Classical text and the digital revolution

‘The future is already here — it’s just not very evenly distributed’
William Gibson, 1999

We are in the middle of a revolution and Classics is included. If not already present when the Digital Classics Association was founded in 2013, by 2015 digital methods had reached its very heart with the appointment of Gabriel Bodard as the first Reader in Digital Classics at the Institute of Classical Studies. Digital Classics, however, began a lot earlier with Roberto Busa, considered the father of Digital Humanities, paving the way with around 11 million Latin words on IBM punch cards as early as the 1940s (published as *S. Thomae aquinatis hymnorum ritualium: varia specimen concordantium* in 1951). Meanwhile, the Perseus Digital Library Project will celebrate its thirtieth birthday in 2017.¹ For classicists researching textual traditions, the future is already here, albeit not yet evenly distributed.

Despite such tremendous advances, most classicists are probably unaware of the increasing scope and rapidity with which data is now created. The digital revolution has seen a massive shift in how we store information; a shift that is even more drastic than the invention of movable letterpress printing in the fifteenth century. In 2012 alone, 2.8 zettabytes of data were created or recreated (a zettabyte is twenty-eight followed by twenty zeros).²

In comparison, Homer’s *Iliad* and *Odyssey* together are around two million bytes in XML. From a data volume perspective, this means that the world

---

¹ Busa 1980; Smith 2007.
² Gantz and Reinsel 2012.
could have produced 1.4 quadrillion XML editions of Homer in 2012; that is, almost 550 editions of Homer per day for every human being on the planet.

Although only a fraction of this massive volume of data is about Classics, it too crunches big numbers. JSTOR currently contains around 200,000 articles, reviews or books drawn from 73 publications in the field, while Google Books, HathiTrust, the Perseus Project and the Open Philology Project at Leipzig (OPP) have digitised a tremendous amount of books on Classics along with editions of ancient sources.\(^3\) In fact, the Perseus Catalogue alone identifies close to 4,000 works in Latin and Ancient Greek.\(^4\) Most editions produced in the last three centuries, for instance all of the Bipontina, and the preponderance of Teubneriana and Oxford Classical Texts, are also now out of copyright and are either already digitised or it is planned that they will be so in the near future.\(^5\) Projects like Europeana, Gallica or the Manuscripta Medievalia have also digitised manuscript catalogues and manuscript folia, many of which are often reusable for research under an open licence.\(^6\)

If we extend this data to include close spatial and temporal neighbours that used and commented on Greek and Roman material extensively (for example, Classical Arabic, Persian, Hebrew, Medieval Latin or Byzantine Greek), then classicists are now faced with several billion words of text; and that does not even include early modern and modern secondary material or non-textual data (a variant matrix of a single textual transmission, for instance, can contain several thousand variants in several hundred textual witnesses). With big data making the inclusion of ‘new’ material possible, now in fact seems like an ideal time to overcome some of the bias in the way Classics favours the study of some ancient languages and cultures over others. Given the

---

3 See for instance, Crane (2016b) and the Open Philology Project (2014).
4 Crane et al. 2014.
5 See for instance, the Open Philology Project (2014).
interconnectedness of the ancient world, it would be appropriate to actively include more languages and cultures; for instance, Classical Arabic, Classical Chinese and Classical Sanskrit. ⁷

No classicist working one century, year, decade, or even one day ago, had access to as much information as we have today. This is essentially a big data problem: how to focus on the ‘interesting bits’ within an ever-growing field. With so much data being produced, one has to process a lot more information in order to produce good Wissenschaft. In response to the challenge of navigating these vast volumes of data, convincing contemporary Classics scholars to use the best available tools and methods for research and education in Classics is my main goal. I should be clear that I am doing this neither directly for these scholars’ sake or my own, but rather for the sake of our field and its creation of knowledge: that is, Wissenschaft in general.

**Wissenschaft and Classics: How to find and use information**

Wissenschaft, a German term often used by classicists to underline the gravitas of their scholarly work, literally translates to ‘knowledge-creation’ and has been part of the Anglophone Classics world since the nineteenth century. More so than the English word ‘science’, Wissenschaft literally switches the focus from knowledge as an entity that is already there and needs to be understood (cf. Lat. scire), to the creation of a new entity based on the scientific method.

In order to do Wissenschaft well, one has to follow the Wissenschaftliche Methode by starting with a hypothesis and then producing research that is reproducible and falsifiable. The latter is of major importance because most published research findings are false or merely a reflection of the prevailing bias. ⁸ It is therefore only possible to evaluate and replicate research results if one can follow the references and thoughts of the author of a study.

