchild of Professor Neel Smith (Department of Classics, Holy Cross College, Worcester, MA USA) with support from the Center for Hellenic Studies (CHS) in Washington, DC USA. The first step was an effort by undergraduate summer interns under his direction to compile a catalogue of texts that the corpus would comprise.[1] The CHS engaged in this effort because access to the existing corpus of Ancient Greek texts, the Thesaurus Linguae Graecae (TLG: http://stephanus.tlg.uci.edu), even though it emanates from a public institution (University of California at Irvine), is fee-based. An individual's subscription to the TLG at this writing costs $140 per year, and institutional site licenses are available on a fee scale that depends on both the number of full-time employees at an institution and the number of anticipated users. Their practice has been to monitor users and adjust their fee, at times drastically, upon renewal if the anticipated number is incorrect. Since they calculate the site license fee based not solely upon usage but also upon the number of full-time employees at an institution, the cost of a license for a large institution even with a small number of students of Ancient Greek can be unaffordable.[2] And even with a subscription, whether individual or institutional, no TLG text (beyond a snippet for fair use) can be downloaded for personal or institutional use. The management of the TLG aggressively enforces its copyright on texts that are at least a thousand years old —it has the obligation to do so on *editions* of those texts that are still under copyright according to US law and for which they have obtained permission to publish, but it enforces a global copyright on the whole corpus.

I dwell on these details because they are so plainly contrary to the desire that burns brightly in many of us who have learned, taught, or studied in this subject area, to encourage and nurture worldwide interest in the literary and cultural achievements of Ancient Greece. It throws up not just a financial obstacle but also an intellectual one, since the only computational tool in this digital age that can be applied to the text or texts in the corpus is the TLG's own search tool.[3] None of the new and promising computational tools for lexicography, topic modeling, stylistic analysis, discourse analysis, text reuse, etc. can be used upon a closed corpus of texts that cannot even be accessed by a computer other than the TLG's own servers. In fact, as we will show, there is undeniable evidence of great energy to work creatively and constructively on a large and open access corpus of Ancient Greek texts, both to construct it and to work with it in surprising, innovative, and generously inclusive ways. My purpose in this article is to document that effort and the amazing interest and energy that is its engine as well as its consequence.

Since 2015, the FF1K has been part of the Open Greek and Latin (OGL) project

[1] For a more complete history and technical account of the Free First Thousand Years of Greek, see Muellner (2019).

[2] In 2003/4 the author inquired of the TLG on behalf of the CHS the cost of a TLG license to support a fledgling classics department at Fudan University in Shanghai (at that point less than ten students and faculty). He was told that the fee for a site license would be 35,000 US dollars per year. I did not know the number of full-time employees at Fudan, but its student population at the time numbered over 30,000.

[3] Curiously, the TLG license (http://stephanus.tlg.uci.edu/individual.php) allows "statistical analysis; the production of indices and concordances when [...] intended as intermediate tools by the scholar [...] engaged in research into the data and are *not intended for wider distribution* [my italics]; and all other text processing and text manipulations activities which are clearly identifiable as scholarly research." Since it is forbidden to download a text, the only way such "intermediate tools" can be created is the old-fashioned way, by hand. Doing anything recognizable as the "text processing" allowed under such a license would be challenging, and the words "not intended for wider distribution" are intended to prevent such a painfully created resource from ever being made public!

(https://www.opengreekandlatin.org), a project whose inclusivity betrays its own name: OGL actually aims to reinvent the notion of 'classical' texts beyond Greek and Latin to include pre-Modern texts in Hebrew, Arabic, Persian, Chinese, K'iche', Mandinka, and more. It is an international cooperative effort by several institutions — the University of Leipzig, Tufts University, Mt. Allison University, University of Virginia, Harvard University Library, and the CHS — to create what will in effect become the next instance of the Perseus Project (http://perseus.tufts.edu), which was the first corpus of Greek and Latin texts that were ever made available free online, starting in 1985.

Beginning in 2016, the Center for Hellenic Studies put out a call for undergraduate students with competence in Ancient Greek to become summer interns and spend June and July learning how to scan, OCR (make a scanned text computer-readable), proofread and verify the format of texts in the FF1K for inclusion into the Github repository of the Open Greek and Latin project. A small team of technicians at CHS had been trained by Thibault Clérice, then a member of the Leipzig team, to teach them the tools and processes required, and we had an overwhelming number of exceptionally qualified and eager applicants to choose among: over 120 people applied for three — the CHS administration found a way to add a fourth — positions. The applicants whom we chose did superb work, as also have the two sets of summer interns who followed. I have heard colleagues and funding agencies voice skeptical reflections on the quality of work done by undergraduates in such circumstances, but it is a grave error to underestimate the commitment, the comfort with technology, the esprit de corps, and the totally immersive nature of the work that young people are capable of, to say nothing of the quality of the results.

