Digital Studies / Le champ numérique

Gauthier, Laurent. 2025. "Digital Humanities and the Notion of Corpus in Ancient History." *Digital Studies/Le champ numérique* 15(1): 1–28. https://doi.org/10.16995/dscn.18169.

Open Library of Humanities

Digital Humanities and the Notion of Corpus in Ancient History

Laurent Gauthier, ESPRI/ArScAn, Université Paris Nanterre, France, and CAC IXXI, Institut rhônalpin des systèmes complexes, France, laurent.o.gauthier@gmail.com

Although ancient historians routinely create and exploit document corpora, and the notion of corpus is recognized as central in historiography, there has been little methodological focus on coming to a unified approach to the design and use of corpora. The massive expanse of digital information and processing capabilities over the past few years has also led to a diversity of approaches. After reviewing the history of the use of corpora in historiography, we examine how ancient historians have taken possession of digital practice, and how it has interacted with the notions and uses of textual corpora: there are many diverse and somewhat incompatible methodological perspectives on historical corpora. Next, we show how the digital textual corpus, as an input into historiography, should not exist anymore as an object and should be seen as a process or a pipeline. Then, its multiple and sometimes opposite perceptions can be unified, at the same time making history more scientific in the sense of Lucien Febvre's definition, in which history is the scientifically elaborated narration of humankind's activities.

Bien que les historiens anciens créent et exploitent régulièrement des corpus de documents, et que la notion de corpus soit reconnue comme centrale en historiographie, il y a eu peu de concentration méthodologique sur l'élaboration d'une approche unifiée de la conception et de l'utilisation des corpus. L'expansion massive de l'information numérique et des capacités de traitement au cours des dernières années a également conduit à une diversité d'approches. Après avoir passé en revue l'histoire de l'utilisation des corpus en historiographie, nous examinons comment les historiens anciens se sont appropriés la pratique numérique, et comment elle a interagi avec les notions et les usages des corpus textuels : il existe de nombreuses perspectives méthodologiques diverses et quelque peu incompatibles sur les corpus historiques. Ensuite, nous montrons comment le corpus textuel numérique, en tant qu'entrée dans l'historiographie, ne devrait plus exister en tant qu'objet et devrait être considéré comme un processus ou un pipeline. Ainsi, ses perceptions multiples et parfois opposées peuvent être unifiées, tout en rendant l'histoire plus scientifique au sens de la définition de Lucien Febvre, selon laquelle l'histoire est la narration scientifiquement élaborée des activités de l'humanité.

It would likely seem obvious to most historians that they know what a corpus is in historiography: they continuously manufacture new ones compiled from all kinds of sources. Still, the logic underlying the design of a corpus is not always explicitly mentioned as such in every research publication, and the term remains quite polysemous; referring to a corpus does not conjure up a single unequivocal definition. It may at times designate a large set of texts from which one will isolate excerpts of interest, or it may designate the excerpts, or it may designate some serialized information derived from such excerpts. Despite the fuzziness surrounding the concept, establishing a corpus in order to be able to rely on a specific set of documentary evidence is deemed central in the historiographical process, according to many fundamental texts (Prost 1996; Marrou [1954] 2016). This tension between necessity and a lack of precise determination is aggravated by the availability of an ever-increasing volume of digitized information, which by its nature demands sorting and categorization, especially in fields where the total amount of documentation was initially limited, such as ancient history, in contrast with contemporary history.

Methodological research on the use of digital tools for defining and exploiting corpora has not converged, if not to say that historians have rarely used the digital environment to its full potential. Here, we will argue that, once we account for digitization, the corpus in historiography should not exist anymore as an object but instead as an abstract process in the sense of information theory and computer science. Indeed, the methodological approaches to corpora applied by historians often remain implicit, making it difficult to fully account for the logic that governs their construction and use. While scholars acknowledge that the constitution of a corpus influences historical analysis, the extent of this influence is rarely examined in a systematic manner. The processes applied to organize and analyze a corpus all shape the results of historiographical inquiry, yet these operations can be left unarticulated in the context of digital humanities. The increasing reliance on structured datasets, search algorithms, and computational methods means that corpora are no longer static entities but dynamic configurations subject to iterative processing. The very act of engaging with a digital corpus involves transformation, whether through data structuring, text mining, or annotation, which makes it impossible to treat the corpus as a fixed object independent of its processing. Underwood and colleagues (Underwood et al. 2022) argue that this methodological transformation has reoriented humanistic inquiry, as large-scale computational analysis may be privileged over traditional close reading. In addition, Audin (Audin 2025) points out that digital infrastructures embed methodological biases that constrain scholarly interpretation, influencing not only what can be analyzed, but also how findings are framed. In this

context, formalizing the corpus as a process rather than an object allows for greater transparency, reproducibility, and methodological rigor in historical research. It also aligns historiography with broader epistemological developments in digital humanities, where the structure and logic of data organization may be as significant as the content itself.

We carry out our analysis in four steps. First, we examine the history of the use of a corpus in history, up to some recent questioning of historians' methods, and, concentrating on textual corpora, we contrast the notion of corpus in history with the notion in linguistics. After addressing the epistemological framework of hermeneutics and humanistic inquiry, we then focus on how corpora are used in history. Next, we investigate the reception of digital methods and their interaction with the notions and uses of corpora in historiography. Finally, we will see that these multiple and sometimes opposed views on the embedding of corpora in the digital context can be aligned if we consider the corpus a process, and not an object. We will see what changing the perspective translates into in terms of practice and propose a few recommendations.

1 A brief history of the corpus

The notion of a corpus is central to historiography, yet its precise definition remains elusive. Historians routinely assemble corpora for their research, but the conceptual and methodological foundations of corpus creation have often been left implicit. This section traces the historical development of the corpus in historiography, from its early association with document-based inquiry to its role in structuring historical analysis, and then examines how historians and others have recently questioned the use of this concept. We begin by proposing a definition for the term.

1.1 A working definition of the corpus

Current perspectives on historiography see a continuum between the archive, the document, the source, and the corpus, and are all connected with the notion of "documents produced by the actors of the history under study" (Offenstadt 2011, 68). Since historians generally have a clear understanding of what a corpus is, one could expect there would be an effort towards defining, clarifying, and systematizing the notion. That has not been the case: for example, the use of the term *corpus* among French medievalists is widely spread, but there has not been any reflection whatsoever about its meaning, nor any effort at conceptualization (Magnani 2017a). Before delving into the history of corpora in historiography, we therefore propose to rely on a simple definition of the term in the context of historiography to avoid any ambiguity: a *corpus* is a set of documents assembled with a specific purpose (a set may naturally

be a singleton, so that a corpus may only contain a single document).[1] In linguistics, the notion of corpus has been an operating concept for many decades, at the core of corpus linguistics. Approaches to corpora in the field are geared towards methods intended to make analyses more systematic (McEnery and Brezina 2022, 4–28). The linguistic corpus is a collection of texts in an electronic database that can be used for linguistic analysis and description (Kennedy 1998, 1–3). In fact, the corpus is seen as a system whose very internal consistency is the subject of study (McEnery and Brezina 2022, 74–78). The use of the notion of corpus is hence narrower in linguistics than in history, where any collection of documents may be a usable corpus, because the logic that defines its construction comes from the historian rather than from the documents themselves. For example, a series of ancient coins in a specifically localized find may constitute a valid historical corpus (Nicolet-Pierre 2002, 58–59), but it may not have any internal consistency beyond the fact that the coins were found together. The difference between mentioning a corpus and mentioning a more generic "set of documents" resides in the former's more direct association with historical methodology. Writing of a corpus conjures up the logic that prevails in determining which documents are included or not.

1.2 The corpus through history

The notion of document and the writing of history have been associated for a long time. Constituting a corpus to carry out historical analysis is part and parcel of historical work and has even been considered to sufficiently define the very fact of writing history. In 1934, in a violent critique of a monograph on royal accounting, which triggered an intense historiographical debate, Lucien Febvre insisted that analysis was more important than erudition. He still recognized that putting together a corpus defined the writing of history, "the most exacting attention to procuring usable documents, unfold them, file them, and sort them in a rational order" (Febvre 1934, 149).

How has history writing come to rely on constituting corpora as the centerpiece of its methodology, "la centralité de l'archive" as Jean Boutier coined it (Boutier 2014, 10–11)? As Jose Carlos Bermejo Barrera pointed out in his historiographical study, there could be no notion of a corpus or document in ancient historiography, because the very witnessing of events served as the foundation for history writing (Bermejo Barrera 2001, 193). Relying on one's direct witnessing of events naturally implies that

[1] Interestingly, recent work interrogating the notion of corpus in history considered its etymology but did not unequivocally define the term: see Treffort (Treffort 2014) and Magnani (Magnani 2017a). This points to a certain unease with the use of the simple definition, stemming from the observation by Magnani that, when asked to define a corpus, many historians offer a description of their own methodology to create a corpus.

history writing was mainly concerned with contemporary history, and texts were held as inferior witnesses (Boutier 2014, 12–14). Boutier also distinguishes the historians, who wrote about contemporary events, from antiquarians such as Varro, who relied on a close study of text and artifacts to examine prior customs and ways of life (Boutier 2014, 15), and both endeavours remained distinct until the end of Antiquity.

