Journal of Data Mining and Digital Humanities

EPIsciences overlay journals

EpiSearch. Identifying Ancient Inscriptions in Epigraphic Manuscripts

Lorenzo Calvelli[1]**, Federico Boschetti**[2]**, Tatiana Tommasi**[3]

[1]DSU and VeDPH, Ca' Foscari University of Venice, Italy
[2]VeDPH and CNR-ILC, Venice, Italy
[3]University of Udine, Ca' Foscari University of Venice and VeDPH, Italy

Corresponding author: Lorenzo Calvelli , `lorenzoc@unive.it`

Abstract

Epigraphic documents are an essential source of evidence for our knowledge of the ancient world. Nonetheless, a significant number of inscriptions have not been preserved in their material form. In fact, their texts can only be recovered thanks to handwritten materials and, in particular, the so-called epigraphic manuscripts. EpiSearch is a pilot project that explores the application of digital technologies deployed to retrieve the epigraphic evidence found in these sources. The application of Handwritten Text Recognition (HTR) to epigraphic manuscripts is a challenging task, given the nature and graphic layout of these documents. Yet, our research shows that, even with some limits, HTR technologies can be used successfully.

Keywords

Epigraphy; Manuscripts; Inscriptions; Digital Humanities; Handwritten Text Recognition

I THE EPIGRAPHIC CULTURES OF THE ANCIENT WORLD, THE HANDWRITTEN TRADITION OF EPIGRAPHY AND THE EPISEARCH PROJECT

1.1 The Epigraphic Cultures of the Ancient World

Epigraphy is the science that studies communication through written media (for two different approaches to defining the term: [Panciera, 2012] and [Grossi, 2016]).[1] In the ancient world, a variety of epigraphic cultures developed across the Near East and the Mediterranean for about four millennia (ca. 3,400 BCE - 600 CE), using different writing systems to reproduce numerous languages, none of which is still spoken today. In particular, from the 9th century BCE onwards, the spread of the alphabetic writing promoted by the Phoenician city-states led to the development of different scripts across the Mediterranean, including Italy, until the Greek and Latin alphabets became predominant in the Hellenistic and Roman times [Ferrara, 2021].

Under the Roman empire, writing became an extremely common social practice, which extended over a vast territory, developing across three continents (Europe, Asia, Africa) and embracing a wide range of areas, including some that until then had not produced specific epigraphic cultures. Thanks to the widespread diffusion of basic mass literacy, between the late 1st century BCE and the early 3rd century CE, the tendency to write both simple and complex messages on an incredible variety of supports (not only wax tablets, papyrus and parchment, but also stone, wood, clay and metal objects, as well as walls, rocks and all sorts of surfaces)

[1] We are deeply grateful to the Director and the Curators of the Marciana National Library in Venice for facilitating our work and granting us the permit to publish reproductions of Astori's epigraphic manuscript.

became a ‘global’ cultural practice and a privileged medium of social communication [Boyes et al., 2021]. In recent years, scholars have labelled this phenomenon as the ‘epigraphic habit’ of the Romans ([MacMullen, 1982] and [Beltrán Lloris, 2015]).

Even if the majority of the epigraphic texts produced in the ancient Mediterranean world have been lost because of the perishable nature of the objects upon which they were written, hundreds of thousands of Greek and Latin inscriptions have survived to date [Mullen and Bowman, 2021], along with numerous documents from Egypt and the Ancient Near East, and a more limited number of texts written in fragmentary languages, such as Gaulish, Raetic, Celtiberian etc.[2] This body of evidence represents a direct legacy of the writing cultures of the ancient world, one which has come down to us without any mediation, unlike the texts of ancient literary authors, which have mostly survived thanks to copies made in the Middle Ages and later.

1.2 The Handwritten Tradition of Epigraphy

Ancient inscriptions are now accessible through a vast collection of printed corpora and online digital resources, which offer information on a multitude of epigraphic sources that constitute a crucial category of documents for the study of ancient history, as well as an immense repository of big data ([De Santis and Rossi, 2018]; [Velázquez Soriano and Espinosa Espinosa, 2021]). Nonetheless, not all the inscriptions that are known to us have been preserved in their material form. In fact, the texts of thousands of them can only be reconstructed from transcriptions that were made in post-classical times. These texts are documented by the so-called epigraphic manuscripts, which represent a rare kind of handwritten source, often very rich in information, since they offer the only record for textual sources whose original supports have not survived ([Buonocore, 2015] and [Calvelli et al., 2019]). Epigraphic manuscripts are also crucial for our knowledge of the lifecycle of ancient inscribed objects and monuments, meaning that they help us reconstruct the individual story of each inscription, from the moment when it was originally produced to the present, or to the time when it was last documented (in cases where it later disappeared or was destroyed). In spite of the invaluable amount of information that they can provide, in recent decades epigraphic manuscripts have been rather neglected in scientific research. Even if studies of these sources have recently been revived, thanks to an increasing interest in the historiography of epigraphic studies and in forgeries (see [Stenhouse, 2005]; [Orlandi, 2019]; [Calvelli et al., 2019] and [Pérez Galván, 2021]), the common assumption is that all critical work on handwritten documents related to epigraphy was already carried out by the founding fathers of the discipline in the second half of the 19th century. The opportunities offered by the Digital Humanities show that this is not the case.

