Digital Classical Philology

Age of Access? Grundfragen der Informationsgesellschaft

Edited by
André Schüller-Zwierlein

Volume 10

Digital Classical Philology

Ancient Greek and Latin in the Digital Revolution

Edited by
Monica Berti

DE GRUYTER
SAUR

ISBN 978-3-11-059678-6
e-ISBN (PDF) 978-3-11-059957-2
e-ISBN (EPUB) 978-3-11-059699-1
ISSN 2195-0210

Library of Congress Control Number: 2019937558

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the Internet at http://dnb.dnb.de.

Typesetting: Integra Software Services Pvt. Ltd.
Printing and binding: CPI books GmbH, Leck

www.degruyter.com

Editor's Preface

Whenever we talk about information, *access* is one of the terms most frequently used. The concept has many facets and suffers from a lack of definition. Its many dimensions are being analysed in different disciplines, from different viewpoints and in different traditions of research; yet they are rarely perceived as parts of a whole, as relevant aspects of one phenomenon. The book series *Age of Access? Fundamental Questions of the Information Society* takes up the challenge and attempts to bring the relevant discourses, scholarly as well as practical, together in order to come to a more precise idea of the central role that the accessibility of information plays for human societies.

The ubiquitous talk of the "information society" and the "age of access" hints at this central role, but tends to implicitly suggest either that information *is* accessible everywhere and for everyone, or that it *should be*. Both suggestions need to be more closely analysed. The first volume of the series addresses the topic of information justice and thus the question of whether information *should be* accessible everywhere and for everyone. Further volumes analyse in detail the physical, economic, intellectual, linguistic, psychological, political, demographic and technical dimensions of the accessibility and inaccessibility of information – enabling readers to test the hypothesis that information *is* accessible everywhere and for everyone.

The series places special emphasis on the fact that access to information has a diachronic as well as a synchronic dimension – and that thus cultural heritage research and practices are highly relevant to the question of access to information. Its volumes analyse the potential and the consequences of new access technologies and practices, and investigate areas in which accessibility is merely simulated or where the inaccessibility of information has gone unnoticed. The series also tries to identify the limits of the quest for access. The resulting variety of topics and discourses is united in one common proposition: It is only when all dimensions of the accessibility of information have been analysed that we can rightfully speak of an information society.

André Schüller-Zwierlein

https://doi.org/10.1515/9783110599572-201

Preface

More than fifty years have passed since 1968, when Harvard University Press published the Concordance to Livy (*A Concordance to Livy* [Harvard 1968]), the first product of what we might now call Digital Classics. In the basement of the Harvard Science Center, David Packard had supervised the laborious transcription of the whole of Livy's *History of Rome* onto punch cards and written a computer program to generate a concordance with 500,000 entries, each with 20 words of context. Fourteen years later, when in 1982 I began work on the Harvard Classics Computing Project, technology had advanced. The available of Greek texts from the Thesaurus Linguae Graecae on magnetic tape was the impetus for my work – the department wanted to be able to search the authors in this early version of the TLG on a Unix system. There was also a need to computerize typesetting in order to contain the costs of print publication. Digital work at that time was very technical and aimed at enhancing traditional forms of concordance research and print publication.

When I first visited Xerox's Palo Alto Research Center in 1985, I also saw for first time a digital image – indeed, one that was projected onto a larger screen. As I came to understand what functions digital media would support, I began to realize that digital media would do far more than enhance traditional tasks. As a graduate student, I had shuttled back and forth between Widener, the main Harvard library, and the Fogg Art Museum library, a five or ten minute walk away. That much distance imposed a great deal of friction on scholarship that sought to integrate publications about both the material and the textual record. It was clear that we would be able to have publications that combined every medium and that could be delivered digitally. My own work on Perseus began that year with a Xerox grant of Lisp Machines (already passing into obsolescence and surely granted as a tax write-off).