It is only when the history of the Church became a domain of inquiry that historians began to consider the existence and necessity of a text corpus, according to Bermejo Barrera. This body of text had a materiality, just like sacred relics did, and the direct testimony from the contemporaries of Christ could not carry as much importance as personal testimony could have in the ancient Greek tradition (Bermejo Barrera 2001, 194). Close reading of the ancient texts developed into hermeneutics, which eventually led to the foundation of a historical method. The first university seminar focusing on the examination of primary sources took place at the university of Göttingen in 1766 (Boutier 2014, 19). According to Gunther Pflug, this prevalence of the document at first prevented any form of deductive operation, and in the eighteenth-century thinker Pierre Bayle's perspective: "The scholar's goal consisted of surveying the factual data, penetrating the historical givens, without attempting to impose any order unless it were for mere purposes of clarity" (Pflug [1954] 1971, 5).

Voltaire and later Turgot pulled history away from straight facts towards scientifically inspired analysis, relying on reason and common sense, thereby making the notion of pure document-based facts less central. Still critical and essential, the corpus now functioned with the application of reason and inserted itself in the context of the question asked by the historian (Pflug [1954] 1971, 9–12, 20–21). As the writing of history became professionalized in the nineteenth century, the methodology of source critique converged towards current practice (Offenstadt 2011, 70). At that juncture, historical knowledge acquired "a new configuration thanks to the introduction of two notions: that of document [...]; and that of the scientific method" (Bermejo Barrera 2001, 198). This perspective effectively established the document and the aspiration to a scientific approach as two facets of the same coin. Indeed, in a Foucauldian approach, analysis that is specifically historical, as well as a more generalized form of analysis common in the social sciences at large, both stem from the same source: "the questioning of the *document*" (Foucault 1969, 13). Bermejo Barrera stresses the primordial place that the corpus holds in current historiography: "History builds its object starting from the constitution of its documentary corpora; it then develops different methods of reading and interpreting the texts, methods that are sometimes contradictory and that are not reducible to a common factor" (Bermejo Barrera 2001, 204). This view is largely connected with Foucault's perception of historiography's

position with respect to the document: it seeks not to interpret it, but to work it from the inside and elaborate it; hence the document should not be seen as inert material. Foucault defines the writing of history as the way a mass of documents is organized (Foucault 1969, 14).

1.3 Questioning the corpus in history

It hence follows that the creation of a coherent corpus is one of the salient issues in historical methodology: determination of the documents to include or not, determination of the level of detail of the analysis that is required, determination of an analytic method for the quantification of various aspects from the documents (Foucault 1969, 19). Recognizing the influence that the presence or absence of a particular document may have on the writing of history, Boutier raised an important issue: one may ask which document should be used for which inquiry, but one should also ask to what extent the historian's questions will drive the gathering of a corpus (Boutier 2014, 10). The use of a corpus in historiography, for Foucault, is strongly related to a necessary serial perspective: it is constituted following a particular and systematic methodology and lends itself to quantitative analyses (Foucault 1969, 19). Collecting similar or comparable elements, by construction, creates information that can be processed in a serialized form.

Some historians have recently sought to put their practices into question as they pertain to the notion of corpus. While there is no question that the document is central in historiography, the issue is raised to precisely establish what a corpus is, and how one should make one. Implicitly following Foucault, Cécile Treffort stresses that a corpus has meaning, an *anima* inspired by the historian, and in that sense would differ from a simple set of documents; it "emanates from, translates and illustrates the researcher's thought" (Treffort 2014; my translation). At the same time, the corpus is tasked with aiming at comprehensiveness or at least representativity for the question under study. It therefore appears that one could distinguish the most comprehensive corpus containing every available document, a universal corpus, from the one specifically formed as a subset of the larger corpus to address a specific question, an oriented corpus. This distinction is largely related with that between a corpus with a collective aim (made available online, for example) and a corpus with a personal aim (for the researcher), drawn by Treffort. Magnani pointed out that the use of the term by French medievalists was very often qualified with a possessive (for example, "mon corpus," "notre corpus," see Magnani 2017a), which stresses the personal and specific way in which the corpus is constituted, from the standpoint of the historian.

To combine perspectives on the corpus in linguistics, literature, and human sciences, Damon Mayaffre proposed a conception of the corpus that reflected the ongoing interaction between the final interpretative act of a corpus and the original act of its constitution (Mayaffre 2002), which in effect internalizes the tension that Boutier pointed out. First, one needs to realize that creating a corpus necessarily implies a serial view, as Foucault had already mentioned. Second, the corpus, which in this context must be an oriented corpus as a research object, is an arbitrary construction whose worth only comes from the questions it raises and the answers it provides (Mayaffre 2002, 3). (We could also express this idea by stressing that it is the corpus's orientation that carries meaning, more than the corpus itself: the corpus's purpose in its definition is its key attribute.) Therefore, the corpus in history must be seen as something flexible and evolving, whose limits are dependent on the historical question and on the researcher's own discretionary decisions. As a mechanical consequence, there cannot be a unique methodological approach to creating a documentary corpus for historical research.

Considering this historical perspective, we conclude by stressing the crucial importance of the dialogue between manufacturing a corpus and effectively exploiting it, and on the fact that this ongoing dialogue renders the making of a corpus heavily dependent on the question at hand. As the power of digital processing advanced, it is natural to wonder how, and to what extent, digitization in general has impacted the making and the handling of corpora.

2 Exploiting a corpus in history

The creation of a corpus is only the first step in historical research; how it is analyzed and interpreted is just as important, all the more as it conditions its creation. While historians have long engaged with corpora through traditional methods, the introduction of digital tools has reshaped the ways in which corpora are constructed and exploited. In this section, we first examine the hermeneutic challenges inherent in historical analysis and how digital methodologies influence the making of historiography. By considering how historians engage with corpora in digital environments, we assess the extent to which new technologies have altered, reinforced, or complicated long-standing practices of historical interpretation.

2.1 The broader question of hermeneutics in the humanities

In the humanities, hermeneutics have long dealt with the problem of how meaning is produced, interpreted, and transmitted through layers of historical context. Considering the evolution from philology to modern textual criticism and, more

recently, to digital humanities, it appears that interpretative practices have been shaped by methodological factors rather than epistemological breakthroughs, unlike in the hard sciences. Hence, the underlying question is not whether digital methodologies alter hermeneutic engagement with texts, but how they reconfigure the conditions under which interpretation takes place.

In his history of the humanities, Bod (Bod 2013) reduces the field to a general form of structured reasoning and argues that the search for underlying principles and structures has historically driven humanistic inquiry, much like in the sciences: "The transition from early modern to modern disciplines is usually seen as a conceptual break. [...] However, [...] contrary to the conventional division between natural sciences and humanities, we find a continuous interaction and even methodological similarities" (Bod 2013, 347). Hence, "the strongest interaction between the sciences and humanities is currently happening in the upcoming field of digital humanities" (Bod 2013, 347). Digital humanities, in this perspective, are a simple extension of the humanities because they rely on the same underlying scientific principles (Bod 2013, 91–92).

Seeing digital humanities as an extension of the humanities is not problematic here, but we take exception about the characterization of this extension. Indeed, one needs to go back to the roots of the humanities. Philology, which dominated the field from Antiquity to the nineteenth century, does not merely consist in the study of texts but in the broader investigation of language and textual transmission. Turner describes it as "the multifaceted study of texts, languages, and the phenomenon of language itself" and traces its development into modern humanities (Turner 2014, 3). He argues that disciplines such as literary studies, history, and linguistics all emerged from the core methods of philology. However, while philology seeks to reconstruct text and uncover its historical meaning, textual scholarship has increasingly recognized that meaning is not merely extracted but also produced through interpretative acts. Indeed, interpretation is an extension of "meaningful action," so that meaning is not simply retrieved from texts, but actively produced through engagement with information (Ricoeur 1971, 211–213).

The traditional dichotomy between lower criticism, focused on establishing reliable readings of texts, and higher criticism, engaged in interpretative analysis, becomes blurred in modern textual scholarship, particularly in digital humanities, where the construction of a digital corpus requires interpretative choices at every stage: "the essential critical components of selection, evaluation, emendation, and annotation of texts still need to be emphasized" (Greetham 2013, 16–17). The assumption that a corpus is a neutral repository of texts is countered by the fact that encoding,

metadata structuring, and selection criteria are themselves hermeneutic decisions. Blanke and Hedges (Blanke and Hedges 2013) also argue that these infrastructures encode epistemic constraints into digital environments, implicitly directing scholarly engagement through preconfigured analytical frameworks. This perspective aligns with the argument that digital textual methodologies should not be assessed merely in terms of their computational capabilities, but in how they mediate interpretative engagement with texts (Aledavood 2024, 14). Aledavood calls for a mixed-method approach that acknowledges both the benefits of systematic processing and the inherently interpretative nature of corpus creation. This suggests that digital humanities should be seen not as a rupture with traditional textual studies but as an adaptation of their underlying hermeneutic principles to digital environments. The notion of "scholarly primitives" describes research activities such as annotating, comparing, and referencing that remain consistent across different methodological paradigms (Unsworth 2000; Pacheco 2022). Digital methodologies transform the scale and efficiency of these activities, but they do not fundamentally change their epistemological status.

