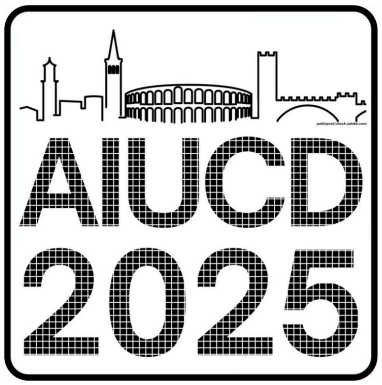

Atti del XIV Convegno Annuale

Diversità, Equità e Inclusione: Sfide e Opportunità per l'Informatica Umanistica nell'Era dell'Intelligenza Artificiale

Verona :: 11-13 giugno 2025

A cura di:
Simone Rebora • Marco Rospocher • Stefano Bazzaco

UNIVERSITÀ di VERONA
Dipartimento di LINGUE E LETTERATURE STRANIERE

daih

ASSOCIAZIONE per l'INFORMATICA UMANISTICA e la CULTURA DIGITALE

ISBN 978-88-942535-9-7

A cura di: Simone Rebora; Marco Rospocher; Stefano Bazzaco (2025). Diversity, Equity, and Inclusion: Challenges and Opportunities for Digital Humanities in the Age of Artificial Intelligence, Proceedings del XIV Convegno Annuale AIUCD, Verona 11-13 giugno 2025, Università di Verona.

Ultimo accesso agli URL in data 8 maggio 2025.

Si prega di notificare all'editore ogni omissione o errore si riscontri: segreteria [at] aiucd.org
Please notify the publisher of any omissions or errors found: segreteria [at] aiucd.org

Il programma della conferenza AIUCD 2025 è disponibile online
https://aiucd2025.dlls.univr.it/detailed-schedule/
The AIUCD 2025 Conference Program is available online
https://aiucd2025.dlls.univr.it/en-gb/detailed-schedule/

I contributi pubblicati nel presente volume hanno ottenuto il parere favorevole da parte di valutatori esperti della materia, attraverso un processo di revisione anonima mediante double-blind peer review, effettuata dai membri del Comitato di Programma sotto la supervisione del Comitato Scientifico di AIUCD 2025.
All the papers published in this volume have received favourable reviews by experts in the field of DH, through an anonymous double-blind peer review, carried out by the members of the Programme Committee under the supervision of the Scientific Committee of AIUCD 2025.

Gli atti del convegno AIUCD 2025 sono pubblicati come raccolta di contributi in formato PDF forniti direttamente dagli autori e dalle autrici. I file sono stati raccolti e assemblati senza interventi redazionali da parte dei curatori.
The proceedings of the AIUCD 2025 conference are published as a collection of PDF contributions provided directly by the authors. The files have been collected and compiled without editorial intervention by the editors..

Il logo di AIUCD 2025 include l'immagine "Verona Dark Line Simple Minimalist Skyline With White Background" di @pabloprat/stock.adobe.com, ottenuta tramite la licenza Adobe Stock dell'Università di Verona.
The AIUCD 2025 logo includes the image "Verona Dark Line Simple Minimalist Skyline With White Background" by @pabloprat/stock.adobe.com, used under the Adobe Stock license of the University of Verona.

Il background della copertina è stato creato con tecniche di AI generativa con lo strumento "Magic Media" disponibile su Canva, usando un prompt con il tema del convegno.
The background of the cover was created using generative AI techniques with the "Magic Media" tool available on Canva, using a prompt based on the conference theme.