From 2016 until now, Professor Bruce Robertson, head of the Classics Department at Mt. Allison University and a long-time member of the Perseus team, has made significant advances in the accuracy, usability, and functionality of LACE, a piece of software that he developed as an Optical Character Recognition tool based on free, open software for making Ancient Greek, with all its diacritical marks and a huge variety of fonts, machine readable. Robertson, who started work on LACE in 2012, has developed a superb interface for the manual correction of texts that were scanned by his software. He has made what appears to be hard and boring work into something pleasant and rewarding — and accurate. Furthermore, he has designed the software so that human correction of the scanning software's results for a few pages generates correction data that can be used computationally to improve the accuracy of succeeding scans, and there is also intelligence built into the software from the beginning: LACE tries to verify every Ancient Greek word that it scans in dictionaries that are accessed by it. If verification fails, it then tries to use simple strategies to correct it, such as replacing double consonants with single ones, or removing spurious spaces between what may be adjacent syllables. Even if the result is lexically correct, such corrected 'words' are color-coded and must be verified by a human user. Here is what LACE looks like at the moment: on the left is an image of a Greek text made with a scanner, and on the right, what happens when the scanned image is 'zoned' (its scannable areas defined electronically) and then 'read' by LACE:

Figure 1. On left, a zoned image, with red borders around it; on right, 'read' image being corrected, with words highlighted in light green (unverified but lexically possible), light blue (lexically verified), olive green (autocorrected and unverified), dark green (numbers), and pink (unverified and lexically incorrect).

When the user puts a cursor on a word of the left-hand page and clicks on it, as has happened with the first pink word on the page, ψιλίππου, the pink word dons a blue outline, a little box appears below it with a magnified image of the word taken from the scanned page on the right, and at the same time the actual word is also highlighted in blue on the scanned page on the right, so the reader can see immediately that the actual word's uppercase Φ was 'misread' as a lowercase ψ. The word inside the pink box outlined in blue is editable, so the user can immediately correct it, for instance, by backspacing over the incorrect letter and typing the correct letter in place of it. With the cursor still in the pink box, the user then types a carriage return to signify that the word is now correct, at which point the pink box turns light blue, and the cursor automatically advances to the next word on the page. It is hard to make a mistake with such a well-designed system, though of course anything is possible. New mistakes or errors not properly corrected may also be automatically corrected or flagged as incorrect on a subsequent scan.

This software, which is free and can be downloaded and installed on any computer,[4] can now convert its corrected output into EpiDoc XML, a standard format for the markup of Ancient Greek texts. XML documents can be verified as to their structure with standard free software tools and then integrated into CTS

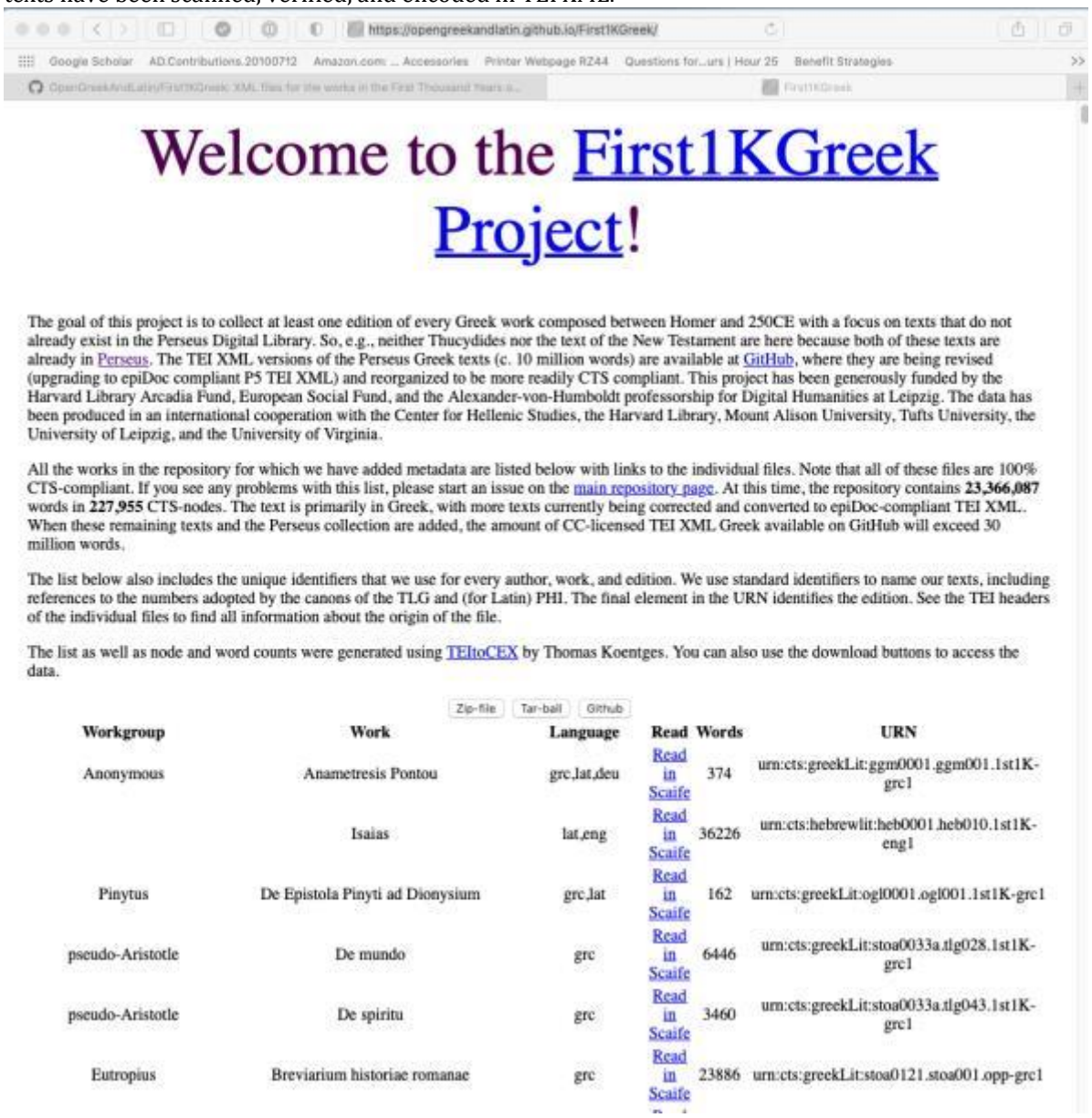
[4] The source code and packages are available at https://github.com/brobertson/Lace2/releases; installation requires familiarity with eXist-db packaging software.

