Università Ca'Foscari Venezia

[ve]dph

Venice Centre for Digital and Public Humanities

MASTER'S DEGREE PROGRAMME IN DIGITAL AND PUBLIC HUMANITIES

Data modelling and visualisation of a digitised manuscript text enhanced through linked data: from text acquisition to data exploration

MASTER CANDIDATE

Giorgia Rubin

Student ID 887760

SUPERVISOR

Prof. Alessandra Melonio

Ca'Foscari University

CO-SUPERVISOR

Dott. Federico Boschetti

CNR-ILC "A. Zampolli"

ACADEMIC YEAR
2022/2023

Abstract

The digitisation of cultural heritage leads to multiple benefits, including the preservation of artworks, the in-depth computational analysis of their characteristics, and their accessible dissemination to a wide public with the application of standards that promote interoperability and reuse of the resources to which they are linked. The implementation of a digitisation project requires a methodical and articulated approach and the application of a workflow specific to the type of cultural object involved.
In the case of the digitisation of a text contained in a mediaeval manuscript, the digitisation paradigm consists of four basic steps: analysis and acquisition of the text (from source to transformation into machine-readable format); modelling of the collected data; standardisation and serialisation of the textual data; and visualisation of the processed data.
This thesis describes the digitisation paradigm applied to the text of the 4th book of the manuscript BnF ita. 590: the first prose vernacularisation of the Aeneid based on a Latin reduction of the ancient poem.
First the text was acquired from images of the manuscript by applying HTR techniques; in the second phase the text was encoded in XML format by applying the TEI guidelines; subsequently a semantic ontology was designed to represent the characteristics of the text and its contents using some of the ontological vocabularies most in use in the field of cultural heritage, enriching the encoded text with data linked to authority files thus expanding the knowledge base of the digitised text; of this ontology, a serialisation in RDF schema has been produced; all the data collected and modelled have been rendered graphically with interactive visualisations implemented from scratch and some appropriate visualisation tools; finally, is proposed the graphic design of a web environment in which to explore and analyse the collected and modelled data of the digitised text and its peculiarities.

Keywords: Manuscript Text Digitisation, Digitisation Paradigm, Text acquisition, Data Modelling, Ontology, Linked Open Data, Data Serialisation, Data Visualisation, Data Exploration

Contents

List of Figures

List of Code Snippets

Introduction

Benefits of digitising cultural heritage

The digital approach to the study of cultural heritage has brought positive changes: its transformation and dematerialisation from physical to digital objects has revolutionised the way we can access, analyse, share and interact with cultural heritage.
The digitisation[1] of works and their online publication in different formats make them accessible to a wide and varied audience. High-resolution images of works of art, conforming to the finest digital facsimile distribution protocols such as the IIIF[2] (International Image Interoperability Framework), can be displayed on museum websites and virtual galleries, enabling enthusiasts from all over the world to explore finely crafted details. The creation of digital twins of physical objects, such as archaeological artefacts, sculptures, fossil replicas or architectural buildings, allows experts to analyse data and information on such objects as never before, sharing these resources with other researchers and the interested public who, through online resources, will be able to explore the objects in their three-dimensional form. The digitisation of ancient manuscripts and their inclusion on specialised web platforms allow scholars worldwide to easily access these valuable documents without having to physically travel to consult them.

Cultural heritage in the web of data

The inclusion of these digitised cultural resources within the latest generation web structure, the semantic web[3] also known as the web of data, also enables the creation of semantic links between different sources, mutually enriching them with meaning. Among the many benefits of this good practice, there is the possibility for users to advance complex searches and serendipitous discoveries, deepened through the association of data. In this way, agile interdisciplinary connections are created that can describe the cultural entity analysed through a network of complex and varied relationships, offering the user a broad and global viewpoint on their object of interest.
Those who benefit from seeing resources placed on the web according to semantic models are not only culture enthusiasts or scholars, but also cultural institutions wishing to index their works and, by adopting infrastructures suited to this type of model, create interoperable catalogues rich in metadata, capable of offering relational search solutions between the various works contained in them. These institutions then take on the task of making the

[1] With the term digitisation we mean the process of transforming tangible artefacts into digital formats.
[2] IIIF (International Image Interoperability Framework) website: https://iiif.io/
[3] According to the W3C, "The Semantic Web provides a common framework that allows data to be shared and reused across application, enterprise, and community boundaries"

digitised catalogues available to all web users, respecting the FAIR[4] (Foundability, Accessibility, Interoperability and Reusability) principles.

GLAM in the web of data

Among the institutions in the GLAM (Galleries, Libraries. Archives, Museums) landscape that stand out for their outstanding achievements in digitisation are: the Prado Museum[5], the Bibliothèque Nationale de France (BnF)[6] and the italian Central Institute for Catalogue and Documentation[7].
Museums, libraries, digital archives and research institutions collaborate and cooperate to create the best infrastructures and the most efficient models to link and semantically enrich their collections.
At the European level, cooperation between these institutions is vibrant and a key role is played by the Europeana Foundation[8] which, as a digital library, aims to collect data from the various institutions containing Europe's cultural heritage, with the objective of disseminating cultural content to a wide public. By adopting the best metadata standards and a refined data model that can describe works of art through all the relevant metadata made available by the institutions that physically preserve them, Europeana catalyses and makes Europe's digitised cultural heritage accessible to all web users through a single online platform.

Digital scholarly editions

The practice of digitising works of art is not only common in large institutions but also in small-scale research projects. In the context of the philological and computational linguistic disciplines, we are observing the gradual emergence of a renewed methodology of data representation. Extensive work is being done on digital scholarly editions (DSE), which, according to the definition of DSE proposed by P. Shale (Shale, 2016, 19-39), are those scholarly editions that, in their theory, method and practice, adopt a digital paradigm, modifying the tradition and methodologies of scholarly editions and adapting them to the digital context. DSEs are useful not only for the interpretation of literary texts, but also for their transmission and dissemination. Some projects that can be mentioned in this context are the Canterbury Tales Project[9] edited by Peter Robinson and collatorators, the Digital Vercelli Book[10] edited by Roberto Rosselli del Turco and collaborators, the digital edition of Francesco da Barberino's Love Documents[11] edited by Tiziana Mancinelli et al.

---

[4] FAIR data principles: The FAIR Data Principles – FORCE11
[5] Prado Museum website: https://www.museodelprado.es/en/the-collection/art-works
[6] BnF website: https://www.bnf.fr/fr
[7] Istituto Centrale per il Catalogo e la Documentazione, website: http://www.iccd.beniculturali.it/
[8] Europeana website: https://www.europeana.eu/it
[9] Canterbury Tales Project, website: https://www.canterburytalesproject.org/
[10] Digital Vercelli Book, website: http://vbd.humnet.unipi.it/beta2/
[11] Information about digital scholarly edition of Francesco da Barberino's Love Documents: https://www.unive.it/pag/23956/

Data-centric digital scholarly editions

Nowadays, the most common structure adopted by online DSEs is the document-centric one, which involves the use of texts as documents, i.e. as information units quite similar to those usable in the analogue typographic environment.
There is, however, an alternative to this document-centric structure, which is that proposed by semantic web technologies, which have led to a rethinking of digital textuality, proposing a data-centric structure capable of enhancing the explicit and implicit relations conveyed by the text. The relationships and interconnections created between textual elements make it possible to move from information to knowledge of the text described through a knowledge graph[12].
These concepts, and those that will be discussed below to understand the necessary logics to be adopted for the purposes of modelling a data centric text for the creation of a data centric structured DSE, were formulated by Francesca Tomasi et. al in "Linked data for digital scientific editions. The publication workflow of the semantic edition of Paolo Bufalini's 78 notebook' (Tomasi, et. al., 2019). Among the projects that Tomasi has been involved in we mention The collection of Vespasiano da Bisticci's Letters[13] and Paolo Bufalini's Notebook[14].

Adopting the data-centric paradigm for the production of DSE means rethinking its structure no longer as a network of documents, but rather as a set of unambiguously identified textual entities interconnected through typed links proper to the RDF[15] structure.
The transformation from a document-centric to a data-centric structure implies a step of modelling the document that constitutes the edition, which is deconstructed to derive uniquely identified information units. From these units, a network of relations can be generated between the data forming part of the edition, thus creating the knowledge graph.
This transformation from the XML/TEI model (document-centric) to the RDF model (data-centric) consists of a few fundamental steps:

1. In the first step it is necessary to implement a meticulous analysis of the text and the information contained in it and its XML/TEI, which implies a view of the text from above, seen as a network of relations and not as a sequence of data. At this stage, it is necessary to distinguish real entities from their instances.
2. The second step consists of choosing the ontological model best suited to the representation of our text. At this stage, it is necessary to reflect on which ontological models to use and whether or not to create new classes and properties. The

[12] According to Ontotext, "The heart of the knowledge graph is a knowledge model: a collection of interlinked descriptions of concepts, entities, relationships and events. Knowledge graphs put data in context via linking and semantic metadata and this way provide a framework for data integration, unification, analytics and sharing."
[13] Digital edition of Vespasiano da Bisticci's Letters, website: https://projects.dharc.unibo.it/vespasiano/
[14] Digital Scholarly Edition of Paolo Bufalini's Notebook, website: http://projects.dharc.unibo.it/bufalini-notebook/
[15] According to Ontotext, "RDF is a standard for data interchange that is used for representing highly interconnected data. Each RDF statement is a three-part structure consisting of resources where every resource is identified by a URI. Representing data in RDF allows information to be easily identified, disambiguated and interconnected by AI systems."

ontological model obtained should be able to convey all the contents of the text, including the most latent ones.

3. Finally, links are established with the control systems of the authorities considering what connections are established and how to ensure a cohesive knowledge network. The need to link data to other valuable resources to improve the knowledge base and meet the different information needs of users is at the core of this phase of selecting relevant datasets and models.

This process enhances the digital text and provides a more enriched information experience for the end user, who can explore beyond the boundaries of the text.

Information visualisation: the importance of effectively representing data

In order to offer the user an effective experience and to best convey the information inherent in the processed knowledge base, it is crucial to adopt effective information visualisation strategies.
For this reason the study related to information visualisation and the choices made on the adoption of best tools plays a crucial role.

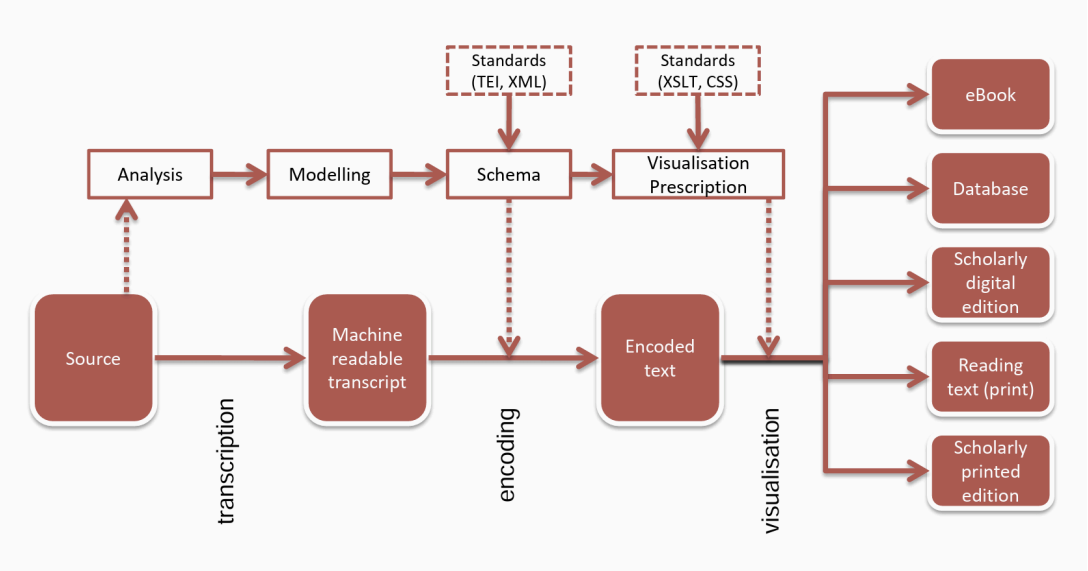
Paraphrasing the definition[16] of Information Visualization proposed by Kard, Mackinlay and Scheiderman, we intend information visualisation not as a mere aesthetic tool for the presentation and enjoyment of data, but as an effective communication and analysis methodology, capable of significantly improving the understanding and interpretation of complex information by simplifying it to the eyes of the user and thus making it more interesting. It allows scholars and the general public to visually examine and interpret data, which can lead to revelations and discoveries that might not be apparent in unprocessed data, bringing out latent information and helping the user to acquire new knowledge.
Through the use of successful data visualisation methods and complying with best practices, such as straightforward design, appropriate use of visual components and colours and accurate data representation, digital humanities projects can increase the accessibility and influence of their conclusions.

Thesis Aims and Methodology

This thesis reports the work conducted on the 4th book of the manuscript BnF ita. 590.
It illustrates the processes of modelling the text from unstructured data, resulting in semi-structured and structured data serialised and stored in various formats and their visualisation.
This is a substantial project, which covers all the phases of the digitisation paradigm (figure 1), starting with the analysis of the digitised document images and the acquisition of the handwritten text, through the encoding and ontological representation of the text enriched by

[16] "Information visualisation is the use of computer-supported, interactive, visual representations of abstract data to amplify cognition".
Stuart T. Kard, Jock D. Mackinlay, Ben Scheiderman in *Readings in Information Visualization Using Vision to Think,* 1999.

Linked Open Data, to the presentation of the modelled data through various interactive and explorable visualisations.



*(**Fig.1** Text digitisation paradigm)*

The phases followed in this project can be summarised as follows:

1. Text acquisition: starting with the high-resolution image of the analysed documents, handwritten text recognition (HTR) techniques were used to acquire the text contained in the manuscript and transcribe it into digital characters;
2. Text encoding: after transcribing the text, standard schemas such as XML/TEI were applied to give the text a hierarchical structure making it machine readable;
3. Ontological representation, from XML/TEI to RDF: an RDF (Resource Description Framework) schema ontology was designed that, by adopting some of the most relevant semantic vocabularies, enriched the contents of the analysed text with meaning; the ontology was then serialised in RDF/XML format, making its contents no longer only machine readable but also machine understandable;
4. Visualisation of information: text modelled and coded in semi-structured and structured formats was represented through text visualisation tools and graphical displays capable of presenting the described content in a more immediate, in-depth and effective manner.
5. Design of a web interface: The interface of a web environment was designed to offer readers the text and visualisations produced. It is an environment in which users can explore content in an engaging system, and experience a digitised manuscript text, which in paper format they would only be able to read in a place that would probably be difficult to reach.

Outline

The thesis is organised as follows. Chapter 1 outlines an analysis of the literary, philological and archival context in which the text under study is set. Furthermore, the choice of this text is justified by providing the context of the early stages of the birth of this thesis project. Chapter 2 contextualises and describes the methodology used for the acquisition of the text and its encoding in XML/TEI and the structure in table format of some of the extracted data. Chapter 3 contextualises and describes the methodology used for modelling a textual content knowledge graph, enriched with semantic vocabularies and returned in the form of RDF triples. Chapter 4 describes the tools used to visualise the information and data extracted from the text, with the aim of proposing interactive visualisations. Chapter 5 shows the prototyped interface design of an environment in which to insert the text and visualisations produced.

Chapter 1
Source: manuscript BnF italien 590

1.1
About Virgil's Aeneid

Classical literature is a treasure of creativity and human reflection that has stood the test of time, profoundly shaping Western culture and thought.
The ancient texts that comprise it, born from time immemorial, come down to us as precious remains of our cultural heritage and represent an extraordinary poetic and ideological legacy that has shaped the collective cultural imagination and ideology of successive generations of readers and writers.
If we think of European classical literature, we cannot avoid thinking of the latin literature and one of its greatest representatives: Publius Virgilius Maron, known simply as Virgil, who stands as a cornerstone in the history of European literature and culture, and his influence has been particularly significant in the Italian sphere.
Among his best known and most important works is the Aeneid, an epic poem composed by the author in the 1st century BC. An epic tale of travel, heroism, destiny, and the founding of a great nation, steeped in complexity and depth, it offers an open window into the challenges and aspirations of ancient Rome, as well as its mythology and worldview.
The tale narrates the exploits of Aeneas, a figure already present in Greek and Roman legends and mythology, who is presented in the Virgilian poem as the Trojan hero.
Aeneas, having escaped from the fall of Troy, sails across the Mediterranean to Latium, thus becoming the progenitor of the Roman people. Virgil's Aeneid constitutes an effective "foundation myth," as well as a national epic that links Rome with Homeric legends, exalts traditional Roman values, and legitimises the Julio-Claudian dynasty as a descendant of the common founders, that is, the heroes and gods of Rome and Troy.
This monumental work, consisting of twelve books, is a priceless treasure trove of wisdom and knowledge ranging from rhetoric to symbolism, from explorations of human nature to meditations on history and society.
It has aroused the admiration and inspiration of generations of readers, scholars and artists throughout the centuries, coming down to us through numerous ancient manuscripts, a network of multiple translations of the text made over the centuries.

We will now adopt some of the reasoning expressed by V. Ricotta and G. Vaccaro in their “Rivolgarizzare e ritradurre. Parole, idee, traduzioni” (Ricotta and Vaccaro, 2018, 133-143) to briefly discuss some technical-philological terms that will help us better understand our subject of study.
By multiple translations we mean independent translations of the same text, thus excluding

from analysis that vast theory of remakes, rehashes and rewritings, which characterise the vernacular[17] as a 'genre'.
Multiple translations represent different translation editions of the same text by different translators. This can occur when a literary work is transposed into multiple languages or when a translation is modified or updated over time. Multiple translations may offer different perspectives, subjective interpretations and cultural adaptations specific to the same literary work.
On the other hand, vernaculars constitute a particular kind of transposition or adaptation of a source text into a vernacular or common language, rather than the more cultured or literary original. Such versions may involve alterations, simplifications or adaptations to make them usable to a wider audience or to comply with specific cultural or linguistic conventions: for the mediaeval learned copyist, in fact, it is "meritorious to correct, to compare with other exemplars and to contaminate, to enrich with his own quizzes or those drawn from elsewhere, but also, in the case of vulgarisation, to remake the text by modernising it, making it more fluent and clearer, retranslating, when given the ability, the opportunity and the desire" (G. Tanturli, 1986, 849).

One of the most interesting groups of vernaculars in the field of multiple translations is precisely that of the *Aeneid* (Ricotta, Vaccaro, 2017, 136).
The first among these is the first Florentine language vulgarisation of the poem, which is attributed to Andrea Lancia in two manuscripts and dates from the last years of the 14th century. This version has a very concise content that in the prologue is attributed not to the translator but to a mysterious author who created the Latin text, namely Friar Anastasius[18]:

---

[17] Vernaculars from the classics can be investigated thanks to a series of tools that can be consulted online (free of charge), produced within the DiVo - Dizionario dei Volgarizzamenti project, directed at the Opera del Vocabolario Italiano-CNR in Florence and the Scuola Normale Superiore in Pisa by Elisa Guadagnini and Giulio Vaccaro: DiVo DB, a philological bibliography of vernacular texts (DiVo DB: http://tlion.sns.it/divo)

[18] Since it is not the intention of the author of this dissertation to investigate and discuss the precise attribution of the manuscript under consideration, an in-depth note of it extracted from the Treccani encyclopaedia under the entry "Anastasius," is given here:
"Anastasius, a friar minor, perhaps of the Florentine convent of Santa Croce, who lived in the first half of the 14th century, compiled a Latin reduction of the Aeneid, on which Ser Andrea Lancia is said to have carried out the first prose vulgarisation of the ancient poem. Nothing else is known of his life, although he plays a role of no small interest in the Virgilian and Italian vernacular traditions, and the news itself about his work is uncertain and not always in agreement. The subscriptions and notes of the codices, the only source to which it is possible to resort, while assigning the reduction to Friar Anastasius, "discreet and literate man" (the name corrupted into Athanasius in the editio princeps and then in a whole erudite tradition, up to Segre), show gaps and diversity in the attribution of the vernacular: to Ser Andrea Lancia in some, but in others, and in greater numbers. to A. himself. It is true that the division of labour between A. and Andrea, noted on a codex by Benci and later validated by Parodi's authority, was accepted without discussion (De Marinis alone leaves it more objectively in suspended), but all the material should be re-examined in order to arrive at more secure proposals and conclusions. Incidentally, no one seems to have noticed a note by Sbaraglia, which seems to identify A. in a friar Anastasio Masini, still alive in Florence in 1389. Thus being the case, and since the Latin translation has not been traced, all inquiry and interest have turned to the Lancia, and from his biography and activities we must infer what may refer to Anastasio. The Lance vulgarisation is dated, on the basis of a chronological note in a codex and a quotation from Villani, to the years 1314-1316, within which and in the same cultural milieu A. is also supposed to have operated; indeed, the two works both seem to have been dedicated to Coppo di Borghese Migliorati, prior in Florence for numerous times in the first half of the fourteenth century. The vulgarisation was soon successful, and copies multiplied: as many as fifteen codices have in fact remained to us, and the editio princeps edited by Levilapide in Vicenza in 1476 was followed by numerous others up to that of Fanfani; but in the same fourteenth century, Angelo di Capua took Lancia's text as the basis for his translation into the Sicilian

"Il quale libro frate Anastasio dell'ordine de frati minori, huomo discreto et litterato cum multa fatica recoduisi in ipsa lassando certa parte sencza laquale li parue che questo libro suffitientemente potesse stare. Io Anastasio poi ad instancia dite et non multo lievemente translatai di grammatica in lingua volgare."