---

⁷ For further reading on this point see Crane (2016a).
⁸ Ioannidis 2005, 700.
If one wants to achieve this with the aid of computational methods, citations have to be precise and machine-actionable. While a trained human mind can normalise fuzzy citations and some imprecision relatively well (very well when there is a page number switched around, less well when there are multiple competing but very similar citation styles for the same work or when crucial information about a printed edition is missing, such as the year of publication), when working with machines we have to make it very clear to what we are referring. Essentially it is a problem of information retrieval, similar to the big data problem mentioned above (that is, finding the ‘interesting bits’ amongst vast volumes of data). However, this problem brings with it a different challenge, relating to the finer, but nevertheless crucial issue of finding the exact information sought. With a robust reference system in place (the specifics of one such system will be discussed further below), one can apply computational quantitative methods in order to narrow down the focus to passages, articles and texts that need close reading.

This article will attempt to introduce some of the tools and data that have been created in the field of text-focused Classics in recent years. Seeing this article as a short introduction to Digital Classics, I will often refer to articles and works that are most likely not known to many researchers in our Classics community, but which will nevertheless deepen points that can only be touched upon here.

Starting from a working definition of Classics as the ‘research and teaching of the ancient world’, all research in Classics can therefore be reduced to a combination of three tasks: creating editions of ancient texts, analysing text and analysing non-text. This article draws on the author’s experience with text-based research to focus on the first two tasks, ultimately arguing that the digital revolution has changed the way we ought to conduct research and educate future researchers.
Task 1: Creating an Edition

Classics has not yet finished creating editions of Classical text. To be honest, it has barely started: more and more manuscript catalogues have been digitised in recent years and it becomes increasingly obvious how much information was unwittingly excluded when editing Classical texts in the nineteenth and twentieth centuries. For instance, Müller's fourth edition of Petronius, published in 1995, used twenty-four manuscripts, although he also mentioned that he excluded twelve manuscripts purposely because he deemed them not valuable for the reconstruction of the text.\(^9\) However, after consulting several manuscript catalogues and collections, I know of at least sixty manuscripts that contain Petronian material.\(^10\) Müller excluded at least twenty-four manuscripts (as many as he used to produce his text), not because he worked carelessly, but because he did not, and probably could not, know of their existence. Given that this example is unlikely to be a rare exception, checking and reworking previously created edited material is fundamentally important for the Wirenschaftlichkeit of our field.

Faced with the task of editing a Classical text, classicists either attempt to produce an edition that resembles, as closely as possible, a non-extant but arguably once-existing original; or they describe the textual transmission of a notional Classical work as precisely as possible and a number of reconstructible or perceivable states of this transmission, thereby regarding every individual state of the transmission as important as an ancient original. Whether one follows the school of New Philology or that of Lachmann, Housman, Maas and Pasquali, the research of textual transmission, that is modelling the mono-directional communication of text-data, graphics (for instance Euclid) and meta-data (for instance Horace’s biography) from the past to the present, and the editing of a Classical text, do not work without precise

---


\(^10\) See the manuscript catalogue at the Open Philology Project (2015).
citations and accurate descriptions of the individual states of transmission.\textsuperscript{11}

When teaching textual transmission, we often oversimplify the matter by showing illustrations representing a hierarchical stemma, that is, as a tree-like model of the textual transmission, with a single, often fictional, entity at the top, which we call the archetype and often address with $\Omega$. As a result, we shouldn’t be surprised when students cannot differentiate between the ‘original’ work and the virtual archetype. Instead, I would argue that we should make the following points very clear in our teaching (see Figure 1).

Firstly, an archetype represents a virtual entity that is the result and summary of all unknown transmission processes and changes to which the ‘original’ text was subject. The archetype usually predates the oldest known, sometimes still extant, manuscripts of that text. Secondly, as an original text can have multiple ‘originals’ (for instance, multi-text, growing texts, canonical texts), it can have more than one archetype (as well as possible transmission-dependent hyperarchetypi of each transmission line). Thirdly, often parts of the transmission are so unclear that the production of a fully hierarchical model would be based on too many assumptions and non-hierarchical sub-stemmata, for instance produced in a joining-neighbours method, have to help shed light on the textual transmission.\textsuperscript{12}

It is the textual scholar’s task to present a summary and interpretation of the raw information they find in manuscripts, editions, inter- and intra-textual references and all other witness material in their edition. This is called text-criticism.

