Leonard Muellner

The Free First Thousand Years of Greek

Abstract: This contribution describes the ideals, the history, the current procedures, and the funding of the in-progress Free First Thousand Years of Greek (FF1KG) project, an Open Access corpus of Ancient Greek literature. The corpus includes works from the beginnings (Homeric poetry) to those produced around 300 CE, but also standard reference works that are later than 300 CE, like the Suda (10th Century CE). Led by the Open Greek and Latin project of the Universität Leipzig, institutions participating in the FF1KG include the Center for Hellenic Studies, Harvard University Libraries, and the library of the University of Virginia.

Ideals and early history of the project

The Free First Thousand Years of Greek (FF1KG), now a part of the Open Greek and Latin Project at the Universität Leipzig, was the brainchild of Neel Smith, Professor and Chair of the Department of Classics at the College of the Holy Cross, with the sponsorship and support of the Center of Hellenic Studies (CHS) in Washington, DC. It started in 2008–2009 from a set of ideals about digital classical philology that Professor Smith and the CHS have been guided by, as follows: 1) digital resources for classical philology should be free and openly-licensed and therefore accessible to all without cost and with the lowest possible technical barriers but the best technology available behind them; 2) software development flourishes long-term in an open environment that uses standardized and free tools and invites collegial participation,[1] as opposed to a closed environment that uses proprietary tools for short- (or even medium-) term gain; 3) in order to survive and thrive in the future, the field of Classics requires and deserves creative, well-designed, and practical digital resources for research and teaching that rigorously implement the two previous principles; 4) rather than presenting a broad spectrum of users with tools that are ready-made without their participation or input, it is best to enable, train, and involve young people, undergraduates and graduate

1 Raymond (1999), originally an essay and then a book, was inspirational for the present author on this point.

Leonard Muellner, Center for Hellenic Studies, Harvard University

Open Access. © 2019 Leonard Muellner, published by De Gruyter. BY-NC-ND This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.
https://doi.org/10.1515/9783110599572-002
Unauthenticated
Download Date | 8/23/19 4:11 AM

students both, in the technologies and the processes that are necessary for the conception, creation, and maintenance of digital resources for classics teaching and research; and 5) the markup of texts, whether primary or secondary, in internationally standard formats, such as TEI XML (http://tei-c.org), is the best way to guarantee their usability, interoperability, and sustainability over time.

The fundamental research and teaching tool that a field like Classics needs is as complete a corpus of open and downloadable texts as possible in each language, Greek or Latin, with a full panoply of ways to read, interpret, search, and learn from them. Building such a corpus from the bottom up is challenging in many obvious ways. Texts in Ancient Greek, which is the disciplinary focus of the Center for Hellenic Studies and the Free First 1K of Greek, present the challenging technical difficulty of an alphabet available in a wide variety of fonts (each standard for a given collection of texts, but there is no overall standard font), and with seven diacritical marks appearing singly and in combinations over and under letters (acute, grave, and circumflex accents; smooth and rough breathings; iota subscript and underdot). That makes it difficult to create machine-readable texts in Ancient Greek from printed texts using basic computational tools for optical scanning and character recognition. As a result, Neel Smith thought it would be wise to begin by making overtures on behalf of CHS to the existing but proprietary and fee-based corpus of Ancient Greek texts, the Thesaurus Linguae Graecae (TLG) in Irvine, CA, in an effort to partner with them in both improving and opening up their collection of texts.

By that time, Smith and his colleague, Christopher Blackwell, Professor of Classics at Furman University, had developed and perfected a protocol that they called CTS (Canonical Text Services, now in its 5th iteration, http://cite-architecture.org) for building, retrieving, querying, and manipulating a digital reference to an item as small as a letter or a chunk as large as anyone might need from a classical text, as long as the text in question is accessible by way of a structured, canonical reference system, and as long as the text is marked up in some form of XML that can be validated. In Smith's and Blackwell's parlance, a canonical reference system is one based on a text's *structure* (chapter and verse, or book and line, for instance) rather than on points in a physical page (like the Stephanus or Bekker page-based references that are normal for citing the works of Plato and Aristotle). They had also developed sophisticated ways of parsing and verifying machine-readable polytonic Greek against a lexicon of lemmatized forms. Both CTS and their verification tools seemed to Smith and Blackwell to offer significant advantages over the existing technologies of the TLG, but their attempt to partner with the leadership of the TLG was not well-received.

