



Decoding the Past: AI Transcription of Handwritten Historical Text for Built Heritage Conservation

Decodificare il passato: IA, trascrizione di manoscritti e conservazione del patrimonio architettonico

Matilde Ridella

University of Genoa, Department of Modern Languages and Cultures |
matilde.ridella@edu.unige.it

Giulio Lucio Sergio Sacco

University of Genoa, Department of Civil, Chemical and Environmental Engineering |
giulioLucioSergio.sacco@edu.unige.it

Citation: M. Ridella, G. L. S. Sacco, Decoding the Past: AI Transcription of Handwritten Historical Text for Built Heritage Conservation, in *Paesaggio Urbano - Urban Design*, 2025, 2, pp. 28–37.
ISSN for printing: 1120-3544
ISSN for online publishing: 2974-5969
Received: October, 2025
Accepted: November, 2025
Published: December 2025
Copyright © 2025 Maggioli Editore.
This is an open access, peer-reviewed article published by *Paesaggio Urbano - Urban Design*, distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

The preservation of architectural heritage and the landscape is an essential part of the conservation and enhancement of our cultural heritage. Achieving this goal requires in-depth knowledge of all aspects of the structure to be preserved: from its shape to the materials used; from the construction techniques employed to its relationship with its surroundings and the historical and cultural background in which it is located. In this context, archives are a primary source of knowledge and yet are often difficult to access and inadequately indexed, placing significant limitations on consultation and dissemination of information among researchers and interdisciplinary working groups. Modern OCR (Optical Character Recognition) techniques can facilitate this process. However, their application to historical documents, especially manuscripts, is still challenging and costly. This paper presents the preliminary results of the application of CHURRO, an open-weight artificial intelligence model for the transcription of historical texts, to a corpus of manuscript documents relating to the Geirato siphon bridge, part of Genoa's historic aqueduct, demonstrating the feasibility of a cost-effective, replicable and locally executed digitisation pipeline.

00.
The Geirato siphon bridge seen from via Geirato (photograph of the authors, 19th March 2023).

La tutela di beni architettonici e del paesaggio è parte fondamentale della conservazione e valorizzazione del patrimonio culturale. Raggiungere tale obiettivo richiede la conoscenza approfondita di tutti gli aspetti che caratterizzano il manufatto da preservare: dalla geometria ai materiali impiegati; dalle tecniche costruttive al rapporto con il contesto e la dimensione storico-storica in cui esso si inserisce. In tale ambito, gli archivi rappresentano una fonte primaria di conoscenza e, nonostante ciò, risultano spesso di difficile accesso e non adeguatamente indicizzati, ponendo limiti significativi alla consultazione e alla diffusione di informazioni tra ricercatori e gruppi di lavoro interdisciplinari. Le moderne tecniche OCR (Optical Character Recognition) possono agevolare tale processo; tuttavia, l'applicazione a documenti storici, soprattutto manoscritti, resta complessa e onerosa.

Il contributo presenta i risultati preliminari dell'applicazione di CHURRO – un modello open-weight di intelligenza artificiale per la trascrizione di testi storici – ad un corpus di documenti manoscritti relativi al ponte sifone sul Geirato, parte dell'acquedotto storico di Genova, dimostrando la fattibilità di un processo di digitalizzazione economico, replicabile ed eseguito localmente.