Comitato Organizzatore / *Organizing Committee*

General Chairs
Simone Rebora (Università degli Studi di Verona)
Marco Rospocher (Università degli Studi di Verona)

Local Chair
Anna Cappellotto (Università degli Studi di Verona)

Registration Chair
Giorgia Pomarolli (Università degli Studi di Verona)

Proceedings Chair
Stefano Bazzaco (Università degli Studi di Verona)

Sponsorship Chair
Matteo Lissandrini (Università degli Studi di Verona)

Publicity Chair
Sabrina Piccinin (Università degli Studi di Verona)

Comitato Scientifico / *Scientific Committee*

Program Chairs
Simone Rebora (Università degli Studi di Verona)
Marco Rospocher (Università degli Studi di Verona)

Digital Humanities e inclusione / *Inclusive DH*
Stefano Bazzaco (Università degli Studi di Verona)
Massimo Salgaro (Università degli Studi di Verona)

Archivi ed Edizioni Digitali / *Archives and Digital Editions*
Elisa Cugliana (Cologne Center for eHumanities)
Christian D'Agata (Università di Catania)

Metodi Computazionali / *Computational Methods*
Rachele Sprugnoli (Università degli Studi di Parma)
Sara Tonelli (Fondazione Bruno Kessler)

Rappresentazione di Dati e Conoscenza / *Data and Knowledge Representation*
Francesco Mambrini (Università Cattolica del Sacro Cuore)
Elena Spadini (Universität Bern)

Preservazione della Memoria e del Patrimonio Digitale / *Preservation of Memory and Digital Cultural Heritage*
Monica Berti (Universität Leipzig)
Daria Spampinato (Istituto di Scienze e Tecnologie della Cognizione-CNR)

Automating XML-TEI Encoding of Unpublished Correspondence: A Comparative Analysis of two LLM Approaches

Marco De Cristofaro, Daniel Zilio
[1] Université de Mons; Université de Namur, Belgium - Marco.DECRISTOFARO@umons.ac.be
[2] Università di Padova, Italy – daniel.zilio@unipd.it

ABSTRACT (ENGLISH)

Encoding texts in TEI-XML format is critical for research projects dealing with copyright and open access issues. The PubCiNET project reconstructs the social network of Italian intellectuals in publishing and film between the 1950s and 1980s. During the three decades, exchanges and collaborations between professionals in these creative industries increased, affecting the convergence of literature and film, film professionals' engagement in publishing, and the perception of publishing as a prestigious field for filmmakers. The project utilizes an XML-TEI encoded corpus of archival correspondence to map the intellectual network. However, key challenges arise, including the complex retrieval of data from vertically structured archives, copyright issues due to the contemporary timeframe, and the sustainability of handling vast volumes of documents. This study proposes a first attempt at these challenges by applying automated text encoding through large language models (LLMs). The research explores automated encoding using ChatGPT-4 and Claude 3.5 Sonnet, analyzing their capabilities in enhancing access to archives and automating the labor-intensive encoding process. Initial findings indicate varying success rates: while both LLMs efficiently extract metadata, they differ in their ability to recognize information in the text of the letters. Improving their efficiency in terms of information recognition and the reliability of reference materials could contribute to more efficient and faster encoding, allowing for greater sustainability in research.

Keywords: XML-TEI Encoding; Intellectual Networks; Automated Text Analysis; Publishing and Film Industries; Large Language Models (LLMs)

ABSTRACT (ITALIANO)

Codifica automatica in XML-TEI della corrispondenza inedita: un'analisi comparativa di due approcci basati su LLM

La codifica XML-TEI dei testi fornisce un buon supporto per i progetti di ricerca che affrontano questioni di copyright. Il progetto PubCiNET ricostruisce la rete intellettuale tra professionisti dell'editoria e del cinema tra gli anni '50 e '80. Questo periodo ha segnato un aumento degli scambi e delle collaborazioni tra le due industrie creative, influenzando la convergenza tra letteratura e cinema. Per ricostruire la rete intellettuale, il progetto utilizza un corpus di corrispondenza d'archivio codificato in XML-TEI. Tuttavia, la complessa raccolta di dati da archivi strutturati verticalmente, le criticità relative al copyright del materiale archivistico e la sostenibilità nella gestione di grandi volumi di documenti presentano problematiche relative all'accessibilità e all'utilizzo dei risultati sia da parte dei ricercatori sia da parte degli utenti finali. Questo studio propone un tentativo di codifica automatizzata di un corpus di testi attraverso Large Language Models (LLMs). La ricerca esplora la codifica automatizzata utilizzando ChatGPT-4 e Claude 3.5 Sonnet, analizzando le loro capacità di migliorare il processo intensivo della codifica. I risultati iniziali indicano tassi di successo variabili: mentre entrambi gli LLMs estraggono efficientemente metadati, dimostrano un grado di affidabilità differente per quanto riguarda la capacità di riconoscere determinate informazioni dei testi. Incrementare la loro efficienza in termini di riconoscimento delle informazioni e di affidabilità dei materiali di riferimento potrebbe contribuire a una codifica più efficiente e rapida permettendo una maggiore sostenibilità della ricerca