This left Smith, Blackwell, and the CHS with one option: to build a free and open corpus of texts from scratch. The initial, modest idea was to create a corpus

Unauthenticated
Download Date | 8/23/19 4:11 AM

of Ancient Greek texts that would answer to the basic needs of students and researchers of texts in the classical language and that would work with the CTS system. Such a scope implied several restrictions: 1) the corpus would include texts attested in manuscript, but not fragments (in other words, texts attested in snippets inside other texts) or inscriptions or papyri, whether literary or documentary, which do not have a canonical reference system; 2) the basic time frame would be from the beginnings of Greek literature up to the end of the Hellenistic period, around 300 CE, to include the Septuagint and the New Testament but not the Church Fathers; 3) some later texts necessary for the study of the basic corpus, such as the Suda, a 10th Century CE encyclopedia of antiquities, or the manuscript marginalia called scholia for a range of classical authors, some of which are pre- and some post 300 CE, would also be included in the collection. Hence the Free First Thousand Years of Greek is in some ways less and in some ways more than its name betokens.

First steps, then a suspension

The first requirement of the project was a catalog of the texts to be included in it, and Smith began the significant task of compiling one with funding from CHS for two student helpers in the summer of 2010; that work continued in the summer of 2011, but then other projects and obligations supervened. An overriding concern for the CHS technical team was the development of software for online commentaries on classical texts, an effort that resulted in the initial publication in 2017 of *A Homer Commentary in Progress*, an inter-generational, collaborative commentary on all the works of the Homeric corpus (more on its sequel and their consequences for the Free First Thousand Years of Greek follow). For Professor Smith, the focus of his energies became the centerpiece of the Homer Multitext Project (http://www.homermultitext.org), the interoperable publication of all of the photographs, text, and scholia of the Venetus A manuscript of the Homeric *Iliad* in machine-actionable, which took place this past spring; it will continue with the similar publication of other medieval manuscripts with scholia, such as Venetus B or the Escorial manuscripts of the Homeric *Iliad*.

Resumption of the FF1KG

But the Free First Thousand Years of Greek was never far from the concerns of either CHS or Professor Smith – in fact, both of these projects are intimately

Unauthenticated
Download Date | 8/23/19 4:11 AM

related to it – and in 2015, with the support of Professor Mark Schiefsky, then chair of the department of classics at Harvard University, we reached out in an attempt to collaborate with our long-term partner, Gregory Crane, editor-in-chief of the Perseus Project, Professor of Classics at Tufts University and Alexander von Humboldt Professor of Digital Humanities at the University of Leipzig. He and his team of colleagues and graduate students at Universität Leipzig and Tufts University had already begun a much more inclusive project that could reasonably subsume it, namely, the Open Greek and Latin (OGL) project.

OGL aims to be a complete implementation of the CTS protocols for structuring and accessing texts in XML documents; it aims to include multiple, comparable versions of a given classical text wherever possible, along with its translation into multiple languages; and it will provide *apparatus critici* (reporting textual variants) where the German copyright law allows them; in addition, it will include POS (part of speech) data for every word in the corpus, with the ultimate goal of providing syntactical treebanks of every text as well. It also will include support for fragmentary texts, such as the digital edition of K. Müller's edition of the fragments of Greek history, the DFHG, http://www.dfhg-project.org, with a digital concordance to the numbering of the fragments in the modern edition of F. Jacoby, which is still under copyright. Developing the infrastructure to include fragmentary texts of this kind has been a major achievement of Monica Berti, the editor-in-chief of the DFHG as well as of Digital Athenaeus, http://www.digitalathenaeus.org, an ancient text that presents canonical reference problems but is also a major source of fragmentary quotations of other texts from antiquity, many of them lost to us otherwise.[2]