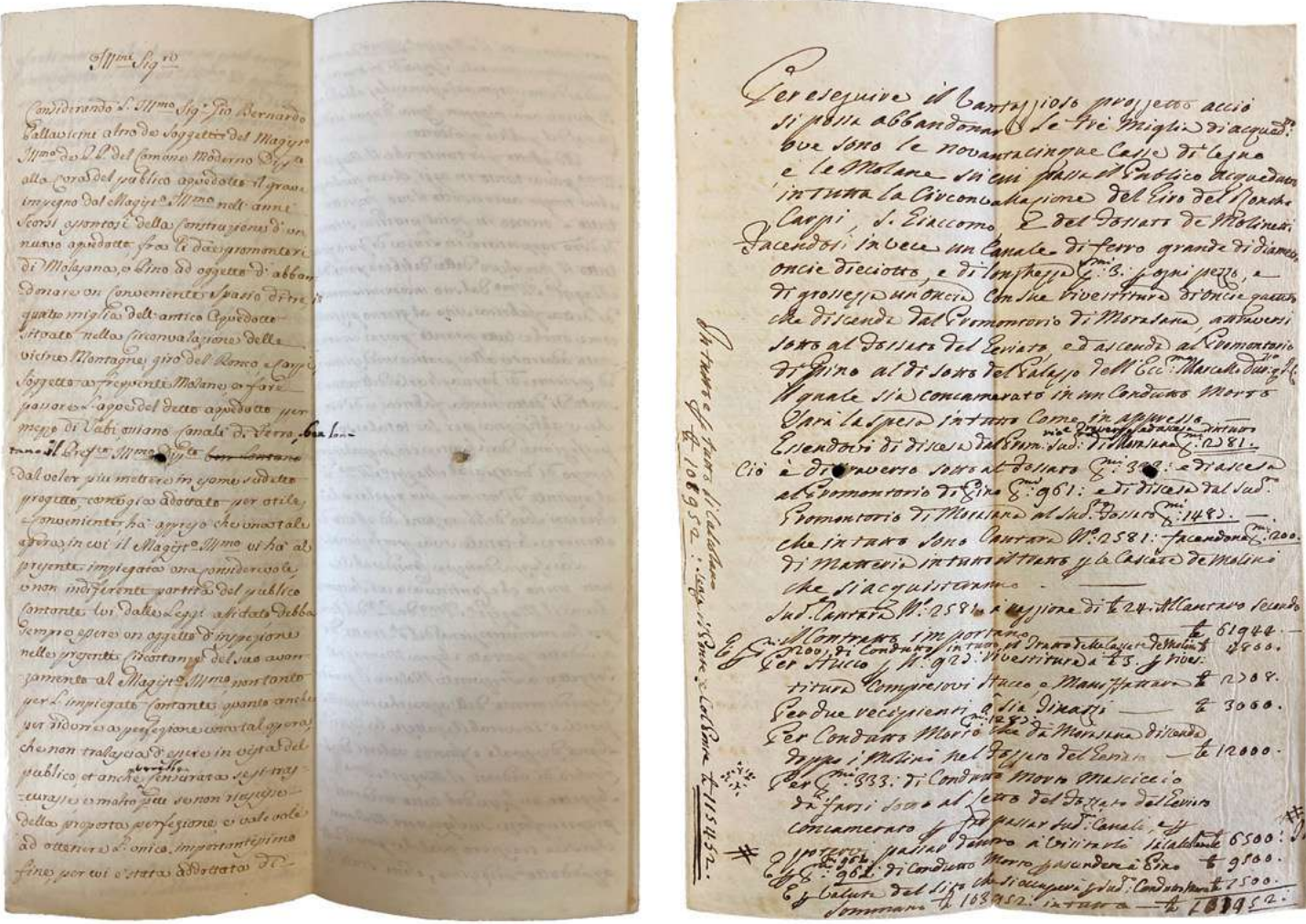
INTRODUCTION

Historical documents are the primary source of knowledge that can bring us closer to the customs and traditions of the past, to the building techniques and the reasoning behind construction materials choices, as well as to the design approaches adopted. This is confirmed in Art. 2 of the Venice Charter drawn up by ICOMOS in 1964 states that “the conservation and restoration of monuments is a discipline that draws on all the sciences and techniques that can contribute to the study and preservation of monumental heritage”.