1.3 The EpiSearch Project

EpiSearch is one of the first research projects that applies a Digital Humanities approach to the field of epigraphic manuscripts. It explores the possibilities offered by state-of-the-art technologies to investigate this kind of document with the aim of creating a system able to link the data recorded in handwritten texts and those registered in the main digital repositories of Greek and Latin inscriptions that are currently available online. These include the Searchable Greek Inscriptions tool promoted by the Packard Humanities Institute (PHI)[3] and the databases of the international federation of epigraphic databases named EAGLE (Electronic Archive of Greek and Latin Epigraphy, now available through the Europeana network of Ancient Greek and Latin

[2] For a recent set of digital publications on these languages see http://aelaw.unizar.es/publications

[3] https://inscriptions.packhum.org

Epigraphy),[4] in particular the Epigraphic Database Roma (EDR),[5] which provides texts, bibliographic citations and descriptive data for Latin and Greek inscriptions from ancient Italy (including Sicily and Sardinia), the Epigraphische Datenbank Heidelberg (EDH),[6] which contains the texts of Latin and bilingual inscriptions from the provinces of the Roman Empire, and the Epigraphic Database Bari (EDB),[7] devoted to inscriptions promoted by the Christian community of ancient Rome. Another important dataset is offered by the Epigraphik Datenbank Clauss-Slaby (EDCS),[8] which supplies texts and bibliographic citations (lemmata of editions) for nearly all published Latin inscriptions.

EpiSearch is conceived as the initial segment of a broader and more ambitious research project, of which it constitutes a proof of concept. As a working example, we identified an epigraphic manuscript written in Venice in the early 1700s by a local antiquarian, named Giovanni Antonio Astori, and kept at the Marciana National Library in Venice.

The project includes three main steps. The first involved the application of Handwritten Text Recognition (HTR) technologies for the automatic acquisition of the manuscript's contents. The second encompassed designing an integrated system, created by collecting data from the main online epigraphic databases; this gave us the possibility to match the inscriptions that are transcribed in the manuscript with their editions in digital resources. The last step, which is still being implemented, will produce a visually annotated version of the manuscript with hyperlinks to the online databases for connecting the transcriptions of inscriptions with their current digital editions.[9]

The EpiSearch team includes Federico Boschetti: Institute for Computational Linguistics "A. Zampolli" – National Research Council of Italy (CNR-ILC), Pisa / Venice Centre for Digital and Public Humanities (VeDPH), Ca' Foscari University of Venice; Lorenzo Calvelli, PI: Department of Humanities and VeDPH, Ca' Foscari University of Venice; Franz Fischer: Department of Humanities and VeDPH, Ca' Foscari University of Venice; Daniele Fusi: VeDPH, Ca' Foscari University of Venice; Silvia Orlandi: Sapienza University of Rome; Thea Sommerschield: Department of Humanities and VeDPH, Ca' Foscari University of Venice; Tatiana Tommasi: University of Udine, Ca' Foscari University of Venice and VeDPH.

[Lorenzo Calvelli]

II ASTORI'S EPIGRAPHIC MANUSCRIPT

2.1 Astori's Profile

The manuscript chosen as a case study was written by Giovanni Antonio Astori (Venice, 1672-1743), a learned ecclesiastical antiquarian. We are informed in detail about Astori's life thanks to a highly reliable biography published by the Italian bibliographer Mazzuchelli [1753] [Cappelletti, 2011, 261-262].

Astori developed an early interest in epigraphic studies. In fact, by the end of the 17th century CE, his contemporaries already described him as skilful at transcribing ancient inscriptions

[4] https://www.eagle-network.eu
[5] http://www.edr-edr.it
[6] https://edh.ub.uni-heidelberg.de
[7] https://www.edb.uniba.it
[8] http://www.manfredclauss.de
[9] The annotated version of the manuscript will be accessible to the public through the ILC4CLARIN platform.

(for example, see [Morelli, 1785, 65, nr. 38]). The letters exchanged within his cultural network are important sources to understand his epigraphic interests. In fact, Astori was in contact with some of the most important intellectuals of his time and had the opportunity to exchange ideas on antiquarian subjects with them. For example, he discussed epigraphic topics with Ludovico Antonio Muratori (1672-1750), the author of the *Novus thesaurus veterum inscriptionum* (1739-1742). Of the epistolary exchange between Astori and Muratori, only Astori's letters have survived. They are 11 in total, dated from 1705 to 1709 and currently preserved in Modena.[10] The letters were all edited by Di Campli and Forlani [1995, 285-291]. A high-resolution digitised copy of 10 of them is also available online through the *Internet Culturale* web portal.[11]

Astori's epigraphic interests are evident in his first antiquarian articles, published around 1697 (see, for example, [Astori, 1697]). In fact, several of these works were dedicated to the analysis of the ancient inscriptions that were then visible in Venice, paying specific attention to both their material and palaeographical aspects. These features, particularly useful to study inscribed monuments that have not survived in their physical form, can also be detected in Astori's epigraphic manuscript.

2.2 Astori's Epigraphic Manuscript

The codex is currently kept in Venice at the Marciana National Library, with the following shelfmark: Marc. Lat. XIV, 200 (4336). A summary description of the manuscript is provided by Zorzanello [1985, 273]. It was already known to Theodor Mommsen, the founding father of the epigraphic discipline, who used it for compiling the *Corpus inscriptionum Latinarum* (*CIL*) and, in particular, for studying the inscriptions attested in Venice for the edition of *CIL* V, *pars* I, *Inscriptiones regionis Italiae decimae* (*Venetia et Histria*), published in 1872.[12] The manuscript gained again the attention of scholars about one hundred years later, when it was mentioned in some eminent studies on Venetian antiquarian collections carried out by Zorzi [1988, 90-91] and Favaretto [1990, 356-357, 384 and 390]. Nonetheless, as of now only one short, yet well-researched, article [Bodon, 1996] has been dedicated exclusively to Astori's epigraphic manuscript.[13]

The codex, rich in relevant information, does not contain a large amount of written text in terms of quantity. It is composed of IV front flyleaves + 14 leaves (of which leaves 3*bis*, 6*ter* and probably also 6 were added by Astori to the original quire) + II back flyleaves; moreover, several leaves are blank.[14] The original front cover of the manuscript, still preserved as front flyleaf IIrv and back flyleaf Irv, contains the title of the work, probably written by Astori himself: *Inscriptiones Graecae et Lat*(*in*)*ae quae Venetiis reperiunt*(*ur*), *aut nondum editae*, *aut correctius*, *si ab aliis vulgatae s*(*un*)*t*, *public*(*atae*) *nunc demum* (*Greek and Latin inscriptions that are to be found in Venice, either not yet published, or now at last better edited, if they were*

[10] Biblioteca Estense Universitaria, Archivio Muratori, 37.10 (1 letter), 49.38 (1 letter), 52.01 (8 letters) and 86.04c (1 letter).