At stake in these debates is the broader question of how interpretation in the humanities is reshaped by changing methodological paradigms. If philology provided the original framework for textual interpretation, and textual scholarship redefined how texts were edited and analyzed, then digital humanities introduce new layers of interpretative mediation. Yet the fundamental hermeneutic challenge remains the same.

2.2 Impact of the digital on the chaîne opératoire of historiography

Today, putting together a corpus in historiography does not in general imply the physical gathering of original manuscripts, archives, or artifacts. It does usually imply, however, gathering paper copies of such material, or notes about them, and organizing them in some manner. In the physical instantiation of a reflective corpus, for example, one may find photocopies of journal articles or book chapters written on the subject of interest, organized in folders along some relevant taxonomy, and with handwritten remarks in their margins. One may also find a recent edition of a primary source filled with numerous and colourful Post-it notes, with some commentaries on them. (Interestingly, Kiewra 1989 provides empirical evidence that active engagement in text organization enhances comprehension and recall, as there are cognitive benefits in note-taking and other manual forms of information processing.) This practical way of making a corpus is only scalable to a point. As the size of the corpus reaches the hundreds of thousands of pages, one must add more and more

meta-information and structure, and, for example, maintain lists of lists of references and extracts. In this process, digitization intervenes at several levels: document access, document exploitation, document organization, and storage and organization of the historian's own work. First, the documents themselves may be identified, accessed, or read in electronic form, from which they still may be printed and become a physical instantiation of a corpus. The electronic format also permits the electronic use of the documents by allowing word searches or the automatic extraction of relevant text excerpts, for example, in the case of textual data (one may also, for example, filter a large corpus of digitized images based on their meta information). The electronic exploitation of documents extends in fact well beyond that of text, and includes all the domains that are related to history, in particular, paleography (see, for example, the range of meta-information that can be added to manuscripts with digital analysis in Andrews and Macé 2015), art history (see, for example, Näslund and Wasielewski 2021), as well as archaeology (the turn towards digital methods in archeology took place early; see, for example, Evans and Daly 2006). The gathering of the documents and their logical organization may be also done in electronic format, using the logic of computer file directories, in its simplest form. The notes, whether on primary or secondary sources, can all be entered and stored into an electronic form as well. Finally, the historian's production itself will typically be carried out in a word processor. When all these steps take place virtually rather than physically, we do not necessarily observe that more information gets produced about how their logic is articulated. This creation and use of corpora in a historical perspective does not contain in itself a systematic description of their construction.

For Philippe Rygiel, historians are indeed "hypertextual polygraphs, who dissimulate most of the inscriptions they produce" (Rygiel 2011, 32). He stresses the fact that most of the historian's work is not visible from the results or the analyses they publish. The practical instance of a corpus we described above would effectively be mostly hidden and could only be marginally inferred from the resulting history work. In Rygiel's view, the historian's annotations, essentially in textual form, constitute the core of their work, their production. In this perspective, the historical inquiry becomes the delineation of a corpus, augmented with these annotations (Rygiel 2011, 34). The production of annotations is organically linked with the material corpus off which it is based, and we can see how this would logically end up in a possessive attribution, as Magnani pointed out: the corpus's orientation is part and parcel of the historian's work. How does the digital impact this framework? It appears the introduction of digital tools has only had a limited impact on historians' practices: Rygiel observes that the historian's production, the annotations, are mostly made in a digital framework and

hence should ideally be made accessible to all, but they are generally not. One reason for this, he argues, is that if this production was systematically made public by providing all with the same raw material, it would end up raising the bar of expectations among all historians (Rygiel 2011, 38). Gibbs and Owens also lamented the fact that the way data is used by historians is rarely well documented and made accessible ("Despite some recent methodological experimentation with data, historians have not been nearly as innovative in terms of writing about how they use it," Gibbs and Owens 2013, 163). Focusing on many of the tools that are nowadays available to historians for textual analysis (such as Google keyword searches or newspapers electronic archives), Tim Hitchcock pointed out that "academic historians did not ask for these resources, and nor for the most part have they been directly responsible for their creation" (Hitchcock 2013, 10). The tools in question, as well as many of the generic tools we mentioned earlier about the practical creation of a corpus, are indeed not specific to the field of history. More recently, Siebold and Valleriani pointed out the fragmentation of historians' approaches: "A close look at digital databases, network analysis, and ML shows that a very diverse and highly specialized set of new research practices has evolved, many of which are still being further developed. However, it must also be noted that the current situation is marked by a certain heterogeneity as to how these new practices are carried out" (Siebold and Valleriani 2022, 174). Crymble provides an overview of the various digital tools that historians are generally trained on, whether through their institutions or through self-learning, and points out that the tools in question are extremely diverse (see Chapter 4, "Building the Invisible College," of Crymble 2021).

Rygiel and Hitchcock's perspectives seem to establish that the evolution towards digital frameworks, tools, and analyses has taken place despite, rather than thanks to, historians. Hitchcock's observations recoup those of Rygiel: he notes that the way historians carry out their work has not changed substantially from the 1980s. Some have argued that the very act of publishing could massively benefit from the new tools afforded by a purely digital framework: Nawrotzki and Dougherty illustrated this idea quite efficiently by using a collaborative and web-based publication process for their book on the subject (Nawrotzki and Dougherty 2013). They stress that, in their experience, the opening and sharing of resources did not increase competition between scholars, but rather led to more collaboration. However, their project did not include putting in common a corpus in electronic form, but rather their own production, that is, what Rygiel described as the historians' annotations.

It appears so far that history aligns rather well with the hermeneutic framework of the humanities even in its extension to the digital: the methods stay seemingly the same, and evolutions take place at the margin.

3 The digital epistemology of historiography

As we can see, in theorizing the use of digital tools by historians, the focus has tended to remain more on the process of writing history than on the objects off which it is written: the historian rather than the corpus. Nevertheless, we will observe that, in fact, these methodologies have changed the perspective of historians, but without a unified epistemological framework. We will examine certain trends in the way digital corpora are created or used, focusing on recent methodological research as well as on some examples of how digital methods were implemented by historians.

3.1 Making digital corpora for history

As Gibbs and Owens stated, "historical scholarship increasingly depends on our interaction with data" (Gibbs and Owens 2013, 159), and indeed such a statement would probably match most historians' intuition.

In her review of the use of the notion of a corpus by medievalists, Magnani took digital processing as a given, and in his definition of the notion of reflective corpus, albeit from a linguistics perspective, Mayaffre proposed that it be in practice structured as hypertext, whereby each text in what we called the oriented corpus would be linked to the its parent texts, in particular using standard XML encoding to account for these connections, implicitly stressing the necessity of an electronic representation of the corpus, in order to make its internal structure apparent (Mayaffre 2002, 8). Corpora and digitization, from the perspective of historians, are more and more linked together.

The large-scale digitization of historical material does not systematically require the formal modelling of complex links between the elements that compose the corpus; the effort sometimes mainly consists of ensuring the quality of the resulting electronic text or data, and in the storage of all relevant metadata pertaining to the original documents. Such electronic corpora are usually constructed to be "simple, multiple, open and free of access," as is the case in the example of a medieval charter corpus (Magnani 2017b, 64). Nevertheless, when the underlying information is more complex, or the orientation of a corpus requires more effort, historians approach the creation of a digital corpus specifically under the guise of an ontology (in the sense of information theory), that is, a formal representation of knowledge, typically in the form of a database, which forms a structured view of the data and of some of its relations. To build a corpus for her dissertation, Ansley Erickson created a database as a way of keeping track of her notes relating to particular sources or material (Erickson 2013). This led her to reflecting upon the role of categorization in storing data for a historiographical use. Attaching attributes to the data that was thus created (as opposed to, for example, a simple alphabetical organization) allowed for more flexible

thought processes pertaining to the matter at hand. Erickson's work produced data (her notes related to various documents), but the underlying data did not have an electronic representation, and, in this instance, the resulting database was effectively disconnected from the sources.

In numerous cases, the sources themselves already are in electronic form, thanks to a prior digitization, and additional data, whether textual, iconographic, or quantitative, resulting from analysis or from other sources, can be attached to it. An example combining electronic sources and historians' annotations is the *Homer Commentary in Progress*, where scholars can contribute detailed analysis attached to any word or sentence in the *Iliad* and the *Odyssey*, so that both the underlying text and the annotations can all be queried, searched, and compared (Crane et al. 2016). This type of relationship between historical documents, a universal or oriented corpus, and historical work associated with them, is indeed conceived of as a set of annotations. This approach is far from systematic among historians, but there are many instances of such practices, in particular in classics. For example, Mugelli and colleagues defined an annotation logic to isolate and categorize references to sacrifice in Greek tragedy (Mugelli et al. 2017). The underlying corpus, in this case, is the Greek text in electronic form, and the annotations inserted within the text contain all the elements contributed by the historians: the specific location in the text where a sacrifice is identified, the categorization of the sacrifice, or the disambiguation of the discourse about the ritual from the ritual itself, for instance. By processing the annotated text, one can generate a database of ritual events. Barker and colleagues (Barker et al. 2013) used a very similar framework to study spatial references in Herodotus. They relied on the English translation of the Greek text and added annotations tracking the geographical information relevant to them; a copy of these annotations with the surrounding text were then stored in a database.