However, the variety of languages, handwriting practices, typefaces, and the evolution of writing and spelling conventions requires the involvement of expert scholars who can study these materials. As a result, access to such information remains limited, posing significant limitations to researchers and practitioners involved in the conservation of architectural heritage. Converting handwritten text into machine-readable encoded formats can mitigate these limitations and have deep implications in various fields, such as the preservation of historical documents, automatic data processing, and accessibility for visually impaired persons.

Handwritten Text Recognition (HTR) still represents a challenge in the field of pattern recognition and machine learning (AlKendi et al. 2024; Garrido-Munoz et al. 2025). Historical documents present significant difficulties as frequently contain features such as marginalia⁰¹, rubrication⁰², and illumination⁰³, seldom found in modern or contemporary writings. The lack of shared orthographic standards and the diversity of fonts used in printed documents further increase these difficulties. HTR is even more challenging due to the multiplicity of handwriting styles and calligraphic traditions, which can differ significantly even among writers of the same language and period. Moreover, scribes frequently employed a wide range of abbreviations to expedite the writing process, producing thousands of distinct shortened forms. Lastly, the physical degradation of documents caused by age and environmental factors adds additional layers of complexity to an accurate text extraction.

Several platforms, such as Transkribus and eScriptorium, provide tools for transcribing and annotating handwritten text, as well as for fine-tuning recognition models (Kahle et al. 2017;



01. On the left, 1st document. First page of the handwritten document *Relazione al Magistrato Illustrissimo de Padri del Comune dell' Illustrissimo Gio Bernardo Pallavicino* che riguarda tutto l'occorso per la fabrica del nuovo acquedotto in canali di ferro fra Molasana e Pino, 16th March 1776 (ASCG, Magistrato dei Padri del Comune, filza 303). On the right, 2nd document. First page of the handwritten document *Conto della spesa vi vorrà in tutto per eseguire il progetto deliberato dal Magistrato Illustrissimo per abbandonare tre miglia di acquedotto con farlo passare ne canali di ferro che discendano dalla montagna di Morasana indi attraversano il fossato del Gerriato ed ascendano a sboccare nel acquedotto su la montagna di Pino*, June 1772 (ASCG, Magistrato dei Padri del Comune, filza 303).

Kiessling et al. 2019). However, these platforms, especially in their free version, require users to grant them ownership or exploitation rights of the data: a practice often incompatible with research ethics or conflicting with the rights of use of corpora. Moreover, software licence costs often exceed the budget of research groups or practitioners, while free research licenses often require institutions to contribute with large corpora of tens of thousands of pages. Conversely, open solutions, such as TrOCR, require programming skills and proper hardware (Romein et al. 2025). This paper presents preliminary results on the application of AI methodologies to a corpus of historical documents concerning the eighteenth-century Geirato siphon bridge, part of the Genoese historical aqueduct. The experimentation, conducted with the CHURRO AI model (Semnani et al. 2025), aimed to assess the feasibility of automated transcription for mixed handwritten documents from raw photographs, and to determine whether it could significantly speed up research ensuring readability and the possibility of querying documents.

CASE STUDY

The corpus related to the Geirato siphon bridge under investigation currently consists of more than 170 documents for a total of about 850 handwritten pages kept in three different archives (i.e. ASCG⁰⁴, BUGe⁰⁵, DocSAI⁰⁶). It comprehends letters, reports, valuation tables, annotated drawings, etc. produced between 1765 and 1816 documenting the design, construction and later technological updates of the bridge. This material, mainly consisting in photographs of the originals, constitutes the portion of the bridge documentations already digitalized and ready to be transcribed. Additionally, some typewritten texts more generally referring to the aqueduct, despite not included in the siphon bridge corpus, were also transcribed as relevant for the research.