(Canonical Text Services). CTS is a CHS-sponsored system for accessing elements as small as a letter of a text in a corpus as long as the text is structured and numbered in the way that almost all Greek texts are, not as pages that reflow and change with any new edition or printing, but as numbered lines or numbered paragraphs in numbered chapters and/or 'books' (scrolls).[5] That system currently requires a human being to edit a couple of lines of simple code in the header of the TEI XML file that identifies the structure and the reference system of the work, but once that happens, the verified, marked up, structured file can be uploaded into the OGL repository on Github, where it will be tested for structural integrity and checked for common encoding errors before formal inclusion in the projects repository of completed texts.

I will return in a moment to what happens next, but there is one crucial element that needs to be discussed first, namely, the canonical list of the works to be included in such a corpus. As mentioned above, the FF1K began with the creation of a catalogue, and the last few months has seen a leap forward in that part of the project. The basis for current work in this area is the massive Perseus catalog of Greek and Latin works through 600 CE developed and maintained since 2013 by Alison Babeu, Lisa Cerrato, and Anna Krohn, available at http://catalog.perseus.edu. Their work offers at least one link to an online edition of each text that it contains, along with metadata for ca. 4000 works in both classical languages. All the metadata fields in the catalog, such as authors' names, titles, Library of Congress catalog numbers, CTS URNs, and more, are searchable. From the Perseus Catalog, Alison Babeu generated a list of works to be included in the FF1K, and from that list, Thomas Köntges of the University of Leipzig and the Center for Hellenic Studies, developed a Github script to update regularly a public list of the works that have been scanned, converted into TEI XML, and uploaded to the project's Github repository (NB: this list does not include works already available as part of the Perseus Project, though eventually the two collections will be combined). This barebones list is available at https://opengreekandlatin.github.io/First1KGreek:

[5] For information on CTS, see http://cite-architecture.org/cts/

Figure 2. List of works currently in the FF1K not including those already in the Perseus collection of online texts. The texts have been scanned, verified, and encoded in TEI XML.



https://opengreekandlatin.github.io/First1KGreek/

Google Scholar AD.Contributions.20100712 Amazon.com: ... Accessories Printer Webpage RZ44 Questions for...urs | Hour 25 Benefit Strategies

Welcome to the First1KGreek Project!

The goal of this project is to collect at least one edition of every Greek work composed between Homer and 250CE with a focus on texts that do not already exist in the Perseus Digital Library. So, e.g., neither Thucydides nor the text of the New Testament are here because both of these texts are already in Perseus. The TEI XML versions of the Perseus Greek texts (c. 10 million words) are available at GitHub, where they are being revised (upgrading to epiDoc compliant P5 TEI XML) and reorganized to be more readily CTS compliant. This project has been generously funded by the Harvard Library Arcadia Fund, European Social Fund, and the Alexander-von-Humboldt professorship for Digital Humanities at Leipzig. The data has been produced in an international cooperation with the Center for Hellenic Studies, the Harvard Library, Mount Alison University, Tufts University, the University of Leipzig, and the University of Virginia.

All the works in the repository for which we have added metadata are listed below with links to the individual files. Note that all of these files are 100% CTS-compliant. If you see any problems with this list, please start an issue on the main repository page. At this time, the repository contains **23,366,087** words in **227,955** CTS-nodes. The text is primarily in Greek, with more texts currently being corrected and converted to epiDoc-compliant TEI XML. When these remaining texts and the Perseus collection are added, the amount of CC-licensed TEI XML Greek available on GitHub will exceed 30 million words.