(BnF ita. 590, f. 3v)

This text was transmitted by thirty witnesses, even reaching print in later centuries and it is precisely from the analysis of a portion of one of these thirty witnesses, the fourth book of the manuscript BnF ita. 590, that the project of this thesis was born, leading to the production of the digitisation and visualisation of the text and its content.

1.2
Gallica digital library, HTRòmance project and ms. Bnf ita. 590

The manuscript BnF ita. 590 is preserved in the Italian manuscripts department of the Bibliothèque Nationale de France (BnF) and accessible online since 2011 thanks to the Gallica Digital Library[19].

Gallica is the digital library of the BnF and its national partners. Available online since 1997, it offers public access to millions of digitised archive documents with Creative Commons Attribution-NonCommercial-ShareAlike (CC BY-NC-SA) licence[20].

It is an international cooperation project that contributes to the feeding of other digital libraries through not only European but also intercontinental collaborations.

Free and constantly expanding, Gallica contains more than 6 million documents, including 144,859 manuscripts[21]. Each document held in this library has descriptive and administrative metadata[22] including a numeric identification code, called ARK (Archival Resource Key)[23].

---

vernacular. And it is such vulgate knowledge a reason that may perhaps account for the disappearance of the Latin text. As far as can be understood from this indirect tradition and as far, therefore, as can be imputed to him, A.'s work would appear to be that of a not ungainly reducer, omitting essential parts of the narrative, even from the point of view of the narrating events, in short, "a hundred or so Virgilian passages connected often badly and roughly"; but he had the merit or the good fortune to leave the Virgilian verse almost intact, "so that, fortunately for him abundant portions of the original stand before Ser Andrea, who renders them in a florid style, studying to preserve all the transpositions and ornaments of the Latin verse" (Folena). In other words, he proposed and formed the basis, perhaps to himself, for one of the most interesting Florentine vulgarizations of classical authors". Claudio Leonardi - Dizionario Biografico degli Italiani - Volume 3, 1961

[19] Reference to Gallica's website: https://gallica.bnf.fr/accueil/en/content/accueil-en?mode=desktop and the declarations of Gallica propos: https://gallica.bnf.fr/edit/und/a-propos

[20] CC BY-NC-SA description: https://creativecommons.org/licenses/by-nc-sa/4.0/deed.en

[21] Updated to June 2023

[22] Descriptive metadata provides information needed for the identification, authentication and description of digital objects. This typology includes title, author, date, place, keywords, cataloguing records, curatorial information, and details about the original.

Administrative metadata helps to manage the resource or the collection of which it is part, like resource type, location information, information about the digital acquisition, legal permissions, licences and the digitization project.

[23] BnF italien 590 ARK:/12148/btv1b8433319z, https://gallica.bnf.fr/ark:/12148/btv1b8433319z

Through this platform, if the BnF partner library has its own digital library, document metadata can be indexed by the BnF and referenced on Gallica. Internet users are referred to the partners' websites to consult these documents. As a result, several hundred thousand documents from over 90 partner libraries are referenced in Gallica.
These resources, mostly in French and free of copyright, offer a wide variety of media (books, journals, newspapers, scores, prints, maps, photographs, sound recordings) and range from antiquity to the first half of the 20th century, with a strong presence of documents published in the 19th century.
Regarding the publication of images of digitised resources, the Gallica Digital Library, uses IIIF (International Image Interoperability Framework)[24] standards to publish high-quality images and their manifest[25].

The collaborations that the BnF carries out through the Gallica project not only involve other libraries but also specialised research centres. It is in this collaborative context that ambitious digitisation programmes are organised around significant corpora, enabling the growth of digital collections.
In recent times, some of the archived documents were subject to optical character recognition (OCR) processes, dedicated to detecting the characters contained in the documents and transferring them into machine-readable digital text, offering an increasing number of documents in both image and text format.
Among these digitisation and OCR projects, we mention HTRomance[26].
A project born out of the collaboration between the BnF and the French research centre INRIA, it focuses on the field of handwriting recognition (HTR). In particular, it aims to evaluate and improve the capabilities of this technology applied to literary manuscripts and public and private archives, in Latin and Romance languages, from the 11th to the 19th century, preserved at the BnF. To this end, the project plans to produce training data and transcription models that are resistant to changes of hand and even language. It also intends to produce language models applicable to documents in ancient languages or ancient language states. The choice of the text corpus is guided by the need to diversify the cursive scripts (hands) and themes of the documents.
Within the corpus of texts chosen for the training of automatic transcription models we can also find some pages from the manuscript BnF ita 590, more precisely those belonging to the fourth book[27].

It is in this latter context that we find the roots of this thesis project.
It is the result of a long process born from the work carried out by the author of this thesis during her curricular internship, which saw her involved in supervising the training phases of

[24] IIIF is a set of open standards for delivering high-quality, attributed digital objects online at scale. It's also an international community developing and implementing the IIIF APIs. IIIF is backed by a consortium of leading cultural institutions. IIIF website: https://iiif.io/
[25] Reference to the declaration of IIIF uses of Gallica: https://iiif.io/guides/guides/gallica.bnf.fr/
[26] Presentation of HTRomance project proposed in the BnF website: https://www.bnf.fr/fr/les-projets-de-recherche-bnf-datalab
[27] The pages are the following (f = folio; r = recto; v = verso): f39_16r; f40_16v; f41_17r; f42_17v; f43_18r; f44_18v; f45_19r; f46_19v; f47_20r.

the HTR (handwritten text recognition) models and the production of corrections to the digital text produced by the process of text recognition and automatic digital transcription of the text.

1.3
Aeneid IVth book: Critical Edition vs. BnF ita. 590

The IVth book of the Aeneid is conceived as a digression from the main project on which the work stands. In this regard, M. Fernandelli expresses himself as follows: "The new 'sentimental' character of Virgil's epic poetry is manifested in a particular way in the IVth book of the Aeneid. The reader is involved in the events so that they acquire for him a 'tragic' value that transcends the needs of epic narration. The IVth book also has a certain autonomy from history, being strongly oriented towards a tragic climax, in which the correlation between sentimental characters, author and reader reaches its apex. This mixture of energies guarantees the individual 'plasticity' of the Virgilian text, a text that is in itself clear and active, and demands a 'mimetic' rather than a coldly intellectual reading" (Fernandelli, 2003).
Pursuing the destiny imposed on him by the gods, exiled from Troy in ruins, Aeneas finds himself temporarily welcomed in Carthage. In books II and III, he recounts the events that led him there; it is here that the complex bond between him and Dido, Queen of Carthage, begins to develop, based on the memory of past events in which the latter had not taken part.
Although essentially a digression, book IVth is instrumental to the unity of the work and the development of the protagonist's character: it is here that Aeneas's pietas, his defining characteristic, is subjected to a difficult examination.
The text contains the tormented story of the love affair between Aeneas and Dido. Born almost as a joke at the behest of the goddess Venus, Aeneas's mother, who simply wanted to make her son's stay in Carthage as safe as possible, this relationship quickly turns into a real tragedy. Aeneas has in fact been forced by the will of fate to leave for Latium, and for this reason Jupiter sends his messenger Mercury to lead the hero back to the destiny he has almost forgotten. But in so doing he turns Aeneas into an unwitting executioner, deaf to all the increasingly heartfelt pleas that Dido, the queen of Carthage, addresses to him. Thus opens the last phase of this relationship. At first the queen thinks only of revenge, quickly moving on to regret for not having killed the Trojans when it was still possible. Dido therefore conjures up several avenging gods and casts a terrible curse on Aeneas himself and his descendants. Finally, distraught with hatred, Dido commits suicide. This last episode concludes book IVth of the Virgilian poem.

We find these themes in the critical editions of Virgil's work as well as in Andrea Lancia's manuscript BnF ita. 590. Since, however, the latter is a translation of a reduction that, as Lancia himself says, lacks some portions of the text: "*Il quale libro frate Anastasio ... lassando certa parte sencza laquale li parue che questo libro suffitientemente potesse stare*...", it is natural to consider comparing the edition of our manuscript with a critical edition of the Virgilian text.
For this purpose, the critical edition edited by M. Geymonat (2008) was used, edited in the

digital edition[28] by M. Gioseffi and I. Canetta (2009) and made available online in the digital archive of Latin poetry "Musisque Deoque" (MQDQ)[29].
This operation of comparing the two texts required a careful analysis of their contents, which first led to the elaboration of a synopsis of the contents of the critical edition book.
For this operation, the synoptic segmentation found in the Italian wikipedia page on the entity "Eneide"[30] was used, carefully verifying its accuracy in the text of the critical edition. The most concise titles were then chosen that could effectively summarise the thematic contents of the segmented paragraphs in a few words.The analysis then developed with the alignment of the verses of the critical edition with the lines of our prose manuscript text, so as to facilitate the parallel reading of the two texts, leading then to the elaboration of the synopsis of the manuscript text and the schematic formulation of the congruences and incongruities between the contents of the two texts, emphasising the reduction operations that Friar Anastasius originally carried out in transcribing the Virgilian text (figure 2).

THEME	MQDQ Critical Edition (from-to)	Ms. BnF ita. 590 (from-to)	CT	ST	MT
1° Burning love of Dido for Aeneas	1-55	1.1.1-1.2.5	x	x	x
2° Offering to the Gods by Dido	56-73	1.2.5-1.2.8	x		x
3° Evening meal	74-79	/			x
4° First night fall	80-89	/			x
5° Juno and Venus plan to unite lovers	90-128	2.3.1-2.3.9	x	x	x
6° Sunrise	129-159	/			x
7° Dido and Aeneas take refuge in the cave	160-172	2.3.10-2.3.12	x	x	
8° Fame flies from city to city causing chaos	173-244	2.3.10-2.3.12	x	x	x
9° Mercury admonishes Aeneas	245-278	3.5.10-3.5.15	x		x
10° Preparation of Aeneas to leave Carthage	279-295	3.6.1-3.6.2	x	x	x
11° Omen of Dido	296-392	3.6.3-5.8.10	x	x	x

[28] Reference to the digital edition of M. Geymonat's critical edition of Virgil's Aeneid: https://www.mqdq.it/texts/VERG|aene|001

[29] Reference to the MQDQ digital archive: https://mizar.unive.it/mqdq/public/
[30] Italian wikipedia page on the entity "Eneide" https://it.wikipedia.org/wiki/Eneide

12° Aeneas comes back to his mates to conclude the preparations	393-407	5.9.1-5.9.3	x		x
13° Dido sends Anna to Aeneas to convince him to stay	408-449	5.9.4-5.9.11	x	x	x
14° Dido prepares herself to die	450-521	5.10.1-6.11.7	x	x	x
15° Night fall	522-553	6.11.8-7.11.14	x	x	x
16° Mercury orders Aeneas to definitely leave Carthage	554-583	7.12.1-7.12.11	x		
17° Suicide of Dido	584-671	7.13.1-9.13.32	x	x	x
18° Iris cuts the hair of Dido leaving her to die	672-705	9.13.33-9.13.43	x		x

*(**Fig. 2** Comparison of synopses of the IVth Book of the Aeneid: MQDQ Critical Edition vs. Ms. BnF ita. 590. The numbers for the BnF manuscript lines are to be readed as page.paragraph.line.*
CT = Corresponding Text; ST = Synthesised Text; MT = Missing Text)

1.4
How to put in evidence the peculiarities of BnF ita. 590 IVth book

The operation of aligning and comparing the two texts has been useful in order to carry out an in-depth analysis of the themes treated within the manuscript text and to correctly trace the differences between this text and that of the critical edition. The objective is to formulate a digital tool for reading the two aligned texts that would allow the reader to easily compare the two editions noting those parts of the text that are congruent, those that are synthesised in the manuscript reduction, and those that are totally absent compared to the "original text".
In fact, this thesis project provides the user-reader with visualisation tools capable of emphasising the peculiarities of the manuscript text that emerge from the analysis of its contents and the data extracted from the text.
It was indeed decided to consider some particularly interesting data from the text, which, given its narrative nature, turned out to be the characters and places mentioned, in order to produce an ontological model capable of describing the social relations between these entities. This ontology also describes their presence within the text by showing in which synoptic themes they are mentioned, based on the synopsis described above.
Moreover, thanks to this ontology, these entities have been linked to external authority files on the web to provide additional information on each of them.
In support of a more complete and explicit presentation of these text peculiarities, it was chosen to propose explorative and interactive data visualisations that would allow the reader to immediately understand what are the contents covered.

Chapter 2
Text acquisition and data modelling

In this second chapter, we define and explain the data modelling techniques and methodologies adopted in the early stages of project production. In particular, we refer to the digitisation paradigm steps required to transform the source data in unstructured format, i.e. the JPEG images of the manuscript, into semi-structured data in XML/TEI format and structured data in table format.
It will then be analysed specifically:

- HTR techniques and useful tools for automatic text acquisition from images;
- The restitution of the text in XML format and the integration of the TEI, with the adoption of the most suitable attributes for our text in accordance with the criteria indicated by the TEI consortium guidelines;
- The use of the Domain Specific Language tool Euporia for the integration of some additional content useful also for the next phase of semantic enrichment of our modelled text;
- The creation of tables in XLSX format to contain certain quantitative data relating to the contents of the analysed text.

Before we go into an in-depth analysis of these work steps, however, it is good to introduce some fundamental theoretical concepts in order to better frame the context and understand what is meant by data modelling, specifically when working with texts.

2.1
What does data modelling mean?

In computer science, modelling data means organising and representing information in structured formats to be easily managed and stored through computational techniques.
For A. Silberschatz et. al., data modelling is "a set of conceptual tools for describing data, relationships between data, data semantics and consistency constraints" (Silberschatz et al., 1996).

In the digital humanities, data modelling research has focused on the best techniques and strategies for expressing the properties and peculiarities of cultural objects, real ones rather than their digital twins, or digital native objects: how these are composed, organised and related to each other.
There are many activities in the digital humanities that require a data modelling process: for instance, the creation of digital editions using hierarchical text encoding schemes, the creation of databases to report characteristic aspects of the subject studied, the creation of relational models to create connections from the resources.

We can therefore understand data modelling as a process of abstraction that finds its beginning in the approach to the object of study and ends with formal, abstract descriptions of it.
Three levels of data models can be distinguished, often also seen as the three stages of the modelling process (Flanders, Jannidis, 2015, 230):

1. Conceptual data modelling: involves identifying the entities to be represented and their characteristics, their relationships with other entities and notation of the results in an entity-relationship diagram. An example of conceptual data models are semantic ontologies structured through Resource Description Framework (RDF).

2. Logical data modelling: focuses on the translation of the conceptual model into a formal abstract data model. Through a normalisation process, it provides a more structured representation of entities and their relationships, eliminating redundancies and ambiguities.

3. Physical data modelling: this involves the optimisation of the database for data collection in its development phase.

To summarise, the conceptual model is concerned with describing the object of study and its peculiarities, identifying the entities and their recurrences, the links and relationships that are present between these entities, taking care to maintain a high level of semantic description, capable of preserving the relational characteristics present in the object of study.
The logical model, unlike the conceptual model, achieves a very high level of abstraction, to the detriment of semantics, which is thus lacking, but allows the information expressed by the conceptual model to be synthesised in a functional manner so that it can be inserted within DBMS (database management systems), enabling the user to use algorithms and computational techniques to work with the data described.

It is now clear therefore that data modelling means organising and representing information in a structured format that can be easily understood and manipulated by a computer system. The main objective of data modelling is to create a data representation that accurately reflects the reality of the domain of interest and efficiently supports data management and analysis operations.

2.2
Categories of data

Modelling processes are applied to various types of data formats. These can be divided into three main categories, each of which has unique characteristics that influence the way the data is modelled:

- Unstructured data: examples of unstructured data include free text, images, audio and video. They do not follow a predefined pattern and are not organised in a specific way. To extract, process and analyse this information in a meaningful way is necessary to use some specialised techniques;
- Semi-structured data: they have a partial or implicit structure, without following a rigid schema. For example, XML documents and JSON files are considered semi-structured. To model them, it is necessary to define a structure that allows the relevant information to be extracted from these formats efficiently;
- Structured data: follow a rigid schema and are organised in a tabular or relational manner. They follow a defined schema, with well-defined rows and columns. To model structured data, tables, relationships and constraints are defined that define the structure and relationships between the data.

Also in digital philology, one of the research activities is the construction of representative models of the object of study, which, as in our case, are textual data.
The text, as a complex system, can be broken down into characters, selected and reorganised into a coherent and formal data model. In this sense, modelling means precisely identifying the components of the text, translating them and transposing them into metadata (Mancinelli, 2021, 71).

2.2.1
Brief overview of some relevant semi-structured data format

XML for TEI and RDF serialisation

In order to reorganise the text through metadata, it is necessary to use markup languages.
A markup language is a text-encoding system, a set of rules defining what markup information may be included in a document and how it is combined with the content of the document in a way to facilitate use by humans and machines and to easily retrieve and display information through automated processing.
W3C[31] (World Wide Web Consortium) has formalised some of the mostly used markup language format that are HTML (Hypertext Markup Language) for data display and XML (Extensible Markup Language) for storing and describe data. The latter has the main purpose in describe textual data for different human languages structurally organising information through schema that defines the necessary metadata for interpreting the contained information.
It is possible to assign grammar to XML through the use of controlled vocabulary. One of the most widely used in the field of digital philology is the TEI[32] standard (Text Encoding Initiative). The TEI guidelines formalise an XML language in which the schema is modelled using a system called ODD (One Document Does it all). The ODD format is a document that contains XML schema fragments and their associated documentation. It also includes

[31] W3C website: https://www.w3.org/
[32] TEI website: https://tei-c.org/

mechanisms to express specific choices and constraints. In the case of TEI, these vocabularies are used to arrive at a formalisation that helps normalise the modalities, criteria and vocabulary of the markup, taking into account the natural language polysemy inherent in literary texts (Mancinelli, 2021, 84).
XML is also widely used to serialise information contained into conceptual models such as Resource Description Framework (RDF). This is used to represent data models in a semantic web context. In RDF, information is organised in the form of triples, consisting of subjects, predicates and objects, which are linked together to form knowledge graphs, enabling a rich and detailed description of information. Serialising these data into RDF/XML format allows to represent data in a readable manner both for machines and humans, facilitating their interpretation and exchange among systems and applications.

JSON for IIIF manifest description

Another markup language used in alternative to XML to describe documents to be shared on the web is the JSON (JavaScript Object Notation) format. This is a data exchange format based on the JavaScript programming language and is widely used due to its easy-to-read and write structure. To describe structured data in the semantic web, JSON adds the possibility of incorporating contexts and links between the resources described through the JSON-LD (JSON Linked Data) syntax. Through this framework, it is possible to describe IIIF (International Image Interoperability Framework) manifests that specify the representation of image metadata. IIIF is a particular framework that originates in a bibliotheconomic context, and aims at the dissemination of images and related data preserved within institutions and their reuse respecting the FAIR principles. The IIIF was created in order to overcome problems related to management, reuse and visualisation of digital images. It consists of a working environment composed of a series of APIs[33] (Application Programming Interfaces) that manage the architecture of image reuse information, ranging from server components to display and search (Mancinelli, 2021, 78), allowing the integration, annotation and exchange of images and their metadata through the web.

Now that we have clarified the meaning of data modelling and listed as examples some of the most widely used textual and graphical standardised data format, we can proceed to the explanation of the techniques and work steps that were adopted for the modelling and transformation of various formats of the data of our text.

2.3
HTR technique: from unstructured (JPG) to semi-structured (XML) data

As previously mentioned, the first of the digitization phases that were worked on for this project involved the automatic acquisition of the text of the 4th book of the manuscript BnF

[33] IIIF APIs: https://iiif.io/api/

ita. 590, from JPEG images provided by the Gallica Digital Library.
Let now briefly explain what is meant by automatic text acquisition through OCR and HTR techniques.
OCR (Optical Character Recognition) and HTR (Handwritten Text Recognition) systems belong to the research field of artificial intelligence, more precisely image and pattern recognition. These are techniques for the detection of characters contained in documents (printed if OCR, handwritten if HTR), and their transfer into digital and machine-readable format. The conversion is returned in ASCII[34] or Unicode[35] characters that can be easily edited through text editors. The latter is a standard that provides an encoding system to represent text characters from almost every world languages. Unicode assigns a unique number, called a code point, to each character, enabling the representation of a wide range of symbols, alphabets and special characters.

OCR and HTR systems require a training period to work properly. This is a crucial training phase if we consider that text analysis elements are neural networks and therefore require training to function.
During this phase, the system receives sample images with corresponding text in ASCII or similar format so that the algorithms can be calibrated to the text they typically analyse.
The latest OCR and HTR software use algorithms that recognise edges and are able to reproduce not only the text but also the formatting of the page.
Text capture and automatic transcription of handwritten and printed documents has long been of interest to the world of research within cultural institutions. The recent increase in IT capabilities and advances in the field of Artificial Intelligence have given rise to new perspectives.

While research campaigns have been implemented for OCR since the 2000s to manage printing, for manuscripts and HTR it is only since 2010 that software available online such as Transkribus[36] (subscription-based) and eScriptorium[37] (open source) began to emerge (Chague, 2022).
The eScriptorium software was used for this thesis project. This is an environment that serves as a workspace for processing and managing the essential phases of a transcription campaign. These phases include:

1. Upload images of the manuscript (it is also possible to extract them from IIIF servers or PDF files),
2. Analysing the layout by identifying groups of text and lines to which a particular tag can be assigned,
3. Finally, transcribe.

The last two steps can be performed using Kraken, an open source automatic transcription software.