[11] These sources, available at https://www.internetculturale.it under a licence CC BY-NC-SA 3.0, have been used by Federico Boschetti for the process of fine-tuning described in section 3.4 of this article.

[12] Nonetheless, Mommsen underlined the limits of the manuscript, probably because of its incompleteness; cf. *CIL* V, p. 205: ≪*perfunctoriam tantum operam dedit nec multum nos iuvit*≫ (≪he produced only a superficial work and not very useful for us≫).

[13] This article offers a preliminary study of Astori's manuscript, but a thorough examination of its contents has not yet been carried out.

[14] The following leaves are blank: front flyleaves Iv, IIIv, IVv; 3*bis*v, 6*bis*r, 6*ter*v, 8r-11v; back flyleaves Ir-IIv.

disclosed by others).[15] This title clearly shows that Astori wanted to collect all the Greek and Latin inscriptions which were visible at his time in Venice and had not yet been published or had already been published but with some mistakes.

The contents of the manuscript are particularly interesting. In fact, Venice is a unique case of study. While the city and the lagoon sites around it did not develop on top of ancient Roman settlements, many ancient monuments, including Greek and Roman inscriptions, were reused as building materials or collected for antiquarian purposes [Calvelli, 2018, 87-89]. Yet, in the course of the ensuing centuries, most of these inscribed monuments were displaced or destroyed and can no longer be seen. For this reason, the information contained in epigraphic manuscripts related to the Venice area is of fundamental value [Calvelli, 2016, 464-466 and 484].

Astori's codex includes the transcriptions of 56 inscriptions, 33 of which are Greek and 23 Latin.[16] Of these epigraphic monuments, 36 are still preserved, while 20 are lost or of unknown location. Astori transcribed inscriptions belonging especially to Venetian private and public collections (43 inscriptions), but also inscribed monuments which had been reused in the city of Venice ('epigraphic *spolia*': at least 9 inscriptions).[17]

For each inscription, Astori first specified its location; he then transcribed the epigraphic text in upper-case, respecting the division of lines and often reproducing some palaeographical features of the letters. For Greek inscriptions, he also wrote a translation in Latin. Moreover, in many cases (more than half of the total), he made an ink drawing of the monument bearing the inscription, sometimes preceded by a preparatory pencil sketch. These drawings were often made on separate leaves and then stuck to the manuscript with sealing wax. The analysis of the characteristics of the manuscript makes it clear that we are dealing with preparatory materials. The codex can be considered as the last phase of an editorial project, which never reached the stage of a printed edition. Nonetheless, the 'work in progress' nature of the manuscript helps us understand how Astori created it. He saw the inscriptions in person and transcribed their texts, often drawing a sketch of the monuments upon which they were carved. Finally, he assembled the materials that he had already collected in the codex, probably choosing all the inscriptions that he wanted to publish, but without giving them a final order.

From the analysis of its contents, the manuscript can be dated approximately between 1706 and 1713. The *terminus post quem* can be fixed thanks to the transcription of *CIL* V 2792 at f. 1v nr. 10, a Latin inscription from Montegrotto, near Padua, discovered not long before 1706 [Vallisneri, 1706, 113]. The *terminus ante quem* is based upon the transcription of *CIL* V 2151 at f. 1r nr. 2, attested by Astori in the house of Bertucci Contarini. Considering that Bertucci died in 1713, it is possible to suggest that the transcription was produced before that year. In spite of that, taking into account other external sources, first of all the letters exchanged between Astori and his contemporaries, one may infer that Astori had already transcribed almost all of the inscriptions of the manuscript between 1700 and 1704 [Morelli, 1785, 89-90 and 222]. In

[15] In front flyleaves Ir, IIv, IIIr and IVr different hands wrote notes and bibliographic references. In particular, at front flyleaf IVr three Latin inscriptions were transcribed, two of which are ancient (*CIL* V 2180 and *CIL* V 2168), while the other is post-classical. Notes written by a hand different from Astori's are clearly visible also in some leaves of the epigraphic collection (for example at f. 2v nr. 16 and at f. 4r nr. 21). They were intended to update the locations where the inscriptions could be seen.

[16] The inscriptions transcribed by a different hand at front flyleaf IVr have been excluded from the count.

[17] For an overview of all the inscriptions transcribed by Astori in his epigraphic manuscript see the Table created by Tatiana Tommasi for the EpiSearch project and accessible online through the following link: https://github.com/vedph/episearch-htr/blob/main/epigraphic_manuscript_by_Astori.csv

the following period Astori rearranged the collected materials and created the manuscript as it is visible today. Yet, he never actually managed to organise his epigraphic collection, which, therefore, remained incomplete.

The analysis of the contents of the manuscript written by Astori confirms its fundamental value for several reasons [Calvelli, 2004, 444]. It allows us to better understand the state of the epigraphic studies in Venice in the early modern period. At the same time, it gives us the possibility to analyse otherwise lost phases of the lifecycles of the inscriptions. Finally, Astori's attention towards the physical monuments makes his work particularly useful for studying ancient inscriptions which are no longer preserved or are still to be located.