In a recent methodological paper, Hoekstra and Koolen generalized these notions of organized data for historical research using *data scopes*, which they defined as "the process through which different views on research data are created that are relevant to a specific research question" (Hoekstra and Koolen 2019, 80). In their perspective, a corpus for historical research should be created with a general data structure in mind, where different related sources of information can be easily paired to better gain new insights. In effect, this corresponds to transforming a general corpus (from the sources that may be available) into an oriented corpus. In their analysis of the interpretation of data by historians, Gibbs and Owens did not focus on a particular way in which data should be represented but concentrated on the general idea of sharing data processing methodology, which can be understood as a more general perspective than data

scoping. Siebold and Valleriani (Siebold and Valleriani 2022) also called for a degree of unification of formats to promote data transparency and sustainability, which stresses that the current situation may not be transparent nor sustainable.

In historiography, the creation of a digital corpus hence appears to be a complex exercise for which, to the dismay of many of the researchers whose work we mentioned above, there is no clear and unique epistemological framework or methodology. Further, the amount of detail provided to readers of historical research, when such work is carried out, varies greatly.

3.2 The digital processing of a corpus

When a digitized corpus is exploited in a way that leverages its electronic form, we may wonder if there is a generally agreed-upon method, a systematic manner to carry out this process. If it appears that it is desirable to make the processing of a corpus automatic, to what extent is this automation done in a unified manner? In the field of classics, it is possible that the relative paucity of textual sources, by making the processing of the entire corpus a reachable objective, encouraged digitization efforts, which led to the widespread use of electronic text search tools. As Barker and Terras (Barker and Terras 2016) pointed out, it has become quite common in ancient history to rely on detailed text searches, and there are a variety of dedicated tools for this purpose. Relying on electronic sources such as the *Thesaurus Linguae Graecae* (*TLG*) or *Perseus* (Pantelia 2020; Crane 2012), for example, is standard practice in Greek history. In these cases, while historians effectively use the algorithms that power these search tools, they do not encode the series of steps they follow in a formal way. The way in which the electronic nature of the underlying corpus is leveraged by the historian through their use of textual research in these examples is not made explicit: the orientation of the corpus, in effect, cannot be exactly reproduced in a systematic fashion.

Do more advanced, or more technical, uses of electronic corpora translate into a fuller documentation of their processes? The use of advanced tools, such as machine learning and text mining, and the tools of big data in general, have made their way into the hands of historians (Graham, Milligan, and Weingart 2015). Some use cases of text corpora effectively involve advanced computational methods, programmed to operate one after the other. Such quantitative analysis of literary texts is not recent, and it has reached a certain maturity (Hoover 2013). The systematic exploitation of textual data has in fact become widespread and standard enough that there are programming manuals focused on this type of exercise, covering all the standard operations one might need. Some domains require specific technical expertise, such as the analysis of the networks stemming from literary texts. These analyses resort to various technical

domains that are not part of traditional linguistics (Kenna, MacCarron, and MacCarron 2017). The analysis of intertextuality through the methods of computational biology is another example (see, for example, Barbrook et al. 1998; Howe et al. 2001; Chaudhuri and Dexter 2017). Still, the tools now available allow their users to program all forms of lexical and semantic analyses within a unified framework, in a process conceived of as a *pipeline*, comparable to the data analytics suites used in the hard sciences. Jockers and Thalken (Jockers and Thalken 2020) cover the practical approaches to this technique. In the case of classics, P. J. Burns (Burns 2019) describes a series of useful operations on classical texts following the same logic. A more general treatment of pipelines from the standpoint of data analytics, can be found in Wickham, Çetinkaya-Rundel, and Grolemund (Wickham, Çetinkaya-Rundel, and Grolemund 2017), pages 261 to 268 in particular. In contemporary French history, Magali Guaresi resorted for instance to logometrics and sentiment analysis, and applied factor analysis on a large corpus of speeches by members of the French Parliament in order to characterize the evolution of these discourses (Guaresi 2019). Classicists and computer science specialists have also collaborated on an empirical study of the quality of automated sentiment analysis in Greek tragedy and shown that automatic processing yielded good results compared with humans (Yeruva et al. 2020). In these examples, one cannot in fact speak of methodological unity, because, while the analyses in question were discussed in these research articles, the full details of the actual tools and algorithms that were used were not made available. If the analytic pipeline is not made explicit, then the logic that prevailed in the assembly of the corpus cannot be made explicit.

It appears that, while various tools and technical approaches have converged towards a form of conceptual pipeline (if not algorithmic), thanks to general technical progress, there is no centralized perspective about how textual corpora may be processed, and there rarely is detailed information on the actual steps undertaken to create the oriented corpus and carry out the analysis, in relation with each other. Then, although the notion of a corpus in historiographical work, stemming from a long history, as a collection of documents, is effectively manipulated by historians, the conjunction of that notion with the massive expanse of digital methods has not resulted in a clear and unified epistemological framework, which would leverage the electronic format in order to make the construction of a corpus and its processing clearly identifiable.

4 The corpus as a process

With so many different approaches to the creation of a corpus, and such a varying degree of digitization at every level, how can we propose a unifying framework? To

address the issue, we first need to draw a distinction between data and operations, after having examined the importance of reproducibility and transparency. Relying on this distinction, we will argue that corpora need to be conceived of as operations, not as data. Then, we will discuss an example in ancient Greek history.

4.1 Corpora and algorithms

It is fundamentally beneficial for the design of a corpus and the processing of its information to be automated and leverage its electronic nature when possible. One can indeed make an argument linking the reproducibility of research, gained by digital processing, to its scientificity. For McGillivray and colleagues, collecting and processing historical material with computational methods "would be a science if we could learn to automate it" (McGillivray, Wilson, and Blanke 2019, 53). In their view, historians should clearly delineate between what they define as "evidence" and what they define as "claims," so that in a positivist perspective one may separate evidence-based findings from other statements, thanks to the systematic analysis of evidence. Example-based analysis would be, in their view, kept separate from quantitative evidence (see the chart in McGillivray, Wilson, and Blanke 2019, 55). Distinguishing evidence and claim in historiography is no trivial affair, as Febvre insisted (Febvre [1952] 1992, 115).

Without trying to reduce history to a simplistic dichotomy between "claims" and "evidence," we can still recognize that there is something to be gained from being able to precisely understand the documents and analysis supporting any statement, when they can be made explicit. Making processes automatic, hence, may not, *per se*, serve a fundamental purpose, but it would have the advantage of making the creation of a corpus and some of its processing reproducible, and open it to critique. Reproducibility makes any statement falsifiable, which is a fundamental feature of any scientific statement according to the Popperian logical framework (see "Falsibiability," Popper [1934] 1992, 78 sq.). We are not arguing for scientificity for the sake of it, but rather as a form of transparency. This would also recoup with the objective laid out by Gibbs and Owens, by inherently providing detailed documentation on the use and processing of a corpus. Hence, keeping in mind the ability of others to reproduce parts of an analysis could help make the method itself more open to critique, without reducing or limiting historians' work. Should the humanities, because they can become digital, strive for the kind of reproducibility generally associated with the natural sciences? Digitization does not negate the hermeneutic challenges attached to textual analysis, so this notion of scientificity must effectively be adapted. Instead, the emphasis should be on transparency and interpretative flexibility, which ensures that the process of digital corpus creation and manipulation is open to critique.

Data, as the etymology tells us, is what is given, and cannot be worked out otherwise. Operations are applied to data to transform it into something more usable or practical. Operations orientate the data. Hence, data is information that cannot be derived from other information using logical operations and requires a human input. In computer science, this distinction between data and operations is better known as that between data and algorithm. Since each needs the other to exist, they are naturally intimately related, but are fundamentally and conceptually different (Wirth 1976). A database, or an ontology, is not simply data, because it combines data and algorithms that describe the way the data may be exploited. Any database can be represented as a combination of raw structured data (typically as tables of numerical or categorical data, because any multi-dimensional dataset can naturally be represented as a two-dimensional array) and algorithms that define potential relationships between the columns in these tables. Seen as a whole, the database may obfuscate the interactions between data and operations, and makes it seem as though it only contained the resulting dataset, organized as its designer intended.

All the examples of digital corpora we have discussed so far effectively are combinations of data and operations: they never were the straight input as provided by a person. These corpora were nevertheless considered as *things*, as data, focusing on the result of a series of operations rather than on the operations themselves. The precise descriptions of these results, without explaining the series of operations, were therefore incomplete and superficial. If we realize that a corpus is not data, but rather a set of operations, it may solve many of the issues we have noted so far. In addition, an operation could at the limit effectively contain any dataset: an algorithm may simply possess the list of values that constitute the data. In this sense, a process is superior to data: it is denser in informational content. We have mentioned the notion of a *pipeline* earlier, in the context of modern data analytics, or in computational linguistics. In such a pipeline, the raw input at each stage is transformed in place and serves as the input for the next stage. From this perspective, an oriented corpus should be seen as a pipeline, that is a series of operations, not as a dataset. The definition of this corpus is the computer code that builds it, whatever the language in which it may be expressed. This code can be analyzed, and it can be run in whole or in part by anyone. Purely seen as a dataset, a corpus cannot be properly analyzed from an external perspective, but once its construction is made entirely explicit, then this process can be fully subject to critique. In practice, historians often follow such a pipeline: for example, obtain some data (in a spreadsheet), transform it, save the clean version, compute some aggregates. However, only the result survives, and the details of all the steps are lost to everyone else. If this entire process is coded as a pipeline or as an algorithm more generally, every single assumption, explicit or not, becomes visible.