Summer interns at CHS and the FF1KG workflow

The subsuming of the Free First Thousand Years of Greek to the Open Greek and Latin project began in earnest in March of 2016, when the CHS hired three summer interns from a pool of over 170 applicants to be trained in the technologies of the OGL and to contribute to the ongoing creation of the corpus of Greek texts. Professor Crane and his team graciously embraced the concept of the Free First Thousand Years of Greek, and because of the extraordinary work of Alison Babeu, a long-time member of the Perseus team, a catalog of works that would include it was already in place, namely, the Perseus Catalog,

2 See her contribution to this collection, entitled "Historical Fragmentary Texts in the Digital Age".

Unauthenticated
Download Date | 8/23/19 4:11 AM

http://catalog.perseus.org. In May of 2016, Crane sent Thibault Clérice, then a doctoral candidate at Leipzig (now MA director of the Master Technologies «Numériques Appliquées à l'Histoire» at the École Nationale des Chartes in Paris) to the CHS in Washington, DC in order to train the CHS year-round publications intern, Daniel Cline, and the author of this article, L. Muellner, in the workflow of the OGL. The idea was that we, in turn, would train the summer interns, who were scheduled to arrive at the beginning of June. Thibault was the right person for the job because he had developed a suite of Python-based tools called CapiTainS (https://github.com/Capitains) to verify that any TEI XML file was valid and in particular compliant with the CTS protocols. But before discussing his tools, we need to go back one step.

The process of generating and verifying files for inclusion in the Free First Thousand Years of Greek begins with high-resolution scans of Greek texts from institutional (for example https://archive.org) and individual sources. These scans are submitted to Bruce Robertson, Head of the Classics Department at Mt. Allison University in New Brunswick, Canada, who has developed a suite of tools for Optical Character Recognition of polytonic Ancient Greek called Lace (http://heml.mta.ca/lace/index.html and for the latest source, https://github.com/brobertson/Lace2). His software is based on the open source Ocropus engine. After its first attempt to recognize the letter forms and diacritics of a Greek text, Lace is set up for humans to check and correct computer-recognized Greek, with the original scanned image on pages that face the OCR version, in order to make verification quick and straightforward.

After someone corrects a set of pages in this interface, Robertson's process uses HPC (High Performance Computing) in order to iterate and optimize the recognition of letters and diacritics to a high standard of accuracy, even for the especially difficult Greek in a so-called *apparatus criticus* "critical apparatus". A critical apparatus is the textual notes conventionally set in small type at the bottom of the page in Ancient Greek and Latin texts (or for that matter of any text that does not have a single, perfect source). It reports both textual variants in the direct (manuscripts, papyri, etc.) and indirect (citations of text in other sources) transmission of ancient texts, along with modern editors' corrections to the readings from both transmissions. Correctly recognizing the letters and diacritics of lexical items in a language is one thing, but it is altogether another thing to reproduce the sometimes incorrect or incomplete readings in the manuscripts (and not to correct them!) that populate a critical apparatus, but Robertson's software can do both. In any case, he is continually optimizing it, and the most recent version uses machine-learning technology to correct its texts. Learning how to edit an OCR text is the first task that the CHS interns learn to do.