The list below also includes the unique identifiers that we use for every author, work, and edition. We use standard identifiers to name our texts, including references to the numbers adopted by the canons of the TLG and (for Latin) PHI. The final element in the URN identifies the edition. See the TEI headers of the individual files to find all information about the origin of the file.

The list as well as node and word counts were generated using TEItoCEX by Thomas Koentges. You can also use the download buttons to access the data.

Zip-file Tar-ball Github

Workgroup	Work	Language	Read	Words	URN
Anonymous	Anametresis Pontou	grc,lat,deu	Read in Scaife	374	urn:cts:greekLit:ggm0001.ggm001.1st1K-grc1
	Isaias	lat,eng	Read in Scaife	36226	urn:cts:hebrewlit:heb0001.heb010.1st1K-eng1
Pinytus	De Epistola Pinyti ad Dionysium	grc,lat	Read in Scaife	162	urn:cts:greekLit:ogl0001.ogl001.1st1K-grc1
pseudo-Aristotle	De mundo	grc	Read in Scaife	6446	urn:cts:greekLit:stoa0033a.tlg028.1st1K-grc1
pseudo-Aristotle	De spiritu	grc	Read in Scaife	3460	urn:cts:greekLit:stoa0033a.tlg043.1st1K-grc1
Eutropius	Breviarium historiae romanae	grc	Read in Scaife	23886	urn:cts:greekLit:stoa0121.stoa001.opp-grc1

This item is about to be updated with a new interface into the metadata of the whole collection also by Köntges. Called *Pinakes*, after the tables of works compiled in the library of Alexandria by the learned poet Callimachus, it will provide access to all the metadata already in the Perseus catalog but will also include the ability to update, correct, and add metadata to existing texts in the collection, an important feature given the time over which the texts have been compiled and the inconsistencies that can result:

Figures 3a. *Pinakes* front end display, here for Plato's *Ion*. The public-facing software is an interface into FF1K and Perseus metadata that includes download options and links to a reader's view (Scaife Viewer) along with a list of related content.