A generation later, the papers in this publication show how far Digital Classics has come. When I began my own work on Perseus in the 1980s, much of Greek and Latin literature had been converted into machine readable texts – but the texts were available only under restrictive licenses. The opening section of the collection, *Open Data of Greek and Latin Sources*, describes the foundational work on creating openly licensed corpora of Greek and Latin that can support scholarship without restriction. Scholars must have data that they can freely analyze, modify and redistribute. Without such freedom, digital scholarship cannot even approach its potential. Muellner and Huskey talk about collaborative efforts to expand the amount of Greek source text available and to begin developing born-digital editions of Latin sources. Cayless then addresses the challenge of applying the methods of Linked Open Data to topics such as Greco-Roman culture.

https://doi.org/10.1515/9783110599572-202

Cataloging and Citing Greek and Latin Authors and Works illustrates not only how Classicists have built upon larger standards and data models such as the Functional Requirements for Bibliographic Records (FRBR, allowing us to represent different versions of a text) and the Text Encoding Initiative (TEI) Guidelines for XML encoding of source texts (representing the logical structure of sources) but also highlights some major contributions from Classics. Alison Babeu, Digital Librarian at Perseus, describes a new form of catalog for Greek and Latin works that exploits the FRBR data model to represent the many versions of our sources – including translations. Christopher Blackwell and Neel Smith built on FRBR to develop the Canonical Text Services (CTS) data model as part of the CITE Architecture. CTS provides an explicit framework within which we can address any substring in any version of a text, allowing us to create annotations that can be maintained for years and even for generations. This addresses – at least within the limited space of textual data – a problem that has plagued hypertext systems since the 1970s and that still afflicts the World Wide Web. Those who read these papers years from now will surely find that many of the URLs in the citations no longer function but all of the CTS citations should be usable – whether we remain with this data model or replace it with something more expressive. Computer Scientists Jochen Tiepmar and Gerhard Heyer show how they were able to develop a CTS server that could scale to more than a billion words, thus establishing the practical nature of the CTS protocol.

If there were a Nobel Prize for Classics, my nominations would go to Blackwell and Smith for CITE/CTS and to Bruce Robertson, whose paper on Optical Character Recognition opens the section on *Data Entry, Collection, and Analysis for Classical Philology*. Robertson has worked a decade, with funding and without, on the absolutely essential problem of converting images of print Greek into machine readable text. In this effort, he has mastered a wide range of techniques drawn from areas such as computer human interaction, statistical analysis, and machine learning. We can now acquire billions of words of Ancient Greek from printed sources and not just from multiple editions of individual works (allowing us not only to trace the development of our texts over time but also to identify quotations of Greek texts in articles and books, thus allowing us to see which passages are studied by different scholarly communities at different times). He has enabled fundamental new work on Greek. Meanwhile the papers by Tauber, Burns, and Coffee are on representing characters, on a pipeline for textual analysis of Classical languages and on a system that detects where one text alludes to – without extensively quoting – another text.

At its base, philology depends upon the editions which provide information about our source texts, including variant readings, a proposed reconstruction of the original, and reasoning behind decisions made in analyzing the text. The

section on *Critical Editing and Annotating Greek and Latin Sources* describes multiple aspects of this problem. Fischer addresses the challenge of representing the apparatus – the list of variants traditionally printed at the bottom of the page. Schubert and her collaborators show new ways of working with multiple versions of a text to produce an edition. Dué and Hackney present the Homeric Epics as a case where the reconstruction of a single original is not appropriate: the Homeric Epics appeared in multiple forms, each of which needs to be considered in its own right and thus a Multitext is needed. Berti concludes by showing progress made on the daunting task of representing a meta-edition: the case where works exist only as quotations in surviving works and an edition consists of an annotated hypertext pointing to – and modifying – multiple (sometimes hundreds) of editions.