Unauthenticated
Download Date | 8/23/19 4:11 AM

Once a Greek text is made machine-readable by an iterated Lace process, OGL requires that it be marked up in EpiDoc TEI XML (for the EpiDoc guidelines, schema, etc., see https://sourceforge.net/p/epidoc/wiki/Home/; for TEI XML in general, see http://www.tei-c.org/). TEI XML endows the text with a suite of metadata in the TEI.header element as well as a structural map of the document (using Xpath) that is a requirement for the CTS protocol. Up to now, that encoding process has been carried out by Digital Divide Data (DDD), https://www.digitaldividedata.com, a third-world (Cambodia, Kenya, Indonesia) company employed by corporations and universities in the first world that trains and employs workers in digital technologies. This step is painstaking and not inexpensive, but by the time that the FF1KG joined them, the OGL team had already generated a large corpus of Greek and Latin texts with funds from multiple sources, including the NEH, the Mellon Foundation, the Alexander von Humboldt Stiftung, and others (see more below on new funding sources for further digitization expenses of this kind). Once an Ancient Greek text in the FF1KG has been marked up in EpiDoc by DDD, it is installed by the OGL team in the GitHub repository of the FF1KG, a subset of the OpenGreekandLatin repository, at http://opengreekandlatin.github.io/First1KGreek/.[3] The directory structure of the installations in that repository are consistent with the structure and numbering schemes of the Perseus catalog for authors and works, and the infrastructure files, such as dot-files like the .cts_xml files, are also consistent with the requirements of CTS.

These newly marked-up and installed sources were the subject of the majority of the work carried out by the CHS interns in the summers of 2016 and 2017; they also received year-round attention from members of the Leipzig team. Thibault Clérice had developed a verification tool called Hooktest (available in the previously cited CapiTainS GitHub directory) that could be run on all of the files in the repository to detect errors in them – flaws in the TEI headers within each XML file, flaws in the structural information specified for CTS compliance, and a host of other small but critical details that could go wrong in the process of generating EpiDoc XML that is CTS-compliant. In training Cline and Muellner in the spring of 2016, Clérice spent most of the time teaching us how to understand and correct and then rerun Hooktest in response to its error messages. Hooktest itself has been updated several times since then, and it now runs on a different system (originally ran on Docker, https://www.docker.com now the online server, Travis, https://travis-ci.org), and over the past three summers, the CHS interns have developed documentation that consolidates its accumulated wisdom on that

3. All files in this repository and the other OGL repositories are backed up at https://zenodo.org (last access 2019.01.31).

Unauthenticated
Download Date | 8/23/19 4:11 AM

process. In the past summer, there was a dearth of newly digitized files from DDD for the FF1KG, so the (now) *four* interns turned to the conversion and verification, again via Hooktest, of the XML files of the Perseus collection to CTS compliance as their major task. In addition to that work and further OCR work training Lace, the CHS summer interns have learned how to contribute to the DFHG (Digital Fragmenta Historicorum Graecorum) and the Digital Athenaeus projects mentioned above. Like the FF1KG, both are openly licensed projects that benefit from hearty participation by anyone who wants to add to and learn from them.

Funding sources and in-kind contributions to the FF1KG and the OGL

As mentioned above, the OGL has been funded over its development by a broad range of sources, including the NEH, the Mellon Foundation, the IMLS, and others. In 2016, the CHS committed $50,000 to fund steps in the digitization of Ancient Greek texts for the FF1KG, with the idea that it would be matched by other funding obtained by OGL. That sum of money has been earmarked and set aside for digitization of the FF1KG since 2016, and the expectation is that it will be spent and matched in 2019 as part of a grant to the OGL by the DFG (Deutsche Forschungsgemeinschaft, or German Research Association). The CHS also earmarked funds for the development of a user interface into the texts of the FF1KG; more about that in a moment. The CHS funds were not from the CHS endowment, but from revenue generated by the CHS publications program, its printed books, in particular the so-called Hellenic Studies Series. In the Fall of 2016, when she heard about renewed progress with the FF1KG, Rhea Karabelas Lesage, the librarian for Classics and Modern Greek Studies at Harvard University Library, applied for $50,000 of funding through the Arcadia Fund, and she succeeded in her application. That sum paid for the digitization and mark-up in EpiDoc by DDD of 4,000,000 words of Greek. In addition, in 2017, Rhea used funds from her budget as Classics librarian to digitize and include in the FF1KG a series of scientific texts for a course being given at Harvard University by Professor Mark Schiefsky, the Classics chair. Another Classics librarian, Lucie Stylianopoulos of the University of Virginia (UVA), became an enthusiastic supporter of the project, and every year since 2016, she has been successful in acquiring funding from the UVA library for a group of four to six interns during the Fall and Spring terms to learn the technologies and to contribute significantly to the conversion and verification of texts in the FF1KG repository. The UVA team originally (in 2016) trained at CHS, but this