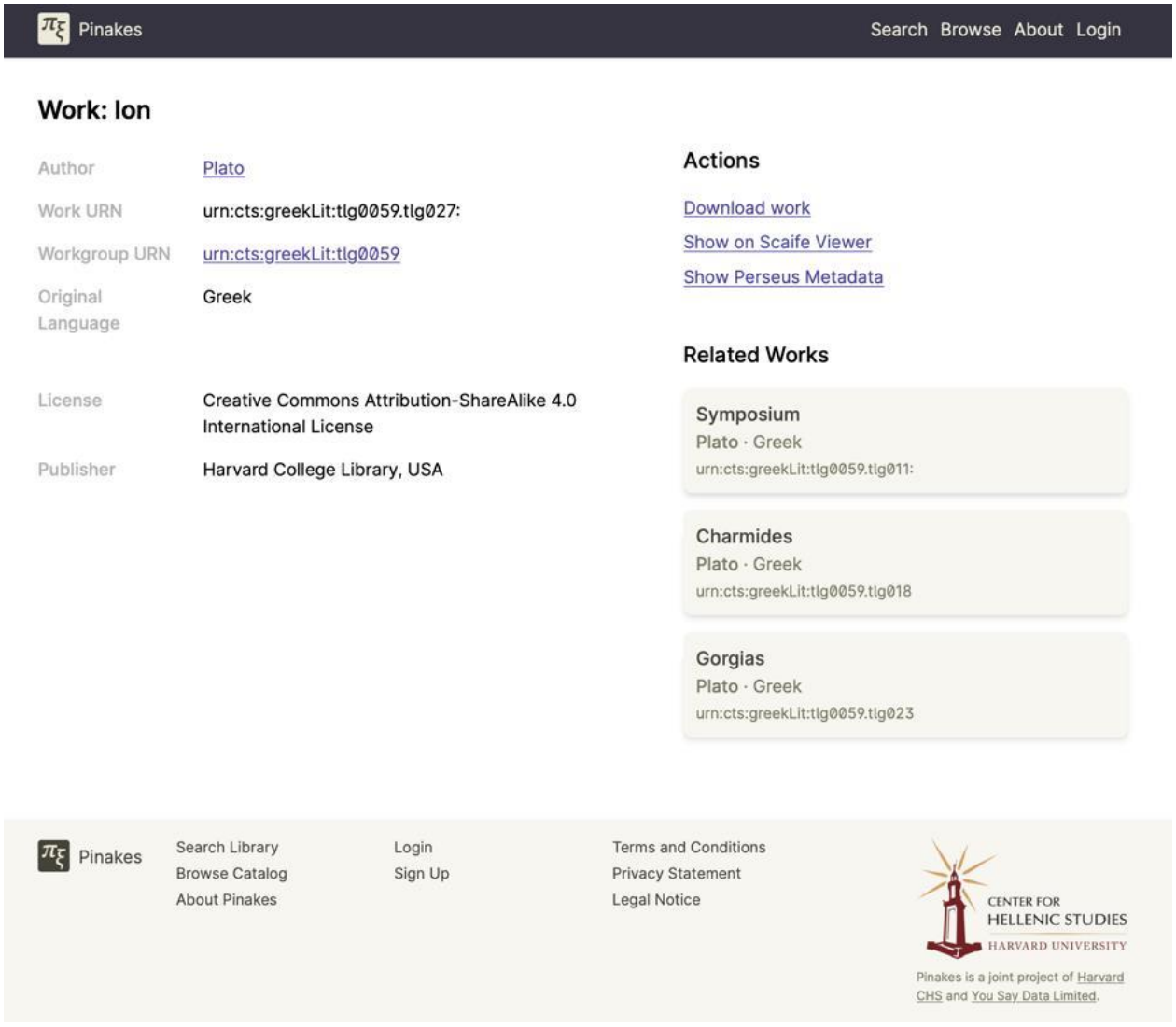
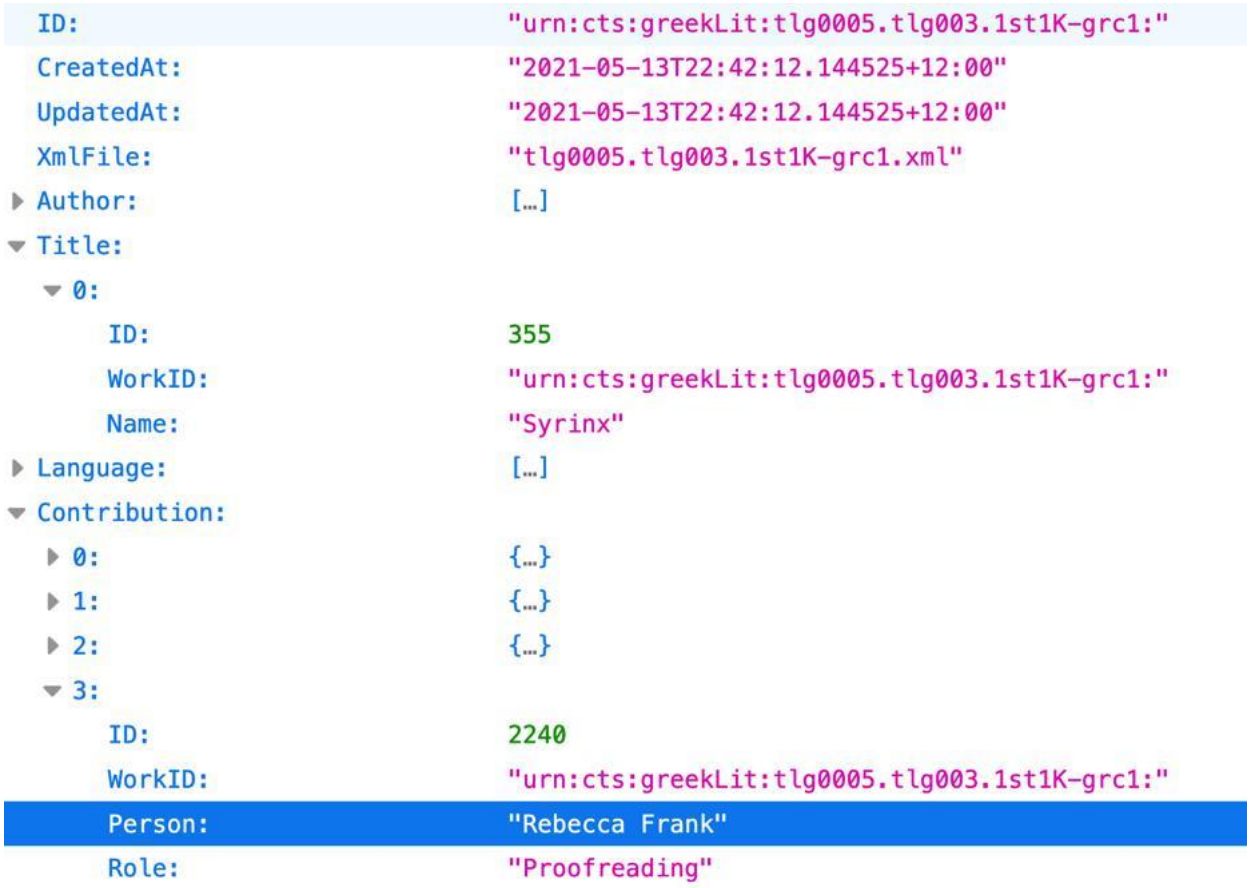


Figure 3b. *Pinakes* backend display for the *Syrinx* of Theocritus. The backend is an interface through which source files' metadata can be aggregated and batch edited, in this example, the name of the person responsible for proofreading the machine-readable text.