[34] ASCII definition by Wikipedia: https://en.wikipedia.org/wiki/ASCII
[35] Unicode definition by Wikipedia: https://it.wikipedia.org/wiki/Unicode
[36] Transkribus website: https://readcoop.eu/transkribus/
[37] eScriptorium website: https://www.sofer.info/

At the end of the process, the triplet consisting of an image, the coordinates of the text lines and their transcription can be exported to standard formats such as XML ALTO and PAGE, to create e.g. digital editions as in our case.
These triplets can also be used to create templates with Kraken. The templates compose an abstract representation of the information Kraken has learnt from the transcription examples, enabling the software to recognise text elements and generate transcripts from the image analysis (Chague, 2022).

The Kraken model that, during the early stages of the thesis project, was used and trained on the eScriptorium platform to transcribe the pages of the BnF ita. 590 manuscript is the CREMMA Medii Aevi model[38].
However, before applying the CREMMA Medii Aevi model and transcribing the text, the images loaded into eScriptorium were segmented to delineate the portions of text, i.e. the rows and columns, by assigning them specific tags following the guidelines offered by SegmOnto[39].

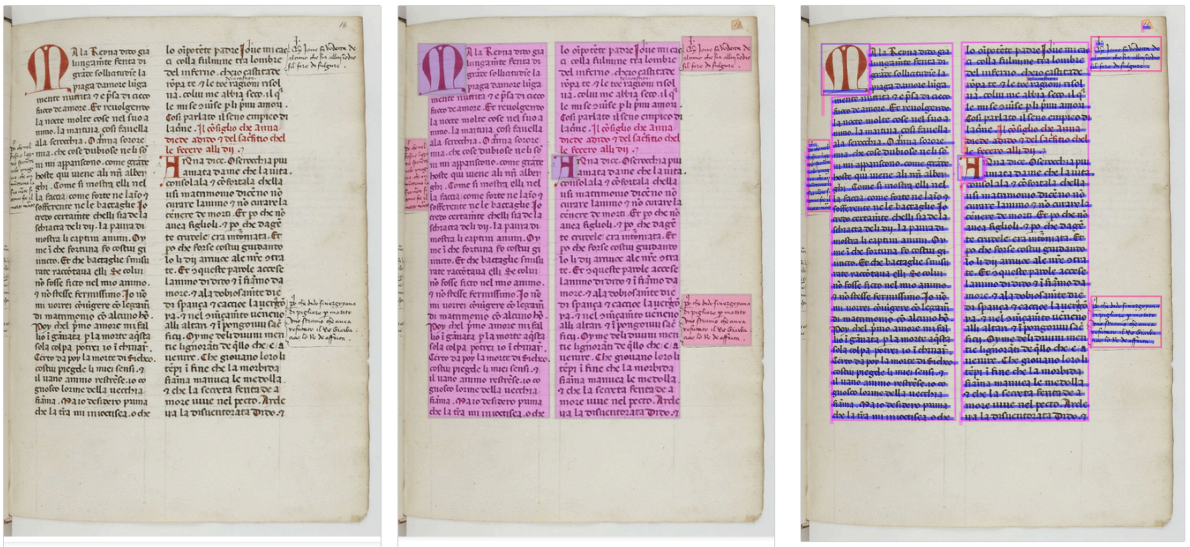
SegmOnto is a controlled vocabulary used to describe the contents of handwritten documents. It is designed to offer a general scheme for describing text in order to homogenise the data required by layout analysers, group data to train more robust models on different layouts and design a standardised process for extracting text from page scans to structured documents.
To analyse the layout, SegmOnto performs two classification tasks:

- Areas of the page such as the main text, title or page number are highlighted as zones,
- Instead, text lines are indicated as lines and can be associated with the text zones to which they belong.

In the segmented pages of the manuscript BnF ita. 590, the following zones were identified: DropCapitalZone for capital letters; MainZone for main text; MarginZone for notes in the margins of the page; NumberingZone for page numbers.
As for lines, the labels were used: DefaultLine for text lines; DropCapitalLine for capital letter lines; HeadingLine for heading lines.
In figure 3 we can distinguish in order from left to right:

- The image of page f39_16r (folio 39_16 recto) of the manuscript BnF ita. 590;
- The areas of the same image segmented in eScriptorium following SegmOnto's guidelines;
- The lines of the same image segmented in eScriptorium following SegmOnto's guidelines

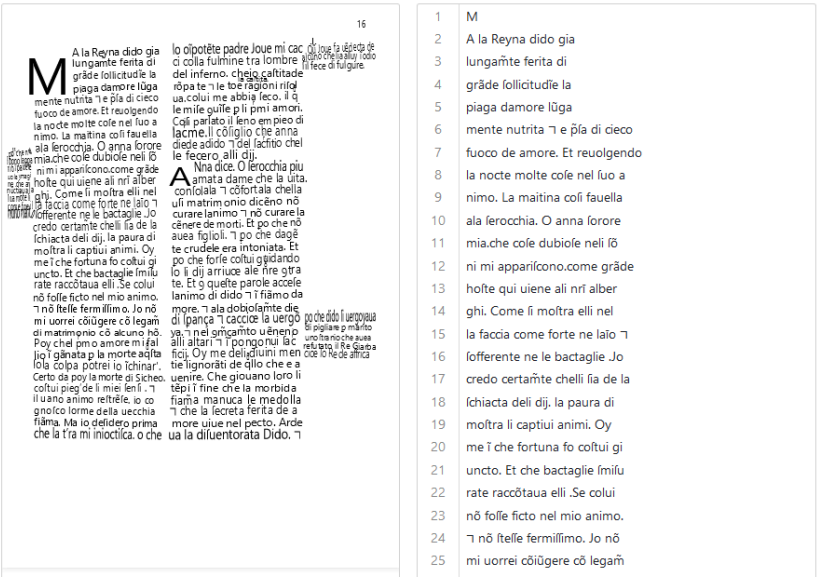
[38] for more informations about CREMMA Medii Aevi model: https://enc.hal.science/hal-03828353
[39] for more informations about SegmOnto: https://segmonto.github.io/



*(**Fig. 3** Example of segmentation performed with SegmOnto on eScritorium)*

Once the text segmentation phase was completed, the CREMMA Medii Aevi model was applied, which automatically transcribed the segmented text, returning it in Unicode characters (figure 4).

The transcription processed by the model was supervised by the author of this thesis, who corrected the errors caused by the imprecision of the model at the time of the transcription useful for this project. As briefly described above, models are nothing more than neural networks, which must be trained extensively in order to return accurate results. This transcription project served precisely to contribute to the training of CREMMA Medii Aevi.



1	M
2	A la Reyna dido gia
3	lungam̃te ferita di
4	grãde ſollicitudſe la
5	piaga damore lũga
6	mente nutrita ⁊ e p̃ſa di cieco
7	fuoco de amore. Et reuolgendo
8	la nocte molte coſe nel ſuo a
9	nimo. La maitina coſi fauella
10	ala ferocchia. O anna ſorore
11	mia.che coſe dubioſe neli ſõ
12	ni mi apparíſcono.come grãde
13	hoſte qui uiene ali nrí alber
14	ghi. Come ſi moſtra elli nel
15	la faccia come forte ne laïo ⁊
16	ſofferente ne le bactaglie .Jo
17	credo certam̃te chelli ſia de la
18	ſchiacta deli dij. la paura di
19	moſtra li captiui animi. Oy
20	me ĩ che fortuna fo coſtui gi
21	uncto. Et che bactaglie ſmiſu
22	rate raccõtaua elli .Se colui
23	nõ foſſe ficto nel mio animo.
24	⁊ nõ ſtelſe fermiſſimo. Jo nõ
25	mi uorrei cõiũgere cõ legam̃

*(**Fig. 4** Example of automated transcription done with eScriptorium)*

This same process applied to page f39_16r was applied to all nine pages that make up the fourth book of the manuscript BnF ita. 590.

The use of HTR techniques and tools made the transcription operation considerably faster than it would have been had the characters of the entire manuscript text been recognised and transcribed manually. After checking that the entire transcription was correct, the XML/ALTO file prepared by eScriptorium was downloaded, containing the coordinates of the text areas and lines and their transcriptions in Unicode characters. The ALTO file was then transformed by adapting it to the structure of the XML/TEI model in order to proceed with the creation of the digital edition of our manuscript book.

2.4
XML/TEI encoding: modelling of semi-structured data with TEI standards

In order to adapt the file to the structure of the TEI model, the XML editor Oxygen[40] was used, which via a pre-installed TEI framework enabled an automatic transformation from XML/ALTO to XML/TEI structure.
Furthermore, specific adjustments were made using some python scripts in order to retrieve from the ALTO file only those elements useful for the desired version in TEI, excluding from the elements of interest the spatial coordinates automatically recorded by eScriptorium and considering only the narrative elements of the text, thus also excluding marginalia.

Thanks to its vast controlled vocabulary of elements and attributes, TEI allows the XML encoding to be adapted to the specific peculiarities of a text, be it in poetic or prosaic form, while guaranteeing a high degree of standardisation and precise handling of even complex and varied texts. In spite of this fluidity, there are two main sections that must be included in the encoding, in the following order:

1. The section of the Header that provides important information about the encoded document. It includes metadata such as the document title, author, date of creation, and other relevant information. This part is essential to accurately catalogue and describe the content of the text, making it an indispensable reference point for scholars and digital archivists.

2. The section of the Text that represents the core of the coded text. This contains the details of the actual content of the document, divided into elements such as paragraphs, sentences, footnotes, quotations and more, depending on the specific needs of the text. This section captures the essence of the content, allowing researchers to analyse and manipulate the text in a structured manner.

In the following, we will describe the two sections in more detail, using the encoding of our text as an example.

[40] Oxygen website: https://www.oxygenxml.com/

Within the Header section we find other optional ones if useful for the project, and the <fileDesc> section which is mandatory and contains the main text description elements.
Inside the fileDesc element we find other three mandatory elements. One of them is the <titleStmt>, which collects details about a work's title and the people who created its content. Nested in the title statement, there can also be the element <respStmt>; when the specialised elements are insufficient or do not apply, the statement of responsibility provides a statement of responsibility for the intellectual content of a text, edition, recording, or series. In the TEI model realised we find <resp> (responsibility), incorporating an indication of the type of intellectual responsibility a person has or the function that an organisation plays in the creation or dissemination of a document.

```
<TEI xmlns="http://www.tei-c.org/ns/1.0">
 <teiHeader>
  <fileDesc>
   <titleStmt>
    <title>Aeneid, IV book. XIV century vernacular reduction</title>
    <author>Andrea Lancia</author>
    <respStmt>
     <resp>Digitally encoded by</resp>
     <name>Giorgia Rubin</name>
     <name>Federico Boschetti</name>
    </respStmt>
   </titleStmt>
```

*(**Snippet 1** TEI Header with FileDesc section)*

Another required part of the fileDesc element is the <publicationStmt> element. Its purpose is to identify the organisation that makes a resource available and to provide any extra details about how it is made available, such as licensing terms and identification numbers. The tag <availability> provides details on the accessibility of a text, such as any limitations on its use or distribution, its copyright status, and any applicable licences.

```
 <publicationStmt>
  <publisher>Ca' Foscari University of Venice, VeDPH</publisher>
  <date>February 2024</date>
  <availability>
   <licence target="https://creativecommons.org/licenses/by-nc-sa/4.0/">
    Distributed under a Creative Commons Attribution-NonCommercial-ShareAlike (CC
BY-NC-SA) licence</licence>
  </availability>
 </publicationStmt>
```

*(**Snippet 2** TEI Header with PublicationStmt section)*

The last mandatory part in the fileDesc element is the <sourceDesc>. It is used to list the sources from which a digital file was derived. This could be a written text or manuscript that has been printed. Nested in <sourceDesc> there are many elements describing the analogue archival documents. For instance, the <msDesc> (manuscript description) contains descriptions related specifically to the manuscript. The tag <msIdentifier> contains the identification number of the manuscript and the place in which it is settled. The tag <msContents> can contain the title of the manuscript and the author. In the following example, we can see the elements:

```
  <sourceDesc>
   <msDesc>
    <msIdentifier>
        <settlement>Paris, France</settlement>
        <repository>BnF, Bibliothèque Nationale de France</repository>
        <idno>BnF italien 590</idno>
    </msIdentifier>
    <msContents>
     <msItem>
        <title>Reduction of the original latin text of the fourth book of Virgil's Aeneid, translated
into the vernacular</title>
        <author>Friar Anastase (Copyst Reducer)</author>
      </msItem>
    </msContents>
   </msDesc>
  </sourceDesc>
 </fileDesc>
</teiHeader>
```

*(**Snippet 3** TEI Header with SourceDec section)*

Within the Text section, which, as we have said, represents the core of the codified text, we find a structure totally dependent on the contents of the text.
Depending on the nature of the latter, whether in verse or prose, we can find different elements useful for describing the structure of the text and its macro and micro divisions.
These containers may be represented with the markup <milestone> rather than and are placed within the subsection of text <body>. Attributes may be added to these elements to indicate specific characteristics, e.g. @unit for the unit of text they are intended to represent (page, column, etc.), rather than @xml:id to indicate with a unique identifier code their presence in the coded text.
Within the milestone and div containers we can then find other sub-elements that explicitly indicate the nature of the text, e.g. to indicate a paragraph in prose we use the marker <p> (paragraph) and to indicate a sentence we use <s> (sentence).
It is possible to section the text into smaller and smaller parts according to the needs of the encoder in accordance with its design, describing smaller and smaller elements down to the single character.

For this project, it was decided to split the text down to the word level using the <w> marker. In addition, the @norm (normalised) attribute associated with the word element was used to indicate the extended word if it was present in the actual text as split and wrapped.
The @corresp attribute inserted within the milestone serves to indicate the correspondence between the line of text encoded in the XML/TEI file and its respective one present within the XML/ALTO file, which was maintained during the transformation of the XML model.

```
<TEI xmlns="http://www.tei-c.org/ns/1.0">
 <teiHeader>
          …
 </teiHeader>
 <text>
  <body>
   <milestone unit="folio" n="16r"/>
   <milestone unit="column" n="1"/>
    <div xml:id="p1">
     <p>
      <s>
       <milestone unit="line" corresp="71bd4a80"/>
       <w norm="Ma">M <milestone unit="line" corresp="ee32882e"/> A</w>
       <w>la</w>
       <w>Reyna</w>
       <w>dido</w>
       <w>gia</w>
         …
```

*(**Snippet 4** TEI Body section)*

The TEI not only has the advantage of offering a certain fluidity in the encoding of textual elements, but also allows the integration of elements other than the text to interact with external resources for the benefit of a more complete digital edition.
In our case, it was possible, for example, to insert images in IIIF standard of the facsimile, offered by the Gallica Digital Library.
To do this, the <facsimile> element, useful precisely to contain the indications of a primary source, was inserted and placed between the Header and Text sections.

```
<TEI xmlns="http://www.tei-c.org/ns/1.0">
 <teiHeader>
        …
 </teiHeader>
 <facsimile>
   <surface>
        <graphic
        url="https://gallica.bnf.fr/iiif/ark:/12148/btv1b8433319z/f39/full/full/0/default.jpg"/>
        <graphic
        url="https://gallica.bnf.fr/iiif/ark:/12148/btv1b8433319z/f40/full/full/0/default.jpg"/>
        <graphic
        url="https://gallica.bnf.fr/iiif/ark:/12148/btv1b8433319z/f41/full/full/0/default.jpg"/>
         …
    </surface>
 </facsimile>
 <text>
```

*(**Snippet 5** TEI Facsimile section)*

The <facsimile> element contains the <surface> element, which corresponds to a map of the manuscript, which in turn contains a <graphic> element that links directly to the URL of the IIIF via the @url attribute: <facsimile> <surface> <graphic url="http://exemple.jpg"/>. Each contains references to image files representing a manuscript map. The IIIF API identifies an image and its specifications through a URI conveyed via the HTTP or HTTPS protocols. This means that a client can connect to an image available on the Web that the server itself provides with explicitly expressed formal parameters of the same URI according to the syntax (Mancinelli, 2021, 80).

2.4.1
Euporia: a DSL for annotating TEI document

After configuring the TEI encoding as outlined above, the document was enriched with annotations using the domain-specific language tool Euporia[41], an application for eXist-db[42], developed by the Collaborative and Cooperative Philology Laboratory (CoPhiLab) of the Institute of Computational Linguistics 'A. Zampolli' of the CNR (ILC-CNR) in Pisa.

Domain-specific languages (DSL) are formal languages or a set of rules and conventions designed to address a specific field of knowledge (such as classical philology). Unlike generic programming languages, DSL are highly specialised and optimised for solving narrow, well-defined problems in a specific domain.
Thanks to this peculiarity, they can significantly improve efficiency, precision and

[41] Euporia presentation: https://github.com/CoPhi/euporia/blob/main/README.md
[42] eXist-db website: http://exist-db.org/exist/apps/homepage/index.html

comprehensibility when dealing with sector or domain specific problems, thus providing a significant advantage to users. In fact, they are often more readable and comprehensible than generic programming languages and allow users to concentrate on domain specifics without having to worry about superfluous technical details.

The Euporia method was created precisely to attempt to combine the needs of traditional scholars with those of digital humanists.
A DSL is defined by a Context-Free Grammar that formally describes its syntax and reserved words. The DSL behind Euporia consists of two main modules: a module for referencing the text to be annotated and a module for structuring the content of the annotation (Boschetti et al., 2020).
What the Euporia method therefore proposes is a textual annotation system with a very small graphic interface, familiar with the text analysis and commentary practices in use in schools and the academy, with innovative elements based on a rigorous and unambiguous text referencing system, which give the annotator the possibility of creating his or her own annotation system that is as simple, compact, effective for the specific purpose and at the same time fully machine actionable and convertible to standard formats (Boschetti, Mugelli, 2021).

The objective of this annotation phase using the Euporia method is to integrate the XML/TEI encoding of our text with the information described in chapter 1 paragraph 3 of this thesis, i.e. to make explicit the comparison between the contents of the IVth book of BnF ita. 590 and those of the book of the critical edition of the Virgilian text provided by the MQDQ.
As described in the first chapter, the manuscript BnF ita. 590 presents a reduced version of the original text. The digital edition of our manuscript is intended to bring out through selected visualisations those variations of the text that originated from the reduction produced by the author, i.e. those parts of the text that are missing rather than synthesised or corresponding to the text of the critical edition.
Furthermore, with the Euporia system, we wish to annotate the themes outlined by the philological analysis that has been carried out on the contents of book IV of the Aeneid, the instances present in the manuscript text (characters and places mentioned) and their occurrences[43].

In order to proceed with the annotation through Euporia of our XML/TEI encoding, it was necessary to prepare some useful elements for the encoded explication of the correspondence between the parts of the manuscript text and those of the critical edition.
This operation is necessary in order to set up the alignment of the two texts within the annotation panels of the Euporia (figure 5) platform, which will then be able to provide us with a simple and orderly graphic interface to facilitate the annotation of the two texts readable in parallel.
To do this, the attributes @xml:id and @corresp were added within the TEI encoding within the <s> (sentence) element. The @xml:id attribute determines a unique code associated with the sentences in our manuscript, a different one for each sentence. The numbers that make it

[43] See fig. 2 in chapter 1

up should be read as page.column.sentence. Through the @corresp attribute, each of these sentences is made to explicitly match the line of text of the critical edition to which it corresponds.

```
<s xml:id="s1.1.1" corresp="#l1">

<s xml:id="s1.1.2" corresp="#l5">

<s xml:id="s1.1.3" corresp="#l9">

<s xml:id="s…" corresp="#l…">
```

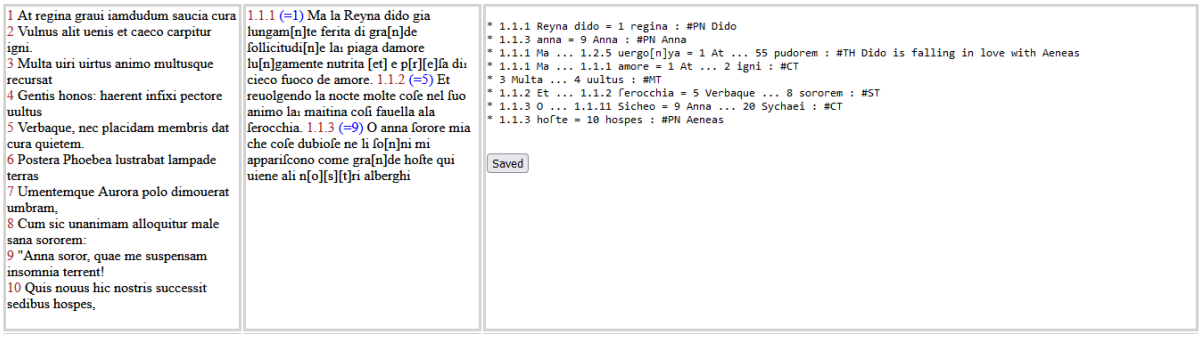
*(**Snippet 6** TEI addition of @corresp attribute for textual matches)*

The DSL grammar system used for text annotation follows the principle of familiarity dear to Euporia, making use of simple hashtags.
The resulting language allows for the expression of precise word granular text references, closed-value annotations and open-value annotations.
Each annotation type is prefixed by an hashtag[44]:

- #TH to identify the themes dealt with in the text,
- #CT to delineate the corresponding parts between the two texts,
- #ST for the parts summarised in the manuscript text,
- #MT for the missing parts,
- #PN to classify names of persons (Dido, Anna, Sychaei, Iulo, Aeneas, Anchisae, Iarba, Numidio, Sacerdos, Nutrix),
- #RG to identify roman gods (Iuno, Venus, Mercurius, Apollo, Iris, Proserpina, Iuppiter)
- #AEN for abstract entities names (Castitade, Fama),
- #GPN to identify groups of people (Phrygios, Tyrios, Mates of Aeneas, Mates of Dido),
- #PL to classify places (Libya, Carthage, Sidonia, Italy).