Unauthenticated
Download Date | 8/23/19 4:11 AM

past September a CHS trainer, the publications intern Angelia Hannhardt, visited Charlottesville and worked with the new interns *in situ*. The same two Classics librarians, Lucie and Rhea, worked together with members of the Tufts team, especially Lisa Cerrato and Alison Babeu, along with David Ratzan and Patrick Burns of the Institute for the Study of the Ancient World (ISAW), to set up a workshop on the OGL and the FF1KG that was held at Tufts University a day before the annual meeting of the Society for Classical Studies (SCS) in Boston in January this year (2018). A large group (over sixty) of librarians, undergraduates, graduate students, and classics professionals came early to the conference in order to attend hands-on demonstrations of the technologies in FF1KG and OGL. Our hope was that they could begin to learn how to participate and also, how to teach others. The workshop was publicized and supported by the Forum for Classics, Libraries, and Scholarly Communication (http://www.classicslibrarians.org), an SCS-affiliated group that has advocated for and worked with the FF1KG team since it resumed development in 2016. Lastly, in response to outreach from Lucie Stylianopoulos, Rhea Lesage, and the librarians at CHS, a memorandum of understanding is about to be (in November, 2018) signed between the reinvigorated National Library of Greece (NLG) in its beautiful new location (see https://transition.nlg.gr) and the OGL/FF1KG team at Leipzig, to train staff and students in Athens in the processes of the development of the corpus. We expect that training and new work will begin there in the very near future.

New developments from an Open Access corpus of texts

Building a corpus of texts takes time, money, and dedicated workers like those from Leipzig, CHS, UVA and soon the NLG, but their work is invisible until there is a way to access it. The current list of texts in the FF1KG is visible and downloadable here: http://opengreekandlatin.github.io/First1KGreek/. There are now over 18 million words of Greek, with about 8 million to come for the "complete" FF1KG. Given that all the texts in the corpus are open access, anyone can download them and build software around them. The CHS leadership, with the agreement of the Leipzig team, wished to inspire an early "proof-of-concept" access system that would highlight the existence and some of the functionality that the new corpus could eventually provide. After an RFP, in July of 2017, CHS financed a design sprint orchestrated by a team from Intrepid (https://www.intrepid.io) headed by Christine Pizzo. They spent three

Unauthenticated
Download Date | 8/23/19 4:11 AM

intense days with the OGL team in Leipzig talking with the staff and connecting in the morning with CHS personnel stateside as well. The goal was to understand the conception of the whole OGL and to develop a design template for the functionality that an access system for the corpus might use. They produced a set of designs, and that fall, after another RFP, Eldarion (http://eldarion.com), and its CEO, James Tauber, were chosen by Gregory Crane to implement the design; funding came from Crane's budget, and the result was made public in March of 2018, namely, the Scaife Viewer (https://scaife.perseus.org). Named for Ross Scaife, an early evangelist for digital classics who was a dear friend to the Perseus team and CHS and whose life was tragically cut short in 2008, the Scaife Viewer is a working prototype for accessing the Greek and Latin texts now in the corpus, along with some Hebrew and Farsi texts. The Viewer currently deploys much (but not all) of the technology that the project teams have envisioned: multiple editions and aligned multiple translations of classical texts, with tools to help learners read the original language and to understand the texts, but also tools to help researchers search within the texts in the corpus in multiple and complex ways. New texts in both languages are being added to the repository at varying rhythms, and the Scaife Viewer is set up to incorporate new sources on a weekly basis. Its software will also soon undergo further development with funding from a grant by the Andrew Mellon Foundation directed by Sayeed Choudhury, Associate Dean for Data Management and Hodson Director of the Digital Research and Curation Center at the Sheridan Libraries of the Johns Hopkins University.