The importance of the contents of the codex and its graphic characteristics, rich in drawings, have guided the selection of it as a suitable case study for the EpiSearch project, with the objective of extracting the precious data contained in the manuscript and, at the same time, making them more accessible through digital technologies.

[Tatiana Tommasi]

III HTR APPLIED TO ASTORI'S MANUSCRIPT

3.1 HTR Solutions

Although in the early 2000s the acquisition by Optical Character Recognition (OCR) of printed texts in ancient Greek was challenging, nowadays not only the application of OCR to critical editions is satisfactory [Romanello et al., 2021], but also the application of HTR to Greek manuscripts is yielding promising results [Perdiki, 2022].

Among the solutions for Handwritten Text Recognition, Transkribus [Kahle et al., 2017] is one of the most used applications in the community of social sciences and humanities [Nockels et al., 2022] (for a recent comparison between Transkribus and eScriptorium, see [Thompson, 2021]), but valid competitors populate the scene, such as OCR for all [Reul et al., 2019],[18] Arkindex,[19] Project PERO,[20] eScriptorium,[21] and others (for example a new tool illustrated in [Cascianelli et al., 2021]).

eScriptorium [Kiessling et al., 2019] is a collaborative web environment based on Kraken (version 3.0.0).[22] The application is open source and encourages the sharing of open HTR models, which can be refined to fit the specific needs of researchers. In this spirit of collaboration the project HTR-United [Chagué et al., 2021] uses open ground-truth transcriptions to create HTR models and provides it with rich metadata.

[18] https://www.ocr4all.org

[19] https://teklia.com/solutions/arkindex

[20] https://pero.fit.vutbr.cz/about

[21] https://escriptorium.inria.fr/

[22] https://kraken.re/master/index.html

3.2 Astori's Manuscript on eScriptorium

An epigraphic manuscript is a limit case study for layout analysis and HTR techniques, because a) text regions are included in the image regions representing the epigraphic monuments; b) texts are multilingual and rendered in different scripts; c) alphabetical signs may be fragmented. As shown in Figure 1,[23] the complex layout of a typical page is constituted by the following kinds of regions: location (light blue), numbering (light red), drawing of the epigraphic monument (orange), Latin inscription (magenta), Greek inscription (purple), and translation from Greek to Latin (lime green).

Astori's manuscript consists only of 14 leaves (17 written pages), which do not provide a sufficient amount of text to create a new training set from scratch and successfully apply it to the rest of the manuscript. But the layout analysis and HTR techniques can be used even on a few pages already entirely transcribed by hand, at least for mapping the text glyph by glyph on the facsimile and testing the fine-tuning of a model with a minimal amount of data.

The detached research unit of the CNR-ILC at the VeDPH provides the scholars and students affiliated to the Centre with an instance of eScriptorium[24] installed on the servers maintained by ILC4CLARIN.[25] The digital facsimile of Astori's epigraphic manuscript was kindly provided by the Marciana Library and was uploaded on the platform (an open access version of it should soon be available through the *Internet Culturale* web portal).

The regions of interest (ROIs: see Figure 1)[26] have been manually identified and marked according to the SegmOnto guidelines [Gabay et al., 2021]. Due to the peculiarity of epigraphic manuscripts, we defined the following subtypes: CustomZone:provenance for the location; NumberingZone:inscriptionNumber for the numeric identifier of each inscription; GraphicZone:textBearingObject for the image of the epigraphic monument; CustomZone:greekInscription and CustomZone:latinInscription for the transcriptions; and CustomZone:translation for the Latin translation of Greek inscriptions.

3.3 Mapping the Transcription on the Facsimile

The manual digitisation of Greek and Latin inscriptions transcribed by Astori has been performed by Tatiana Tommasi, according to two criteria: a) all the allographs of the Greek or Latin alphabetic signs have been represented by the corresponding upper-case letter;[27] b) fragmentary characters have been ignored; non alphabetic signs have been ignored.

We performed the recognition of text baselines by Kraken, the layout analyser and HTR engine behind eScriptorium. The result is highly accurate (as shown in Figure 2)[28] even when the text of the inscriptions is inside the drawing of the epigraphic monument. The aforementioned manual digitisation of Astori's transcriptions of the inscriptions has been used as input into the ALTO-XML file downloaded from eScriptorium after the layout analysis, in order to map the text to the facsimile line by line. The fine-grained mapping, glyph by glyph, has been obtained by overfitting the HTR engine. In normal conditions, overfitting must be prevented to avoid a biased recognition of samples absent in the training set. But in our case, we had at our disposal the complete manual transcription of the inscriptions (which are a small amount of text) and our

[23] See p. 8, Figure 1.

[24] Version 0.11.0 available at https://gitlab.inria.fr/scripta/escriptorium

[25] https://ilc4clarin.ilc.cnr.it

[26] See p. 8, Figure 1.

[27] For example, both diamond-shaped and circular O have been encoded as upper-case omicron.

[28] See p. 8, Figure 2.