Parole chiave: XML-TEI; Reti intellettuali; Analisi testuale automatizzata; Editoria e industria cinematografica; Large Language Models (LLMs)

1. INTRODUCTION

Encoding new texts in XML-TEI format plays a central role in research projects that focus on automatically analyzing texts with issues related to copyright, as is the case with the PubCiNET project. In the PubCiNET project, we aim to reconstruct the social network that emerged between the 1950s and 1980s among intellectuals from two creative industries: publishing and film industry. As it has been suggested and demonstrated (Brunetta, 2004; Ivaldi, 2001; Simonetti, 2018), starting from the post-war period,

intellectual exchanges between professionals in the publishing world and those in the film industry intensified and strengthened. The intensification of the exchange of ideas has led to three main consequences. First of all, several authors admit the influence of cinema on their stylistic choices. Secondly, there is a growing interest in the publishing market among film professionals such as directors, screenwriters, and producers. They contribute by making editorial proposals related to their own cinematic works or by publishing books based on their creative experience. Finally, film professionals—particularly directors—regard the publishing world as a space for achieving recognition beyond the boundaries of their specific cultural domain. To make this crucial intellectual network accessible and visible, one can draw on correspondence between the various professionals involved: publishers, directors, publishing house consultants, and producers. This approach seeks to map, for the first time, for what concerns the Italian cultural field, the process of transmitting ideas from the film industry to publishing and vice-versa. The overall goal of PubCiNET is to identify the key nodes through which new forms of cinematic expression circulate within the publishing sphere. To support this objective, the encoding of both published and unpublished letters helps to map the points of contact between the film and publishing industries. Given the large number of letters collected from seven different archives and transcribed manually, automated encoding can provide substantial support for the project's success. This paper focuses on the initial attempts to automatically encode the extensive corpus of collected letters.

The study of correspondence has proven to be a valuable tool for exploring criticism. Correspondence between professionals in the cultural industries makes it possible to investigate the relationship between the private dimension—that is, the initial intentions of the various agents involved—and the public dimension, meaning the final outcome accessible to the public. The role of correspondence in analyzing the relationship between initial intentions and final outcomes has been demonstrated in both the publishing (Cadioli, 2012, 2021; Italia, 2006, 2013; Pischedda, 2022) and film fields (Guerra & Martin, 2019; Mariani & Venturini, 2017; Noto, 2019; Noto et al., 2020; Rigola, 2021). Nevertheless, the structure of publishing and cultural archives makes it difficult to study interactions between multiple figures rather than focusing on just one.

The three main critical issues that emerged during the study of the project are as follows: i) the vertical structure of personal or publishing house archives makes direct data retrieval challenging: while the workflow for collecting materials from paper archives, both in publishing and cinema, has been well defined and outlined(Cesana, 2006; Guerra & Martin, 2019), the involvement of multiple archives with different consultation and publication practices makes data collection significantly more difficult; ii) given the selected time frame, 1950–1980, considerations on the copyright status of each document are necessary: the need to request authorization from all rights holders to first consult and then publish the materials raises questions about the sustainability of projects dealing with contemporary archives; iii) a third critical issue again concerns research sustainability: the volume of documents consulted and collected, which would need to be encoded to enable future consultation and analysis, necessitates the exploration of automated processes for managing data in light of advancements in encoding by LLMs (DeRose, 2024).

Building on an initial manual encoding in XML TEI of the correspondence collected, we will compare automated encoding techniques using two different LLMs: a widely available LLM, ChatGPT4[1], selected based on previous experiments specifically involving the automated encoding of correspondence(Pollin et al., 2023), and another LLM up-and-comer, Claude 3.5 Sonnet[2] chosen to explore the potential for creating a ready-to-use chatbot for educational purposes.