Corpus construction often involves nested layers of selection rather than a single sub-setting operation. Consider, for example, a study of certain interjections in nineteenth-century dialogues (Gauthier 2025): one moves from the complete corpus of digitized texts to a subset of novels and plays, then to direct speech, then to interrogative sentences, and finally to those containing interjections. Each stage structures the corpus in a way that preconditions analysis, making the creation of the corpus itself a dynamic and relational process rather than a static object. One might argue that corpus processing pertains more to analysis than to its construction, vindicating the notion of a corpus as an object. However, corpus formation itself is a structured, iterative process that encodes methodological choices at every stage. Selection, categorization, and filtering are not neutral operations but interpretative acts that define analytical possibilities, even more so in a digital environment.

Hence, once defined as computer code, or as a pipeline, the corpus contains and makes explicit all the decisions, small and large, made by the historian in cleaning, filtering, completing, or arranging the raw information, in the most concise manner possible. By construction, this perspective on the corpus as *anima*, that is, the process which orientates it, also directly expresses the dialogue between the research question and the setup of the corpus. Modern data analysis and text edition tools are converging, so that there is not such a strong distinction between the two anymore: the text a researcher produces (the historian's annotation work in Rygiel's words) and the computer code that gathers and processes data can now exist in the same document. Considering the data organization, processing, analysis, and write-up of a historical corpus as a continuum in a seamless process naturally leads to reproducible research. The RMarkdown language, for example, combines the data and statistical modelling infrastructure of the R language with the edition capabilities of the Markdown syntax and LaTeX system; it has been suggested as a good framework for reproducible research (Calero Valdez 2020).

We can relate this view of the corpus back to the *Annales*' co-founder's definition of history: for Lucien Febvre, history may be viewed not as a science, but as the scientifically elaborated narrative of the activities and creations of humankind (Febvre [1952] 1992, 19). Febvre's perspective is in line with that of Raymond Aron, for whom a philosophy of history cannot be simply positivist.[2] We argue that the notion of scientificity here, at the core, is about transparency and openness to critique. The idea

[2] See Aron: "notre livre conduit à une philosophie historique qui s'oppose au rationalisme scientiste en même temps qu'au positivisme" (Aron [1938] 1991, 13). This is also consistent with Marrou's perspective: "ni objectivisme pur, ni subjectivisme radical" (Marrou [1954] 2016, 221).

of carrying out historical research not as a science, but with a scientific aspiration, is clearly furthered by making the historiographical process more replicable, at least for the part that concerns the setup of a corpus.

4.2 An example in ancient history

To illustrate this logic, let us consider an example with ancient Greek theater, where we would want to study the occurrence of decisions. This would require the creation of a dataset of these decisions, as a subset of the entire set of theater texts, so that they can be closely read and examined.[3] The traditional way of tackling this project could benefit from the fact that these texts are available in electronic form: one could easily copy and paste all the Greek text of interest, maybe accompanied with translations, into a word processing software. For each excerpt, one could give a categorization for the decision at hand (whether it is an acceptance or refusal, for example). If instead we followed a more digital-based logic, comparable to the one described by Mugelli and colleagues (Mugelli et al. 2017), we would first create a copy of the Greek text (from *Perseus*, for example) in some standard TEI format. Then, inside this text, we would add markers that designate decisions, with some specific coding logic to distinguish between different sorts of decisions. We may also correct the text, if there are issues in how it has been established. Processing the annotated text could then result into a database, containing all the text excerpts with their categorization, amounting to a much more structured and easily exploitable corpus than a list of quotes in a word processor. Researchers would presumably query the database in question, and obtain, for example, tables showing which kind of character expresses decisions the most often; or they may query all decisions expressed by male characters who are slaves. The resulting tables could be pasted into the researcher's text. This later approach, nevertheless, by creating a new object, severs the link between the original textual source and the resulting corpus.

A process-driven, rather than object-driven, perspective on this corpus would consider the text source as an input that should not be changed in place or copied. If corrections are needed on the text, they are made explicit in the pipeline, as a set of overrides. The historian's work in defining the corpus would be embedded in a separate dataset, simply containing the decisions' location identification (in the form of line

[3] We are not addressing here the complex issue of defining decision in a historical context, nor whether one would consider decisions in theater from the perspective of the author's construction of the play and management of dramatic tension, or from the perspective of their value as realia, as snapshots of actual decisions. We will treat decisions as a generic example of moments of interest in a textual source.

and word number, for example) and their categorization. From the text input and the categorization data, by merging the two, one can automatically generate the equivalent of the annotated text if needed. Any textual analysis on the text of the oriented corpus, the decisions themselves, can be easily compared with the same analysis applied to the rest of the text, excluding decisions. It may be useful to define categories as a function of the characters' social status, gender, age, or any set of characteristics that may be relevant when examining decisions, and one may also distinguish sections such as the prologue, or sung parts in the text, for example. If one has indeed carried out this categorization, then all these taxonomies directly and automatically percolate through to the distinction of decision related in opposition to non-decision related text. If a researcher wanted to categorize certain parts of the text differently, leaving all else the same, this would be possible, and allow them to directly obtain the corresponding results by running the pipeline from beginning to end. In addition, if the underlying electronic Greek text's edition is improved and some words are corrected, then the decision corpus immediately benefits from these improvements.

Figure 1 shows, as an example, the proportion of words pronounced and the proportion of wills or decisions expressed by different types of characters in a selection of ancient Greek tragedies. This can be commented and studied in many ways, but this is not our purpose here. This data, as is, cannot be further explored. A full algorithmic traceability of how this data is constructed is necessary. **Figure 2** shows (for a handful of plays) one of the inputs, which can be independently examined and questioned: the manner in which each character is categorized. **Figure 3** shows another one of the inputs, that is, the categorization of expressions of will, decisions, or acceptations/refusals attached to specific verses and word numbers in the verses, here for *Philoctetes*. Odysseus wants Neoptolemos to do various things, and Neoptolemos agrees or reformulates some of them. The precise references to verses have to be mapped to an electronic reference for the text itself, in the example here, Perseus's text shown in **Figure 4**. Now, **Figure 1** does not depend on a very close analysis of the text, so one could argue that the data from **Figures 2** and **3** is almost enough by itself. **Figure 5** shows another example of analysis that looks much more closely into the way decisions are expressed and relies on the full details of the text. in this case, one must closely map each decision to the words that are related to it. Every input in the process described here and every step in the algorithms that put together the data that can then be analyzed must be open to scrutiny. Just saying that the underlying data is the electronic text in Perseus is clearly far from sufficient. In addition, a simple "data dump" of the results would not allow any inquiry into these steps.

Category Type	Category	Nb Characters	Nb of Words	Pct of Words	Nb Wills	Freq of Wills
Gender	Female	57	51,397	32.6	155	3.0
	Male	123	106,127	67.4	281	2.6
Age	Child	3	225	0.1	0	0.0
	Young	38	34,459	21.9	112	3.3
	Mid	102	73,321	46.5	203	2.8
	Old	37	49,519	31.4	121	2.4
Xenos	No	149	127,978	81.2	307	2.4
	Yes	31	29,546	18.8	129	4.4
Royal (Elite)	No	110	82,971	52.7	151	1.8
	Yes	70	74,553	47.3	285	3.8
Slave	No	158	143,910	91.4	399	2.8
	Yes	22	13,614	8.6	37	2.7
Divine	No	161	143,921	91.4	383	2.7
	Yes	19	13,603	8.6	53	3.9

Figure 1: Relationship between character categorization, scenic presence, and expressions of will across a large sample of tragedies.