The transcription procedure was carried out with CHURRO, a 3B-parameter open-weight Vision-Language Model, fine-tuned from Qwen 2.5 VL, and specialized in historical text recognition, with extensive support for handwritten sources. The model was trained on CHURRO-DS, a database which unifies more than 155 historical corpora comprising about 100,000 pages, spanning 22 centuries of textual heritage across 46 language clusters. The processing was performed locally on a desktop workstation equipped with an Intel Core i7 12700 CPU, 64 GB RAM, and an Nvidia RTX 3060 GPU, operating in a Windows environment and WSL2. CHURRO entire setup is deployed as a Docker container and is accessed through a Python-based command-line interface. The syntax is intentionally minimal, only requiring the specification of image files or directories for transcription and the ones for returning the transcribed document as .xml files. Although CHURRO provides optional preprocessing utilities such as page cropping, multipage detection, PDF-to-image conversion, binarization, etc. these modules were not used for this preliminary work, and the raw photos of the original handwritten documents were processed without alterations.

Three samples were chosen to capture the variety of the handwritten text within the corpus and to illustrate the range of legibility and structural organization of the documents: from a highly regular and easy to read script [Fig. 01. left], to an intermediate example [Fig. 02.], and finally to a more complex document presenting marginal notes, a disordered layout, and a less clear handwriting [Fig. 01. right]. Lastly, two typeface examples related to the research project are included as a reference in [Fig. 04.].

The page in [Fig. 02.], representing the intermediate case, was used as input for CHURRO without any pre-processing. The raw output is reported on the following page, alongside an interpretation drawn up manually in accordance with transcription standard. In both, **bold** indicates transcriptions error that introduce relevant alteration to the original meaning of the text, but that could be plausible and hard to identify without the original text as a reference, while underlined text highlights minor inaccuracies that can be easily identified and corrected. These issues are also reported highlighting or underlining in red the corresponding location on the original document in [Fig. 02.]. Based on the original manuscript, we identified five possible types of error: Abbreviation expansion errors (Abbr): incorrect interpretation or expansion of scribal abbreviations or contractions; Hallucination (Hall): words or phrases introduced by the model that are not present in the source text; Hyphenation errors (Hyph): incorrect or missing line-break hyphens as well as interpretation errors due to the presence of the hyphenation; Modernization (Mod): unmotivated normalization or archaic spelling, grammar, or diacritics; Source corruption (Src): incorrect transcription caused by illegible source either due to script deterioration or ambiguous calligraphy rather than model misunderstanding; General transcription error (Gte): ambiguous or mixed cases. The two significant holes, caused by the archival storage method “*in filze*”, are highlighted in green. Although interrupting words, they had no effect on the transcription.

In addition to the transcription, CHURRO also outputs a brief description of the text based on the information it contains. In this case, the description provided is: “*A letter from an unidentified*

Human transcription:

Illustrissimo Signore

*Questa letera doverà copiar^{si} la prima **per mandarla** poi al capo d'opra Claudio Storace acciò / la sottoscriva acciò possa trasmettersi **a chi già** si dieda l'**incommodo** di mandarla.*

*Avendo ponderato il Magistrato Illustrissimo de Signori **Padri** / del Commune della Serenissima Repubblica di **Genova** / l'**erudita** perizia di Vostra Signoria Illustrissima contenuta in lettera / di 19 scorso **maggio** diretta a Sua Eccellenza il Signore Prencipe / di **Giraci**, e responsiva al progetto / rimessole della diversione delle acque nel publico / acquedotto. Ha stimato di adottare nell'alternativa / richiesta il sentimento di Vostra Signoria Illustrissima cioè **di far** tras- / corere ne tubi di ferro con inclinazione di linea / retta da prepararsi e non gibbosa, et **iregolare** / tutta l'acqua del publico acquedotto dalla montagna / di Morasana all'altra oposta di **Pino** non sotto / del **Geriato** torrente, ma bensì sopra del piccolo ponte / da costruersi a norma del disegno inviatole. / Prima però di determinarsi a quanto sopra il Prefetto / Illustrissimo Magistrato a fatta mattura considerazione, ad ambe- / due i progetti occasionalmente rilevati da Vostra Signoria Illustrissima / cioè di conservare la continuazione del corso delle / acque nel giro del vecchio acquedotto che si tratta / di abbandonare, mediante la restaurazione e ria- / dattamento e parziale riedificazione dello stesso / oppure di formare il passaggio coperto dell'acquedotto / da Morasana a Pino con una retta continuata areata / sostenuta dallo sottoposto ponte, ma ha incontrate [...]*