2. The XML-TEI Encoding

The structure of the data model was inspired by four different projects: the edition of *Vespasiano da Bisticci. Lettere* (Tomasi, 2013); the Darwin Correspondence Project (University of Cambridge, 2022); the Van Gogh Letters(Van Gogh Museum & Huygens ING, 2018); the Bellini Digital Correspondence(Del Grosso & Spampinato, 2023).

The proposed markup model aims to translate the core elements of the edition into a formal system. The core elements are: the people mentioned in the letters; the occupation of the people; works; book series; the organizations; the places; the name of the awards and bibliographic references. We represent all the data in a macro.xml file through appropriate elements and attributes. The data encoded in the specific file of each letter are linked to the macro.xml file. This choice will prove particularly useful in the

[1] https://openai.com/index/gpt-4/.
[2] https://www.anthropic.com/news/claude-3-5-sonnet.

phase of encoding automatization. Most of the elements considered are therefore based on a pre-existing approach, and we refer to De Cristofaro(2024) for further details.
The most interesting element, however, is the encoding of the works. They are not always identified by a specific title in the letters, as they often refer to drafts that precede the final published version. In fact, the same book may be mentioned at different stages of its development. In our encoding, we chose to treat each draft as a separate work, since they have distinct characteristics and the final version is not always published. Annotating drafts separately also supports further philological and editorial research. Another goal of the project is to actively describe the publication process. To encode the works, we used a generic <rs> tag to define the references to each work as they appear in the letters. The @type attribute specifies the type of work: bibliographic or cinematic. We encoded each phase of a work as a different work to define the various moments of the publication process. These will be the main challenges in the automatization of the process through LLMs: i) how to recognize a work that is quoted without its specific published name; ii) how to encode a pre-publication step of a specific volume. The prompt and the provided referenced encoded texts would be crucial as well as the macro.xml file.

3. The XML-TEI Encoding automatization: ChatGPT4

Given the vast amount of material underpinning the project, in this work, we are experimenting with an automated encoding process using the LLMs. We will compare automated encoding techniques using two different large language models (LLMs): a widely available LLM, OpenAI ChatGPT-4, selected based on previous experiments specifically involving the automated encoding of correspondence (Pollin et al., 2023); and an LLM developed by the Anthropic Claude 3.5 Sonnet. We decided to use ChatGPT in chat mode because it has demonstrated promising capabilities at the current state of the art. In contrast, we opted for Claude 3.5 via API due to its timing performance compared to the chat mode. For future tests, our planned approach will involve using both models through their API.
Our initial attempt to apply automated XML TEI encoding using LLMs will be grounded in the research by Pollin et al. (2023). Two main findings from their study were: i) the prompt is crucial for success; and ii) human evaluation is becoming increasingly important. However, there is one key difference between our project and the primary goal of Pollin's experiment: they applied the automated LLM strategy solely to the text of the letters and only to a single text. In contrast, given our objective to automatically encode a corpus of hundreds of letters, we have decided to conduct our experiment by also including the <teiHeader> and considering from the outset encoding more than one letter through a single prompt. Building on this difference and considering the crucial role of the prompt, we developed our prompt based on the example provided by Pollin et al. (2023).
The prompt has been divided into two parts. The first part gathers all the information regarding the <TeiHeader>. Since LLMs cannot deduce archival metadata solely from the text, we decided to adopt the following strategy: providing the letters organized by archive in order to automatically encode the section related to each repository. This approach would also facilitate the automation of the metadata section. In this initial part of the prompt, we provide a detailed map of all the information of the <TeiHeader>. Here, we present an example of key information to be included in the TeiHeader, as outlined in the prompt.:

> You will act as a skilled expert automaton that is proficient in transforming unstructured text, specifically multilingual letters, into well-formed TEI XML. Analyze the provided text based on the mapping rules I have shared and then execute the transformation to produce TEI XML, ensuring you adhere to the TEI guidelines and annotate if certain.
> Mapping rules:
> * <teiHeader>: contains metadata;
> fill the <teiHeader> using all the tags that are relevant for the project metadata as follows:[…]
> * The first <correspAction> nest in <correspDesc> has @type attribute "sent"
> * The <persName> nest in the first <correspAction> has @ref attribute that links the sender to the right person who sends the letter in the macro.xml file. You can understand who sends the letter from the title and the signature. This <persName> also has a @role attribute that refers to the role attributed to the person who sends the letter in the macro.xml file

The second section of the prompt has been structured specifically in relation to the content of the letter. Here is an example of how we structured this second part of the prompt:

> * <opener> Opening of the letter

> * <salute> Salutations within the letter
> * <closer> Closing of the letter
> * <signed> Signature section
> * <postscript> Represents a postscript
> * <persName> Person with a @ref attribute that links the person to the corresponding person in the macro.xml file
> * <orgName> Organisation with a @ref attribute that links the organization to the corresponding organization in the macro.xml file

Since we are communicating via ChatGPT4, we have provided the macro.xml file containing all the information about the individuals, organizations, volumes, and films mentioned. We also provided nine sample files of the encoded letters collected from the same archive. The final part of the prompt refers to the files provided as examples:

> I provide you with nine examples of encoded letters collected in the same archive.
> I provide you with the macro.xml file where you can find all the information you need to encode all the data of the letters.

Finally, we provided the plaintexts of the four letters to be encoded.
The first and fundamental issue encountered with encoding the four letters together is that the result is invalid encoding. In particular, validation fails due to three recurring errors in the TeiHeader:

- Invalid characters: the inclusion of unrecognized characters "©", which represent the copyright symbol "©" present in the example letters.
- Incorrect placement of the <langUsage> tag: the <langUsage> tag should be nested within <profileDesc>. However, in all four cases, GPT closes the <profileDesc> tag first and then places the <langUsage> tag, which is incorrect.
- Improper use of the <p> tag: within the <closer> tag, which contains the closing and signature of the letter, GPT inserts a <p> tag that is not allowed.

Since these were recurring errors, we specifically requested corrections via chat. GPT successfully corrected the three recurring issues, subsequently providing valid files.
Once valid files were obtained, a second verification was conducted on ChatGPT's ability to encode the text of the letters within the . In this case, it demonstrated the ability to encode the various parts that characterize the letter: <opener>, <closer>, <signed>, and <postscript> were correctly identified. The letter was divided into paragraphs: our encoding does not require a strict <lb> tag since visualization of the archival document is not the ultimate goal of the research. A division into <p> is sufficient for the project's objectives.
Regarding the ability to identify relevant information within the text, performance recorded an average success rate of 85%. The success rate was calculated based on a comparison between the elements recognized and correctly encoded by the LLM and those identified and manually encoded in the same letter. In this initial phase of experimentation, manual supervision is necessary for the evaluation metrics. Almost all names were correctly identified with the <persName> tag and the @ref attribute linking to the correct person in the macro.xml file (Fig. 1). Once again, a recurring shortcoming in the encoding was noted: the person included in the <signed> tag was never encoded in any of the 4 letters. However, a brief prompt specifying the inclusion of that figure in the encoding allowed for the correction of the omission in all four letters.
As for organizations, publishing houses were correctly identified and encoded using the <orgName> tag. In this case, it does not appear that the macro.xml file was used as a reference for encoding the organization. In the manual encoding, indeed, the publishing houses are encoded according to the following scheme: <orgName ref="macro.xml#cappelli-edizioni">Cappelli</orgName>. The publishing house's name is supplemented with the term "edizioni." ChatGPT's encoding, on the other hand, adopts the following scheme: <orgName ref="macro.xml#longanesi">Longanesi</orgName>.
Two considerations are necessary. On the one hand, the support of the macro.xml file does not always seem relevant and, in some cases, proves counterproductive. On the other hand, when the LLM does not rely on the information from the macro.xml file, it performs much better in recognizing names and organizations. The encoding of books is more complex. As previously specified, the objective here is to trace the publication process. We aim to encode published works and works in the development stages. In

the stages before publication, works may have different or not always explicit titles. In the macro.xml file, we have encoded each stage with a corresponding unique identifier. ChatGPT demonstrates strong performance in encoding published works but fails when encoding the pre-publication stages. This result demonstrates that the macro.xml file was not considered the main reference point, while it relies on its own information for encoding works.