In text-criticism, scholars are often divided into two camps: those opting for emendations and those opting for the manuscript tradition as the guide to

\textsuperscript{11} For New Philology see Nichols (1990) and for a clinical Lachmannian approach, see Maas (1927). For further differentiation between open and closed recension, see Pasquali (1934) and West (1973).

\textsuperscript{12} For instance Tara Andrews’ Stemmaweb (2015).
what extent we may omit supposed interpolations or change the text depends more on what we know about the transmission than on our personal opinion. The only demand that is applicable to every text is that we have to mark any change precisely and give as much information as possible (for instance, in the critical apparatus). Unfortunately, as editors of ancient texts were limited to the space provided by the paper form of an edition, they often decided to give only as much information as was necessary to present the reader with a satisfactory text. However, because we are dependent on the examinations of others when reading a Classical text, this information can be too little to provide a clear picture of the transmission, if there is one, and we might have a different opinion on how much information is appropriate.

All of this changes when we move to a digital format: the editor is now freed from the constraints of a paper edition. More importantly, textual data and the visualisation of the textual data are no longer inseparable. Using mark-up languages, for example XML or TeX, one can enrich any text with an

---

13 It is difficult to name examples without offending scholars of either camp, but an obvious example may be the two Horaces of Klingner (1959) and Shackleton Bailey (1985) or even Fraenkel’s influence on Müller’s first edition of Petronius, published in 1961.
abundance of additional information, but how much of this information one chooses to visualise is independent from the process of collecting and structuring the textual information and does not necessarily require an editorial decision. The same textual data can be visualised in multiple ways and an editor of a Classical edition can now let their readers decide just how much information is necessary for them to understand the text. This does not mean that the editor has to fear that they will overwhelm the reader with too much information, given that to some extent they still control the default initial view on their edition.

Furthermore, with the book format no longer providing an unavoidable page break and limited scope for mark up, we are now faced with the challenge of precisely dividing an edition into citable logical units of text. To this end, a handy reference system that enables scholars to structure the raw information they acquire is the CTS/CITE architecture, developed by Neel Smith and Chris Blackwell. One of the strengths of this architecture is that it differentiates between the notional canonical work, editions and versions of it (CTS), and entities that are not text (CITE), and gives both of them a unique reference (URN). Using CTS, a scholar can compare and align different versions of a work; in fact, they can even connect precise portions of the text with precise portions of another text or object.¹⁴ For instance, a reference to the word οὐλομέν in the second line of Homer’s Iliad in the version of the Venetus A manuscript published in a specific edition of that manuscript could be: urn:cts:greekLit:tlg0012.tlg001.msA.thosJeff:1.2@οὐλομέν, with tlg0012 referring to the textgroup ‘Homer’ and tlg001 referring to the work. This makes it possible to precisely refer to a chapter, paragraph, line, word or even character of any version of any canonical text (see also Figure 2). Due to this precision CTS is used by several leading Digital Classics projects (for example the Alpheios Project, the Homer Multitext Project, the Leipzig Open Philology Project and the Perseus Digital Library Project).

¹⁴ For the early beginnings of the CTS/CITE architecture see Smith (2009). For a more recent overview see Smith and Blackwell (2015).
Task 2: Analysing Text

In what follows I would like to introduce the reader to two modern computational methods of analysing Classical text. One method is a very close analysis of the morpho-syntactic structure of a sentence; this is called treebanking. The other is a distant-reading method based on statistical inference and an attempt to classify texts automatically according to their content; this method is called topic modelling.

A treebank is a morpho-syntactically parsed corpus of text.\textsuperscript{15} The process of generating these treebanks is called treebanking. Although some of the text can be parsed automatically using machine-learning and morphological parsers (for example, the Perseids Morphological Webservice API), generating the training set and resolving complex sentences depends on a very close reading and grammatical understanding of the text. As such, it is increasingly used for teaching morphologically complex languages like Ancient Greek.\textsuperscript{16} Building treebanks of Latin and Greek is not new: David Bamman and Gregory Crane began building treebanks for those languages back in 2006. Since then, those treebanks have grown steadily and the tools have become more sophisticated.\textsuperscript{17} For instance, treebanking has become a

\textsuperscript{15} Nivre 2008, 225–6.
\textsuperscript{16} Palladino 2015.
\textsuperscript{17} Bamman and Crane (2006). For an introduction to this area see the list of sample publications at http://sites.tufts.edu/perseids/publications/treebank-publications/.
central part of OPP’s eLearning prototype *AncientGeek* and Giuseppe GA Celano has combined the corpus linguistic approach with Smyth’s grammar of Classical Greek in an overhaul of Bamman and Crane’s guidelines.\(^{18}\)

Once treebanks of Classical Latin and Greek are in place, scholars can search precisely for grammatical constructions and compare the morpho-syntactic structure of one sentence with a group of others. When this is combined with a robust citation system like CTS, automatic processing of Classical text becomes even easier. For instance, sentences used in the exercises in *AncientGeek* are automatically shortened by the machine using a link to Smyth’s grammar topics and treebanking data, so learners can focus on a certain grammatical problem based on the text they are currently reading without being exposed to artificial Greek invented by a teacher.