CHURRO raw transcription:

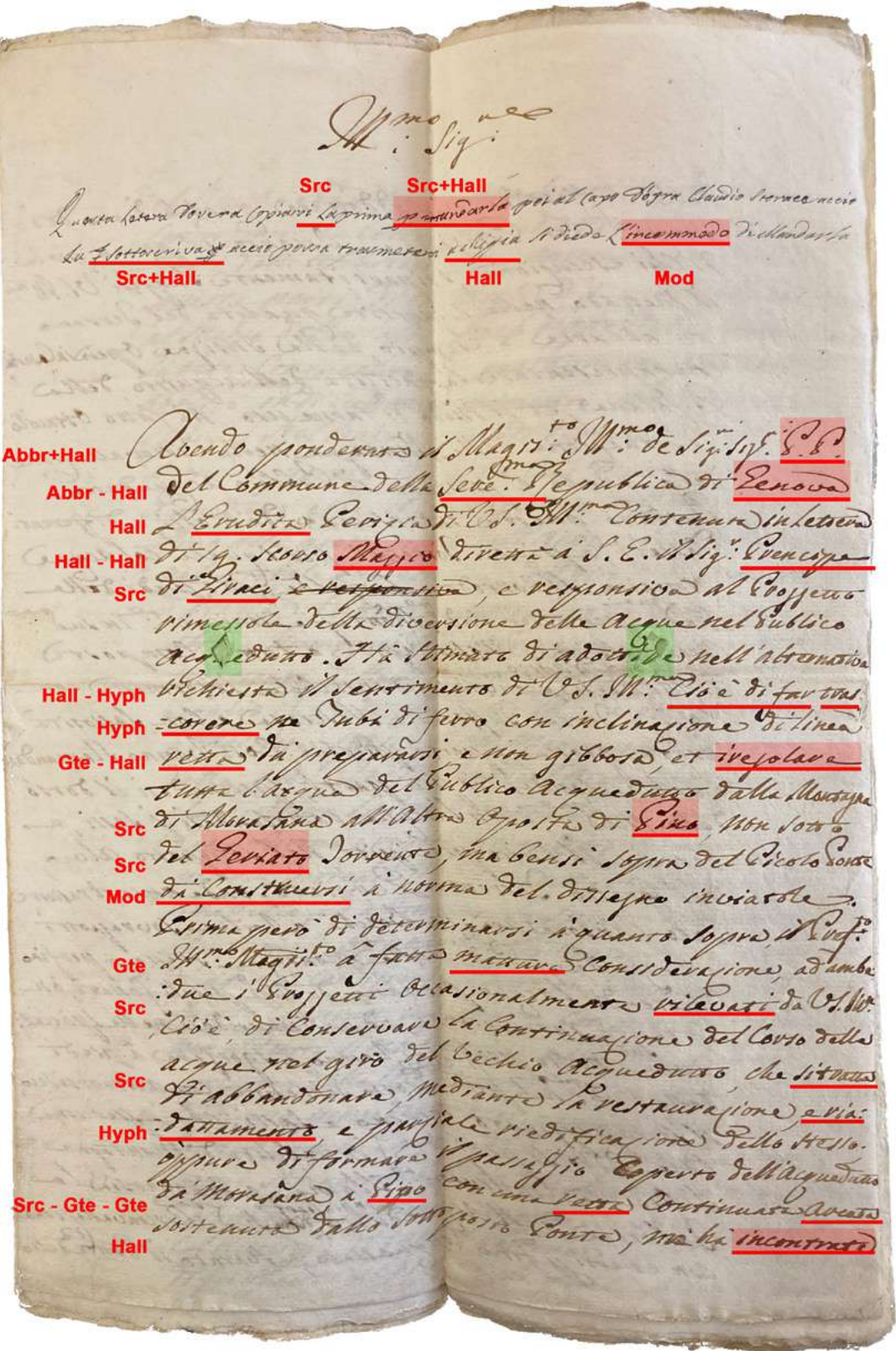
Ill.mo Sig.r

*Questa lettera dovera copiarvi la prima **promuo varla** poi al capo d'opera Claudio Storace accio / la trasporeriva accio possa trasmettervi il **siglia** si diede l'incarico di mandarla.*

*Avendo ponderato il Magist.tò Ill.mo Sig.r **S.E.** / del Comune della Seva Republica di **Leva** / L'**Eredità** Perizia di V.S. Ill.ma contenuta in lettera / 19. scorso **Magist.to** diretta à S.E. il Sig.r Provvo / di Tiraci, e responsiva al Progetto / rimessole della diversione delle Acque nel Publico / Acquedotto, Ha stimato di adottare nell'alternativa / richiesta il Sentimento di V.S. Ill. ma Ciò è **difatto** as / corere ne tubi di ferro con inclinazione di linea / rette da prepararsi e non gibbosa, et **ivi solare** / tutta l'acqua del Publico Acquedotto della Montagna / di Morasana all'Altra Opposta di **Tino**, Non Sotto / del **Leviato** Torrente, ma Bensì sopra del Piccolo Ponte / da. Costruirsi a norma del disegno inviato. / Prima però di determinarsi à quanto sopra, il Ref.to / Ill. mo Magist.ro à fatto Matture considerazion, ad ambi / due i Progetti Occasionalmente vilavari da V.S. Ill.m / Ciò è, di Conservare la Continuazion del Corso delle / acque nel giro del Vecchio Acquedotto, che sitvatta / L'abbandonare, mediante la restaurazion, e via / dattamento, e parziale riedificazion dello stesso. / oppure di formare il passaggio Coperto dell'acquedotto / di Morasana à Tipo con una Verda Continua a vece / Sostenuta dallo sopposto Ponte, mi hi **incontrast** [...]*

author to an illustrious recipient (Il Sig.re), concerning a proposal by the Commune of Leva regarding the continuation and improvement of the Morasana aqueduct. The author mentions that the Magistrato della Repubblica di Leva has considered two alternative plans and is now tasked with making a decision based on the Consiglio's opinion. The author also expresses a willingness to contribute further to the project if it is deemed necessary". The information provided is not entirely correct, but the result is remarkable considering that the model was only provided with a single page and no further context. The raw output is provided as an .xml file formatted according to the HistoricalDocument XML schema, which allows for the codification of the original structure of the document, comprehensive of headings, paragraphs, and lines, as well as marginal notes, figures, and other features. A sample of the output is reported in [Fig. 03].

3rd document. First page of the handwritten document Lettera al Magistrato Illustrissimo anzi dell'architetto Claudio Storace al Vanvitelli di Napoli, 12th June 1772 (ASCG, Magistrato dei Padri del Comune, filza 303). Manually annotated to show CHURRO transcription error types and severity.