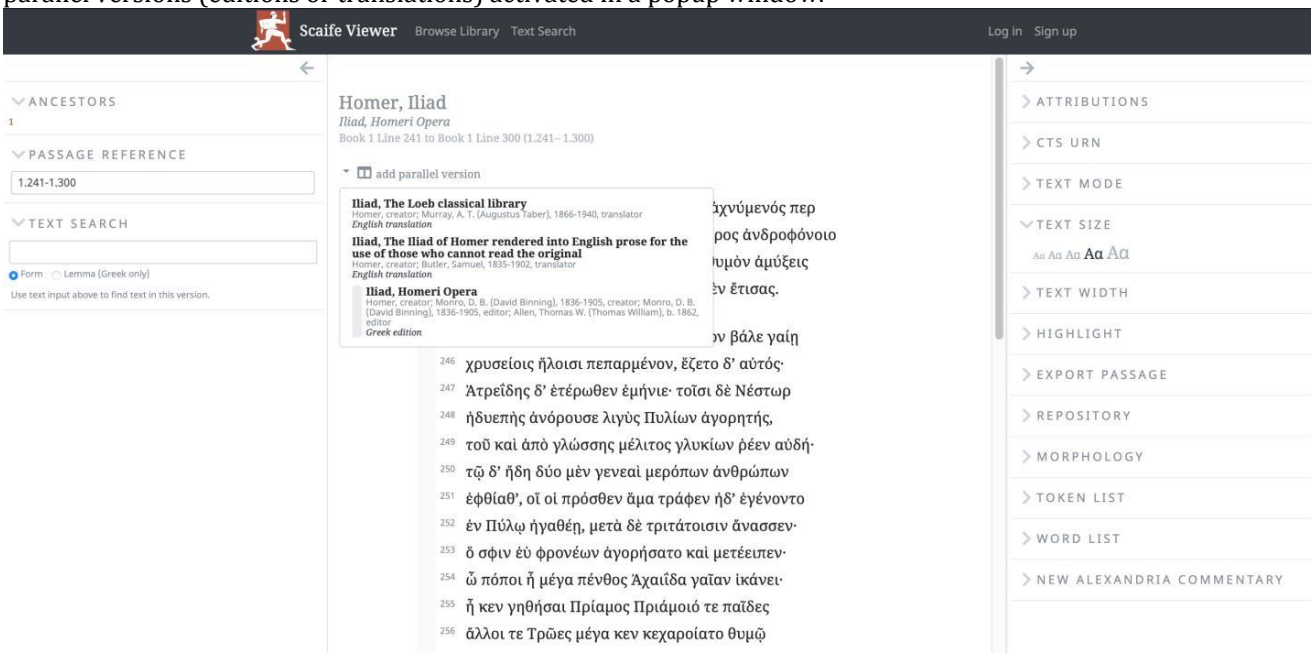


In the screen shot (Figure 3a) of the front end of *Pinakes*, new features are the options available to any user to download the specific text being viewed and the appearance of related texts (based on topic modeling technology). The backend (Figure 3b), which will be available to authenticated users, in other words, an expanding group of qualified users, goes further than anything before. In Köntges' own words, it "enables us to consolidate and edit metadata from different sources and then send amendments to each data source. The backend can be used independently as a web service."[6]

Another sign of the growing value that can be realized in a world of open data is the Scaife Viewer, links to which are provided in both of Köntges' projects. The Scaife Viewer, named after Ross Scaife, a dynamic and inspiring pioneer in digital classics whose life was cut short by cancer, is a website in progress for both searching and reading texts in the whole OGL corpus, not just the FF1K, for readers at levels of competence with the original languages ranging from none whatever to *ne plus ultra*. The current version is a harbinger of what is to come, but here is what a sample page in its reading environment looks like:

[6] Thomas Köntges, *per litteras* 05/14/21.