We end with a glimpse into born-digital work. *Linguistic annotation and lexical databases* extends practices familiar from print culture so far that they become fundamentally new activities, with emergent properties that could not – and still cannot fully – be predicted from the print antecedents. Celano describes multiple dependency treebanks for Greek and Latin – databases that encode the morphological and syntactic function of every word in a text and that will allow us to rebuild our basic understanding of Greek, Latin, and other languages. Passarotti's paper on the Index Thomisticus Treebank also brings us into contact with Father Busa and the very beginning of Digital Humanities in the 1940s. With Boschetti we read about the application of WordNet and of semantic analysis to help us, after thousands of years of study, see systems of thought from new angles.

I began my work on (what is now called) Digital Classics in 1982 because I was then actively working with scholarship published more than a century before and because I knew that my field had a history that extended thousands of years in the past. Much has changed in the decades since, but the pace of change is only accelerating. The difference between Classics in 2019 and 2056 will surely be much greater than that between 1982 and 2019. Some of the long term transformative processes are visible in this collection.

One fundamental trend that cuts across the whole collection is the emergence of a new generation of philologists. When I began work, few of us had any technical capabilities and fewer still had any interest in developing them. What we see in this collection of essays is a collection of classical philologists who have developed their own skills and who are able to apply – and extend – advances in the wider world to the study of Greek and Latin. This addresses the existential question of sustainability of Greek and Latin in at least two ways.

First, I was very fortunate to have five years of research support – 1.000.000 EUR/year – from the Alexander von Humboldt Foundation as a Humboldt

Professor of Digital Humanities at Leipzig. I also have been able to benefit from support over many years for the Perseus Project from Tufts University. Both of those sources contributed to a number of these papers, both directly (by paying salaries) and indirectly (e.g., by paying for people to come work together). But what impresses me is how rich the network of Digital Classicists has become. We were able to help but the system is already robust and will sustain itself. We already have in the study of Greek and Latin a core community that will carry Digital Classics forward with or without funding, for love of the subject. In this, they bring life to the most basic and precious ideals of humanistic work.

Second, we can see a new philological education where our students can learn Greek and Latin even as they become computer, information or data scientists (or whatever label for computational sciences is fashionable). Our students will prepare themselves to take their place in the twenty-first century by advancing our understanding of antiquity. Our job as humanists is to make sure that we focus not only on the technologies but on the values that animate our study of the past.

Gregory R. Crane
(Perseus Project at Tufts University and Universität Leipzig)

Contents

Data Entry, Collection, and Analysis for Classical Philology

Critical Editing and Annotating Greek and Latin Sources

Linguistic Annotation and Lexical Databases for Greek and Latin

Introduction

Many recent international publications and initiatives show that *philology* is enjoying a "renaissance" within scholarship and teaching. The digital revolution of the last decades has been playing a significant role in revitalizing this traditional discipline and emphasizing its original scope, which is "making sense of texts and languages". This book describes the state of the art of digital philology with a focus on ancient Greek and Latin, the classical languages of Western culture. The invitation to publish the volume in the series *Age of Access? Grundfragen der Informationsgesellschaft* has offered the opportunity to present current trends in digital classical philology and discuss their future prospects.

The first goal of the book is to describe how Greek and Latin textual data is accessible today and how it should be linked, processed, and edited in order to produce and preserve meaningful information about classical antiquity. Contributors present and discuss many different topics: Open data of Greek and Latin sources, the role of libraries in building digital catalogs and developing machine-readable citation systems, the digitization of classical texts, computer-aided processing of classical languages, digital critical analysis and textual transmission of ancient works, and finally morpho-syntactic annotation and lexical resources of Greek and Latin data with a discussion that pertains to both philology and linguistics.