4. The XML-TEI Encoding automatization: Claude 3.5 Sonnet

We developed a Python script to directly interact with the model through the provided API[3]. We decided to use it instead of the dedicated interface, maintaining the same logic as the previous approach. Indeed, this script follows a two-step process to produce the final result. In the first step, the prompt comprises three components: a list of instructions, an XML file containing all directives, and the plain text of the letter. The instructions have been articulated in a manner that aligns more closely with the requirements of the model while still conveying the same essential demands as those outlined for ChatGPT. The output from this first prompt then serves as the input for the second step, where the directives in the macro.xml file are re-evaluated to identify the names of people, movies, and other relevant terms.

The results of the first prompt demonstrate its ability to produce valid files. The metadata is correctly encoded, no invalid characters are present, and the nesting order is properly executed.

At the textual level, within the , two substantial aspects distinguish the automated encoding performed by Claude Sonnet 3.5. The first is that there is perfect adherence to the plaintext file: the encoding of the letter's parts is correct, and if the plaintext, as happens in some cases, lacks the letter's signature, the <closer> is also absent from the encoding. A second substantial element is the greater alignment with the macro.xml file. The information is primarily drawn from the reference file, as confirmed by the encoding of <persName>, all of which include the @ref attribute with the correct identifier from the macro.xml file, and the encoding of <orgName>, which also features the corresponding @ref attribute linking to the correct identifier in the macro.xml file.

Here, the main issues are found in the encoding of works. Published works are once again correctly recognized and linked to the macro.xml file. However, the stages before publication are not encoded.

```
purtroppo negativa. Come lei forse sapeva dal comune amico <persName
  ref="macro.xml#renzo-renzi">Renzo</persName>, io avevo ricevuto lusinghiere e
considerevoli offerte da case editrici con i cui direttori editoriali ero e sono da tempo
in ottimi rapporti: ma per l'antica amicizia che mi lega a <persName
  ref="macro.xml#renzo-renzi">Renzo</persName>, avevo declinato le proposte concrete e
pressanti della <orgName ref="macro.xml#longanesi">Longanesi</orgName>, di <orgName
  ref="macro.xml#mondadori">Mondadori</orgName> e di <orgName ref="macro.xml#rizzoli"
  >Rizzoli</orgName>, per privilegiare, la <orgName ref="macro.xml#cappelli"
  >Cappelli</orgName> alla quale mi sento vicino per il lavoro di tanti anni.</p>
```

Fig. 1 ChatGPT4 encoding of a Letter from Federico Fellini to Mario Musso, 21 January 1990

```
risposta purtroppo negativa. Come lei forse sapeva dal comune amico <persName
    ref="macro.xml#renzo-renzi">Renzi</persName>, io avevo ricevuto lusinghiere
e considerevoli offerte da case editrici con i cui direttori editoriali ero e
sono da tempo in ottimi rapporti: ma per l'antica amicizia che mi lega a
    <persName ref="macro.xml#renzo-renzi">Renzo</persName>, avevo declinato le
proposte concrete e pressanti della <orgName ref="macro.xml#longanesi-edizioni"
    >Longanesi</orgName>, di <orgName ref="macro.xml#mondadori-edizioni"
    >Mondadori </orgName> e di <orgName ref="macro.xml#rizzoli-edizioni"
    >Rizzoli</orgName>, per privilegiare, la <orgName
    ref="macro.xml#cappelli-edizioni">Cappelli </orgName> alla quale mi sento
vicino per il lavoro di tanti anni. </p>
```

Fig. 2 Claude encoding of a Letter from Federico Fellini to Mario Musso, 21 January 1990