Another example of the potential of an open-access corpus is not yet functional, but there is again a working prototype that makes concrete what can and will be done. This project, funded by the CHS and under development by Archimedes Digital (https://archimedes.digital), is called New Alexandria, and its purpose is to provide a platform for the development of fully-featured, collaborative online commentaries on texts in classical languages around the world – not just the Ancient Greek and Latin texts in the OGL/FF1KG, but also the 41 other languages in the corpus being developed by the Classical Language Toolkit (https://github.com/cltk; the principals of CLTK are Kyle Johnston, Luke Hollis, and Patrick Burns). Current plans are to provide a series of curated commentaries by invitation only but also an open platform for uncurated commentaries by individuals or groups that wish to try to provide insight into a text in a classical language as the CLTK defines it. The working prototype for such an online commentary is *A Homer Commentary in Progress*, https://ahcip.chs.harvard.edu, a collaborative commentary on all the works in the Homeric corpus by an intergenerational team of researchers. This project, which is permanently "in progress", is intended to provide an evergreen database of comments by a large and

Unauthenticated
Download Date | 8/23/19 4:11 AM

evolving group of like-minded specialists. The comments they produce are searchable by canonical reference, by author, and also by semantic tags that the author of a comment can provide to each comment; the reader of comments always sees the snippet of text being commented upon and can opt to see its larger context in a scrolling panel, and there are multiple translations as well as multiple texts on instant offer for any text. Every canonical reference within a comment to a Homeric text is automatically linked to the Greek texts and translations, and every comment also has a unique and stable identifier that can be pasted into an online or printed text.

As a last example of what can happen when the ideals with which this presentation began are realized, we point to one further development: the last two projects, the Scaife Viewer and the New Alexandria commentaries platform, are interoperable and will in fact be linked, because both are implemented in compliance with the CTS protocols. Even now, a reader of Homer in the Scaife Viewer can already automatically access comments from *A Homer Commentary in Progress* for the passage that is currently on view; the right-side pane of the viewer simply needs to be expanded in its lower right-hand corner to expose scrolling comments. Further linkage, such as to Pleiades geospatial data on ancient sites (https://pleiades.stoa.org) and to the *Lexicon Iconographicum Mythologiae Classicae* (*LIMC*, headquarters in Basel) encyclopedia of ancient iconography, are in the pipeline for the New Alexandria project and the Scaife Viewer as well.

Bibliography

Berti, M. (ed.): "Digital Athenaeus". http://www.digitalathenaeus.org (last access 2019.01.31).
Berti, M. (ed.): "Digital Fragmenta Historicorum Graecorum (DFHG)". http://www.dfhg-project.org (last access 2019.01.31).
Clérice, T.: "Capitains". https://github.com/Capitains (last access 2019.01.31).
Crane, G.: "First 1000 Years of Greek". http://opengreekandlatin.github.io/First1KGreek/ (last access 2019.01.31).
Elliott, T.; Bodard, G.; Cayless, H. (2006–2017): "EpiDoc: Epigraphic Documents in TEI XML. Online material". https://sourceforge.net/projects/epidoc/ (last access 2019.01.31).
Frame, D.; Muellner, L.; Nagy, G. (eds.) (2017): "A Homer Commentary in Progress". https://ahcip.chs.harvard.edu (last access 2019.01.31).
Johnston, K.; Hollis, L.; Burns, P.: "Classical Language Toolkit". https://github.com/cltk (last access 2019.01.31).
Perseus Digital Library (2018): "Scaife Viewer". https://scaife.perseus.org (last access 2019.01.31).
Raymond, E. (1999): The Cathedral and the Bazaar. Sebastopol, CA: O'Reilly Media.

Unauthenticated
Download Date | 8/23/19 4:11 AM

Robertson, B.: “Lace: Polylingual OCR Editing”. http://heml.mta.ca/lace/index.html (last access 2019.01.31).
Smith, N.; Blackwell, C. (2013): “The CITE Architecture: Technology-Independent, Machine-Actionable Citation of Scholarly Resources”. http://cite-architecture.org (last access 2019.01.31).

Unauthenticated
Download Date | 8/23/19 4:11 AM

Unauthenticated
Download Date | 8/23/19 4:11 AM