The selection of these topics has been guided by challenges and needs that concern the treatment of Greek and Latin textuality in the digital age. These challenges and needs include and go beyond the aim of traditional philology, which is the production of critical editions that reconstruct and represent the transmission of ancient sources. This is the reason why the book collects contributions about technical and practical aspects that relate not only to the digitization, representation, encoding and analysis of Greek and Latin textual data, but also to topics such as sustainability and funding that permit scholars to establish and maintain projects in this field. These aspects are now urgent and should be always addressed in order to make possible the preservation of the classical heritage. Many other topics could have been added to the discussion, but we hope that this book offers a synthesis to describe an emergent field for a new generation of scholars and students, explaining what is reachable and analyzable that was not before in terms of technology and accessibility. The book aims at bringing digital classical philology to an audience that is composed not only of Classicists, but also of researchers and students from many other fields in the humanities and computer science. Contributions in the volume are arranged in the following five sections:

https://doi.org/10.1515/9783110599572-001

Open data of Greek and Latin sources

This section presents cataloging and publishing activities of two leading open access corpora of Greek and Latin sources: the Free First Thousand Years of Greek of the Harvard's Center for Hellenic Studies that is now part of the Open Greek and Latin Project of the University of Leipzig, and the Digital Latin Library of the University of Oklahoma. The third paper describes principles and best practices for publishing and sustaining Linked Ancient World Data and its complexities.

Cataloging and citing Greek and Latin authors and works

The first paper of this section describes the history of the Perseus Catalog and its use of open metadata standards for bibliographic data. The other two papers describe digital library architectures developed for addressing citations of classical scholary editions in a digital environment. The first contribution describes CITE (Collections, Indices, Texts, and Extensions), which is a digital library architecture originally developed for the Homer Multitext Project for addressing identification, retrieval, manipulation, and integration of data by means of machine-actionable canonical citation. The second contribution presents an implementation of the Canonical Text Services (CTS) protocol developed at the University of Leipzig for citing and retrieving passages of texts in classical and other languages.

Data Entry, collection, and analysis for classical philology

The four papers of this section discuss practical issues about the creation and presentation of digital Greek and Latin text data. The first paper explains the technology behind recent improvements in optical character recognition and how it can be attuned to produce highly accurate texts of scholarly value, especially when dealing with difficult scripts like ancient Greek. The second paper presents an overview of character encoding systems for the input, interchange, processing and display of classical texts with particular reference to ancient Greek. The third paper introduces the Classical Language Toolkit that addresses the desideratum of a complete text analysis pipeline for Greek and Latin and other historical languages. The fourth paper addresses the phenomenon of viral intertextuality and demonstrates how current digital methods make its instances much easier to detect.

Critical editing and annotating Greek and Latin sources

The four papers of this section present different topics concerning critical editions and annotations of classical texts. The first paper describes current challenges and opportunities for the critical apparatus in a digital environment. The second paper gives a short description of the software tool e-Comparatio developed at the University of Leipzig and originally intended as a tool for the comparison of different text editions. The third paper describes the Homer Multitext Project and its principles of access within the long history of the Homeric epics in the centuries through the digital age. The fourth paper describes how the digital revolution is changing the way scholars access, analyze, and represent historical fragmentary texts, with a focus on traces of quotations and text reuses of ancient Greek and Latin sources.

Linguistic annotation and lexical databases for Greek and Latin

This section collects papers about morpho-syntactic annotation and lexical resources of Greek and Latin data. The first paper is an introduction to the dependency treebanks currently available for ancient Greek and Latin. The second paper is a description of the Index Thomisticus Treebank based on the corpus of the Index Thomisticus by father Roberto Busa, which is currently the largest Latin treebank available. The third paper investigates methods, resources, and tools for semantic analysis and thematic annotation of Greek and Latin with a particular focus on lexico-semantic resources (Latin WordNet and Ancient Greek WordNet) and the semantic and thematic annotation of classical texts (Memorata Poetis Project and Euporia).

I would like to thank all the authors of this book who have contributed to the discussion about the current state of digital classical philology. I also want to express my warmest thanks to the editors of the series *Age of Access?* and to the editorial team of De Gruyter for their invitation to publish the volume and for their assistance. I'm finally very grateful to Knowledge Unlatched (KU) for its support to publish this book as gold open access.