In contrast to close work on individual sentences, topic modelling is done using statistical inference to process and classify large amounts of texts. Topic modelling is ‘a method for finding and tracing clusters of words (called “topics” in shorthand) in large bodies of texts’.\(^{19}\) A topic can be described as a recurring pattern of co-occurring words.\(^{20}\) Topic models are probabilistic models that are often based on the number of topics in the corpus being assumed and fixed. The simplest and probably one of the most frequently applied topic models is the latent Dirichlet allocation (LDA).\(^{21}\) LDA is a method of statistical inference deducing the properties of an assumed underlying Dirichlet distribution by analysing the words of a corpus. LDA is based on a simplification model of the creation of text in which the word order, for instance, does not matter. It is highly effective in tracing recurring clusters of co-occurring words that may reveal a lot about text-reuse, recurring topics or the transmission of ideas throughout time and genre.

\(^{18}\) For a description of the methods used in *AncientGeek* and Celano (2014) for the treebanking guidelines for Ancient Greek, see Moritz et al. (2014).
\(^{19}\) Posner 2012.
\(^{20}\) Brett 2012.
\(^{21}\) Blei 2012.
The success and results of LDA rely on a number of a priori-set variables, including the number of topics assumed in the corpus, the number of iterations of the modelling process, the decision for or against morpho-syntactic normalisation of the research corpus, and how stop-words are implemented in the process. Furthermore, its interpretation is often influenced by how the topics are graphically represented and how the words of each topic are displayed. It is an incredibly helpful method to use when surveying large amounts of texts and it can synchronically or diachronically track topics and inter- and intratexual references. A draft of a functioning code that generates topic-models for Greek, Latin, English and Arabic texts can be found on my github code repository: https://github.com/ThomasK81/TopicModellingR (also see Figure 3). A release of a topic-modelling toolbox for historical languages is scheduled for the DH2016 conference in Krakow, so that researchers, even those who do not have deep scripting knowledge, can be enabled not only to understand but also to modify the topic-modelling process, adopt it to their corpus- or language-specific needs and then use the results of this process for further qualitative research.

Figure 3: Visualisation of topic-models in Thucydides using R’s LDA and LDavis library.
Both treebanks and topic models derived from any given edition are simply different tokenisations of the same text. Computationally, they can be treated as their own version of a notional Classical work. With a robust citation system in place, we can easily connect those versions to one another, thereby enriching the data instantly and reducing the code-complexity of new digital tools. This leads to cleaner code that is easier to understand and sharable with other scholars. Ultimately, less time is spent coding and more time is spent analysing and enjoying Classical texts.

Closing remarks

This article attempted to give a brief glimpse into what Classics can do in the twenty-first century. I actually hope that it generated more questions than it gave answers. After all, nobody can foresee the exact nature of the changes to society and learning in this truly revolutionary time. One thing is clear though: there is immense institutional pressure for Classics departments to become more efficient in teaching languages, publishing articles and in handling increasing administrative loads. As classicists we do not want to be hurried, rather, we want to understand every minute aspect of every word in a sentence. The best seminars I have visited as a student were those in which the smallest part of a text was read with the greatest diligence. With this in mind, digital methods and tools and a robust citation scheme help to retrieve and identify the passages one wants and needs to read slowly. If we can combine slow-reading with the best digital methods for finding and processing texts, and if we can employ the best practices of open data and creative commons in order to enable the sharing of research outcomes (including the data I have used), then the future will be more evenly distributed.

\footnote{Smith in conversation with the author, 2015.}
Bibliography


Celano, GGA. 2015, 1 December. Guidelines for the Ancient Greek Dependency Treebank 2.0 (Pre-Release). https://sites.google.com/site/giuseppegacelano/treebanking/agguidelines


Köntges, T. 2015, 30 August. ‘Visualisation of Topics in Thucydides’ Peleponesian War’. http://thomask81.github.io/Greek_vis/#topic=4&lambda=0.6&term=