[44] For #TH, #CT, #ST, #MT correspondence see fig. 2 in chapter 1

Verg., *Aeneid*



*(**Fig. 5** Euporia annotation panel interface)*

Euporia's annotation tools are based on the ANTLR5[45] compiler (i.e. compiler compiler). Starting from the grammar of the annotation language used, expressed in a variant of the EBNF format (Extended Backus-Naur Form), ANTLR generates a parser for the language itself, which is used both for the validation of the annotations and for serialisation in XML, with a proprietary schema conforming to the rules of the DSL grammar.
Once the XML has been generated from the DSL, XSLT transformation sheets are used to convert the file towards the standard XML/TEI encoding (F. Boschetti,G. Mugelli, 2021:90), and thus transfer, through precise indications dictated by the encoder, the annotations made within the previously structured TEI file.
For our project, the result is the following:

```
</teiHeader>
<standOff>
   <listPerson type="person">
   <person xml:id="Dido">
   <persName xml:lang="latin">Dido</persName>
   </person>
   <person xml:id="Anna">
   <persName xml:lang="latin">Anna</persName>
   </person>
        …
   </listPerson>
   <listPerson type="ancient-roman-god">
   <person xml:id="Iuppiter">
   <persName xml:lang="latin">Iuppiter</persName>
   </person>
   <person xml:id="Iuno">
   <persName xml:lang="latin">Iuno</persName>
   </person>
```

[45] ANTLR website: https://www.antlr.org

```
            …
    </listPerson>
    <listPerson type="group-of-people"/>
    <listPerson type="allegory-god"/>
    <listPlace type="old-place"/>
    <listPlace type="current-place"/>
</standOff>
<facsimile/>

<text>
  <body>
    <milestone unit="folio" n="16r"/>
    <milestone unit="column" n="1"/>
    <div xml:id="p1" n="1">
      <p xml:id="Burning-love-of-Dido-for-Aeneas">
       <s xml:id="s1.1.1" corresp="#l1">
          <milestone unit="line" corresp="71bd4a80" n="1"/>
          <w norm="Ma">M <milestone unit="line" corresp="ee32882e" n="2"/> A</w>
          <w>la</w>
          <w>Reyna</w>
          <persName ref="#Dido">dido</persName>

                    …

          <persName ref="#Anna">anna</persName>
       </s>
      </p>
    </div>
  </body>
</text>
```

*(**Snippet 7** TEI including Euporia annotations)*

References to elements annotated in Euporia found in the <text> section are inserted within the specific elements, for example in the case of person names the element is <persName>. Within these elements is the @ref attribute. It corresponds to a specific @xml:id attribute different for each annotated entity, which is found inserted within the <standOff> section. This section, located between the header and the facsimile section, allows to list all the elements present within the text and trace them back to specific entities represented by their xml:id. In this way, whenever there is an occurrence of an instance within the text, it will be identified as the entity it represents.

2.5 From semi-structured annotated data to structured (XLSX) data

Annotation via Euporia platform was very useful due to its clear and straightforward interface, to compose two XLSX documents containing the annotated information.
The XLSX document is a representation of structured data in table format. It allows information to be organised and analysed in a rigid manner, which is ideal for rapid identification and analysis of recorded data.
The table is organised in rows and columns, with each row representing an object and the columns representing various attributes associated with each object.
The first of the two XLSX files (figure 6) records the number of occurrences of each cited instance within the text, also recording its location. The first column (Type) of the table contains the type of entity, i.e. the category or classification, registered in Euporia through the use of hashtag, to which the entity belongs. The second column (Name) contains the name of the entity, an attribute that provides a unique identification for each entity. The subsequent columns are dedicated to the various thematic paragraphs of the text, with each column containing the number of times the entity is mentioned within the corresponding paragraph.

TYPE	NAME	Burning lo	Offering to	Juno and	Dido and	Fame flies	Mercury a	Preparatio	Omen of D	Aeneas co
Person	Dido	2	2	5	1	2	0	1	7	1
Person	Anna	3	0	0	0	0	0	0	0	0
Person	Sychaei	1	0	0	0	0	0	0	0	0
Person	Iulo	0	0	0	0	0	2	0	1	0
Person	Aeneas	2	0	3	0	4	1	2	4	1
Person	Anchisae	0	0	0	0	0	0	0	1	0
Person	Iarba	0	0	0	0	2	0	0	0	0
Person	Numidio	0	0	0	0	0	0	0	0	0
Person	Sacerdos	0	0	0	0	0	0	0	0	0
Person	Nutrix	0	0	0	0	0	0	0	0	0
Roman God	Iuno	0	0	4	0	0	0	0	0	0
Roman God	Venus	0	0	4	0	0	0	0	0	0
Roman God	Mercurius	0	0	0	0	1	0	0	1	0
Roman God	Apollo	0	0	0	0	0	0	0	2	0

*(**Fig.6** XLSX file containing entity categories and occurrences inside the IV book of the ms. BnF ita. 590)*

The second XLSX file (figure 7) is the result of a careful quantitative analysis of the lemmas belonging to the portions of the two texts compared and annotated in Euporia as corresponding, synthesised or missing. This analysis led to the recording of percentage data showing for each thematic paragraph and for the entire book, in which portion the original text of the Virgilian poem was respected, synthesised or omitted in the manuscript reduction.

THEME	Burning love of Dido for Aenea	Offering to the Gods by Did	Evening mea	First night fal	Juno and Venus plan to unite love
Corresponding Text	45,45454545	33,33333333	0	0	41,46341463
Synthesised Text	47,27272727	0	0	0	39,02439024
Missing Text	7,272727273	66,66666667	100	100	19,51219512

*(**Fig. 7** XLSX file containing percentage data of the comparison of the BnF ita. 590's text and the one of the MQDQ critical edition)*

The structured data files generated from these tables will be used to develop quantitative visualisations. Using data analysis tools, it will be possible to generate graphs, charts and other visual representations to explore and understand relationships and patterns in the data.

Chapter 3
Representing Data Meaning

The result that emerged from the data modelling process described above leads to the generation of an annotated XML/TEI document that, if properly modified, would lead to the realisation of a document-centric, i.e. document-oriented, digital edition of the text.

In this chapter, we continue our discussion of semi-structured data modelling. We will illustrate the process and methodology used to design an ontology, i.e. a conceptual model, representing the knowledge base of our manuscript text. We will analyse the method of integrating a series of useful attributes into the XML/TEI file to enrich the knowledge base and create an RDF model capable of describing the elements of the text with a machine-readable and comprehensible semantic syntax.
The semantic representation of the contents of a text can lead to the transformation of a document-centric edition into a data-oriented one. This approach has the advantage of leading to the explication of underlying meanings based on the disambiguation of data and entities in the text, related to each other and linked to their representation on the web defined by authority files.
This is made possible by the adoption of the renewed methodologies of information representation that the renewal of web languages brought with it after the birth of the Semantic Web, which adopts a data-centric approach (Tomasi, 2017).

Before describing the semantic transformation process implemented on our XML/TEI document, we will shed some light on the nature of the Semantic Web and its logic of abstraction and information representation.

3.1
Semantic Web and Linked Data

The original purpose of the World Wide Web was geared towards human use. Despite being machine-readable, the data within it lacks machine comprehension. An essential goal of the Semantic Web is to enhance the web's content with ontology annotations, enabling machines to understand it effectively.
The Semantic Web, a project presented in 2001 by Tim Barners-Lee[46], represents an advanced paradigm in the evolution of the World Wide Web, in which data is treated with a deep semantic perspective. In the initial phases of the Web (Web 1.0 and Web 2.0), the focus was on the visualisation of HTML documents through URLs (Uniform Resource Locators), conceived as unitary blocks of information for users. However, in the context of the Semantic

[46] Berners-Lee, T., Hendler, J., & Lassila, O. (2001). A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. *Scientific american*, *284*(5), 34-43.

Web, also known as Web 3.0, the focus shifts from the simple visual presentation of documents to the essence of the data contained within them. In this perspective, the essential thing is to make this data machine-interpretable and distinct through URIs (Uniform Resource Identifiers) so as to allow interoperability between different systems.
This conceptual metamorphosis is made possible by the widespread use of the HTTP (Hypertext Transfer Protocol). This protocol facilitates the interchange of data across the Internet and, thanks to its universality, allows for easy integration of data that are processed by applications and services, creating a network of interconnected information that transcends the limits of individual websites or isolated applications, thus contributing to the creation of a semantic environment that allows users to access deep and interdisciplinary knowledge across a wide variety of sources and domains.
The semantic representation of data and their relationships is enabled by the relational structure of the conceptual models with which it is represented. For example, the representation scheme consisting of RDF (Resource Description Framework) triples. These triples follow a relation schema of the subject-predicate-object type. The subject represents the main entity, the object is another entity or a specific value, while the predicate defines the relationship between the subject and object. These entities and their predicates are described through the use of controlled vocabularies that define concepts and relationships used to create ontologies, conceptual models that organise information in a specific domain. Labels, dictated by these vocabularies, serve as a standardised mechanism for attributing machine-understandable meaning to data, thus facilitating the interpretation and sharing of information across heterogeneous systems. Some of the most used vocabularies in the cultural heritage domain are: CIDOC CRM, DCTERMS, SKOS, FOAF.
The RDF triples that make up the ontologies can be queried through SPARQL (Simple Protocol And RDF Query Language) queries, a flexible query language that allows detailed exploration of the data network, enabling advanced queries that retrieve specific information from the RDF data, enabling users to obtain highly relevant results.

This change in knowledge representation methodologies and procedures has consequently led to a new rethinking of digital textuality (Mancinelli, 2021, 82).
Adopting the paradigms of the semantic web, cultural data are described by ontologies and vocabularies belonging to philological-literary domains, rather than archival or iconographic and whatever else, capable of representing the knowledge expressed in works.
Indeed, it is the vocabularies that constitute the semantic glue that transforms raw data into meaning-rich data.
Reframing the concepts expressed by LOV[47] (Linked Open Vocabularies), we understand how data are structured through the use of properties, commonly referred to as predicates, and classes, also identified as types, in order to delineate in detail individuals, places, events and every possible category of objects.
An illustrative example of this practice could be the following statement: 'Dido is a person, her sister is Anna, and she lives in the city of Carthage'.
In this context, 'Person' represents the class to which Dido belongs, while 'City' represents the

[47] LOV website: https://lov.linkeddata.es/dataset/lov/

class to which Carthage belongs. Furthermore, the properties 'has a sister' and 'lives in' are exploited to describe a person, with the latter also acting as a connection between the individual and the place where she lives.
The classes and properties are organised in such a way as to create a collection of definitions that form the terms of a vocabulary.
Through the linked data system, entities described with these vocabularies can be associated with corresponding content in authority files. These are organised and standardised academic and library resources containing authoritative and reliable information on uniquely identified entities. Authority files serve as reference points to establish the authenticity and accuracy of information expressed in ontologies, thus guaranteeing data integrity in the context of academic and cultural research.
It is clear, therefore, how the semantic web and related data play a crucial role in transforming document-centric digital editions into semantic environments for exploring the knowledge base expressed in texts.

We now turn to the description of the conceptual model designed to describe the contents of our manuscripted text, and the transformation of the previously encoded XML/TEI file into serialised RDF triples expressed in RDF/XML, a machine-readable format.

3.2
Design of the source's RDF data model

The ontological model devised for this thesis project aims to briefly describe the provenance of the text following the FRBR[48] (Functional Requirements for Bibliographic Records) schema, and to outline and analyse the interconnections that exist between the entities identified within our text. This model also aims to examine the salient characteristics of these entities, distinguishing them between individuals and places, associating each profile with the corresponding one provided by selected authority files, and investigating their positioning within the thematic context of the text. This positioning is determined by their association with each of the themes in the text outlined above, in which they are mentioned.
The selection of authority files was carried out, which led to the choice of Wikidata[49] and DBpedia[50] due to their remarkable ability to provide a vast repertoire of accurately documented and described entities. These entities range from the general to the specific, including those relevant to the content of the IVth book of the Aeneid.
Wikidata and DBpedia are widely recognised semantic resources that adhere to the guidelines established by the W3C (World Wide Web Consortium) regarding the representation and connectivity of data on the web. Both contribute significantly to the enrichment of our

[48] FRBR description by IFLA (International Federation of Library Associations and Institutions) https://www.ifla.org/references/best-practice-for-national-bibliographic-agencies-in-a-digital-age/resource-description-and-standards/bibliographic-control/functional-requirements-the-frbr-family-of-models/functional-requirements-for-bibliographic-records-frbr/
[49] Wikidata website: https://www.wikidata.org/wiki/Wikidata:Main_Page
[50] DBpedia website: https://www.dbpedia.org/

ontological perspective, enabling the integration of data from reliable sources, thus offering a more complete view of the characteristics of the entities within our ontology.

The vocabularies used for the semantics of our ontology are as follows:

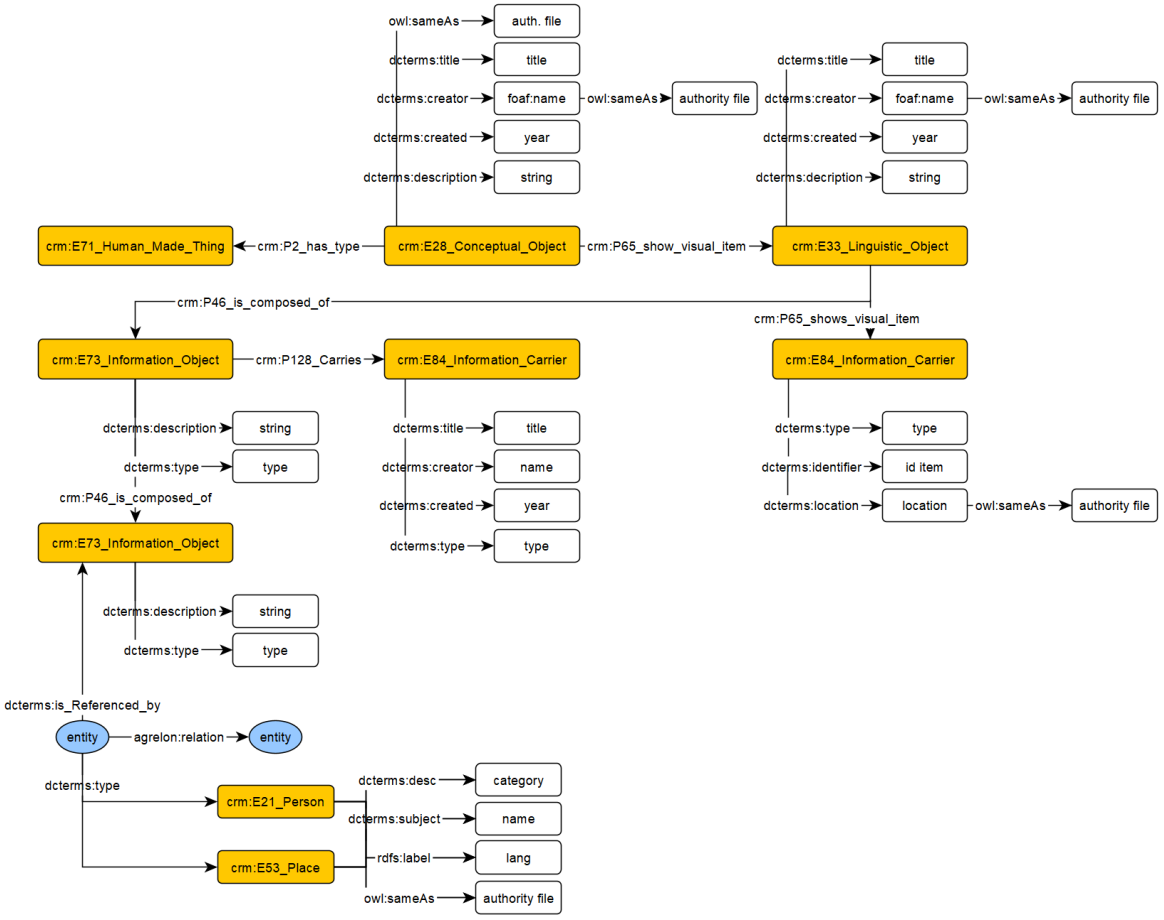
- rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
- rdfs: <http://www.w3.org/2000/01/rdf-schema#>
- owl: <http://www.w3.org/2002/07/owl#>
- dcterms: <http://purl.org/dc/terms/>
- agrelon: <https://d-nb.info/standards/elementset/agrelon#>
- crm: <http://www.cidoc-crm.org/cidoc-crm/>
- foaf: <http://xmlns.com/foaf/0.1/>

In order to describe the analysed text according to the FRBR scheme, which outlines the relationships between work, expression, manifestation and item in the library context, the following reasoning was developed, considering only the first three elements:

- Work: The work is Virgil's Aeneid. This term represents the abstract concept of the Aeneid as a literary work, independent of its translations, interpretations or formats. In the ontology it has been identified with the class *crm:E28_Conceptual_Object*, which through the property *crm:P2_has_type* has been defined as *crm:E71_Human_made_thing*. Within the ontology, the type of work (*dcterms:description* → epic poem), the year of creation (*dcterms:created* → year), the title (*dcterms:title* → Aeneid), the author (*dcterms:creator* → *foaf:name* → Virgil), and the title (*dcterms:title* → Aeneid) were also made explicit. The latter elements have been disambiguated and associated with authority files through the *owl:sameAs* property.
- Expression: The translation of the Aeneid manuscript into a new language represents an Expression of the Work. This level concerns the intellectual and artistic realisation of the work, which in this case is the specific translation of the Latin text of the Aeneid into the new language by Andrea Lancia. It has been represented in the ontology with the class *crm:E33_Linguistig_Object*. For this too, as for the work, description, year of creation, author and title have been made explicit.
- Manifestation: Here, we have two different manifestations:
  a. The physical manuscript of the translation of the Aeneid preserved at the BnF. This manifestation is the manuscript itself, a unique physical object representing a specific realisation of the expression, i.e. the translation.
  b. The digital edition of the manuscript whose encoder is me, Giorgia Rubin. This is a different manifestation, produced in digital format. While it represents the same expression (the same translation), the format in which this expression is presented is different (digital instead of physical).
  These two manifestations were indicated in the ontology through the class *crm:E84_Information_Carrier* and as for the previous elements, author, year of publication, etc. were made explicit.

From the expression, the fourth book identified through the class *crm:E73_Information_Object* was inserted into the ontology, and the same was done for the various thematic paragraphs into which the book was divided during analysis.
In each of the paragraphs, through the property *dcterms:isReferencedBy*, the entities corresponding to the characters and places mentioned and linked to the authority files have been inserted through the property *owl:sameAs*, also declaring in which language the translation of the name provided is proposed through the property *rdfs:label*.
In addition, each entity is further described as belonging to the class *crm:E21_Person* rather than *crm:E53_Place* and described through the property *dcterms:desc* in more detail by specifying with regard to the class person, whether it is a person, a group of persons, an ancient Roman god rather than an abstract entity, with regard to places whether they are ancient places or correspond to cities still existing today. Each of these properties is specified uniquely through the link with authority files enabled by the *dcterms:subject* property. The relationships that make up the social network present between the entities is expressed thanks to the *agrelon:relation* property.

The conceptual model just described in natural language can be read in its graph form designed specifically for this thesis project (figure 8).



*(**Fig. 8** RDF data model describing the FRBR structure of the IV book of the ms. BnF ita. 590, the entities mentioned inside it and the relations between them)*

This ontological model is capable of expressing and augmenting the knowledge contained in our manuscript text through a series of triples of subjects, objects and predicates that are linked together according to the RDF schema.

In figure 9, we can read some examples, which can be translated into natural language such as: the subject of the story Dido corresponds to the subject that Wikidata disambiguates through the unique code Q905162. It is found in the paragraph entitled "Burning love of Dido for Aeneas", which is a fragment of the text of our manuscript book.