Monica Berti (Universität Leipzig)

Bibliography

Apollon, D.; Bélisle, C.; Régnier, P. (eds.) (2014): Digital Critical Editions. Urbana, Chicago, and Springfield: University of Illinois Press.

Bod, R. (2013): A New History of the Humanities. The Search for Principles and Patterns from Antiquity to the Present. Oxford: Oxford University Press.

Lennon, B. (2018): Passwords. Philology, Security, Authentication. Cambridge, MA: The Belknap Press of Harvard University Press.

McGann, J. (2014): A New Republic of Letters. Memory and Scholarship in the Age of Digital Reproduction. Cambridge, MA: Harvard University Press.

Pierazzo, E. (2015): Digital Scholarly Editing. Theories, Models and Methods. Farnham: Ashgate.

Pollock, S.; Elman, B.A.; Chang, K.K. (eds.) (2015): World Philology. Cambridge, MA: Harvard University Press.

Turner, J. (2014): Philology. The Forgotten Origins of the Modern Humanities. Princeton, NJ: Princeton University Press.

Open Data of Greek and Latin Sources

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

https://doi.org/10.1515/9783110599572-002

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

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

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".

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.

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).

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

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

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

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.

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).

Samuel J. Huskey

The Digital Latin Library: Cataloging and Publishing Critical Editions of Latin Texts

Abstract: The Digital Latin Library has a two-fold mission: 1) to publish and curate critical editions of Latin texts, of all types, from all eras; 2) to facilitate the finding and, where openly available and accessible online, the reading of all texts written in Latin. At first glance, it may appear that the two parts of the mission are actually two different missions, or even two different projects altogether. On the one hand, the DLL seeks to be a publisher of new critical editions, an endeavor that involves establishing guidelines, standards for peer review, workflows for production and distribution, and a variety of other tasks. On the other hand, the DLL seeks to catalog existing editions and to provide a tool for finding and reading them, an effort that involves the skills, techniques, and expertise of library and information science. But we speak of a "two-fold mission" because both parts serve the common goal of enriching and enhancing access to Latin texts, and they use the methods and practices of data science to accomplish that goal. This chapter will discuss how the DLL's cataloging and publishing activities complement each other in the effort to build a comprehensive Linked Open Data resource for scholarly editions of Latin texts.

Introduction

Although Latin texts have been available in electronic form for decades, there has never been an open, comprehensive digital resource for scholarly editions of Latin texts of all eras. In the era before the World Wide Web, collections such as the Packard Humanities Institute's (PHI) Latin Texts, Perseus, or Cetedoc made collections of texts available on CD-ROM, but those collections were limited by era (e.g., PHI and Perseus covered only Classical Latin texts) or subject (e.g., Cetedoc covered Christian Latin texts).[1] Matters improved with the wide

1 Cetedoc (sometimes known erroneously as CETADOC) was originally developed by the Centre Traditio Litterarum Occidentalium (CTLO). The full name of the database was "Cetedoc Library of Christian Latin Texts."

Samuel J. Huskey, University of Oklahoma

https://doi.org/10.1515/9783110599572-003

adoption of networked computing, but for many years collections of Latin texts were limited to a particular era (e.g., Perseus[2]), behind a paywall (e.g., Cetedoc, which became part of Brepolis' Library of Latin Texts[3]), or offline (e.g., PHI, which did not publish its texts online until 2011[4]). Sites such as The Latin Library and Corpus Scriptorum Latinorum are more expansive, but they have not kept pace with developments in technology, and since it is not always clear what the source of their texts is, they are of limited use for scholarly purposes.[5]

The Open Greek and Latin Project, however, promises to publish millions of words of Greek and Latin from all eras, along with robust resources for analyzing and reading the texts. As of this writing they have made significant progress toward that goal. Aside from the scale, what separates the Open Greek and Latin project from others is the focus on creating an open scholarly resource, with rich, citable metadata on the sources for the texts. But even the Open Greek and Latin project has established a boundary of 600 CE, which means that much of Medieval and Neo-Latin will be excluded.