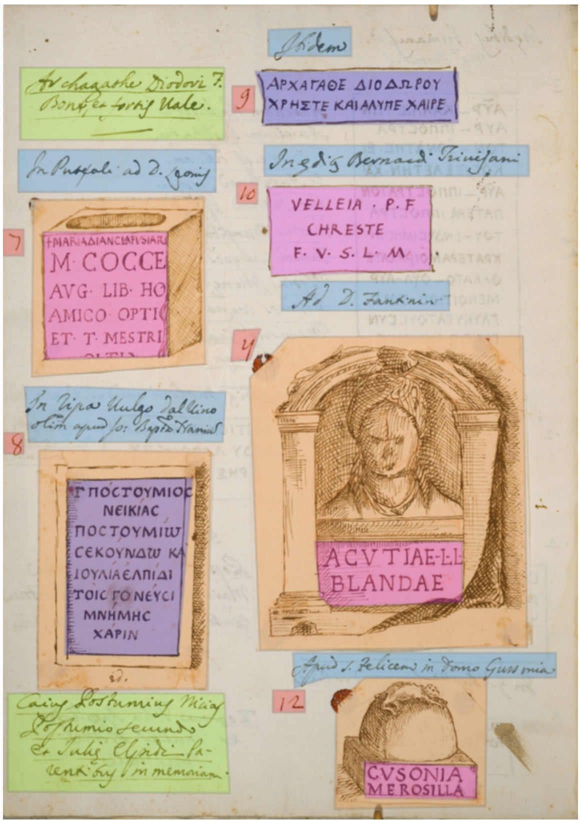


Figure 1: Venice, Marciana National Library, Marc. Lat. XIV, 200 (4336), f. 1v: ROIs coloured by type. By concession of the Italian Ministry of Culture - Marciana National Library; reproduction is forbidden.

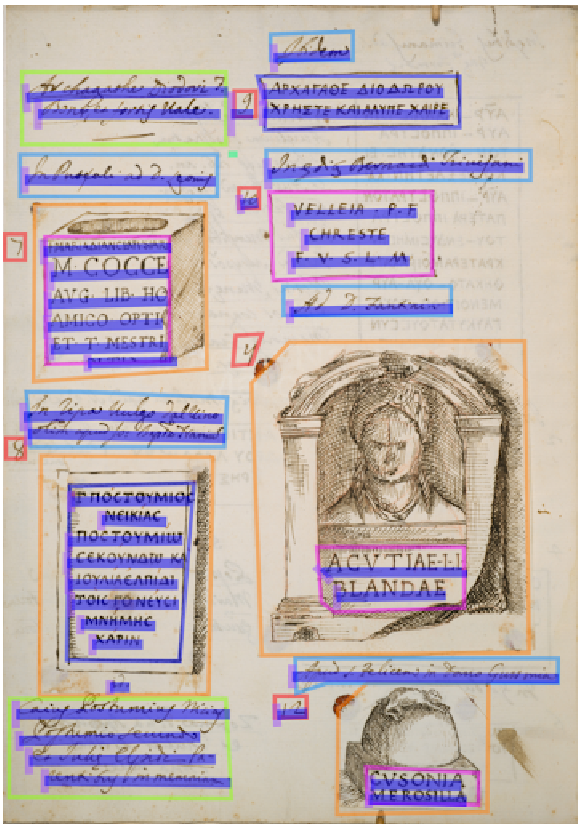


Figure 2: Venice, Marciana National Library, Marc. Lat. XIV, 200 (4336), f. 1v: baseline recognition. By concession of the Italian Ministry of Culture - Marciana National Library; reproduction is forbidden.

task was exclusively to identify the coordinates of each glyph on the facsimile (this aspect is relevant for digital palaeographical projects, such as DigiPal: see [Brookes et al., 2015]). Thus, the training set entirely corresponds to the text: the recognition approximates 100% of accuracy and, as the desired side effect, provides the coordinates of each glyph. Only fragmentary letters, such as the *vestigia* in the last line of the inscription number 7 (Figure 2),[29] must be treated by hand and represented by dotted characters or *lacunae*. The ALTO-XML file downloaded from the current version of eScriptorium, even if based on Kraken, contains only the coordinates of text lines and words, not of glyphs. Thus, for this operation we used Kraken through the command line on a local computer, outside the eScriptorium environment.

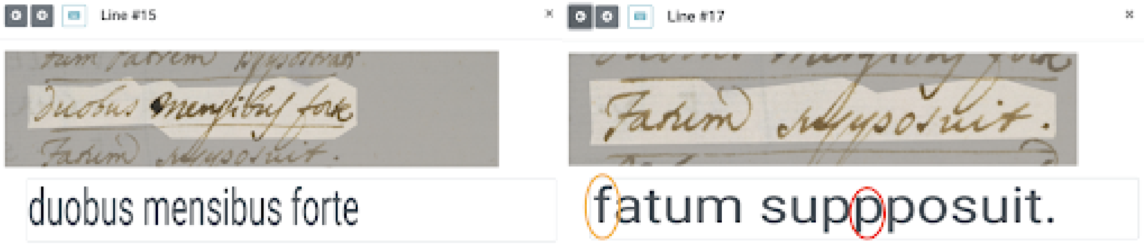
3.4 HTR Applied to Locations and Translations

Latin and Greek inscriptions were faithfully transcribed by Astori in capital letters. As a consequence, the variation among the glyphs is rather limited. On the other hand, Astori wrote his notes about the locations of the inscribed monuments, particularly useful for reconstructing their lifecycle, and the Latin translations of Greek inscriptions in cursive script, with a variety of glyphs, allographs and ligatures. Tatiana Tommasi also manually digitised Astori's notes and translations. As a proof of concept, we applied HTR to these parts in cursive. Due to the small amount of cursive text in Astori's epigraphic manuscript, we searched for other documents written by Astori, in a similar script. For this purpose, we used Astori's letters addressed to

[29] See Figure 2 in this page.

Muratori.[30] We acquired by OCR the transcription published by Di Campli and Forlani [1995, 285-291]. We used Tesseract[31] as the OCR engine and we edited the result, constituted by the interpretative printed edition of the letters, to obtain a faithful diplomatic transcription glyph by glyph, according to the following guidelines: a) maintenance of lower- and upper-case letters; b) maintenance of the original punctuation marks; c) maintenance of abbreviations. The transcription,[32] mapped on the facsimile line by line, was used for fine-tuning an existing model created by Chagué and Clérice [2022].[33] Figure 3[34] shows a text line accurately recognised in Astori's epigraphic manuscript and typical errors of recognition, to demonstrate that even 18 pages of cursive text (in this case from epistolary exchanges of the author) are enough to fine-tune a robust model provided by the HTR-United project.[35]

[Federico Boschetti]



duobus mensibus forte

fatum suppposuit.