Author	LatinizedTitle	LatinizedSpeaker	PlayType	Chorus	Abstract	Generi	Religious	Slave	Xenos	Divine	Royal	Warrior	Genre	Age	Male	Aged
Sophocles	Oedipus at Colonus	Angelos	Tragedy	No	No	Yes	No	No	No	No	No	No	Male	Mid	Yes	Yes
Sophocles	Oedipus at Colonus	Antigone	Tragedy	No	No	No	No	No	Yes	No	Yes	No	Female	Young	No	No
Sophocles	Oedipus at Colonus	Choros	Tragedy	Yes	No	Yes	No	No	No	No	No	No	Male	Old	Yes	Yes
Sophocles	Oedipus at Colonus	Ismene	Tragedy	No	No	No	No	No	Yes	No	Yes	No	Female	Young	No	No
Sophocles	Oedipus at Colonus	Kreon	Tragedy	No	No	No	No	No	Yes	No	Yes	No	Male	Old	Yes	Yes
Sophocles	Oedipus at Colonus	Oidipous	Tragedy	No	No	No	No	No	Yes	No	Yes	No	Male	Old	Yes	Yes
Sophocles	Oedipus at Colonus	Polyneikes	Tragedy	No	No	No	No	No	Yes	No	Yes	Yes	Male	Young	Yes	No
Sophocles	Oedipus at Colonus	Theseus	Tragedy	No	No	No	No	No	No	No	Yes	Yes	Male	Mid	Yes	Yes
Sophocles	Oedipus at Colonus	Xenos	Tragedy	No	No	Yes	No	No	No	No	No	No	Male	Mid	Yes	Yes
Sophocles	Oedipus Tyrannus	Angelos	Tragedy	No	No	Yes	No	No	Yes	No	No	No	Male	Old	Yes	Yes
Sophocles	Oedipus Tyrannus	Choros	Tragedy	Yes	No	Yes	No	No	No	No	No	No	Male	Old	Yes	Yes
Sophocles	Oedipus Tyrannus	Exangelos	Tragedy	No	No	Yes	No	No	No	No	No	No	Male	Mid	Yes	Yes
Sophocles	Oedipus Tyrannus	Hiereus	Tragedy	No	No	Yes	Yes	No	No	No	No	No	Male	Old	Yes	Yes
Sophocles	Oedipus Tyrannus	Iokaste	Tragedy	No	No	No	No	No	No	No	Yes	No	Female	Mid	No	Yes
Sophocles	Oedipus Tyrannus	Kreon	Tragedy	No	No	No	No	No	No	No	Yes	No	Male	Mid	Yes	Yes
Sophocles	Oedipus Tyrannus	Oidipous	Tragedy	No	No	No	No	No	No	No	Yes	Yes	Male	Mid	Yes	Yes
Sophocles	Oedipus Tyrannus	Teiresias	Tragedy	No	No	No	Yes	No	No	No	No	No	Male	Old	Yes	Yes
Sophocles	Oedipus Tyrannus	Therapon	Tragedy	No	No	Yes	No	Yes	No	No	No	No	Male	Old	Yes	Yes
Sophocles	Philoctetes	Choros	Tragedy	Yes	No	Yes	No	No	No	No	No	No	Male	Mid	Yes	Yes
Sophocles	Philoctetes	Emporos	Tragedy	No	No	Yes	No	No	No	No	No	No	Male	Mid	Yes	Yes
Sophocles	Philoctetes	Herakles	Tragedy	No	No	No	No	No	No	Yes	Yes	No	Male	Mid	Yes	Yes
Sophocles	Philoctetes	Neoptolemos	Tragedy	No	No	No	No	No	No	No	Yes	Yes	Male	Young	Yes	No
Sophocles	Philoctetes	Odysseus	Tragedy	No	No	No	No	No	No	No	Yes	Yes	Male	Mid	Yes	Yes
Sophocles	Philoctetes	Philoktetes	Tragedy	No	No	No	No	No	No	No	No	No	Male	Mid	Yes	Yes
Sophocles	Trachiniae	Angelos	Tragedy	No	No	Yes	No	No	No	No	No	No	Male	Mid	Yes	Yes
Sophocles	Trachiniae	Choros	Tragedy	Yes	No	Yes	No	No	No	No	No	No	Female	Young	No	No
Sophocles	Trachiniae	Deianeira	Tragedy	No	No	No	No	No	No	No	Yes	No	Female	Mid	No	Yes
Sophocles	Trachiniae	Hemichorion 1	Tragedy	Yes	No	Yes	No	No	No	No	No	No	Female	Young	No	No
Sophocles	Trachiniae	Hemichorion 2	Tragedy	Yes	No	Yes	No	No	No	No	No	No	Female	Young	No	No
Sophocles	Trachiniae	Herakles	Tragedy	No	No	No	No	No	No	No	Yes	Yes	Male	Mid	Yes	Yes
Sophocles	Trachiniae	Hyllos	Tragedy	No	No	No	No	No	No	No	Yes	Yes	Male	Young	Yes	No
Sophocles	Trachiniae	Lichas	Tragedy	No	No	No	No	No	No	No	No	Yes	Male	Mid	Yes	Yes
Sophocles	Trachiniae	Presbys	Tragedy	No	No	Yes	No	No	No	No	No	No	Male	Old	Yes	Yes
Sophocles	Trachiniae	Therapaina	Tragedy	No	No	Yes	No	Yes	No	No	No	No	Female	Mid	No	Yes
Sophocles	Trachiniae	Trophos	Tragedy	No	No	Yes	No	Yes	No	No	No	No	Female	Old	No	Yes

Figure 2: Greek theater character categorization example.

Author	LatinizedTitle	ID	LatinizedSpeakerAgent	Type	LatinizedSpeakerSu	Superset	ConseqFrom	ImportanceForAger	Lie	KnownBySubj	Error	Address	DisplayBegin	DisplayEnd
Sophocles	Philoctetes	1	Odysseus	W	Neoptolemos			HI		Yes			13 1	14 -1
Sophocles	Philoctetes	1	Odysseus	W	Philoktetes			HI		No			13 1	14 -1
Sophocles	Philoctetes	2	Odysseus	W	Neoptolemos			LI		Yes			15 1	15 -1
Sophocles	Philoctetes	3	Odysseus	P	Neoptolemos	2		LI		Yes			16 1	17 1
Sophocles	Philoctetes	4	Odysseus	P	Neoptolemos	3		LI		Yes			17 2	21 -1
Sophocles	Philoctetes	5	Odysseus	P	Neoptolemos	2		LI		Yes			22 1	23 -1
Sophocles	Philoctetes	6	Odysseus	P	Neoptolemos	2		HI		Yes			24 1	25 -1
Sophocles	Philoctetes	7	Neoptolemos	DAS			2	LI		Yes		Name	26 1	26 -1
Sophocles	Philoctetes	8	Neoptolemos	DAA			3	LI		Yes			27 1	27 -1
Sophocles	Philoctetes	9	Odysseus	W	Neoptolemos		8	LI		Yes			30 1	30 -1
Sophocles	Philoctetes	10	Neoptolemos	DAA			9	LI		Yes			31 1	31 -1
Sophocles	Philoctetes	11	Odysseus	W	Neoptolemos		1	LI		Yes			45 1	45 -1
Sophocles	Philoctetes	12	Odysseus	R	Neoptolemos	1		HI		Yes			46 1	46 5
Sophocles	Philoctetes	13	Neoptolemos	DAA			11	LI		Yes			48 1	48 -1
Sophocles	Philoctetes	14	Neoptolemos	W	Odysseus		6	HI		Yes			49 1	49 -1
Sophocles	Philoctetes	15	Odysseus	W	Neoptolemos		6	HI		Yes		Name	50 1	53 -1
Sophocles	Philoctetes	16	Odysseus	DAA			14	HI		Yes			54a 0	54a 0
Sophocles	Philoctetes	17	Odysseus	W	Neoptolemos		6	HI		Yes			54a 1	55 -1
Sophocles	Philoctetes	18	Odysseus	P	Neoptolemos	17		HI		Yes			77 1	78 -1
Sophocles	Philoctetes	19	Neoptolemos	DRS			17	HI		Yes		Name	86 1	87 -1
Sophocles	Philoctetes	20	Neoptolemos	P		19		HI					88 1	89 -1
Sophocles	Philoctetes	21	Neoptolemos	DLS		19	17	HI		Yes			90 1	91 3
Sophocles	Philoctetes	22	Neoptolemos	DRS			17	HI		Yes		Qualifier	94 3	95 -1
Sophocles	Philoctetes	23	Odysseus	R	Neoptolemos	17		HI		Yes			101 1	101 -1
Sophocles	Philoctetes	24	Odysseus+Neoptolemos	E			17 19						96 1	119 -1
Sophocles	Philoctetes	25	Odysseus	P	Neoptolemos	17		HI		Yes			113 1	113 -1

Figure 3: Categorization of decisions and expressions of will.

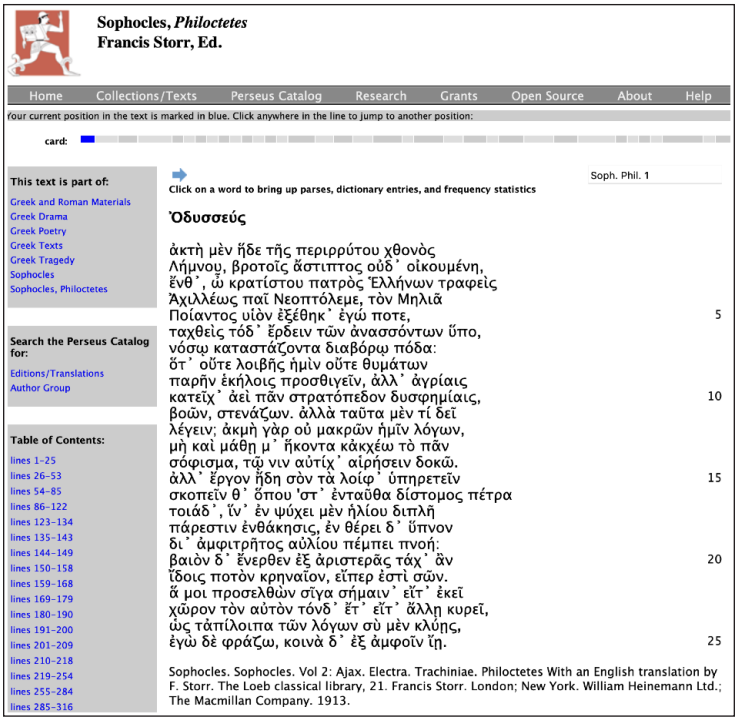

Figure 4: Sophocles's *Philoctetes* on Perseus.