But one thing that all of these resources have in common is that they omit the features that distinguish scholarly critical editions. That is, their texts lack an editorial preface that explains the history of the text and its sources, a bibliography of previous scholarship on the text, a critical apparatus with variant readings and other useful information, or any of the other items necessary for serious study. Whether the omission is because of copyright restrictions, the technical difficulty of presenting the information in a digital format, or the needs of the site's intended readership, it means that, with some exceptions, scholars must still consult printed critical editions for certain kinds of information.[6]

That is not to say that existing digital collections are useless for scholarship. After all, the goal of the Open Greek and Latin Project is not to publish critical editions, but to increase the amount of human-readable and machine-actionable Greek and Latin available online, and it promises to be an invaluable resource for a wide range of scholarship, from traditional literary and historical studies to

2 http://www.perseus.tufts.edu (last access 2019.01.31). It should be noted that the Perseus Digital Library expanded its Latin holdings to include authors from later eras, but on a limited basis. Its latest version (https://scaife.perseus.org, last access 2019.01.31) promises to be more expansive in terms of both texts in its library and tools available for studying them.

3 http://www.brepolis.net (last access 2019.01.31).

4 http://latin.packhum.org (last access 2019.01.31).

5 Corpus Scriptorum Latinorum: A Digital Library of Latin Literature: http://forumromanum.org/literature/index.html (last access 2019.01.31); The Latin Library: https://thelatinlibrary.com (last access 2019.01.31).

6 Kiss (2009–2013) is a notable exception. The catalog edited by Franzini et al. (2016–) contains details on other resources, but truly critical editions on the internet are still rare.

the latest developments in natural language processing. Rather, the point of this brief survey has been to define the space that the Digital Latin Library (DLL) means to fill: the collection and publication of critical editions of Latin texts from all eras, and the materials associated with them.[7]

To accomplish its objective, the DLL has two main initiatives: the DLL Catalog and the Library of Digital Latin Texts (LDLT). The purpose of the former is to collect, catalog, and provide an interface for finding Latin texts that have been digitized or published in digital form. The purpose of the latter is to publish new, born-digital critical editions of Latin texts from all eras. The rest of this paper will discuss these two wings of the DLL and their complementary goal of supporting new work in Latin textual criticism.

The DLL Catalog

As with other elements of the DLL, the "D" stands for "Digital" in a number of different ways. First and foremost, all of the items in the DLL Catalog are digital in some respect, either as digitized versions of printed materials or as digital texts.[8] Second, the catalog itself is digital, built with and operating entirely on open source technology. Most people will use the DLL Catalog via the web interface, but the datasets will be serialized in JSON-LD and available for downloading and reuse, in keeping with the best practices known as Linked Open Data.[9] Third, owing to the abundance of materials and the limited resources of the DLL, leveraging digital technology to ingest, process, and publish data is essential. Accordingly, building applications to facilitate those tasks is part of the scholarly endeavor of the DLL Catalog.[10]

Another way in which the DLL Catalog is digital is in its use of data modeling. Taking a cue from the Perseus Catalog[11] and using concepts from the

7 The Digital Latin Library project has been funded by generous grants from the Andrew W. Mellon Foundation's Scholarly Communications division from 2012 to 2018, and by ongoing institutional support from the University of Oklahoma.

8 See Sahle (2016) for an extended discussion of the difference between "digitized" and "digital." In short, a digital scan of a book may be referred to as "digitized," but not "digital," since it merely represents an object that exists in a non-digital format. To qualify as "digital," an edition must have distinct characteristics that would cease to function outside of the digital realm.

9 The repository is available at https://github.com/DigitalLatin (last access 2019.01.31).

10 See https://github.com/DigitalLatin/dllcat-automation (last access 2019.01.31).

11 http://catalog.perseus.org (last access 2019.01.31).