SUBJECT	PREDICATE	OBJECT
https://aeneidbnfms.org/person/Dido	owl:sameAs	https://www.wikidata.org/wiki/Q905162
https://aeneidbnfms.org/person/Dido	dcterms:isReferencedBy	https://aeneidbnfms.org/text/Burning-love-of-Dido-for-Aeneas
https://aeneidbnfms.org/text/Burning-love-of-Dido-for-Aeneas	frbroo:R15_is_fragment_of	https://aeneidbnfms.org/IV_aeneid_bnf590

*(**Fig. 9** Some of the RDF triples described within our ontology graph)*

3.3
RDF serialisation (RDF/XML) with LIFT

In order to provide serialisation, i.e. transformation into machine-readable text, of the RDF triples described in our model, LIFT has been adopted. The following explains how it works and how it has been applied to our model.
LIFT is an open source tool created in the context of the University of Bologna. It is written as a set of Python scripts for generating linked data from TEI-encoded texts and proposes a method of organising and publishing cultural knowledge and, specifically, digital scholarly editions on the web in a perspective of data integration (Giovannetti, Tomasi, 2022).
LIFT involves a data modelling process that is divided into two basic phases:

1. preparation of the XML/TEI document with the inclusion of certain attributes in the coded elements of interest, useful for the creation of the semantic network;
2. adoption of Python scripts to parse the annotated TEI document and to extract the data of interest for the generation of serialised RDF files. LIFT uses two libraries to perform these actions:
    - lxml.etree: a Python library for XML processing
    - RDFLib: a Python library for working with RDF data

First of all, in accordance with the rules for the functioning of the RDF schema, it is necessary to assign a URI to each of the entities that will take part in constituting the serialised RDF graph. To do this LIFT uses the @xml:id attribute, previously inserted in our XML/TEI document to disambiguate the characters and places present within our text,

concatenating it to the @xml:base attribute of the <TEI> section.
For example, the element below representing a person:

```
<TEI xmlns="http://www.tei-c.org/ns/1.0" xml:base="https://aeneidbnfms.org">
        …
        <person xml:id=”Dido”>...</person>
        …
</TEI>
```

*(**Snippet 8** LIFT preparation: assignment of a URI to entities)*

The URI is assigned: <https://aeneidbnfms.org/person/Dido>

Next to the @type attribute inserted in lists containing characters and places in the text, the @corresp attribute is inserted to which the value of the URI provided by an authority file is given to represent the class indicated by type:

```
<listPerson type="person" corresp="https://www.wikidata.org/wiki/Q5">
        <person xml:id="Dido">
```

*(**Snippet 9** LIFT preparation: assignment of the authority files URI to the classes)*

As with classes indicated by type, entities are likewise disambiguated, through the @sameAs attribute:

```
<person xml:id="Dido" sameAs="https://www.wikidata.org/wiki/Q905162">
```

*(**Snippet 10** LIFT preparation: disambiguation of entities through @sameAs attribute)*

The Python scripts prepared by LIFT for the generation of RDF triples were adapted for the purposes of this thesis project, adopting the vocabularies presented in the previous section (paragraph 3.2). The result obtained from the application of these scripts led to the serialisation of the triples in RDF/XML schema.
An extract is given below:

```
<?xml version="1.0" encoding="utf-8"?>
<rdf:RDF
  xmlns:agrelon="https://d-nb.info/standards/elementset/agrelon#"
  xmlns:dcterms="http://purl.org/dc/terms/"
  xmlns:frbroo="http://iflastandards.info/ns/fr/frbr/frbroo/"
  xmlns:owl="http://www.w3.org/2002/07/owl#"
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
  xmlns:fabio="http://purl.org/spar/fabio/" >
```

```
<rdf:Description rdf:about="https://aeneidbnfms.org/person/Dido">
<rdf:type rdf:resource="http://www.cidoc-crm.org/cidoc-crm/E21_Person"/>
<owl:sameAs rdf:resource="https://www.wikidata.org/wiki/Q905162"/>
<rdfs:label xml:lang="latin">Dido</rdfs:label>
<dcterms:isReferencedBy
rdf:resource="https://aeneidbnfms.org/text/Burning-love-of-Dido-for-Aeneas"/>
<dcterms:isReferencedBy
rdf:resource="https://aeneidbnfms.org/text/Offering-to-the-Gods-by-Dido"/>
<dcterms:isReferencedBy
rdf:resource="https://aeneidbnfms.org/text/Juno-and-Venus-plan-to-unite-lovers"/>
<dcterms:isReferencedBy
rdf:resource="https://aeneidbnfms.org/text/Dido-and-Aeneas-take-refuge-in-the-cave"/>
        ….
<dcterms:description>person</dcterms:description>
<dcterms:subject rdf:resource="https://www.wikidata.org/wiki/Q5"/>
<agrelon:hasSister rdf:resource="https://aeneidbnfms.org/Anna"/>
<agrelon:hasLover rdf:resource="https://aeneidbnfms.org/Aeneas"/>
<agrelon:hasHusband rdf:resource="https://aeneidbnfms.org/Sychaei"/>
<agrelon:hasKingdom rdf:resource="https://aeneidbnfms.org/Carthage"/>
<agrelon:hasMates rdf:resource="https://aeneidbnfms.org/Mates-of-Dido"/>
</rdf:Description>
```

*(**Snippet 11** RDF/XML result obtained adopting LIFT)*

Chapter 4
Data Visualisation

The data extraction and modelling process, illustrated in chapters two and three, traces the initial and intermediate stages of the digitisation paradigm adopted. It resulted in structured and semi-structured data that provide a detailed and semantic organisation of the contents of our text.
In this chapter we explore the final stage of the digitisation paradigm: the process of data visualisation. This phase leads us to realise the graphical representation of data and the visualisation of textual content in various graphical forms. To achieve this, we embrace the principles and best practices of Information Visualisation.

4.1
Information Visualisation and computational philology

Information visualisation is based on cognitive processes and visual perception, both during the creation (encoding) and utilisation (decoding) phases. Successful visualisation is determined by the effectiveness of decoding, with meaning conveyed through clear, reliable, recognisable and easily comprehensible correspondences (Pinker, 1990). The literature dealing with Cognition and Information Visualisation expounds on the cognitive underpinnings of graphic visualisations, including aspects such as recording information, communicating meaning, expanding operational memory, supporting search and discovery, and supporting perceptual inference by enhancing the detection and recognition of patterns and patterns.

Analytical and visualisation tools become invaluable when examining linguistic or semantic aspects of the text. Indeed, there are several reasons why information visualisation plays a key role in the analysis and understanding of data extracted from digitised historical documents and manuscripts.
Transforming complex and abstract data into visualisations and graphical representations greatly improves the clarity and accessibility of such resources. Consider how graphics, concept maps or other visual tools accompany and lighten the reading of such data, greatly facilitating the understanding of complex structures or relationships. Visualisation, in fact, allows for multidimensional representations. As in our case, for instance, information derived from manuscripts can span multiple dimensions, including text, images, quantitative data and semantic relationships. Integrating these dimensions facilitates the understanding of text content and the identification of interconnections.
Moreover, using visualisation tools, it is possible to synthesise data from different sources and integrate them with data already contained in the text under analysis, contributing to a deeper understanding of that text.
In philology projects, visualisation can help scholars make informed decisions about the

structure of the text, the distribution of information and the interpretation of documents, while also facilitating decision making in digitising them. Patterns, trends or anomalies can be more easily identified by means of graphs and diagrams, which can reveal information not easily discernible through simple numerical or textual analysis.
Visualisation tools also often offer direct interaction with data. This allows users to explore specific data and modify visualisation parameters by focusing on the most relevant ones to deepen a customised analysis.
Finally, the ability to visually synthesise complex information offers the possibility of effectively communicating the analysed data even to an audience with limited technical expertise, thus broadening the target audience.

S. Sinclair and G. Rockwell, developers of many visualisation tools as Voyant Tools[51], say: "we read texts we enjoy, we then explore and study them with analytic tools and visualisation interfaces, which then brings us back to rereading the texts differently. This is what we call the agile interpretive cycle" (Sinclair and Rockwell, 2015). The offer of text visualisation tools allows readers to implement this agile interpretative cycle: reading, exploring and studying texts with analytical tools and visualisation interfaces, leading to a nuanced reinterpretation of texts.

4.2
Requirements to be fulfilled

For the reasons just explained, visualisation is an essential component in the ecosystem of a manuscript digitization project, significantly contributing to enhancing understanding, analysis, and dissemination of extracted data.

The selection of tools and visualisation methods to be adopted is inherently tied to the specific objectives one aims to achieve. Not all situations and types of data necessitate the same visualisation tools, for this reason, it is important to meticulously choose tools that are best suited for representing particular data types and are capable of effectively conveying the intended messages.
In the context of this thesis project, four distinct visualisation methods have been implemented. In certain instances, existing tools deemed particularly suitable were employed, while in other cases, bespoke visualisations were designed from scratch for seamless integration within a web environment.
The specific requirements identified within our project are as follows:

1. High-resolution display of images from manuscript BnF ita. 590 using the IIIF manifesto authored by Gallica: For this task, the IIIF viewer Mirador was chosen. Widely adopted within the IIIF community, this tool is characterised by an intuitive interface designed for easy integration into web environments.

[51] Voyant Tools website: https://voyant-tools.org/

2. Visualisation of text modelled in XML/TEI format: To achieve this, an HTML code was crafted. Through the use of an XSLT parser, this code can transform the XML stylesheet into a visual representation of the text enriched with specific graphical features.

3. Visualisation of quantitative data extracted from the text and collected in excel files: These files contain information regarding the number of occurrences per thematic paragraph for each entity described in the ontology (characters and locations). Additionally, they provide data on the percentage of matches and differences between the two texts: that of manuscript BnF and the critical edition Mqdq.

4. Visualisation and exploration of the semantic ontology serialised in RDF/XML: This operation was carried out using the visual graph tool provided by the GraphDB database. It is an interactive graphical instrument enabling serendipitous exploration of entities expressed within the RDF model and the explicit understanding of the relationships inherent among them.

A final requirement identified in the project, which would necessitate the use of APIs and the development of backend code for implementation, is as follows:

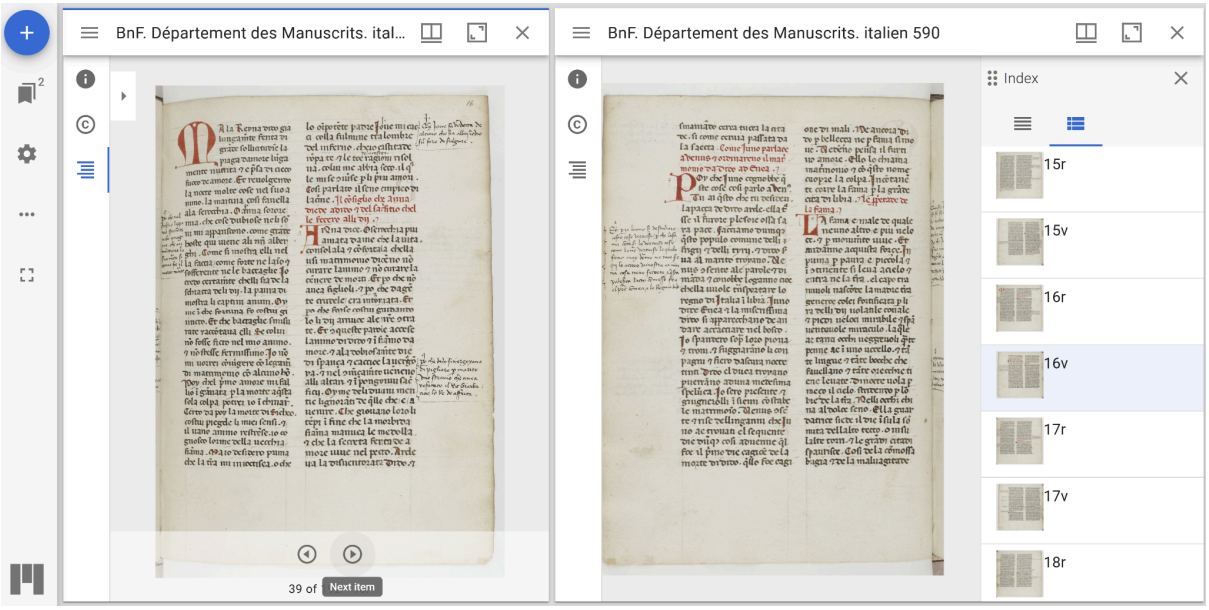
5. Access additional details about entities explored in the ontological graph: For this requirement, a proposal has been made, although not implemented, suggesting the creation of an HTML visualisation of a technical sheet detailing the characteristics of each entity. This process would involve querying the authority file to which the entity is linked in the RDF schema, namely Wikidata. The objective is to display some of the most relevant specifications found on the Wikidata page of the entity.

4.3 Mirador: IIIF viewer

Mirador[52] is an open source web-based image viewer. It allows scholars and the public to visualise, compare and annotate digital objects without having to download any files, PDF or JPG, by doing the entire work on the web. Designed specifically for the exploration of two-dimensional digital images, Mirador is particularly useful and suitable for the visualisation of IIIF-compliant images. Starting from IIIF manifests in JSON format, containing the metadata and structure of digital objects, it visualises images based on the information in the manifest. It is indeed possible to direct Mirador to an IIIF manifest simply by entering its URL. If the viewer is then directed to a location such as https://gallica.bnf.fr/iiif/ark:/12148/btv1b8433319z/manifest.json, it will find all the information it needs to start sending to the user the images of the manuscript BnF ita. 590.

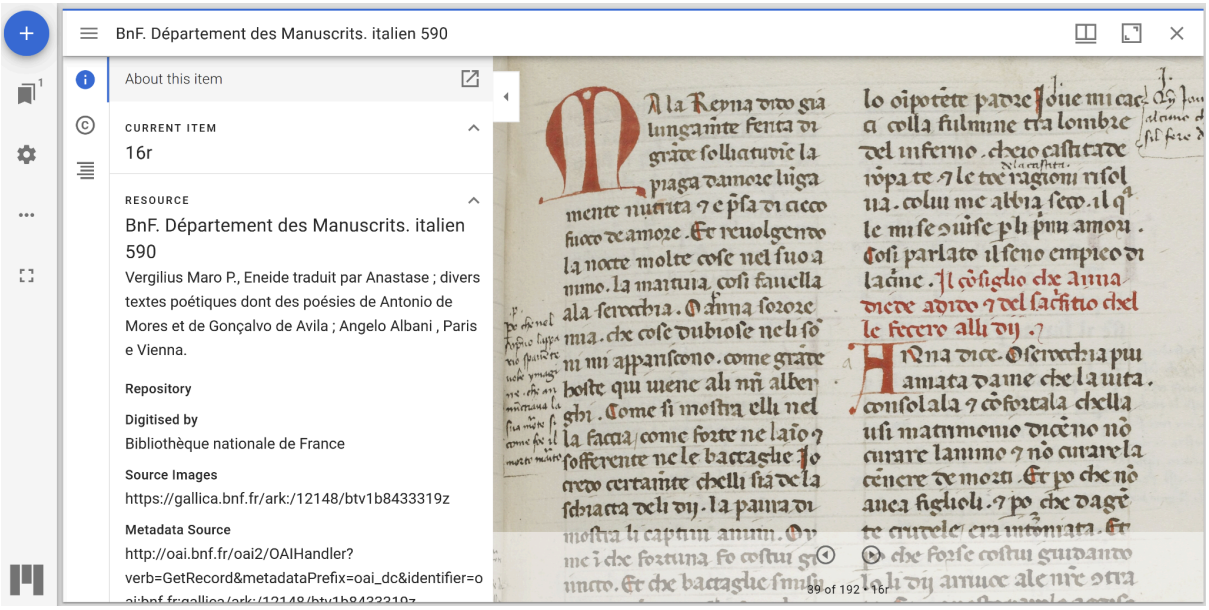
[52] Mirador website: https://projectmirador.org/

There are several versions of Mirador, and in this project the most up-to-date version is proposed: Mirador 3.0[53]. This tool integrates several existing software libraries, attempting to combine these resources to create a versatile image viewer. In particular, it incorporates the OpenSeadragon[54] viewer, known for providing viewing, zooming and moving images. Using OpenSeadragon ensures a smooth, high-quality viewing experience.
Another central component is the window manager Isfahan.js[55]. It makes it possible to open, resize and move windows on the computer screen. Mirador reuses this component to allow users to open and organise several image viewers within the same browser window. This is one of the main functionalities of this tool. If a collection of images is displayed, intuitive controls are provided directly in the user interface for exploring it: left and right pages, and a sidebar with thumbnails of other pages.
The latter can be compressed to save screen space for displaying the actual image (figure 10). If a collection has appropriate metadata, Mirador will also display an index or list of the contents, provided it is configured correctly, and the information contained in the IIIF manifest (figure 11).



*(**Fig. 10** Mirador, usage of multiview windows open for the alignment of 2 folios, 16r on the right and 16v on the left, and sidebar with other images on the left)*

[53] Mirator Github page: https://github.com/ProjectMirador/mirador/wiki
[54] OpenSeadragon Github page: https://openseadragon.github.io/
[55] Isfahan.js Github page: https://github.com/aeschylus/Isfahan



*(**Fig. 11** Mirador, high quality zoom and info from the IIIF manifest)*

The comparison of codex and manuscripts is a common activity in philology, which is why it is advantageous to have multiple codexes simultaneously displayed on the screen.
The developers themselves put it this way: "Users, such as scholars, researchers, students, and the general public, need to compare images hosted in multiple repositories across different institutions. They want a best-in-class experience with deep zoom capabilities, and viewing modalities optimised for single images, books and manuscripts, scrolls, or museum objects. End users want to create and view image annotations, comments, and transcriptions within a single user interface, regardless of the system in which they were originally created or hosted" (Sanderson et al., 2015).

For these reasons, Mirador is well established in the field of digital humanities and widely used by museums, digital archives and cultural institutions wishing to offer their online audience a tool to visualise their collections at very high quality levels.

4.4
Text visualisation

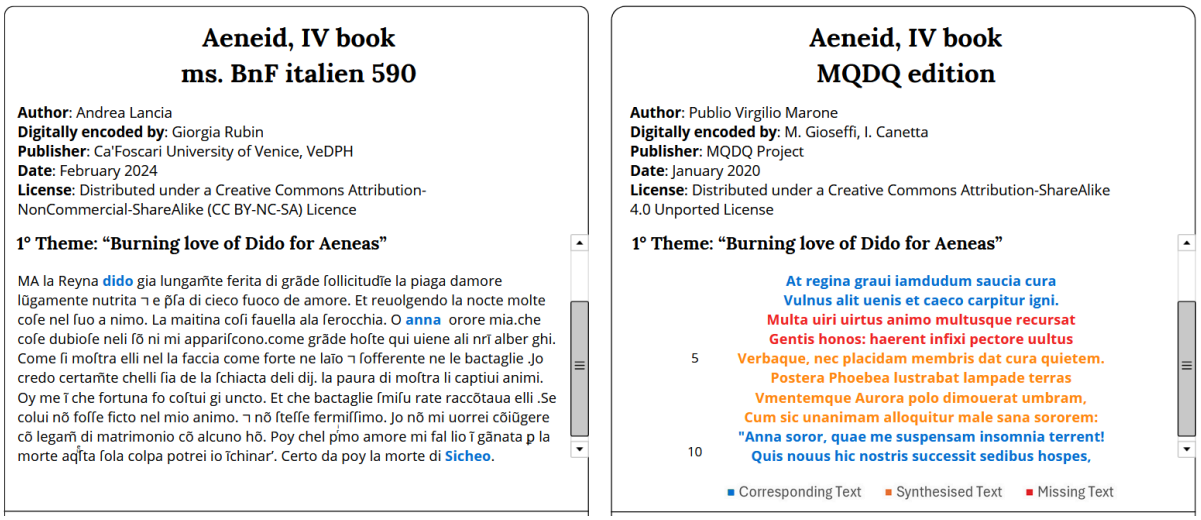
“Both print and digital text is represented visually for reading, and typography is about the graphical representation of characters in a particular medium. In this simple sense, text is already a type of visualization, an instantiation of a more notional text that is not concerned with specificities like page numbers or scrolling position. Emphasizing displayed text as visualization has the benefit of allowing us to take into account a full spectrum of text visualizations. Consider a text with only slight stylistic changes, such as having all adjectives displayed in green. Is this a text or a visualization? It is both.”
This is how Sinclair and Rockwell express themselves talking about text analysis and visualisation in their article "Text Analysis and Visualisation: Making Meaning Count" (Sinclair and Rockwell, 2015). Adopting information visualisation techniques strategically in

order to bring out the characteristics of a text means transforming that text into a new visualisation. This principle has been applied to the formulation phase of the visualisation of encoded text. We refer both to the XML/TEI coding of the manuscript text developed ad hoc for this project and to that provided by the MQDQ digital archive.
There are a considerable number of reusable text visualisation software tools that originated within academic contexts in the field of digital humanities. These tools were developed with the aim of producing digital editions and assisting philologists in the coding and implementation of text visualisations. Among these we can mention Anastasia (Robinson, 2002), a project that was later transformed into a publishing house called SDE Publisher, and EVT (R. del Turco, 2019), designed within the context of the University of Turin. In both cases, these tools are widely used in the context of the development of digital scholarly editions and visualisation of XML/TEI encoding. Tools such as these help in the production of serial digital editions. This term, re-proposing the concept expressed by Elena Pierazzo in "Quale infrastruttura per le edizioni digitali? Dalla tecnologia all'etica" (Pierazzo, 2019, 9), identifies those digital editions that exploit the practicality of tools such as those just mentioned, characterised by an iterative and continuous nature, i.e. by the presence of serial and cyclical features in their structure and graphics. On the contrary, Pierazzo, identifies specialised editions as those that adopt specific scientific or editorial requirements in both their structure and graphics. The work conducted in this project on the visualisation of the modelled text is similar in its characteristics to those belonging to specialised editions.
In the context of this thesis, the visualisation is presented in its unrealised conceptual form. A prototype is provided that faithfully reflects the graphical aspects of the final product.
This is a transformation via XSLT[56] of the XML/TEI style sheet into an HTML document. In the transition to HTML, the information to be displayed is selected, optimising the transformation process and focusing only on the layout and display aspects. This leads to the specialised display of the text of the manuscript BnF ta. 590 and that of the codification proposed by the MQDQ critical edition. This visualisation can be considered specialised because the visualised contents and the way they are displayed, enriched with graphical features, have been specially selected for this specific project.

OBJECTIVES AND CHARACTERISTICS OF VISUALISATIONS

Among the information displayed first is that encoded within the TeiHeader. In fact, the title of the text we are going to read is immediately indicated, pointing out that it is the 4th book of the Aeneid and that the two texts belong to two different versions of that book, one being the ms. BnF ita. 590, the other is the critical edition MQDQ.
The title is then followed by those data considered of particular relevance: these include the author of the text, the catalogue identification number and its location in the case of the manuscript, the publication licences of both texts and the author of their digital encoding.