DISCUSSION AND CONCLUSIONS

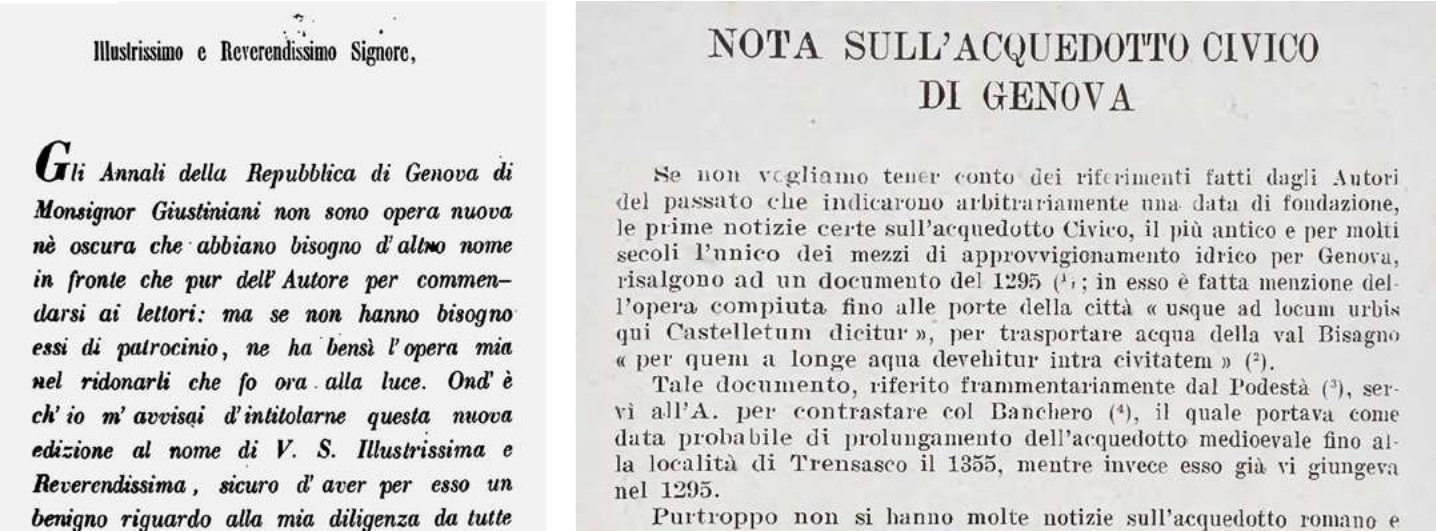
In most cases, the transcriptions provided by CHURRO were faithful to the handwritten text; however, we occasionally observed adaptation of the Italian to modern usage, without preserving the peculiarities of 18th-century writing. These peculiarities take on greater importance since, during this period, Italian underwent a series of linguistic developments that can be traced precisely between the lines of the documents and may assist in establishing a hypothetical chronological sequence between undated documents. The model showed difficulties in correspondence of hyphenations and shortenings, where recurring errors were observed. However, most of the errors appear to be related to difficulties related to the script itself. Overall, the raw output provides a solid basis for a correct and rapid transcription, while the brief description could be useful for indexing and searching. The typed documents tested (e.g. [Fig. 04.]), on the other hand, presented no transcription errors during preliminary testing. This, as expected according with CHURRO accompanying paper, confirmed a higher capability in recognising printed characters. For similar reasons, manuscripts with cleaner handwriting produced better transcriptions compared to examples such as the one presented in [Fig. 01.] right, which yielded the highest number of errors within the sample.

```
1 <HistoricalDocument xmlns="http://example.com/historicaldocument">
2   <Metadata>
3     <Language>Italian</Language>
4     <WritingDirection>ltr</WritingDirection>
5     <PhysicalDescription>Handwritten letter on a single sheet of paper, recto.</PhysicalDescription>
6     <Description>A letter from an unidentified author to an illustrious recipient (Il Sig.re), concerning a proposa
7     <Script>Latin</Script>
8   </Metadata>
9   <Page>
10    <Body>
11      <Heading type="main">
12        <Line>Ill.mo Sig.r</Line>
13      </Heading>
14      <Paragraph>
15        <Line>Questa lettera dovera copiarvi la prima promuo varla poi al capo D'opera Claudio Storace accio</Line>
16        <Line>La trasporteriva accio possa trasmettervi il siglia si diede l'incarico di mandarla</Line>
17      </Paragraph>
18      <Paragraph>
19        <Line><Initial type="decorated">A</Initial>vendo ponderato il Magist.to Ill.mo de Sig.r S.E.</Line>
20        <Line>del Comune della Seva Republica di Leva</Line>
21        <Line>L'Eredità Perizia di V.S. Ill.ma contenuta in lettera</Line>
22        [...]
40        <Line>oppure di formare il passaggio Coperto dell'acquedotto</Line>
41        <Line>di Morasana à Tipo con una Verda Continua a vece</Line>
42        <Line>sostenuta dallo sopposto Ponte, mi hi incontrast,</Line>
43      </Paragraph>
44    </Body>
45  </Page>
46 </HistoricalDocument>
```