5. Conclusions. For a comparison between the two models

We acknowledge the limited scope of our current evaluation, which analyzed only four letters as a preliminary test case. This small sample size was chosen for an initial proof-of-concept to establish baseline methodological approaches and identify key challenges before scaling to larger datasets. Furthermore, at this initial stage, the evaluation of the success rate of the encoding process had to be carried out manually. To assess the accuracy of the encoding, we compared the letters automatically encoded by the two LLMs with the same letters manually encoded in XML-TEI by the project lead. The comparison focused particularly on the ability of the automatic encoding to identify all relevant elements and annotate them correctly, as in the manual encoding. The evaluation metrics were primarily based on a quantitative calculation of the number of elements recognized and correctly annotated automatically compared to those annotated manually. Once a satisfactory success rate is achieved, the automatic encoding will be extended to a larger corpus, organized by archive. From the previous evaluations, we can attempt a comparison between the two models. If ChatGPT initially shows difficulties in producing a valid file, this is not the case with Claude, which immediately encodes the four letters and returns valid files. At the same time, Claude demonstrates greater adherence to the input files provided: both to the examples of letters and to the macro.xml file from which it extracts all the most relevant information to encode. ChatGPT, by deviating from the input information, proves to be more effective in independently

[3] The model used was *anthropic.claude-3-5-sonnet-20241022-v2:0*, running on Amazon Bedrock.

recognizing textual data, showing potential for improvement in encoding elements that are not previously included in the macro.xml file. It is worth noting that encoding must be carried out on a corpus of letters to demonstrate its actual efficiency and its real contribution to the sustainability of the research. The main challenges of the project lie precisely in the encoding and manual standardization of a large corpus of texts originating from different archival locations. Manual encoding and verification would not only be complex, requiring constant review, but also time-consuming, without always ensuring uniformity in the encoding process.

The implementation of a reference database, to which the texts of the letters would link with the respective identifiers for each element to be encoded, could offer additional advantages in automated encoding. In that case, Claude's greater adherence to input files would appear more efficient. On the other hand, to expand the project, we foresee a hybrid approach that also takes into account ChatGPT's superior capabilities in encoding elements not present in the input files and flagging them. We propose to explore a hybrid approach that integrates Large Language Models (LLMs) with complementary techniques in both the pre-processing and post-processing phases. We plan to implement a key initial technique called Named Entity Recognition (NER). This technique will enhance our model's ability to identify and classify specific entities within texts, including technical terms, and domain-specific concepts. We anticipate that this capability will significantly improve the contextual understanding and relationship mapping in our analysis pipeline. In the post-processing phase, we aim to establish an automatic feedback system based on current patterns of encoded letters. This will create a self-improving mechanism that continuously evaluates output quality against established benchmarks. This system will serve as a foundational tool for training the model to refine its prompts effectively. Additionally, our next research stage will focus on comprehensive fine-tuning of the parameters for both models. Our goal is to optimize their performance, particularly in terms of response relevance and contextual precision.

To address the methodological limitations identified in this preliminary phase of our research, we plan to adopt a more quantitative evaluation framework in the next stages. This will involve metrics to assess the performance of LLM models in XML-TEI encoding, including: (1) precision in identifying named entities (such as people, organizations, and works), along with analyses of false positives and false negatives; (2) accuracy in associating the correct references from the macro.xml file; (3) the structural validity of the generated markup; and (4) computational efficiency. This systematic approach will enable a more objective evaluation of the capabilities of different LLMs and offer concrete guidance for optimizing prompts and implementation strategies in similar archival contexts.

While manual verification still appears indispensable, as argued by Pollin (2023) and DeRose(2024), encoding a corpus rather than a single text would still benefit the project's development, resulting in substantial savings in both human and economic resources. From the perspective of research sustainability, involvement of LLMs has already demonstrated their potential(Schmidgall et al., 2025). Considering the archival challenges we have identified, from the plurality of figures involved to the diverse origins of archival materials, the possibility of automated encoding of an extensive corpus of letters would greatly support the achievement of the project's objectives.

ACKNOWLEDGEMENTS

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska Curie grant agreement No 101034383

REFERENCES

Brunetta, G. P. (2004). *Gli intellettuali italiani e il cinema*. Bruno Mondadori.