Figure 3: Venice, Marciana National Library, Marc. Lat. XIV, 200 (4336), f. 1r: example of accurate recognition (on the left) and typical errors (on the right). The errors are highlighted using different colours: in orange the inexact rendering of an upper-case letter with the corresponding lower-case character; in red the erroneous addition of a letter. By concession of the Italian Ministry of Culture - Marciana National Library; reproduction is forbidden.

IV DATASETS AND MODELS

We have applied different recognition models according to the various parts of Astori's epigraphic manuscript.

For the Latin inscriptions, transcribed by Astori in upper-case letters, we have used the *Modèle imprimé 16-18e Fra+Lat*. The model comes from the CREMMA project and combines French and Latin training data (such as that in this repository: [Clérice, 2021]).

For the sections of the manuscript in Latin alphabet and in cursive script (locations of the epigraphic monuments and translations in Latin of the Greek inscriptions) we have produced a new

[30] Facsimiles are provided by https://www.internetculturale.it with the following identifiers: MO0089_A.M-37.10, MO0089_A.M-49.38, MO0089_A.M-52.01. For each of these sources the IIIF (International Image Interoperability Framework) manifest is available on the *Estense Digital Library* online platform (https://edl.cultura.gov.it/home/index.aspx): https://edl-jarvis.cultura.gov.it/meta/iiif/040bab16-3e0b-40ae-8fd5-284392af19b4/manifest, https://edl-jarvis.cultura.gov.it/meta/iiif/3781bd07-e7cb-43f4-97df-90dfcca1b846/manifest, https://edl-jarvis.cultura.gov.it/meta/iiif/80984208-a7a2-4494-b8e1-c9f99df5b854/manifest.

[31] https://github.com/tesseract-ocr/tesseract

[32] https://github.com/vedph/episearch-htr/blob/main/astori_letters.txt

[33] https://htr-united.github.io

[34] See Figure 3 in this page.

[35] The estimated accuracy is 89.7%. The percentage is not particularly high in general, but in our specific case it can be considered a quite good result. At the same time it demonstrates the potentialities offered by epistolary exchanges for studying epigraphic manuscripts through a digital approach.

HTR model by fine-tuning the pre-existing *HTR-United - Manu McFrench V1 (Manuscripts of Modern and Contemporaneous French) (1.0.0)*, created by Chagué and Clérice [2022]. The base for our model's fine-tuning was the dataset derived from the transcriptions of Astori's letters to Ludovico Antonio Muratori. Both our new model, indexed by HTR-United as *EpiSearch HTR* [Calvelli et al., 2023], and its dataset can be accessed online via our project's github repository.[36]

V CONCLUSION

Our analysis of Astori'codex gave us the opportunity to understand the possibilities and limits of the application of HTR to epigraphic manuscripts. The layout of the document is complex with texts inside drawings, different languages (Greek and Latin) and scripts both in cursive and upper-case. The small amount of written text in terms of quantity (14 leaves, corresponding to 17 written pages) has constituted a limit for the automatic acquisition of the transcription; in spite of that, we saw that HTR technologies can be used successfully to map legacy manual transcriptions on the manuscript facsimile and to improve the layout analysis of the document. These results will be used for one of the final goals of the EpiSearch project, that is the design of an Open Access web application to allow users to browse the manuscript, which will be visually annotated and connected to the main online resources for digital epigraphy.

As a proof of concept we have also tried to face the problem of the small amount of written text, using different techniques for the different parts of the manuscript that have been identified and marked manually. For the recognition of the transcription of the Latin inscriptions (CustomZone:latinInscription), realised in upper-case letters, we used the *Modèle imprimé 16-18e Fra+Lat* which gave quite good results. The Greek inscriptions (CustomZone:greekInscription) were more problematic. Therefore, we limited our scope to the identification of the coordinates of the words and the glyphs on the facsimile, using an overfitted model. The results will be useful for the future visual annotation of the manuscript. For the cursive parts in Latin (CustomZone:provenance and CustomZone:translation) we used other documents written by the same author in the same period: Astori's letters to Muratori, which were digitised at high resolution and accessible online via *Internet Culturale*. The transcriptions of Astori's letters in XML-ALTO format, the transcription of the epigraphic manuscript in TXT format, and the HTR recognition[37] model created for EpiSearch and indexed by HTR-United [Calvelli et al., 2023] are available on the github of the project.[38]

These experiments confirm that even small amounts of ground-truth transcriptions can be useful to refine large models available through HTR-United for specific needs. Within the EpiSearch project, we now plan to apply HTR to other epigraphic manuscripts.

[36] https://github.com/vedph/episearch-htr

[37] We did not create custom segmentation models for the project.

[38] https://github.com/vedph/episearch-htr

VI LIST OF ABBREVIATIONS

CIG = *Corpus inscriptionum Graecarum*, I-IV, Berlin 1828-1856.

CIL = *Corpus inscriptionum Latinarum*, Berlin 1862-

CNR-ILC = Institute for Computational Linguistics "A. Zampolli", National Research Council of Italy.

EAGLE = Electronic Archive of Greek and Latin Epigraphy: http://www.eagle-eagle.it

EDB = Epigraphic Database Bari: https://www.edb.uniba.it

EDCS = Epigraphik Datenbank Clauss-Slaby: http://www.manfredclauss.de

EDH = Epigraphische Datenbank Heidelberg: https://edh.ub.uni-heidelberg.de

EDR = Epigraphic Database Roma: http://www.edr-edr.it

IG = *Inscriptiones Graecae*, Berlin 1873-

ILS = H. Dessau, *Inscriptiones Latinae selectae*, Berlin 1892-1916.