[56] XSLT description made by W3C: https://www.w3.org/TR/xslt-30/



*(**Fig. 12**, on the left: visualisation of ms. BnF 590 text; on the right: visualisation of the critical edition MQDQ text)*

With regard to the visualisation of the text contained in the manuscript (figure 12 on the left), the features that were highlighted during the previous annotation and enrichment phase of TEI encoding, using the Euporia tool, are displayed graphically. These are the annotations relating to the entities mentioned within the text and their reporting. In this visualisation they are highlighted with the use of specific colours, enabling the reader to immediately identify the characters and places mentioned and their position within the text. The colours assigned to each entity correspond to their category described with Euporia. They are: blue for person, orange for Roman god, green for abstract entity, red for group of people and purple for place. Furthermore, the text is divided into thematic paragraphs, outlined during the analysis and synthesis of the text. Each paragraph is preceded by a title assigned during this process. This subdivision does not faithfully reflect the original text, as the purpose of the visualisation is not to accurately reproduce the graphic characteristics of the manuscript. Rather, the aim is to highlight salient features of the text, identified during analysis and noted during encoding and modelling. This allows the reader to effectively grasp the thematic content of the text while reading.

In a similar manner, the text of the critical edition was elaborated (figure 12 on the right). Following the same synoptic structure, the Virgilian text was also subdivided into thematic paragraphs, each preceded by its title. However, the objective of this textual visualisation is different from that of the manuscript visualisation. Here, the aim is to show the reader which parts of the original text have been fully, partially or omitted in the reduced version of the manuscript. This comparison is made possible through the use of colours: blue for the corresponding portions of text, red for those omitted and yellow for those summarised.

This approach aims to provide the reader with an agile method of comparing the two texts, enabling him to grasp the peculiarities of the reduction of Book IV of the Aeneid in the manuscript text. A further advantage of this visualisation lies in the ease of reading the manuscript text, which greatly facilitates the reader in this process. Otherwise, the reader

would have to make an effort to understand the contents of the manuscript text by interpreting the author's handwriting by observing the images of the text displayed by the Mirador tool.

4.5
Quantitative data visualisations

The detailed analysis provided on the text of the 4th book of the manuscript BnF ita. 590 compared to the one of the same book of the MQDQ critical edition of the Virgilian poem led to the compilation of several Excel files containing quantitative data extracted from the text. These data allow us to answer two research questions, namely:

- How many times is each entity mentioned within each of the thematic paragraphs with which the manuscript text has been divided?
- In what percentage did the manuscript text, being a reduction of the Virgilian text, faithfully reproduce, synthesise or omit the text of the original epic poem?

The aim of this phase of the project is to implement visualisations based on these data that will enable the reader of the text to easily understand the results of the Exploratory Data Analysis (EDA) produced. According to Prazad Patil's definition (Patil, 2018) in Towards Data Science, EDA "refers to the critical process of performing initial investigations of data in order to uncover patterns, detect anomalies, test hypotheses, and verify hypotheses with the help of summary statistics and graphical representations".

To produce these visualisations, thought to be displayed subsequently to the text visualisations discussed above, the seven stages for data visualisation listed by Benjamin Fry in his text Visualizing Data (Fry, 2007) were taken into account.
The first four (Acquire, Parse, Filter, Mine) were fulfilled in the data modelling stage extensively discussed in chapter two, the last three were instead processed in this stage of the project:

- Represent: choose a basic visual model and draw the data;
- Refine: improve the basic representation to make it clearer, more meaningful and more appealing;
- Interact: add methods to manipulate data or control which features are visible.

PLOTLY CHART STUDIO

To implement the visualisations, the tool offered by Plotly Chart Studio[57] was used, selected for its ease of use.
It is in fact an online environment specifically designed for the creation and customisation of interactive charts. Based on the Plotly[58] visualisation library, the platform framework

[57] Plotly Chart Studio website: https://chart-studio.plotly.com/
[58] Plotly library: https://plotly.com/python/

incorporates Python libraries, including Plotly and Dash[59], to facilitate the generation of advanced data visualisations. It uses a no-code approach, an extremely advantageous feature that allows Chart Studio users to develop complex visual representations of data without requiring advanced programming skills. It features an intuitive graphical user interface that offers drag-and-drop functionality, allowing users to modify crucial aspects of charts such as colours, labels and styles directly in the platform.
A further advantage of this platform is its ability to support a wide range of data formats for import and visualisation, including Excel files (directly loadable from .xsl or .xlsx spreadsheets), CSV files and JSON files. It is also possible to connect Plotly Chart Studio directly to databases or integrate it with external data APIs.
Finally, after saving the produced work, it is possible to download it either as an image in PNG, PDF, SVG, EPS formats, or as an HTML file or zip archive to view and explore offline the produced chart with all its interactive functions.
All these reasons have made Plotly Chart Studio an excellent candidate for the implementation of our visualisations.

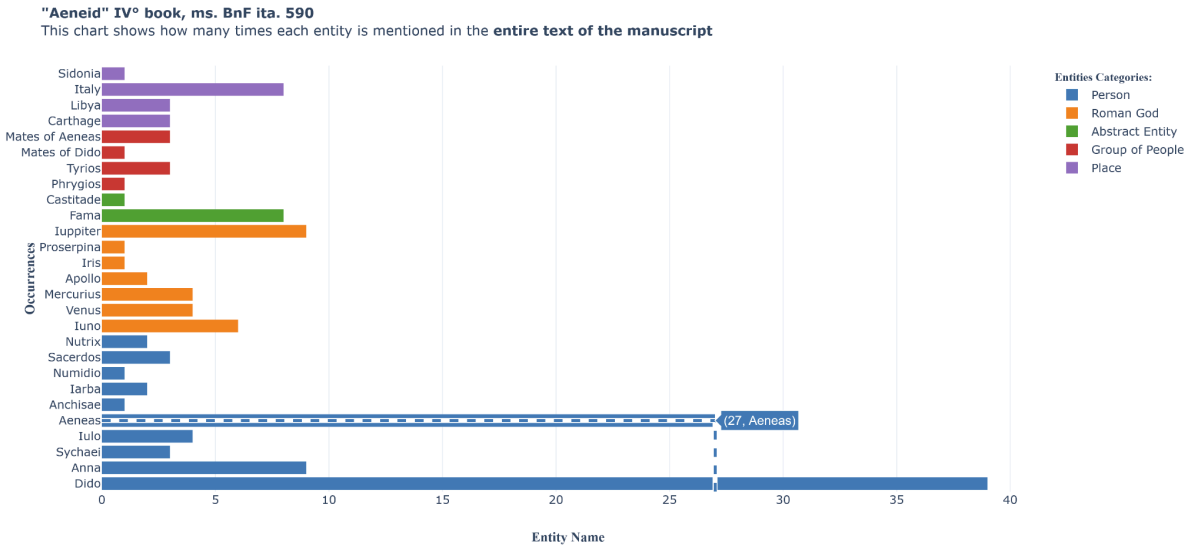
In addition, the numerous possibilities for customising graphics offered by this platform made it possible to produce visualisations that comply with the three principles of good visualisation described by Andy Kirk in Data Visualisation: a handbook for Data Driven Design (Kirk, 2019):

- Trustworthy: A data visualisation must be reliable and accurate, ensuring that data is correctly represented without distortion or manipulation. Users must be able to trust that the visualisation accurately reflects the information in the underlying data. Transparency regarding data sources, analysis methodology and clarity in annotations help to build this trust.
- Accessibility: Visualisation must be accessible to a wide range of users, including those with disabilities. This principle emphasises the importance of ensuring that visualisation is understandable by everyone, regardless of any sensory or cognitive limitations. To guarantee the principles of accessibility, guidelines have been defined, the WCAG (Web Content Accessibility Guidelines)[60], developed by the WAI (Web Accessibility Initiative) of the W3C (World Wide Web Consortium).
  In the area of visibility, the use of contrasting colours, clear labelling, and an understandable structure are key elements for improving accessibility.
- Elegance: Although simplicity may be an important aspect, the principle of elegance is not only about the visual form of the visualisation, but also about clarity and effectiveness in conveying the message. A clean, aesthetically pleasing design can help make the visualisation more engaging and memorable.

[59] Dash library: https://dash.plotly.com/
[60] Web Content Accessibility Guidelines (WCAG): https://www.w3.org/TR/WCAG21/

BAR CHART

In order to answer the first of the two questions (how many times is each entity mentioned within each of the thematic paragraphs with which the manuscript text has been divided?), multiple graphs were produced: each one suitable to represent the number of occurrences of each entity in each of the thematic paragraphs and a graph showing the amount of citations of each entity within the entire IV° Book of the ms. BnF ita. 590 (figure 13).
Since they are quantitative elements that need to show a comparison between them, the best solution is to use a bar chart. Possible alternatives to this chart can be the pie chart and the tree map. However, these are less immediate visualisations to understand in this situation.



*(**Fig. 13** This bar chart shows how many times each entity is mentioned in the entire text of the manuscript)*

In the barcharts produced, the principles of good visualisation were satisfied as follows.

With regard to trustworthiness, a title and a subtitle have been inserted to make explicit what the reader is observing in the graph; concise and effective titles have also been attributed to each of the axes and to the legend; finally, vertical parameters have been added to highlight references to x-axis values.
As far as accessibility is concerned, the graphic has been oriented horizontally to allow the reader to easily read the name of the entities displayed on the y-axis; considering the Web Content Accessibility Guidelines, bright, high-contrast colours and an associated legend have been used to associate each entity with its category (there are five categories in question, namely: Person, Roman God, Abstract Entity, Group of People, Place). In addition, the graphs are composed of an interactive component, which not only makes the visualisations more engaging, but also contributes to increasing accessibility: by passing the cursor over the bar relating to each entity, a descriptive label appears that reiterates the name of the selected entity and the precise numerical value with which it is cited within the text. It is also possible to deactivate and reactivate the display of the various categories by simply selecting them

from the legend, so that the interested reader can either explore only the entity categories in which he or she is interested or view the complete picture.
In addition, as far as elegance is concerned, the style with which the graphs are presented is minimal and extremely clear, displaying only those elements that are strictly necessary for the reader to understand the content without the risk of being distracted by superfluous elements.

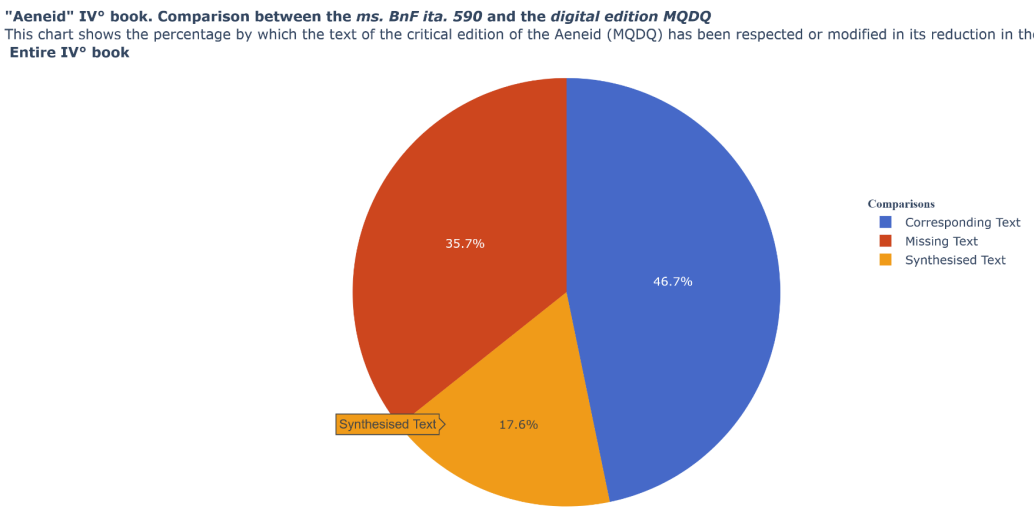
RESULTS DISCUSSION

Looking at the resulting bar charts, stored in a Drive folder[61] through which they can be downloaded, viewed and explored interactively in html format, it is therefore possible to know which characters are mentioned within the fourth book of the Aeneid, noting among them those who are the protagonists. It is immediately understandable how, in this book of the Virgilian poem, the central character of the story is not the protagonist of the poem but Dido. In fact, this is the book of the Aeneid in which the love story between Aeneas and the queen of Carthage takes centre stage, stealing the scene from the protagonist of the Aeneid and becoming the central character of the book.
Moreover, by looking at the graphs, it is possible to understand at first glance the places where the story described is set, without having to read the entire text. Also through a quick reading of these graphs, it is possible to understand in which theme each character and place is mentioned, thus studying their presence within the text.

PIE CHART

Also to answer the second question (in what percentage did the manuscript text, being a reduction of the Virgilian text, faithfully reproduce, synthesise or omit the text of the original epic poem?), multiple charts were produced: each corresponding to a thematic paragraph and one dedicated to the representation of the overall picture of the entire fourth book of the Aeneid (figure 14). Since the data to be displayed are percentage data, it was decided to use the pie chart. Again, the data could have been visualised with bar charts, but pie charts were preferred as they are better suited to displaying percentage data.

[61] Public access drive folder from which to download the produced bar charts:
https://drive.google.com/drive/folders/15hcW0OYuthu7cO7zUrEo3uzyqCbg9H3a?usp=drive_link



*(**Fig. 14** This pie chart shows the percentage by which the text of the critical edition of the Aeneid (MQDQ) has been respected or modified in its reduction in the ms. BnF ita. 590)*

Again, all three principles of good visualisation were respected, similarly to what was done for bar charts.
Thus, with regard to trustworthiness also for pie charts, explicit and comprehensive titles have been added to the chart and legend. In addition, precise percentage values have been inserted within each of the pie chart slices.
With regard to accessibility, bright, high-contrast colours were used, respecting the chromatic semantics also used in the previously described text visualisation: the colour blue for data relating to the corresponding text, the colour red for data relating to the missing text, the colour yellow for data relating to the synthesised text. Here too, an interactive proposal has been implemented that allows the reader to deselect and re-select the categories of interest from the legend and to display a descriptive slice label within the pie chart to show the category to which it refers.
Finally, as far as elegance is concerned, a minimalist style was used here as well, allowing the reader to understand the content of the displayed graph without the risk of being distracted by superfluous elements.

RESULTS DISCUSSION

Looking at the resulting pie charts, stored in a Drive folder[62] through which they can be downloaded, viewed and explored interactively in html format, it is evident that much of the Virgilian text represented by the MQDQ edition was not adhered to by the author of the manuscript BnF ita. 590, who produced a major reduction on the original text.
As can be seen, the themes related to the "Evening meal" (3rd theme), the "First night fall" (4th theme) and the "Sunrise" (6th theme), have been completely omitted in the manuscript reduction. On the other hand, the theme related to "Mercury orders Aeneas to definitely leave

[62] Public access drive folder from which to download the pie charts produced:
https://drive.google.com/drive/folders/1S8o_4KluMU-SgkP3aqwnj_zYdLeK9jj-?usp=drive_link

Carthage" (16th theme) has been completely carried over, while the theme "Dido and Aeneas take refuge in the cave" (7th theme) has been largely synthesised as has the theme "Dido sends Anna to Aeneas to convince him to stay" (13th theme). As for all the other thematic paragraphs, each of them has been reduced in more or less substantial parts by eliminating, synthesising and reporting portions of the text.
The expert scholar who delves philologically and literally into the contents of the manuscript text, from the implemented analyses and visualisations thus made available to him, will be able to deduce some rather interesting conclusions, wondering about the possible reasons why the authors of the manuscript text preferred to omit or summarise some of the parts of the Virgilian text, perhaps out of mediaeval censorship, perhaps to save precious ink.

This is the precious result of the collaboration between the digital humanist who elaborates the useful tools for the purposes of computational analysis and the domain expert humanist who can adopt them to formulate interesting hypotheses that enrich the understanding and value of the objects that the past has left us, consequently enriching the image we have of it.

4.6
GraphDB visual graph tool

Information visualisation and the Semantic Web can mutually enhance each other in addressing crucial aspects related to the structuring and retrieval of extensive information repositories.
RDF schema is one of the most popular formats for exchanging semantic data because it offers a powerful structure for representing information in a structured and relational way but its verbose representation can be complex for humans to fully understand. To overcome this challenge, it is important to translate this complex information into a visual format that is intuitive and easy to interpret. It is for this reason that a highly dynamic domain within the Semantic Web involves the creation of diverse tools dedicated to authoring, extraction, visualisation, and RDF inference.
Among the various RDF data visualisation techniques, the graph format proves particularly effective. This approach uses nodes to represent entities and directed arcs to denote the relationships between them. It is a visualisation that provides a clear and visually intuitive representation of the complex connections between different entities and their attributes, allowing users to easily understand the structure of the knowledge graph and the interactions between various parts of this system.

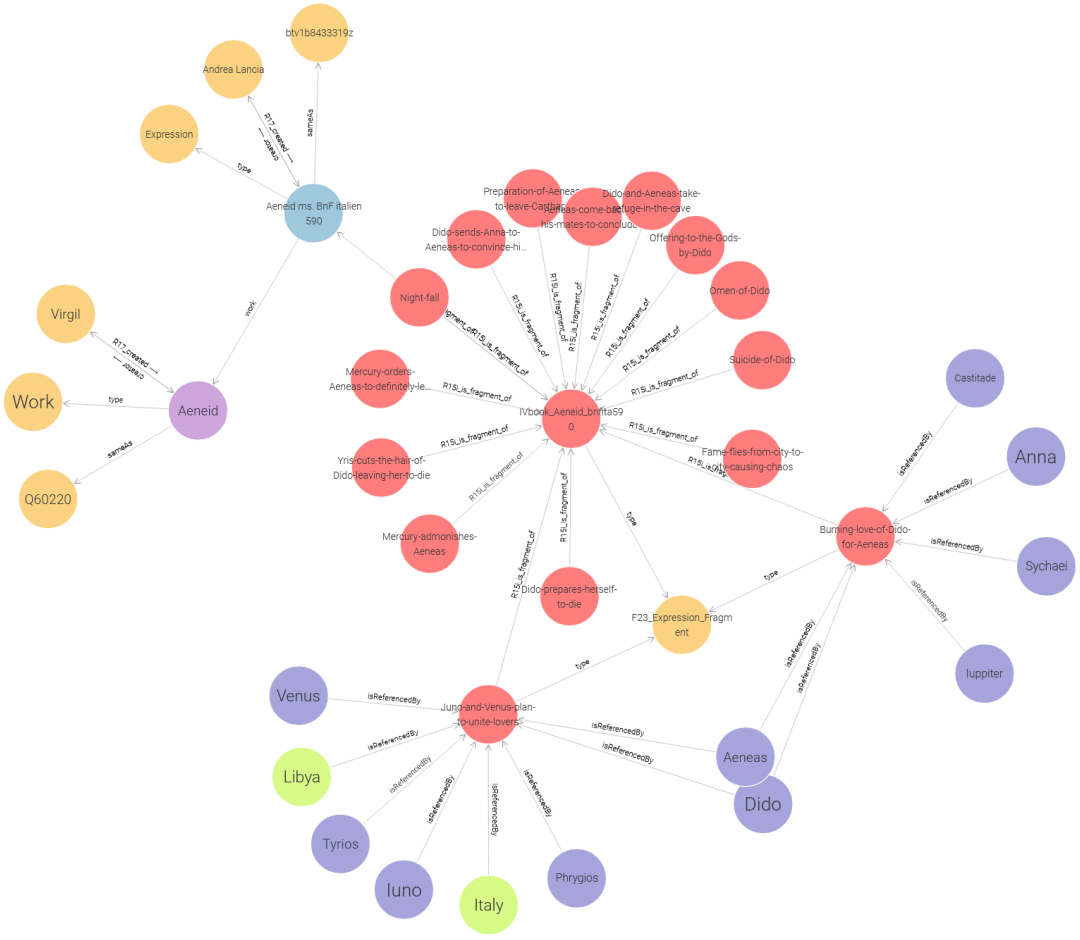
GRAPHDB

After considering several open source tools for the management and representation of our ontology model, it was decided to adopt the GraphDB[63] system. This is an RDF graph database management system developed by Ontotext[64], a company specialising in semantic

[63] GraphDB website: https://graphdb.ontotext.com/
[64] Ontotext website: https://www.ontotext.com/

technologies. This software offers a powerful environment for storing, querying and analysing complex RDF data and allows users to store and manage large amounts of semantic data in an efficient and scalable manner. Furthermore, by supporting the SPARQL language, GraphDB allows users to query RDF data and extract the information of their interest from the system.

This software offers a graph visualisation tool that allows the data to be explored visually in an intuitive manner. RDF models can thus be represented like the graph in figure 15, with nodes representing the entities, each of a different colour according to the class to which it belongs, and arcs representing the relationships between them, which can be interpreted by the direction of the arrows indicating the order in which they are read, making it easy to understand the complex connections in the model.



*(**Fig. 15** Part of the RDF ontology visualised with the GraphDB visual graph tool)*

This visualisation tool also offers interactive exploration capabilities.
It is possible to execute SPARQL queries to extract selected data from the serialised model and see the results displayed as a graph.