Cadioli, A. (2012). *Le diverse pagine. Il testo letterario tra scrittore, editore e lettore*. Il saggiatore.

Cadioli, A. (2021). *«La sana critica». Pubblicare i classici italiani nella Milano di primo Ottocento*. Firenze University Press.

Cesana, R. (2006). La memoria bibliografica: Storia dell'editoria e archivi editoriali. *Bibliologia : an International Journal of Bibliography, Library Science, History of Typography ant the Book*, *1*, 175–197.

De Cristofaro, M. (2024, October 10). *From Cinema to Publishing, there and back. Encoding a Corpus of Letters Between Filmmakers and Publishers in XML TEI*. TEI 2024. Texts, languages and communities.

Del Grosso, A. M., & Spampinato, D. (2023). *Bellini Digital Correspondence* (digitale). Cnr Edizioni. https://bellinicorrespondence.cnr.it

DeRose, S. J. (2024). Can LLMs help with XML? *Proceedings of Balisage: The Markup Conference 2024*, *29*. https://doi.org/10.4242/BalisageVol29.DeRose01

Guerra, M., & Martin, S. (2019). La cultura della lettera. La corrispondenza come forma e pratica di critica cinematografica. *Cinergie – Il cinema e le altre arti*, *8*(15), 1–3. https://doi.org/10.6092/issn.2280-9481/9661

Italia, P. (2006). Le «penultime volontà dell'autore». Considerazioni sulle edizioni d'autore del Novecento. *Ecdotica*, *3*, 175–186.

Italia, P. (2013). *Editing Novecento*. Salerno.

Ivaldi, F. (2001). *Effetto rebound. Quando la letteratura imita il cinema*. Felici.

Mariani, A., & Venturini, S. (2017). L'archivio e lo studioso. *Bianco e Nero*, *78*(588–589), 40–51.

Noto, P. (2019). Quale "mestiere del critico"? Un'intrusione nella corrispondenza di Guido Aristarco. *Cinergie – Il Cinema E Le Altre Arti*, *8*(15), 55–67. https://doi.org/10.6092/issn.2280-9481/9357

Noto, P., Malvezzi, J., & Mariani, A. (2020). Spostamenti progressivi nella critica cinematografica tra 1930 e 1970: Spazi, relazioni, movimenti. *Cinergie – Il Cinema E Le Altre Arti, 12*(23), 1–4. https://doi.org/10.6092/issn.2280-9481/17672

Pischedda, B. (2022). *La competizione editoriale. Marchi e collane di vasto pubblico nell'Italia contemporanea (1860-2020)*. Carocci.

Pollin, C., Steiner, C., & Zach, C. (2023). *New Ways of Creating Research Data: Conversion of Unstructured Text to TEI XML using GPT on the Correspondence of Hugo Schuchard with a Web Prototype for Prompt Engineering*.

Rigola, G. (2021). I fondi archivistici personali, la corrispondenza e la ricerca sul cinema. Il caso del carteggio tra Elio Petri e Leonardo Sciascia. In *Scrivere la Storia, costruire l'archivio. Note per una storiografia del cinema e dei media* (pp. 55–67). Meltemi.

Schmidgall, S., Su, Y., Wang, Z., Sun, X., Wu, J., Yu, X., Liu, J., Liu, Z., & Barsoum, E. (2025). *Agent Laboratory: Using LLM Agents as Research Assistants* (arXiv:2501.04227). arXiv. https://doi.org/10.48550/arXiv.2501.04227

Simonetti, G. (2018). «Quasi verità». Scrittori italiani e cinema. In *La letteratura circostante. Narrativa e poesia nell'Italia contemporanea*. il Mulino.

Tomasi, F. (2013). *Vespasiano da Bisticci, Lettere* (Digitale). CRR-MM, Università di Bologna. http://vespasianodabisticciletters.unibo.it/

University of Cambridge. (2022). *Darwin Correspondence Project*. https://www.darwinproject.ac.uk

Van Gogh Museum, & Huygens ING. (2018). *Vincent van Gogh. The Letters* (digitale). https://vangoghletters.org/vg/letters.html