Mesure	Fréquence relative au corpus entier			Corpus entier
	Décision acceptation	Décision de refus	Toutes volontés	
Nombre de mots	1201	1046	6775	114691
Verbe	119 %	100 %	107 %	21,3 %
Nom	73 %	71 %	83 %	17,0 %
Nom propre	56 %	41 %	59 %	2,8 %
Ô	128 %	147 %	85 %	0,7 %
Négation *me*, *ou*	159 %	210 %	147 %	2,1 %
ei	211 %	52 %	208 %	0,6 %
Question	19 %	120 %	63 %	1,7 %

Figure 5: Comparison of syntactic forms used in the expression of decisions, relative to occurrences in entire tragic corpus (in French).

Disambiguating the universal corpus, in which one seeks excerpts or categorizations, from the added information produced by the historian, we open the entire process to a much better understanding and to document critique. The overriding principle is that any input by a historian related to the constitution of a corpus, whether in terms of raw data or processing logic, should be reflected as a step in an algorithm or pipeline, and not as a physical operation, such as clicking on some instruction on a

piece of software. In Bod's (Bod 2013) history of the humanities, hermeneutics were commingled with the hard sciences. Here, the perspective aligns much more closely with Turner's (Turner 2014) history of philology, in that making the creation of digital corpora explicit allows a philological critique of this very process, which is impossible if one is just given an object.

5 Conclusion

Despite the essential role played by document corpora in the writing of history, and despite the availability at a large scale of electronic sources and processing capacities, historiography has not fully taken stock of the fundamental change brought about by their combination. We have seen that many researchers complained that historians did not fully embrace the digital and did not fully exploit the tools at their disposal. This is, to some extent, missing the forest for the trees. The issue is not whether one should use such and such keyword search on some database rather than spend a few hours with a large dictionary. The issue is that the corpus has been considered as a body, as data, rather than what it has fundamentally become, a process. Once this distinction is made, and once historians realize that this corpus as a static object is dead, they will be able to fully attain the scientific aspiration that Febvre had in mind, where scientificity is understood in a humanistic and non-reductive way. Recognizing the corpus as a process rather than an object also sets historiography in a structuralist perspective, where meaning emerges through the structured relation between selected elements and the broader dataset. The ability to refine corpora dynamically, through hierarchical selection or filtering, turns their creation into an interpretative workflow that must be explicit and reproducible.

Many research journals focused on the more quantitative aspects of history nowadays ask that contributors provide the underlying data, if any, supporting their analyses. This requirement would be more beneficial to all if it was instead phrased as a request for the code or the algorithms that supported the analysis.

Acknowledgements

I wish to thank Daniel O'Donnell and an anonymous reviewer for their fruitful recommendations about this article.

Competing interests

The author has no competing interests to declare.

Contributions

Editorial

Section Editor

Frank Onuh, The Journal Incubator, University of Lethbridge, Canada

Copy Editor

Christa Avram, The Journal Incubator, University of Lethbridge, Canada

Copy and Layout Editor

A K M Iftekhar Khalid, The Journal Incubator, University of Lethbridge, Canada

References

Aledavood, Parham. 2024. "Taking the Middle Road: Reflections on Mixed Methodology within the Digital Humanities." *Digital Studies/Le champ numérique* 14 (1): 1–19. Accessed May 11, 2025. https://doi.org/10.16995/dscn.11069.

Andrews, Tara L., and Caroline Macé, eds. 2015. *Analysis of Ancient and Medieval Texts and Manuscripts: Digital Approaches*. Studies in the Transmission of Texts and Ideas 1. Brepols.

Aron, Raymond. (1938) 1986. *Introduction à la philosophie de l'histoire. Essai sur les limites de l'objectivité historique*. Tel 58. Gallimard.

Audin, Yann. 2025. "For a General Theory of Scholars–Digital Tools Interactions." *Digital Studies/Le champ numérique* 15 (1): 1–20. Accessed May 11. https://doi.org/10.16995/dscn.11058.

Barbrook, Adrian C., Christopher J. Howe, Norman Blake, and Peter Robinson. 1998. "The Phylogeny of *The Canterbury Tales*." *Nature* 394 (6696): 839–840. Accessed May 11, 2025. https://doi.org/10.1038/29667.

Barker, Elton, Leif Isaksen, Nick Rabinowitz, Stefan Bouzarovski, and Chris Pelling. 2013. "On Using Digital Resources for the Study of an Ancient Text: The Case of Herodotus' *Histories*." In *The Digital Classicist 2013*, edited by Stuart E. Dunn and Simon Mahony, 45–62. Bulletin of the Institute of the Classical Studies 122. University of London Press. Accessed May 11, 2025. https://uolpress.co.uk/book/the-digital-classicist-2013/.

Barker, Elton, and Melissa Terras. 2016. "Greek Literature, the Digital Humanities, and the Shifting Technologies of Reading." *Oxford Handbooks Online* 1–25. Accessed May 11, 2025. https://doi.org/10.1093/oxfordhb/9780199935390.013.45.

Bermejo Barrera, Jose Carlos. 2001. “Making History, Talking about History.” *History & Theory* 40 (2): 190–205. Accessed May 11, 2025. https://doi.org/10/b9d829.

Blanke, Tobias, and Mark Hedges. 2013. “Scholarly Primitives: Building Institutional Infrastructure for Humanities e-Science.” *Future Generation Computer Systems* 29 (2): 654–661. Accessed May 11, 2025. https://doi.org/10.1016/j.future.2011.06.006.

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

Boutier, Jean. 2014. “L’usage historien des archives.” In *Corpus, sources et archives*, by Jean Boutier, Jean-Louis Fabiani, and Jean-Pierre Olivier de Sardan, 9–22. Études et travaux de l’IRMC. Institut de recherche sur le Maghreb contemporain. Accessed May 11, 2025. http://books.openedition.org/irmc/776.

Burns, Patrick J. 2019. “Building a Text Analysis Pipeline for Classical Languages.” In *Digital Classical Philology: Ancient Greek and Latin in the Digital Revolution*, edited by Monica Berti, 159–176. Age of Access? Grundfragen der Informationsgesellschaft 10. De Gruyter. Accessed June 5, 2025. https://www.degruyterbrill.com/document/doi/10.1515/9783110599572-010/html.

Calero Valdez, André. 2020. “Making Reproducible Research Simple Using RMarkdown and the OSF.” In *Social Computing and Social Media. Design, Ethics, User Behavior, and Social Network Analysis*, edited by Gabriele Meiselwitz, 27–44. Lecture Notes in Computer Science 12194. Springer International Publishing. Accessed May 11, 2025. http://link.springer.com/10.1007/978-3-030-49570-1_3.

Chaudhuri, Pramit, and Joseph P. Dexter. 2017. “Bioinformatics and Classical Literary Study.” In “Computer-Aided Processing of Intertextuality in Ancient Languages,” edited by Marco Büchler and Laurence Mellerin, special issue, *Journal of Data Mining & Digital Humanities*, 1–8. Accessed May 11, 2025. https://doi.org/10.46298/jdmdh.1386.

Crane, Gregory. 2012. *Perseus Digital Library*. Tufts University. Accessed May 11, 2025. http://www.perseus.tufts.edu.

Crane, Gregory, Casey Due, Mary Ebbott, David Elmer, Douglas Frame, Angelia Hanhardt, et al. 2016. “A Homer Commentary in Progress.” Accessed June 6, 2025. https://homer.oc.newalexandria.info.

Crymble, Adam. 2021. *Technology and the Historian: Transformations in the Digital Age*. Topics in the Digital Humanities. University of Illinois Press.

Erickson, Ansley T. 2013. “Historical Research and the Problem of Categories: Reflections on 10,000 Digital Note Cards.” In *Writing History in the Digital Age*, edited by Jack Dougherty and Kristen Nawrotzki, 133–145. Digital Humanities. University of Michigan Press. Accessed June 5, 2025. https://muse.jhu.edu/pub/166/oa_edited_volume/chapter/1030713.

Evans, Thomas L., and Patrick T. Daly. 2006. *Digital Archaeology: Bridging Method and Theory*. Routledge.

Febvre, Lucien. 1934. “Comptabilité et Chambre des Comptes.” *Annales d’histoire économique et sociale* 6 (26): 148–153. Accessed May 12, 2025. https://www.jstor.org/stable/27573281.

———. (1952) 1992. *Combats pour l’histoire*. L’ancien et le nouveau 12. Armand Colin.

Foucault, Michel. 1969. *L’archéologie du savoir*. Gallimard.

Gauthier, Laurent. 2025. “Les interjections *hein* et *non* dans l’oral représenté (XVIe–XXe s.).” *Langue française* 225 (1): 43–57. Accessed May 12, 2025. https://doi.org/10.3917/lf.225.0043.