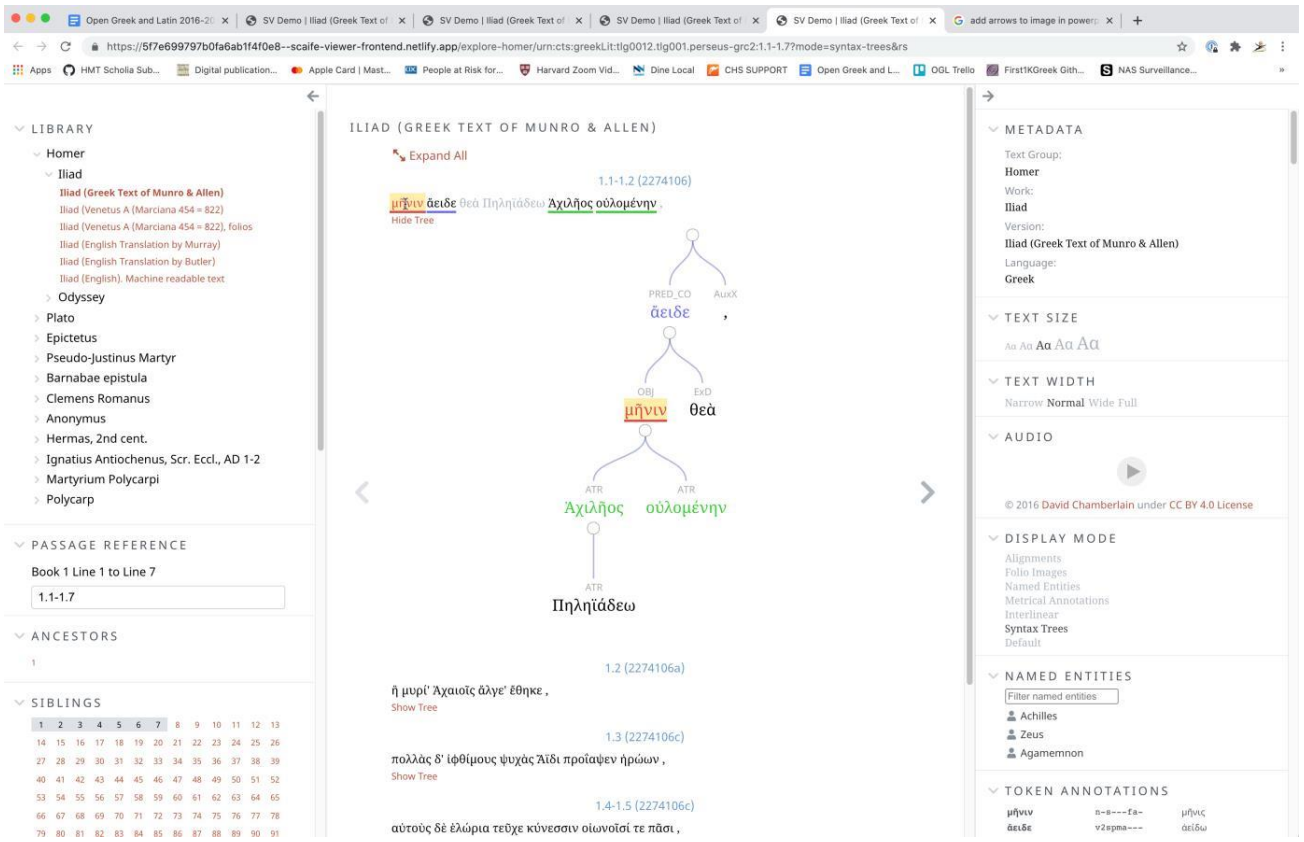
Figure 4. Scaife Viewer, showing a portion of Iliad 1.241-300 in middle column, with the option to choose among parallel versions (editions or translations) activated in a popup window.



In this screen shot, the left panel provides navigation information on the viewable text and the ability to select either a lemmatized search or one for an inflected form, both within the whole of the text being viewed, not just the lines displayed. In the right panel, a series of widgets are available to the reader for navigation, download (export passage), vocabulary (token list), grammatical parsing (morphology), as well as display (text size and width) and exegetical commentary.

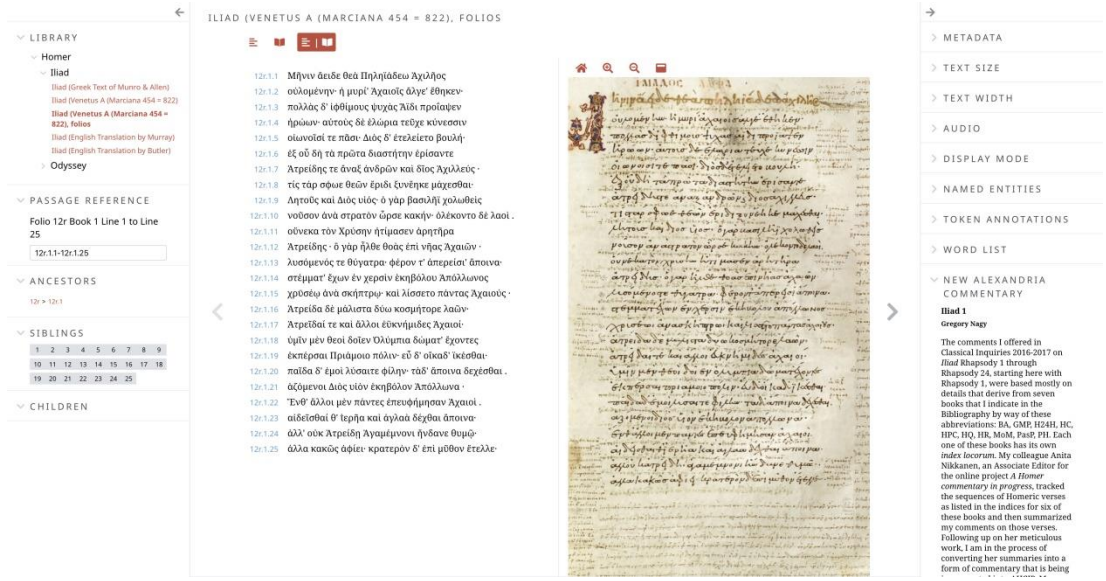
Here are some additional features being developed for this reading environment:

Figure 5a. Scaife features in development: display of syntax trees.