InscrIt = *Inscriptiones Italiae*, Roma 1931-

Pais, *SupplIt* = E. Pais, *Corporis inscriptionum Latinarum supplementa Italica*, I, *Additamenta ad volumen V Galliae Cisalpinae*, Roma 1888.

PHI = Packard Humanities Institute: https://inscriptions.packhum.org

SEG = *Supplementum epigraphicum Graecum*, Leiden-Amsterdam 1923-

SupplIt = *Supplementa Italica*. Nuova serie, Roma 1981-

Ubi erat lupa = http://lupa.at

VeDPH = Venice Centre for Digital and Public Humanities, Ca' Foscari University of Venice: https://www.unive.it/pag/39287

References

G.A. Astori. Commentariolum in antiquum Alcmanis poetae Laconis monumentum. *La galleria di Minerva*, 2(5): 145–155, 1697.

F. Beltrán Lloris. The "Epigraphic Habit" in the Roman World. In C. Bruun and J. Edmondson, editors, *The Oxford Handbook of Roman Epigraphy*, pages 131–148. Oxford University Press, New York, 2015.

G. Bodon. Vicende di epigrafi greche tra Venezia e l'Europa attraverso la lettura di un codice marciano. In M. Fano Santi, editor, *Venezia, l'archeologia e l'Europa = Atti del Congresso Internazionale (Venezia, 27-30 giugno 1994)*, number 17 in Rivista di Archeologia. Supplementi, pages 34–38, Roma, 1996. G. Bretschneider.

Ph.J. Boyes, Ph.M. Steele, and N.E. Astoreca, editors. *The Social and Cultural Contexts of Historic Writing Practices*. Oxbow Books, Oxford – Philadelphia, 2021.

S. Brookes, P. Stokes, M. Watson, and D. Marques De Matos. The DigiPal Project for European Scripts and Decorations. In A. Conti, O. Da Rold, and P. Shaw, editors, *Writing Europe 500-1450. Texts and Contexts*, number 68 in Essays and Studies, pages 25–58. D.S. Brewer, Cambridge, 2015.

M. Buonocore. Epigraphic Research from Its Inception: The Contribution of Manuscripts. In C. Bruun and J. Edmondson, editors, *The Oxford Handbook of Roman Epigraphy*, pages 21–41. Oxford University Press, New York, 2015.

L. Calvelli. *CIL* V, 2262: un'epigrafe urbana da espungere dal *Corpus* di *Altinum*. *Aquileia Nostra*, 75: columns 429–456, 2004.

L. Calvelli. Iscrizioni esposte in contesti di reimpiego: l'esempio veneziano. In A. Donati, editor, *L'iscrizione esposta = Atti del Convegno Borghesi (2015)*, number 37 in Epigrafia e antichità, pages 457–490, Faenza, 2016. Fratelli Lega Editori.

L. Calvelli. "Li marmi segatti che incrostatto havevano li muri della chiesa vecchia". Il reimpiego di epigrafi di epoca romana nella cattedrale di San Pietro di Castello. In G. Guidarelli, M. Hochmann, and F. Tonizzi, editors, *La chiesa di San Pietro di Castello e la nascita del patriarcato di Venezia*, pages 87–109. Marcianum Press, Venezia, 2018.

L. Calvelli, G. Cresci Marrone, and A. Buonopane, editors. *"Altera pars laboris". Studi sulla tradizione manoscritta delle iscrizioni antiche*. Edizioni Ca' Foscari, Venezia, 2019. URL http://doi.org/10.30687/978-88-6969-374-8.

L. Calvelli, T. Tommasi, and F. Boschetti. EpiSearch HTR, 2023. URL https://htr-united.github.io/share.html?uri=c2cf58d8f.

C. Cappelletti. Biografia e autobiografia per lettera: l'epistolario Mazzuchelli come fonte degli "Scrittori d'Italia". In R. Bertazzoli, F. Forner, P. Pellegrini, and C. Viola, editors, *Studi per Gian Paolo Marchi*, pages 249–269. Edizioni ETS, Pisa, 2011.

S. Cascianelli, M. Cornia, L. Baraldi, M.L. Piazzi, R. Schiuma, and R. Cucchiara. Learning to Read *L'Infinito*: Handwritten Text Recognition with Synthetic Training Data. In N. Tsapatsoulis, A. Panayides, T. Theocharides, A. Lanitis, C. Pattichis, and M. Vento, editors, *Computer Analysis of Images and Patterns: 19th International Conference, CAIP 2021 (Virtual Event, September 28–30, 2021). Proceedings, Part II*, pages 340–350, Cham, 2021. Springer.

A. Chagué, T. Clérice, and L. Romary. HTR-United: Mutualisons la vérité de terrain! In *DHNord2021 - Publier, partager, réutiliser les données de la recherche: les data papers et leurs enjeux. (Lille, November 15-19, 2021)*, 2021. URL https://hal.science/hal-03398740.

A. Chagué and T. Clérice. HTR-United - Manu McFrench V1 (Manuscripts of Modern and Contemporaneous French), June 2022. URL https://doi.org/10.5281/zenodo.6657809.

T. Clérice. CREMMA 16-18 Prints, a French and Latin Dataset for HTR and Segmentation of Printed Books (Version 0.0.1), August 2021. URL https://doi.org/10.5281/zenodo.5235144.

A. De Santis and I. Rossi, editors. *Crossing Experiences in Digital Epigraphy: from Practice to Discipline*. De Gruyter, Warsaw – Berlin, 2018.