Gibbs, Fred, and Trevor Owens. 2013. “The Hermeneutics of Data and Historical Writing.” In *Writing History in the Digital Age*, edited by Jack Dougherty and Kristen Nawrotzki, 159–170. Digital Humanities. University of Michigan Press. Accessed June 6, 2025. https://muse.jhu.edu/pub/166/oa_edited_volume/chapter/1030715.

Graham, Shawn, Ian Milligan, and Scott Weingart. 2015. *Exploring Big Historical Data: The Historian’s Macroscope*. Imperial College Press.

Greetham, David. 2013. “A History of Textual Scholarship.” In *The Cambridge Companion to Textual Scholarship*, edited by Neil Fraistat and Julia Flanders, 16–41. Cambridge University Press.

Guaresi, Magali. 2019. “La logométrie en histoire, une herméneutique numérique. Exploration d’un corpus de professions de foi électorales de député-e-s (1958–2007).” *Digital Studies/Le champ numérique* 9 (1): 1–15. Accessed May 12, 2025. https://doi.org/10/gkmqzz.

Hitchcock, Tim. 2013. “Confronting the Digital: Or How Academic History Writing Lost the Plot.” *Cultural and Social History* 10 (1): 9–23. Accessed May 12, 2025. https://doi.org/10/gc3gsg.

Hoekstra, Rik, and Marijn Koolen. 2019. “Data Scopes for Digital History Research.” *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 52 (2): 79–94. Accessed May 12, 2025. https://doi.org/10/gmdj4t.

Hoover, David L. 2013. “Quantitative Analysis and Literary Studies.” In *A Companion to Digital Literary Studies*, edited by Ray Siemens and Susan Schreibman, 517–533. Blackwell Companions to Literature and Culture. Blackwell Publishing. Accessed May 12, 2025. https://doi.org/10.1002/9781405177504.ch28.

Howe, Christopher J., Adrian C. Barbrook, Matthew Spencer, Peter Robinson, Barbara Bordalejo, and Linne R. Mooney. 2001. “Manuscript Evolution.” *Trends in Genetics* 17 (3): 147–152. Accessed May 12, 2025. https://doi.org/10.1016/S0168-9525(00)02210-1.

Jockers, Matthew L., and Rosamond Thalken. 2020. *Text Analysis with R: For Students of Literature*. 2nd ed. Quantitative Methods in the Humanities and Social Sciences. Springer International Publishing. Accessed May 12, 2025. https://doi.org/10.1007/978-3-030-39643-5.

Kenna, Ralph, Máirín MacCarron, and Pádraig MacCarron, eds. 2017. *Maths Meets Myths: Quantitative Approaches to Ancient Narratives*. Understanding Complex Systems. Springer International Publishing.

Kennedy, Graeme. 1998. *An Introduction to Corpus Linguistics*. Studies in Language and Linguistics. Longman.

Kiewra, Kenneth A. 1989. “A Review of Note-Taking: The Encoding-Storage Paradigm and Beyond.” *Educational Psychology Review* 1 (2): 147–172. Accessed May 12, 2025. https://doi.org/10.1007/BF01326640.

Magnani, Eliana. 2017a. “Qu’est-ce qu’un corpus?” *Journées d’étude*, October 2. Les carnetiers de l’IHRT. Accessed May 12, 2025. https://irht.hypotheses.org/3187.

———. 2017b. “Un corpus structuré et hétérogène de textes latins médiévaux (Bourgogne, Ve–XVe siècle).” *Bulletin du CERCOR* 41: 59–65. Accessed May 12, 2025. https://shs.hal.science/halshs-01529451/document.

Marrou, Henri-Irénée. (1954) 2016. *De la connaissance historique*. Histoire H21. Seuil.

Mayaffre, Damon. 2002. “Les corpus *réflexifs*: entre architextualité et hypertextualité.” *Corpus* 1:1–15. Accessed May 12, 2025. https://doi.org/10/gktbfv.

McEnery, Tony, and Vaclav Brezina. 2022. *Fundamental Principles of Corpus Linguistics*. Cambridge University Press.

McGillivray, Barbara, Jon Wilson, and Tobias Blanke. 2019. “Towards a Quantitative Research Framework for Historical Disciplines.” In *Proceedings of the Workshop on Computational Methods in the Humanities 2018* 2314: 53–58. CEUR Workshop Proceedings. Accessed May 12, 2025. https://doi.org/10.17863/CAM.36281.

Mugelli, Gloria, Andrea Bellandi, Federico Boschetti, and Anas Fahad Khan. 2017. “Designing an Ontology for the Study of Ritual in Ancient Greek Tragedy.” In *Language, Ontology, Terminology and Knowledge Structures (LOTKS 2017)*, edited by Francesca Frontini, Larisa Simeunović, Špela Vintar, Fahad Khan, and Artemis Parvisi, 96–105. Association for Computational Linguistics. https://aclanthology.org/W17-7011/.

Näslund, Anna, and Amanda Wasielewski. 2021. “The Digital U-Turn in Art History.” *Konsthistorisk tidskrift/Journal of Art History* 90 (4): 249–266. Accessed May 11, 2025. https://doi.org/10.1080/00233609.2021.2006774.

Nawrotzki, Kristen, and Jack Dougherty. 2013. “Introduction.” In *Writing History in the Digital Age*, edited by Kristen Nawrotzki and Jack Dougherty, 1–18. Digital Humanities. University of Michigan Press. Accessed June 6, 2025. https://muse.jhu.edu/pub/166/oa_edited_volume/chapter/1030699.

Nicolet-Pierre, Hélène. 2002. *Numismatique grecque*. Les Outils de l’histoire. Armand Colin.

Offenstadt, Nicolas. 2011. “Archives, Documents, Sources.” In *Historiographies. Concepts et débats*, edited by Christian Delacroix, François Dosse, Patrick Garcia, and Nicolas Offenstadt, 1: 220–231. Folio Histoire 179. Gallimard.

Pacheco, André. 2022. “Digital Humanities or Humanities in Digital: Revisiting Scholarly Primitives.” *Digital Scholarship in the Humanities* 37 (4): 1128–1140. Accessed May 12, 2025. https://doi.org/10.1093/llc/fqac012.

Pantelia, Maria. 2020. *Thesaurus Linguae Graecae*. University of California, Irvine. Accessed June 6, 2025. http://stephanus.tlg.uci.edu.

Pflug, Gunther. (1954) 1971. “The Development of Historical Method in the Eighteenth Century.” *History and Theory* 11: 1–23. Accessed May 12, 2025. https://doi.org/10/ccrnm3.

Popper, Karl R. (1934) 1992. *The Logic of Scientific Discovery*. 2nd ed. Routledge.

Prost, Antoine. 1996. *Douze leçons sur l’histoire*. Histoire. Seuil.

Ricoeur, Paul. 1971. “What Is a Text? Explanation and Interpretation.” In *Mythic-Symbolic Language and Philosophical Anthropology*, edited by David M. Rasmussen, 135–150. Springer. Accessed May 12, 2025. http://link.springer.com/10.1007/978-94-011-9327-6_7.

Rygiel, Philippe. 2011. “L’enquête historique à l’ère numérique.” *Revue d’histoire moderne et contemporaine* 58-4bis (5): 30–40. Accessed May 12, 2025. https://doi.org/10/gkmrf7.

Siebold, Anna, and Matteo Valleriani. 2022. “Digital Perspectives in History.” *Histories* 2 (2): 170–177. Accessed May 12, 2025. https://doi.org/10.3390/histories2020013.

Treffort, Cécile. 2014. “Le corpus du chercheur, une quête de l’impossible ? Quelques considérations introductives.” *Annales de Janua*, April 2. Accessed June 6, 2025. https://annalesdejanua.edel.univ-poitiers.fr/index.php?id=725.

Turner, James. 2014. *Philology: The Forgotten Origins of the Modern Humanities*. Princeton University Press.

Underwood, Ted, Laura McGrath, Richard Jean So, and Chad Wellmon. 2022. “Culture, Theory, Data: An Introduction.” *New Literary History* 54 (1): 519–530. Accessed May 12, 2025. https://doi.org/10.1353/nlh.2022.a898319.

Unsworth, John. 2000. “Scholarly Primitives: What Methods Do Humanities Researchers Have in Common, and How Might Our Tools Reflect This?” Presented at Humanities Computing: Formal Methods, Experimental Practice, King’s College, London, May 13. Accessed May 12, 2025. https://people.brandeis.edu/~unsworth/Kings.5-00/primitives.html.

Wickham, Hadley, Mine Çetinkaya-Rundel, and Garrett Grolemund. 2017. *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. 2nd ed. O’Reilly.

Wirth, Niklaus. 1976. *Algorithms + Data Structures = Programs*. Prentice-Hall Series in Automatic Computation. Prentice-Hall.

Yeruva, Vijaya Kumari, Mayanka ChandraShekar, Yugyung Lee, Jeff Rydberg-Cox, Virginia Blanton, and Nathan A Oyler. 2020. “Interpretation of Sentiment Analysis in Aeschylus’s Greek Tragedy.” In *Proceedings of the 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, edited by Stefania DeGaetano, Anna Kazantseva, Nils Reiter, and Stan Szpakowicz, 138–146. International Committee on Computational Linguistics. Accessed June 6, 2025. https://aclanthology.org/2020.latechclfl-1.17/.