Displayed in Figure 5a, the middle panel, is a syntactical tree for the first line of the Homeric *Iliad*, whose words are also parsed in the "Token Annotations" widget on the right. That is one among several display modes that the reader can choose among on the right panel, the others being Alignments, in which the individual words of a text and its translation are linked (when possible; words added in a translation are signaled); Metrical Annotations, for poetry, in a lucid visualization that makes rhythmical reading easy; Named Entities, where persons and places specified in the text are identified and mapped; and Folio Images, as here:

Figure 5b. Scaife features in development: Folio images



In this screenshot, the machine-readable text of the first twenty-five lines of the Homeric Iliad on the left side of the middle panel is synchronized with their folio in the 10th Century MS. of the poem held in the Marciana library and known as Venetus A on the right. Also visible on the lower right of the left panel is a bit of the collaborative running commentary on the Iliad in progress on the New Alexandria platform. The platform itself, which is currently in beta (https://beta.newalexandria.info), will soon be public (June -July 2021, at https://newalexandria.info). It will provide a series of curated commentaries on select works but also an open platform for the public to generate and publish either individually authored or collaborative commentaries on the Greek and Latin texts of the Perseus Project and the FF1K but also works in Classical Hebrew, Chinese, Arabic, and more open data from the Classical Languages Toolkit (http://cltk.org).

The Scaife Viewer and the New Alexandria platform are but two examples of the burgeoning new reading and publishing tools that the open data of OGL and FF1K make possible. There are more are in development as well: *Tesserae* (https://tesserae.caset.buffalo.edu/), now in Version 5, a web tool that implements algorithms for exploring text reuse between portions of texts or whole works from the OGL corpus; or *Greek WordNet* (https://greekwordnet.chs.harvard.edu/), a new way of doing semantics that is based on cognitive as well as historical linguistics and that will soon far surpass the standard tools of lexicography in scientific quality. Both of these projects are thriving because of instantaneous, free access to a large corpus of texts that came to light over a long period of time.

One last example of the creativity and excitement that is being unleashed is the new effort to complete the Free First Thousand Years of Greek. Up to now, the work of converting scanned texts into machine readable form and then into TEI XML markup has been carried out by a consortium of institutions training and monitoring the work of graduate students and selected undergraduate interns (for the current group, see https://opengreekandlatin.org/). As of December 2020, there were still around 8 million words of Greek to be added to the 21 million already in Perseus and the FF1K repositories in order to complete the corpus. So, the OGL team decided to develop a pilot project to crowdsource the verification step, the most labor-intensive one, of the remaining texts. An intergenerational team at the Center of Hellenic Studies under the leadership of Angelia Hannhardt developed a plan to train and supervise volunteers. They chose as the first

group of trainers the twelve summer interns who had worked on the FF1K during June and July at the Center for the three years prior to the COVID-19 pandemic. To our surprise, eight of the twelve former interns were eager to take a chance and take part. Hannhardt and Robertson trained these trainers in the new features of LACE that Robertson had implemented over the previous year, and an open call for volunteers was launched in the beginning of February 2021.

Within three days, 250 people from around the world had signed up, which was many more than we expected, so the call was capped, and the working of training and supervising them began. Using free tools for tracking work and communicating like Slack and Trello, the trainers and volunteers quickly turned into a lively community that helped and taught each other. By the end of May, in four months' time, they will have verified approximately 450,000 words (or perhaps more: https://chs.harvard.edu/programs/oglvp/)! Once verified, the text files still need to be converted automatically to TEI XML, tweaked for CTS compliance, and tested for common errors, but the majority of the work for the digitization of those texts will be complete. Before the end of May, certificates of participation will be issued to volunteers, with specific mention of their achievements, and a new session including existing volunteers who wish to continue and new ones responding to an open call to join them will be made at the end of the month. They will be trained by the existing trainers as well as the CHS interns who were unable to do any work last summer. In addition, they will be joined by two CHS/Greece interns from the University of Athens. The project will continue with this refreshed team through the summer, and then a new call and a new training team will begin a new session in the fall. It would be difficult to invent a clearer indication of the world-wide appeal of the study of classical languages and of the excitement that flows from and with the creation of open data for it.

References

Muellner, L. (2019). The Free First Thousand Years of Greek. In: M. Berti (ed.), *Digital Classical Philology: Ancient Greek and Latin in the Digital Revolution* (pp. 55-89). Boston and Berlin: De Gruyter. https://doi.org/10.1515/9783110599572