In the example below, a query has been expressed that returns as a result all attributes and links present for the Dido entity:

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX dcterms: <http://purl.org/dc/terms/>
PREFIX agrelon: <http://example.org/agrelon/>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
SELECT ?predicate ?object
WHERE {
```

*(**Snippet 12** SPARQL query to retrieve data from RDF/XML file)*

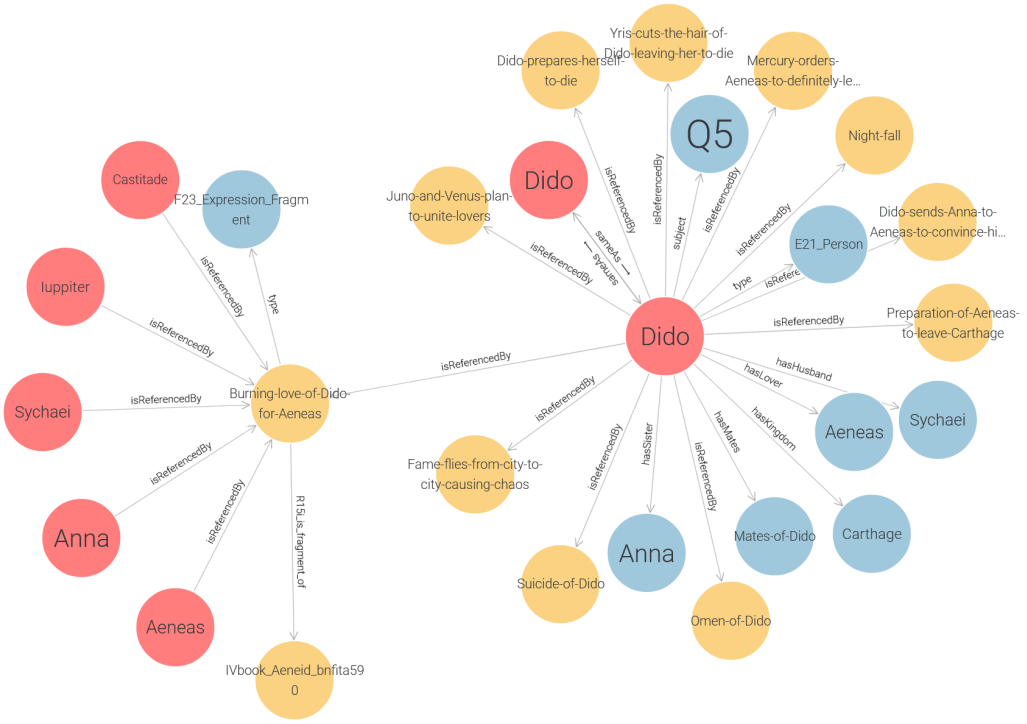
The RDF/XML serialisation that expresses these properties is the one that can be read in chapter three, paragraph three of this thesis.

The result displayed graphically with GraphDB is as follows (figure 16):

*(**Fig. 16** In this figure the visual graph displays the relationships that the entity "Dido" has with others, as described in the RDF ontology)*

Interactivity with the graph is also enabled by the function of exploring node properties through the selection of the node of interest and the automatic return of links. This is an operation that favours the serendipitous search of the contents expressed in the ontological model.

In the example below (figure 17), starting from the result obtained from the research described above, the relationship between elements and classes described by the RDF triples shown in paragraph 3.3 was explored, i.e.: Dido is present as an element in the class 'Burning love of Dido for Aeneas' which in turn is an element of the class 'IVbook_Aeneid_bnfita590'.



*(**Fig. 17** In this figure is explored the relation between the entity "Dido", the paragraph in which it's mentioned "Burning love of Dido for Aeneas" and the relation that the latter has with the entity "IVbook_Aeneid_bnfita590")*

Through this tool, the reader of the new edition of our manuscripted book will be able to explore the social relationships inherent between the characters known from previous visualisations. They will also be able to visualise and understand the FRBR structure (discussed in Chapter 3) that characterises this text and the manuscript in which it is contained.
In short, he will be able to understand through this network of connections the knowledge base describing the main features of the text he has read and analysed.

4.7
Entities datasheets

An additional visual representation was conceived with the idea of further exploiting the ontological scheme outlined. The latter in fact, thanks to the use of Linked Open Data, extended the knowledge base of the digitised text, going beyond the boundaries of the textual content through the links between the entities described in it and the authority files present online containing relevant information on their characteristics.
This section presents a potential graphical structure of this visual representation. Its actual realisation would require the implementation of a software structure capable of making SPARQL queries via HTTP requests to the relevant databases in order to display the content obtained in response. Even though this visualisation has not been implemented, its structure will be described below as it has been imagined.

This is an info box designed as a container of information, which is intended to be displayed in response to the stimuli given as input when navigating the visual graph described above.
The entities outlined in the graph presented have, among other properties, a direct link to the respective pages on Wikidata or DBpedia.
In the latter, a wide range of information is available, including biographical data in the case of individuals and geographical data in the case of places. In addition, it is possible to find images depicting the described entities, provided by cultural institutions.
By selecting from the information proposed only those of greatest relevance and inserting them into the visualisation, it will be possible to propose a summarised datasheet, a sort of identity card on the entity selected in the graph.

Figure 18 gives an example of how this datasheet might be displayed. The example illustrated goes deeper into the entity "Dido".

(***Fig. 18*** *Datasheet of entity "Dido"*)

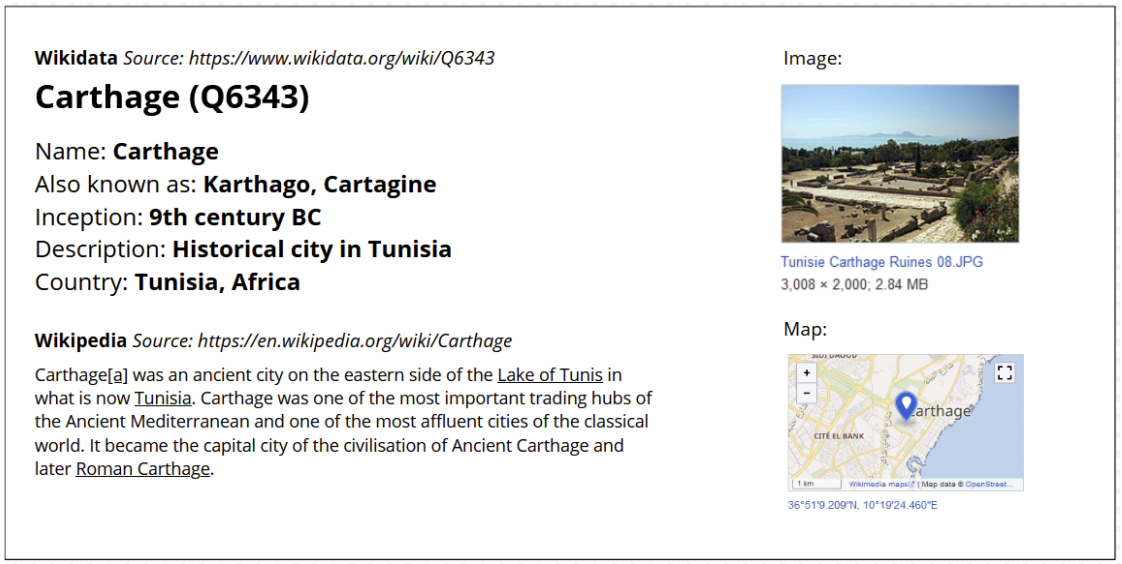
As can be seen, the source from which the collected information came is immediately reported. Together with the source, the unique code with which the entity was disambiguated and identified in the authority file is reported; in this specific case, the entity Dido is present on Wikidata with the code Q905162[65].
Next, the properties of the entity are reported in a list structure, which in this case, since it is a person, include:

- the name of the entity
- the possible variants of the name in different languages or textual sources;
- the sex and gender of the character;
- a brief description that can briefly outline the profile of the entity;
- its provenance.

[65] Dido Wikidata page: https://www.wikidata.org/wiki/Q905162

The entity's profile is then deepened with an excerpt from Wikipedia.
In addition, as can be seen, an image depicting the entity from Wikidata is included, along with any associated metadata.

In the case where the datasheet delves into a geographical entity, this could present itself as in figure 19. Here is presented the entity "Carthage".



(***Fig. 19*** *Datasheet of entity "Carthage"*)

Also in this case, the source from which the collected information came, Wikidata, is reported, and together with it the unique code with which the entity was disambiguated and identified in the authority file, Q6343[66].
The properties that can be listed in the case of geographical entities are:

- the name of the city,
- the possible variants of the name in different languages or textual sources,
- time when the entity begins to exist,
- a short description that can briefly outline the profile of the entity,
- the country and state in which the entity is located.

Again, the entity's profile is deepened with an excerpt from Wikipedia and an image depicting the entity from Wikidata is included, along with any associated metadata.
Considering the geographical nature of the entity, a geographical map provided by Wikidata is included. This tool allows the user to orient themselves geographically and gain a visual understanding of the location of that entity.

In this way, the user would have the opportunity to enrich the information explored in the ontology graph, deepening interesting aspects related to the entities investigated. Such an approach would contribute to a more complete interpretation of the ontology's contents and,

[66] Carthage Wikidata page: https://www.wikidata.org/wiki/Q6343

consequently, of the text's knowledge base.
Picking up on the concept of the agile interpretative cycle proposed by Sinclair and Rockwell, the user, after having explored the proposed visualisations of the strictly text-related contents and integrated the visualised information with that coming from external authoritative file, would acquire the ability to reread and reinterpret the text with a considerably improved maturity.

Chapter 5
Interface design for data exploration

In this final chapter is presented the design of the web interface conceived to host the previously described visualisations. It is a prototype that aims to provide a faithful preview of the aesthetics of an application capable of offering users a web environment composed of a set of tools useful for consulting the digitised text of the fourth book of the manuscript BnF ita. 590 and the exploration of its contents.
Wireframe and mockup are presented to illustrate the proposals regarding the layout of the elements and the visual structure of the interface.

5.1 User Interface and User Experience

The interface represents the point of convergence between users and technology, therefore, its design emerges as a recurring necessity in digital humanities projects. Occupying a central position, it acts as a mediator, aimed primarily at facilitating communication between people, technology, content and the abstract concept being studied and visualised.
Gino Roncaglia, in his text entitled "La quarta rivoluzione, sei lezioni sul futuro del libro" (Roncaglia, 2011), identifies two different types of interface: the physical interface, which, in the electronic sphere, refers to the hardware as the carrier of information; and the logical interface, which is the space in which this information is organised.
The logical interface materialises on the screen through information design. This aspect is commonly known as the User Interface[67] (UI), whose task is to organise information efficiently, taking into account the peculiarities of the physical medium, in order to make it accessible and usable to users.

It is therefore clear that the design of the web interface is not merely an aesthetic component of a digital humanities project, but a strategic component that facilitates access to and interaction with the information contained in an environment that fosters understanding and exploration of cultural resources. When designing a web interface, it is therefore crucial to consider not only the objectives one intends to achieve with the final design, but also the needs of the user in order to guarantee an effective and satisfying User Experience[68] (UX).
To be effective, the interface as an information access system must be characterised by certain fundamental qualities. It has to be intuitive and easy to navigate, presenting a logical organisation of information. Content and text must be clear and readable, with a visual presentation that utilises appropriate colours and typography to aid understanding of the information. Visual design is also important to create an engaging user experience.
These characteristics are properties that belong to the sphere of usability. The latter is defined

[67] Definition of User Interface (UI) proposed by the Interaction Design Foundation, https://www.nngroup.com/articles/definition-user-experience/
[68] Definition of User Experience (UX) proposed by Don Norman and Jakob Nielsen in "The definition of User Experience (UX)", 1998, https://www.nngroup.com/articles/definition-user-experience/

by the international standard ISO 9241-11[69] as the "extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use." According to the guidelines defined by this standard, usability must fulfil the characteristics of effectiveness, efficiency and satisfaction according to the context of use.

The characteristics described so far were considered for the design of the web interface of the project developed in this thesis.

5.2 VIVA interface: objectives and design

In order to give personality to the presented project, a meaningful name was devised with which to call it: VIVA. It is an acronym that plays with the characteristics of the modelled text, combining letters that refer to the following words: Vernacular, IV (fourth book), Aeneid. Furthermore, the word 'viva' has a semantic character in the Italian language. It can have two meanings: it can be translated as 'alive' but can also be interpreted as an exclamation, an expression of support and pride. As the author of the project, I consider both meanings appropriate: on the one hand thinking of the analysed text as a living and dynamic work that through this project acquires new life thanks to digitisation and the multiple visualisations that its new virtual readers could enjoy; on the other hand, the conclusion of a substantial and laborious project such as this one leads me to exclaim a joyful VIVA!

As already outlined, the interface of VIVA, which will soon be presented, aims to show the way in which the visualisations produced following the modelling of the contents of the digitised manuscript text have been organised.
VIVA is therefore aimed at the potential readers of this text and intends to present itself as an environment in which one can move easily to explore in a serendipitous and engaging manner the contents organised according to an ordered logical structure.
It presents a vertical structure: all visualisations have been arranged one after the other within the same page to allow the reader to conduct the exploration through an experience guided by the order in which the visualisations are presented.

WIREFRAME

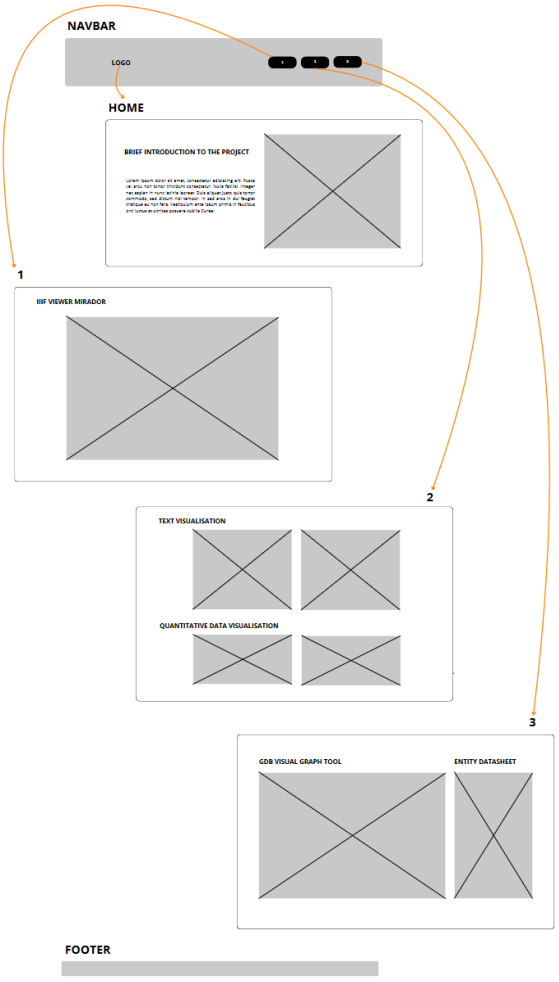
The wireframe presented in figure 20 provides a general overview of how VIVA was imagined. It is shown in its structure consisting of four sequentially ordered sections:

- The first section acts as both the homepage and cover page for this new version of the text and contains a brief introduction to the project;
- The second section, marked by number 1, hosts the Mirador tool (discussed in detail in section 4.3), which is central to displaying high-resolution images of the manuscript pages via the IIIF viewer;

[69] ISO 9241-11 Ergonomics of human-system interaction. Part 11: Usability: Definitions and concepts https://www.iso.org/standard/63500.html

- The third section, marked with number 2, is dedicated to visualising the text of the digitised manuscript and comparing it with the text of the critical edition (visualisations described in detail in section 4.4). In this section, directly following the two texts, the charts showing quantitative analyses (discussed in section 4.5) are displayed;
- The fourth and last section, marked with number 3, hosts the visual graph tool provided by GraphDB (analysed in section 4.6) and the entity datasheets (described in section 4.7).

Each of these sections is also accessible via shortcuts associated with the respective buttons in the horizontal navigation bar at the top of the page. This bar also contains the VIVA logo, which also acts as a button to take the user back to the first section, dedicated to the home. Finally, at the bottom of the page is a bar dedicated to the footer.



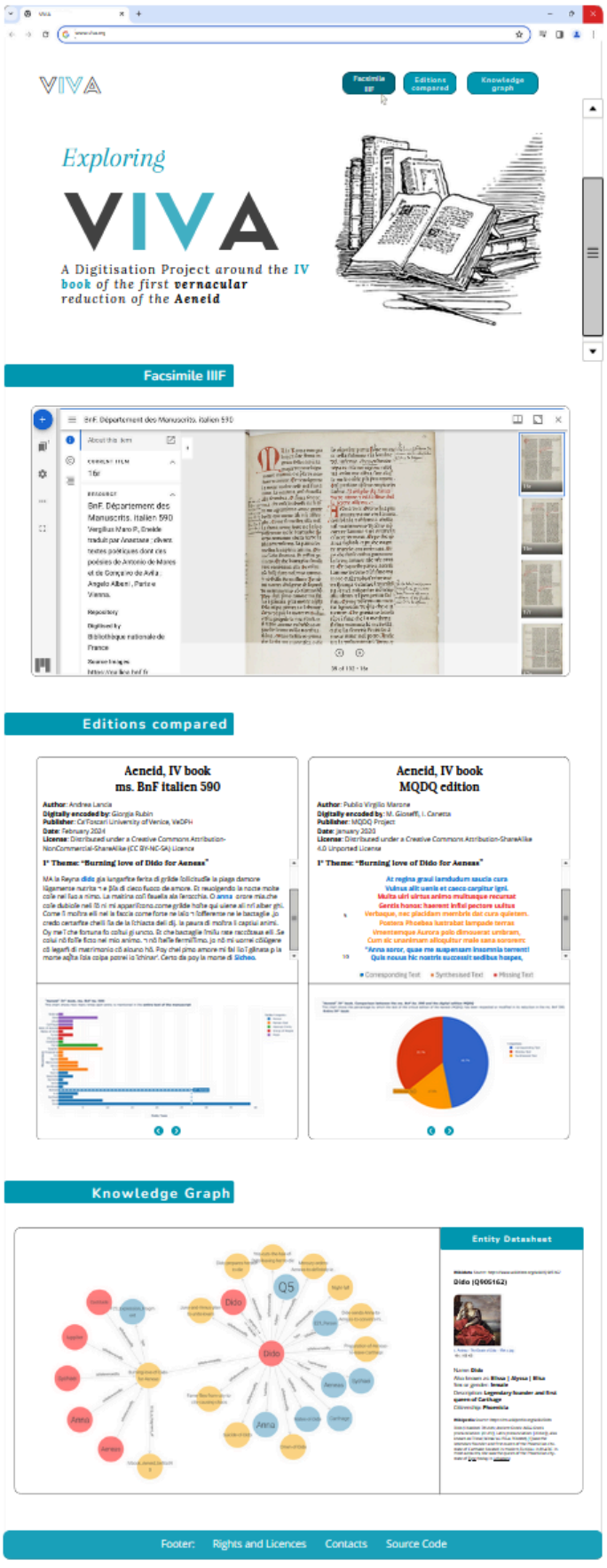
*(**Fig. 20** Low fidelity wireframe of VIVA interface)*

MOCKUP

Following the structure outlined in the wireframe, a high fidelity mockup was developed (figure 21) which aims to provide a more accurate representation of the desired product.

For the surface of the interface skeleton, high-contrast colours were chosen, including white, dark grey and petrol blue: white, dominant and bright, forms the main background, making the graphics light and clean; dark grey is used for most of the text, while petrol blue is used for buttons and section labels. Sections appear as described below:

- The top navigation bar includes buttons entitled "Facsimile IIIF", "Editions compared" and "Knowledge graph". Each title has been chosen to make the content associated with the button easily comprehensible, each called up by the labels placed at the beginning of the respective linked sections. A simple, minimalist logo is also included in the bar, which is subsequently called up on the homepage;
- In the home section, the meaning of the name VIVA and consequently also the content of the application is explicitly disclosed: 'Exploring VIVA. A digitisation project around the IV book of the first vernacular reduction of the Aeneid'. Next to this sentence, a picture of an open manuscript ready to be read is shown;
- The “IIIF Facsimile” section is entirely occupied by the Mirador tool, a choice made specifically to allow first of all an agile exploration of the content displayed thanks to the IIIF viewer, but also to exploit the potential of mirador to display the manuscript images in full screen and in very high resolution;
- The section “Editions compared” is divided into four quadrants: the two upper quadrants show the text visualisations of the two compared editions (on the left, the digitised one of the manuscript BnF ita. 590; on the right, that of the MQDQ critical edition). Both texts can be read in their entirety by scrolling down the vertical sidebar located to the right of each quadrant. The lower quadrants present the visualisations of the quantitative analyses conducted on the contents of the two texts. Their position has been purposely chosen: just below the text of the manuscript, bar graphs containing the occurrences of the entities mentioned in the text are displayed; below the text of the critical edition, instead, pie charts showing in percentages the same data presented in the text, i.e. the characteristics of the portions of the two texts compared (for more details, see the detailed description in section 4.5). The graphs in the two lower quadrants can be browsed and displayed by means of navigation buttons, which allow the user to explore the graphs relating to the specific thematic paragraphs, both previous and subsequent;
- The “Knowledge graph” section is also divided into two quadrangular portions. The rectangle on the left, occupying most of the space in this section, hosts the GraphDB visual graph tool and allows interactive exploration of the knowledge base and the entities present in the text and their reciprocal relationships. The rectangle on the right contains the space dedicated to hosting the data sheet of each of these entities, which can be viewed following the exploration of the node relating to the entity analysed;
- The footer bar contains additional information on the contact details of the VIVA developer, the publication licences adopted and a link to the source code of the application and the files processed for the development of the project.



*(**Fig. 21** High fidelity mockup of VIVA interface)*

Conclusions

With this thesis we have described the text digitisation paradigm applied to the text of the IVth book of the manuscript BnF ita. 590: the first prose vernacularisation of the Aeneid based on a Latin reduction of the ancient poem.
The steps constituting the workflow were described in detail. Starting with the analysis of the primary source, described in the first chapter, the text was first acquired from the images of the manuscript by applying automated manuscript character recognition (HTR) techniques. In the second phase, that of modelling the acquired data described in chapter two, the text was transcribed and encoded in machine-readable XML format applying TEI standards. Then, the third chapter describes the process and logic applied to the design of an RDF conceptual model capable of representing the knowledge base and the main characteristics of the text using some of the ontological vocabularies most in use in the domain of cultural heritage, and the serialisation of this model in RDF/XML format, useful for enriching the encoded text with data linked to authority files, thus expanding the knowledge base of the digitised text. The fourth chapter describes the process of implementing visualisations of all collected and modelled data, how they were rendered graphically with interactive visualisations implemented from scratch and some appropriate visualisation tools. Finally, in the fifth and final chapter is proposed the design of a web environment interface that allows the users to serendipitously and interactively explore the visualisations produced and strategically ordered to enable them to learn relevant content about the analysed text.

It is clear that the thesis project led to a significant transformation of the support of the analysed text. What originally existed as tangible material, in the specific case of the pages of IVth book of the manuscript BnF ita. 590 preserved in the Bibliothèque Nationale de France, is now presented in its new immaterial form through the result obtained in VIVA. It is important to emphasise that this practice is not an end in itself: as has been attempted to demonstrate throughout this thesis, the benefits of digitising text are multiple.