Functional Requirements for Bibliographic Records (FRBR)[12] model as a basis and data gathered from user studies, June Abbas and her team of researchers from the University of Oklahoma's School of Library and Information Studies designed an information behavior model to accommodate the different kinds of data to be stored in the catalog and the different ways in which users would interact with that data.[13] The following sections describe the resulting information architecture of the catalog and how it seeks to cater to the needs identified in Abbas' user studies.

Authority records

Authority records for authors and works provide the foundation for the DLL Catalog's information architecture. Each author of a Latin work has an authority record that identifies that author unambiguously and provides supporting attestations from a variety of sources to confirm the identity. In most cases, several forms of the author's name are recorded, especially the authorized name, which is usually identical to the authorized name in a major research library such as the U.S. Library of Congress, the Bibliothèque nationale de France, the Deutsche Nationalbibliothek, or others. Alpha-numerical or numerical identifiers such as the Virtual International Authority File ID or the Canonical Text Services identifier are also recorded, along with details about relevant dates and places. The purpose of an author authority record is to provide a single point of reference for individual authors. That way, searches for "Vergil", "Virgil", or "Vergilius" lead to the same information. Additionally, the cataloging process is more successful when automated matching algorithms have access to variant name forms.

Similarly, authority records for works support the vital functions of the catalog. Since dozens, if not hundreds, of works are known simply as *Carmina, Historiae*, or simply *fragmentum*, to take just three examples, it is important to have a means of disambiguating them. Accordingly, each work has its own authority record, with an authorized form of the title and any variant titles, along with information about its place in any collections, its author(s), and any abbreviations or other identifiers commonly in use.

Different content types for digitized editions, digitized manuscripts, and digital texts are the DLL Catalog's architectural frame. These content types

12 https://www.ifla.org/publications/functional-requirements-for-bibliographic-records (last access 2019.01.31).

13 See Abbas et al. (2015) for information about the methods and outcomes of the user studies.

store metadata related to specific instances of texts, each one connected to its creator and work through an entity reference so as to be discoverable in a variety of searches. Each record also contains a link to the external resource where the item can be found.

Contents

As of this writing, the DLL team has added authority records for over three thousand authors and nearly five thousand works spanning the time period from the third century BCE to the twentieth century CE. Many of those records were culled from information in standard reference works (e.g., *Clavis Patrum Latinorum*) and dictionaries (e.g., *Oxford Latin Dictionary, Thesaurus Linguae Latinae*), but records are also added nearly every time a new collection is added to the catalog, which is one of the reasons why the quest to catalog all Latin authors and works will be asymptotic.

Several collections are at different stages of being added to the catalog. With regard to digital texts, all of the items in following collections have been processed and cataloged: Perseus, PHI, Digital Library of Late-antique Latin Texts, and Biblioteca Italiana. Items on the related sites Musisque Deoque and Poeti d'Italia are in process and will be added by the end of 2019. These sites were selected because the sources of their texts are clearly identified and the texts themselves are openly available. Collections of texts behind a paywall (e.g., the Loeb Classical Library and Brepolis) are also in process, but since freedom of access is a priority, they will be added to the catalog at a later date.

As for digitized editions, efforts have focused on cataloging items in the public domain at resources such as the HathiTrust Digital Library, the Internet Archive, and Google Books.[14] Two categories in particular have received the most attention: early editions (*editiones principes*) and items in Engelmann's magisterial survey of Latin texts published between 1700 and 1878, *Bibliotheca Scriptorum Classicorum*. As of this writing, eighty-two early editions have been cataloged, including fifty-four editions of Latin texts published by Aldus Manutius. Over time, editions published by other early printers (e.g., Sweynheym and Pannartz, Jodocus Badius Ascensius), will be added to the collection. The survey of Engelmman's bibliography has so far yielded nearly three thousand

14 HathiTrust Digital Library: https://www.hathitrust.org (last access 2019.01.31); Internet Archive: https://archive.org (last access 2019.01.31); Google Books: https://books.google.com (last access 2019.01.31).