M.G. Di Campli and C. Forlani, editors. *Carteggi con Amenta ... Azzi*, volume 2 of *Edizione nazionale del carteggio di L. A. Muratori*. L.S. Olschki, Firenze, 1995.

I. Favaretto. *Arte antica e cultura antiquaria nelle collezioni venete al tempo della Serenissima*. Number 55 in Studia archaeologica. L'Erma di Bretschneider, Roma, 1990.

S. Ferrara. *La grande invenzione. Storia del mondo in nove scritture misteriose*. Feltrinelli, Milano, 2021.

S. Gabay, J.B. Camps, A. Pinche, and N. Carboni. SegmOnto, A Controlled Vocabulary to Describe the Layout of Pages, 2021. URL https://github.com/SegmOnto.

M. Grossi. ΕΓΡΑΨΕΝ ΔΕ ΚΑΙ ΤΙΤΛΟΝ Ο ΠΙΛΑΤΟΣ (Gv 19,19). Verso una nuova definizione di iscrizione. *Zeitschrift für Papyrologie und Epigraphik*, 197:85–95, 2016.

P. Kahle, S. Colutto, G. Hackl, and G. Mühlberger. Transkribus - A Service Platform for Transcription, Recognition and Retrieval of Historical Documents. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR) (Kyoto, November 9-15, 2017)*, volume 4, pages 19–24. IEEE Computer Society, 2017. doi: 10.1109/ICDAR.2017.307.

B. Kiessling, R. Tissot, P. Stokes, and D. Stökl Ben Ezra. eScriptorium: An Open Source Platform for Historical Document Analysis. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW) (Sydney, September 22-25, 2019)*, volume 2, page 19. IEEE Computer Society, 2019. doi: 10.1109/ICDARW.2019.10032.

R. MacMullen. The Epigraphic Habit in the Roman Empire. *The American Journal of Philology*, 103:233–246, 1982.

G.M. Mazzuchelli. Astori, Gio(vanni) Antonio. In *Gli scrittori d'Italia*, volume 1.2, pages 1191–1193. Giambattista Bossini, Brescia, 1753.

J. Morelli, editor. *Lettere di Apostolo Zeno*, volume 1. Francesco Sansoni, Venezia, 1785.

A. Mullen and A. Bowman. *Scripts and Texts*, volume 1 of *Manual of Roman Everyday Writing*. LatinNow ePubs, Nottingham, 2021.

J. Nockels, P. Gooding, S. Ames, and M. Terras. Understanding the Application of Handwritten Text Recognition Technology in Heritage Contexts: A Systematic Review of Transkribus in Published Research. *Archival Science*, 22(3):367–392, 2022.

S. Orlandi. Editing Ligorio's Epigraphic Manuscripts: New Discoveries and New Issues. In F. Loffredo and G. Vagenheim, editors, *Pirro Ligorio's Worlds. Antiquarianism, Classical Erudition and the Visual Arts in the Late Renaissance*, number 293 in Brill's Studies in Intellectual History, pages 39–50. Brill, Leiden – Boston, 2019.

S. Panciera. What Is an Inscription? Problems of Definition and Identity of an Historical Source. *Zeitschrift für Papyrologie und Epigraphik*, 183:1–10, 2012.

E. Perdiki. How to (Auto) Collate Big Manuscript Data with Minimal HTR Training. Preprint, 2022. URL https://hal.science/hal-03880102v1.

P. Pérez Galván. *Not Set in Stone: Epigraphy Between Manuscript and Print in Renaissance Europe, 1521-1603*. PhD thesis, University of Warwick, 2021. URL http://wrap.warwick.ac.uk/153492.

C. Reul, D. Christ, A. Hartelt, N. Balbach, M. Wehner, U. Springmann, C. Wick, C. Grundig, A. Büttner, and F. Puppe. OCR4all—An Open-Source Tool Providing a (Semi-)Automatic OCR Workflow for Historical Printings. *Applied Sciences*, 9(22):1–54, 2019.

M. Romanello, S. Najem-Meyer, and B. Robertson. Optical Character Recognition of 19th Century Classical Commentaries: the Current State of Affairs. In *HIP '21: Proceedings of the 6th International Workshop on Historical Document Imaging and Processing (Lausanne, September 6, 2021)*, pages 1–6, New York, 2021. Association for Computing Machinery.

W. Stenhouse. *Reading Inscriptions and Writing Ancient History. Historical Scholarship in the Late Renaissance*. Number 86 in Bulletin of the Institute of Classical Studies. Supplement. Institute of Classical Studies, London, 2005.

W. Thompson. Using Handwritten Text Recognition (HTR) Tools to Transcribe Historical Multilingual Lexica. *Scripta & e-Scripta*, 21:217–231, 2021.

A. Vallisneri. Breve relazione di quanto ha osservato nelle terme euganee Antonio de' Vallisnieri. *La galleria di Minerva*, 5(4):110–114, 1706.

I. Velázquez Soriano and D. Espinosa Espinosa, editors. *Epigraphy in the Digital Age. Opportunities and Challenges in the Recording, Analysis and Dissemination of Inscriptions*. Archaeopress, Oxford, 2021.

P. Zorzanello. *Catalogo dei codici latini della Biblioteca Nazionale Marciana di Venezia non compresi nel catalogo di G. Valentinelli*, volume 3. Etimar, Trezzano sul Naviglio, 1985.

M. Zorzi, editor. *Collezioni di antichità a Venezia nei secoli della Repubblica (dai libri e documenti della Biblioteca Marciana) = Catalogo della mostra (Venezia, 27 maggio – 31 luglio 1988)*. Istituto poligrafico e Zecca dello Stato. Libreria dello Stato, Roma, 1988.