Digital tools applied to texts allow a proliferation of representations that are extremely useful for exploring their linguistic and semantic characteristics in depth. Thanks to these tools, we are able to generate new interpretations of the text. Even for texts with which we are already familiar, the use of digital tools can reveal elements of interest that previously escaped our attention, or deepen aspects that we had noticed but for which we did not have the appropriate tools to analyse them exhaustively (Sinclair and Rockwell, 2019).
The application of the technologies and logics of the semantic web make it possible to disambiguate the entities mentioned in the text, link them to external authority files and make explicit the relationships between the persons and places mentioned, thus extending the information horizons of the closed system of the text to external sources (Tomasi et al., 2019).
This technological contribution leads us to understand new associations, which in some cases may even generate completely new insights. It is clear, therefore, how digital tools are

intended to facilitate the augmented hermeneutic cycle, enabling navigation between text reading, analysis and visualisation at various scales (Sinclair, Rockwell, 2015).

Limitations & Future works

As has probably been noted, it has always been chosen not to use the term digital edition to refer directly to VIVA, but rather to call it a text digitisation project. This is because it is considered inappropriate to use the term digital edition in this context, which by definition implies a publication of content in the form of a new version of the text. No publication of VIVA has in fact been produced. The production of a digital edition would have entailed a preliminary assessment of the necessary costs and timeframe that would definitely have deserved to be examined in depth, and the selection of multiple methods involving techniques to automate the process would have been beneficial in terms of time and development. Furthermore, an edition that can be defined as not only digital but also scholarly must involve significantly more in-depth philological and literary domain analyses.
These are the main limitations of the presented project implemented so far. They can be met in the future with a more thorough analysis of the text, perhaps even extending the work to the entire manuscript BnF ita. 590, thus producing a DSE of the entire first prose vernacularisation of the Aeneid based on a Latin reduction of the ancient poem, leading to a work of great quality and significance for the study of the Italian language tradition and the transmission of the Virgilian poem.

By definition, projects belonging to the digital humanities domain, especially if public, find their final solution with the public dissemination of the developed content. Although VIVA was not distributed as a digital edition, the files generated for its development, which include modelled data and visualisations, were shared within public folders on Github[70] e Google Drive[71]. However, it should also be considered to publish these materials on other platforms or repositories, such as H2IOSC[72] (Humanities and Heritage Italian Open Science Cloud). Funded by NextGenerationEU & PNRR 'Italia Domani', H2IOSC aims to create an environment to foster collaboration between researchers in the humanities, language technologies and cultural heritage. It aims to create a single and simple access point, providing tools, datasets, and pilot projects to meet research needs, promoting accessibility and the FAIR approach.
National level platforms, such as the service offered by H2IOSC, and international level platforms such as Europeana, are of crucial importance in the dissemination of cultural content and digital humanities projects. They can provide considerable benefits to scholars and stakeholders alike, facilitating collaboration, access and sharing of cultural knowledge and resources locally and globally.
In the context of this thesis project, a text was modelled which, for a large-scale audience, is

[70] Github public repository with VIVA digitisation project files published under CC0-1.0 licence https://github.com/GiorgiaRubin/VIVA.git
[71] Google Drive public repository from which to download and access VIVA digitisation project html visualisations https://drive.google.com/drive/u/1/folders/1Dm7Po9jDZcI5YH5vtEPtrDmI4R9KFalK
[72] H2IOSC Humanities and Heritage Italian Open Science Cloud, website: https://www.h2iosc.cnr.it/

difficult to access due to the characteristics of the delicate manuscript support in which it is embedded, stored in a precise and remote physical location. Exploiting the potential of public access platforms such as those mentioned above, by publishing the work produced in compliance with licences and copyrights, would give a much broader audience the opportunity to consult this resource and explore the processed content.

References

Bibliography

Bambaci, L., Boschetti, F., & Del Gratta, R. (2019). Qohelet Euporia: a Domain specific Language for the Encoding of the critical Apparatus. *International Journal of Information Science and Technology*, *3*(5), 26-37.
https://publications.cnr.it/api/v1/documents/download/154167

Bertin, E. (2014). I tre volgarizzamenti dell''Eneide' in compendio: caratteristiche e rapporti tra i testi secondo le testimonianze antiche. *StEFI. Studi di Erudizione e Filologia italiana*, 3, pp. 5-58
http://www.studierudizionefilologia.it/stefi/it/anno-iii-2014

Börütecene, A (2011). Progettazione di thesauri online. Interaction, Interface, Information Design e case studies.
https://www.researchgate.net/publication/313098153_Progettazione_di_thesauri_online_Interaction_Interface_Information_Design_e_case_studies

Boschetti, F., & Mugelli, G. (2021). Il metodo Euporia per creare nuovi archivi digitali sulla tragedia greca. *FuturoClassico FCl*, (7), 83-113.
https://ojs.cimedoc.uniba.it/index.php/fc/article/viewFile/1381/1192

Card, S. K., Mackinlay, J., & Shneiderman, B. (Eds.). (1999). *Readings in information visualization: using vision to think*. Morgan Kaufmann.
https://hci.ucsd.edu/220/CMSChap1_Using_Vision_to_Think.pdf

Catarci, T., Guercio, M., Santucci, G., & Tomasi, F. (2014, January). Evaluating Cultural Heritage Information Access Systems. In *Bridging Between Cultural Heritage Institutions: 9th Italian Research Conference, IRCDL 2013, Rome, Italy, January 31--February 1, 2013. Revised Selected Papers* (Vol. 385, p. 7). Springer.
https://books.google.it/books?hl=it&lr=&id=d8y5BQAAQBAJ&oi=fnd&pg=PA7&dq=francesca+tomasi+information+visualization&ots=uhWbWloUEm&sig=n0ols596Ekxjt9Xy_7cRWZ-K0xU#v=onepage&q&f=false

Chagué, A. (2022). eScriptorium : une application libre pour la transcription automatique des manuscrits. *Arabesques*, 107.
https://publications-prairial.fr/arabesques/index.php?id=3100

Ciula, A., & Eide, Ø. (2017). Modelling in digital humanities: Signs in context. *Digital Scholarship in the Humanities*, *32*(suppl_1), i33-i46.
https://doi.org/10.1093/llc/fqw045

Ciula, A., & Eide, Ø. (2017). Modelling in digital humanities: Signs in context. *Digital Scholarship in the Humanities*, *32*(suppl_1), i33-i46. https://opus.bibliothek.uni-wuerzburg.de/opus4-wuerzburg/frontdoor/deliver/index/docId/11127/file/flanders_jannidis_datamodeling.pdf

Crucitti, M., Benedetti, M., Mirandola, R., Maneschi, G., Soldani, A., Amato, L., ... & Boschetti, F. (2021). La collaborazione inclusiva: un'esperienza didattica di annotazione tramite Euporia. *Umanistica Digitale*, (11), 145-162. https://aiucd2021.labcd.unipi.it/wp-content/uploads/2021/05/AIUCD2021_BOA-versione3A.pdf

Daquino, M., Giovannetti, F., & Tomasi, F. (2019). Linked data per le edizioni scientifiche digitali. Il Workflow di pubblicazione dell'edizione semantica del quaderno di appunti di Paolo Bufalini. *Umanistica Digitale*, (7). http://doi.org/10.6092/issn.2532-8816/9091

Driscoll, M. J., & Pierazzo, E. (Eds.). (2016). *Digital scholarly editing: Theories and practices* (Vol. 4). Open Book Publishers. https://www.openbookpublishers.com/product/483

Drucker, J. (2015). Graphical approaches to the digital humanities. *A new companion to digital humanities*, 238-250. https://doi.org/10.1002/9781118680605.ch17

Fernandelli, M. (2003). Virgilio e l'esperienza tragica. Pensieri fuori moda sul libro IV dell'Eneide. https://www.openstarts.units.it/server/api/core/bitstreams/80df2ead-1c65-4826-b7df-5341327f4c7b/content

Fiorentini, L. (2022). Congetture sul compendio virgiliano del frate Anastasio di Santa Croce. *Congetture sul compendio virgiliano del frate Anastasio di Santa Croce*, 99-102. https://www.torrossa.com/en/resources/an/5334182

Flanders, J., & Jannidis, F. (2015). Data modeling. *A new companion to digital humanities*, 229-237. https://doi.org/10.1002/9781118680605.ch16

Flanders, J., & Jannidis, F. (2018). Data modeling in a digital humanities context: an introduction. In *The shape of data in digital humanities* (pp. 3-25). Routledge.

Flanders, J., & Jannidis, F. (Eds.). (2018). *The shape of data in digital humanities: modeling texts and text-based resources*. Routledge.

Franzini, G., Mahony, S., & Terras, M. (2016). A catalogue of digital editions. Open Book Publishers. https://www.researchgate.net/publication/264205341_A_Catalogue_of_Digital_Editionsn

Fry, B. (2008). *Visualizing data*. " O'Reilly Media, Inc.".
https://media.espora.org/mgoblin_media/media_entries/1633/Visualizing_Data.pdf

Italia, P., & Bonsi, C. (Eds.). (2016). *Edizioni Critiche Digitali Digital Critical Editions: Edizioni a confronto Comparing Editions* (Vol. 34). Sapienza Università Editrice.
https://www.editricesapienza.it/sites/default/files/5369_Italia_Bonsi_EdizioniCriticheDigitali.pdf

Kirk, A. (2019). Data visualisation: A handbook for data driven design. *Data Visualisation*, 1-328.

Mancinelli, T. (2021). Per l'edizione scientifica digitale dei Documenti d'Amore di Francesco da Barberino: modelli, metodi e strumenti. In S. Bischetti & A. Montefusco (Ed.), *Francesco da Barberino al crocevia: Culture, società, bilinguismo* (pp. 65-90). Berlin, Boston: De Gruyter.
https://doi.org/10.1515/9783110590647-005

Clérice, T., Vlachou-Efstathiou, M., & Chagué, A. (2023). CREMMA Medii Aevi: Literary manuscript text recognition in Latin. *Journal of Open Humanities Data*, *9*, 4.
https://enc.hal.science/hal-03828353

Meirelles, I. (2011). Visualizing data: new pedagogical challenges. In Selected Readings of the 4th Information Design International Conference (Vol. 4).
https://isabelmeirelles.com/pdfs/isabel_SR4-2010.pdf

Norman, D. (2014). *Things that make us smart: Defending human attributes in the age of the machine*. Diversion Books.

Ware, C. (2019). *Information visualization: perception for design*. Morgan Kaufmann.
https://dl.acm.org/doi/book/10.5555/2285540

Pierazzo, E. (2016). *Digital scholarly editing: Theories, models and methods*. Routledge.

Pierazzo, E. (2018). Facsimile and Document-Centric Editing. *Creating a Digital Scholarly Edition with the Text Encoding Initiative*.
https://projects.history.qmul.ac.uk/wp-content/uploads/sites/6/2017/09/05-Digital-Facsimiles-EP.pdf

Pierazzo, E. (2019). Quale infrastruttura per le edizioni digitali? Dalla tecnologia all'etica. *Textual Cultures*, *12*(2), 5-17.
https://www.jstor.org/stable/26821533?seq=1

Pierazzo, E., & Mancinelli, T. (2020). *Che cos' è un'edizione scientifica digitale?*
https://www.carocci.it/prodotto/che-cose-unedizione-scientifica-digitale

Pinker, S. (1990). 4 A Theory of Graph Comprehension. *Artificial intelligence and the future of testing*.
https://www.researchgate.net/profile/Steven-Pinker/publication/213802830_A_theory_of_graph_comprehension/links/572b60ee08ae2efbfdbdd51f/A-theory-of-graph-comprehension.pdf

Reeve, L., Han, H., & Chen, C. (2006). Information visualization and the semantic web. In *Visualizing the Semantic Web: XML-Based Internet and Information Visualization* (pp. 19-44). London: Springer London.
https://link.springer.com/chapter/10.1007/1-84628-290-X_2

Reeve, L., Han, H., & Chen, C. (2006). Information visualization and the semantic web. In *Visualizing the Semantic Web: XML-Based Internet and Information Visualization* (pp. 19-44). London: Springer London.
https://doi.org/10.1007/1-84628-290-X_2

Ricotta, V., & Vaccaro, G. (2018). Rivolgarizzare e ritradurre. Parole, idee, traduzioni. *Annales Universitatis Paedagogicae Cracoviensis| Studia de Cultura*, *9*(237), 133-143.
https://rep.up.krakow.pl/xmlui/bitstream/handle/11716/9489/AF237--13--Rivolgarizzare--Ricotta--Vaccaro.pdf?sequence=1&isAllowed=y

Robinson, P. (2013). Towards a theory of digital editions. In *The Journal of the European Society for Textual Scholarship* (pp. 105-131). Brill.
https://doi.org/10.1163/9789401209021_009

Robinson, P., & Van Vliet, H. T. M. (2001). What is a critical digital edition. *Variants*, *1*, 43-62.
https://zenodo.org/record/6533168

Roncaglia, G. (2011). *La quarta rivoluzione: sei lezioni sul futuro del libro*. Gius. Laterza & Figli Spa.
https://books.google.it/books?hl=it&lr=&id=ZMGKDwAAQBAJ&oi=fnd&pg=PT11&dq=roncaglia+la+quarta+rivoluzione&ots=_Ke8nvZBWE&sig=Q-geyvjgpEC3OsthgJAA5H4AU1o&redir_esc=y#v=onepage&q=roncaglia%20la%20quarta%20rivoluzione&f=false

Ruecker, S. (2015). Interface as Mediating Actor for Collection Access, Text Analysis, and Experimentation. *A New Companion to Digital Humanities*, 395-407.
https://doi.org/10.1002/9781118680605.ch27

Rushmeier, H., Pintus, R., Yang, Y., Wong, C., & Li, D. (2015). Examples of challenges and opportunities in visual analysis in the digital humanities. *Human Vision and Electronic Imaging XX*, *9394*, 397-405.
https://doi.org/10.1117/12.2083342

Sahle, P. (2013). *Digitale Editionsformen*. Teil,1-3.

Eide, Ø. (2014). Ontologies, data modeling, and TEI. *Journal of the Text encoding initiative*,

(8).
https://doi.org/10.4000/jtei.1191.

Sahle, P. (2016). What is a scholarly digital edition?. *Digital scholarly editing: Theories and practices*, *1*, 19-39.
https://library.oapen.org/bitstream/id/9668bc0d-eb07-4b4f-a5e5-0cf335188694/633780.pdf

Sanderson, R., Snydman, S., Winget, D., Albritton, B., & Cramer, T. (2015). Mirador: A Cross-Repository Image Comparison and Annotation Platform. In *OR2015| 10th International Conference on Open Repositories. Indianapolis, IN, June* (Vol. 9).

Silberschatz, A., Korth, H. F., & Sudarshan, S. (1996). Data models. *ACM Computing Surveys (CSUR)*, *28*(1), 105-108.
https://dl.acm.org/doi/pdf/10.1145/234313.234360

Sinclair, S., & Rockwell, G. (2015). Text analysis and visualization: making meaning count. *A new companion to digital humanities*, 274-290.
https://doi.org/10.1002/9781118680605.ch19

Spadini, E., Tomasi, F., & Vogeler, G. (Eds.). (2021). *Graph Data-Models and Semantic Web Technologies in Scholarly Digital Editing* (Vol. 15). BoD–Books on Demand.

Tanturli G. (1986). Volgarizzamenti e ricostruzione dell'antico. I casi della terza e quarta Deca di Livio e di Valerio Massimo, la parte del Boccaccio (a proposito di un'attribuzione). *Studi medievali*, s. 328: 811–888.

Tomasi, F. (2013). Digital editions as a new model of conceptual authority data. *Digital editions as a new model of conceptual authority data*, 21-44.
https://www.jlis.it/index.php/jlis/article/view/240

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. *Scientific american*, *284*(5), 34-43.

Tomasi, F. (2017). L'informazione digitale e il Web semantico. Il caso delle scholarly digital editions. In *Informatica umanistica: risorse e strumenti per lo studio del lessico dei beni culturali* (pp. 157-174). Firenze University Press.
https://cris.unibo.it/handle/11585/611250

Tomasi, F., & Giovannetti, F. (2022). Linked data from TEI (LIFT): A Teaching Tool for TEI to Linked Data Transformation. *DIGITAL HUMANITIES QUARTERLY*, *16*(2), 1-14.
http://www.digitalhumanities.org/dhq/vol/16/2/000605/000605.html#figure04

Van Zundert, J. (2018). On not writing a review about Mirador: Mirador, IIIF, and the epistemological gains of distributed digital scholarly resources. *Digital Medievalist*, *11*(1).
http://doi.org/10.16995/dm.78

Zordan, E. (2023). Reshaping the figure of the Courtesan in a digital archive: a feminist case study on Veronica Franco.
http://hdl.handle.net/10579/23169

Sitography

BnF website [Accessed May 2023]:
https://www.bnf.fr/fr

Canterbury Tales Project, website [Accessed May 2023]:
https://www.canterburytalesproject.org/

Definition of User Experience (UX) proposed by Don Norman and Jakob Nielsen in "The definition of User Experience (UX)" [Accessed on February 2024]:
https://www.nngroup.com/articles/definition-user-experience/

Definition of User Interface (UI) proposed by the Interaction Design Foundation [Accessed on February 2024]:
https://www.nngroup.com/articles/definition-user-experience/

Digital edition of Vespasiano da Bisticci's Letters, website [Accessed May 2023]:
https://projects.dharc.unibo.it/vespasiano/

Digital Scholarly Edition of Paolo Bufalini's Notebook, website [Accessed May 2023]:
http://projects.dharc.unibo.it/bufalini-notebook/

Digital Vercelli Book, website [Accessed May 2023]:
http://vbd.humnet.unipi.it/beta2/

eScriptorium How to use [Accessed May 2023]:
https://lectaurep.hypotheses.org/documentation/escriptorium-tutorial-en

eScriptorium model CREMMA MEDII AEVI [Accessed May 2023]:
https://zenodo.org/record/6669508#.Y-yoBK3MLEZ

Europeana website [Accessed May 2023]:
https://www.europeana.eu/it

FAIR data principles [Accessed May 2023]:
The FAIR Data Principles – FORCE11

IIIF (International Image Interoperability Framework) website [Accessed May 2023]:
https://iiif.io/

FRBR description by IFLA [Accessed June 2023]:
https://www.ifla.org/references/best-practice-for-national-bibliographic-agencies-in-a-digital-age/resource-description-and-standards/bibliographic-control/functional-requirements-the-frbr-family-of-models/functional-requirements-for-bibliographic-records-frbr/

HTR-UNITED [Accessed May 2023]:
https://htr-united.github.io/models.html

Information about digital scholarly edition of Francesco da Barberino's Love Documents [Accessed May 2023]:
https://www.unive.it/pag/23956/

Introduction to IIIF, Rutges, Digital Humanities Initiative [Accessed on January 2024]:
https://dh.rutgers.edu/introduction-to-iiif/

ISO 9241-11 Ergonomics of human-system interaction. Part 11: Usability: Definitions and concepts [Accessed on February 2024]:
https://www.iso.org/standard/63500.html

Istituto Centrale per il Catalogo e la Documentazione, website [Accessed May 2023]:
http://www.iccd.beniculturali.it/

Leonardi, C. (1961). Anastasio. *Dizionario Biografico degli Italiani*. Volume 3. Disponibile su Treccani.it al sito [Accessed May 2023]:
https://www.treccani.it/enciclopedia/anastasio_%28Dizionario-Biografico%29/#:~:text=ANASTASIO%20Frate%20minore%2C%20forse%20del%20convento%20fiorentino%20di,il%20primo%20volgarizzamento%20in%20prosa%20dei%20poema%20antico.

LOV website [Accessed June 2023]:
https://lov.linkeddata.es/dataset/lov/

Mirador Github page [Accessed on January 2024]:
https://github.com/projectmirador/mirador

Mirador website [Accessed on January 2024]:
https://projectmirador.org/

Patil, P. (2018). What is exploratory data analysis. *Toward Data Science*. [Accessed in January 2024]:
https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15

Prado Museum website [Accessed May 2023]:
https://www.museodelprado.es/en/the-collection/art-works

SEGMONTO guidelines sintassi eScriptorium [Accessed May 2023]:
https://segmonto.github.io/gd/syntax/

Web Content Accessibility Guidelines (WCAG) 2.1 [Accessed on January 2024]:
https://www.w3.org/TR/WCAG21/

Wikipedia, Riconoscimento ottico dei caratteri [Accessed May 2023]:
https://it.wikipedia.org/wiki/Riconoscimento_ottico_dei_caratteri