Sample of the raw output of the transcription of the 3rd document in the HistoricalDocument XML schema, showing metadata and the structured transcription of the handwritten source.

This preliminary experimentation conducted using CHURRO showed that AI methodologies can offer a reliable and cost-effective support for the transcription of historical handwritten and typewritten documents in the context of architectural heritage preservation, providing a solid starting point for the digitalization of large handwritten archives. Even without specific training or adaptation to our corpus, CHURRO demonstrated to be capable of producing consistent results, confirming its applicability to a wide range of archival sources. Moreover, the possibility of local execution on consumer hardware allowed compliance with the aforementioned legal and ethics issue, also assuring economic sustainability. Automating the transcription of handwritten and typewritten documents significantly speeds up the work of studying and deepening the knowledge that forms the basis of a good restoration project. Furthermore, the ability to query documents facilitates comparison between them and the systematization of the corpus based on dates, places, people, etc. This systematized information

guarantees researchers easier access to historical memory and its use in the study of artefacts. At present, some trial documents have been transcribed using CHURRO, but the methodology developed so far is a work in progress. For the purposes of the ongoing research on the siphon bridge and, in general, on the Genoa's historical aqueduct, further documents could be manually transcribed and used to further fine-tune the model. A further development, in addition to refining the training, could be to work on more complex documents that contain not only words and numbers, but also drawings, sketches, annotations and calculations in the margins (Cheng *et al.* 2024). Conversely, the use LLMs for self-correction and transcription refinement, despite being available in CHURRO, has shown limited effectiveness for non-English corpora (Crosilla *et al.* 2025).



04. Two examples of typewritten texts samples on which we tested the CHURRO model, achieving excellent results. The first, on the left, is taken from Giustiniani A. (1537). *Annali della Repubblica di Genova*, while the second one, on the right, is taken from Pesce G. (1940). *Nota sull'Acquedotto Civico di Genova*.

ACKNOWLEDGEMENTS

We would like to thank the Archivio Storico del Comune di Genova for granting access to the historical documents used in experimentation and allowing publication, and Sina J. Semnani from Stanford University for his kind support in setting up CHURRO.

NOTES

- 01| Annotations and comments in the margins of a text.
- 02| Adding initial letters and titles, usually in red, to highlight certain sections.
- 03| Decoration with gold, silver or brilliant colours, elaborate designs or miniature pictures.
- 04| Archivio Storico del Comune di Genova.
- 05| Biblioteca Universitaria di Genova.
- 06| Centro di Documentazione per la Storia, l'Arte e l'Immagine di Genova.

BIBLIOGRAPHICAL REFERENCES

AlKendi W., Gechter F., Heyberger L. and Guyeux C. (2024). *Advancements and Challenges in Handwritten Text Recognition: a Comprehensive Survey*. In "Journal of Imaging", 2024, 10 (8). <https://doi.org/10.3390/jimaging10010018>

Cheng L., Frankemölle J., Axelsson A. and Vats E. (2024). *Uncovering the Handwritten Text in the Margins: End-to-End Handwritten Text Detection and Recognition*. In "Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024)", St. Julian's (Malta), pp. 111–120.

Crosilla G., Klic L. and Colavizza G. (2025). *Benchmarking large language models for handwritten text recognition*. In "Journal of Documentation", 81 (7), pp. 334–354. <https://doi.org/10.1108/JD-03-2025-0082>

Garrido-Munoz C., Rios-Vila A. and Calvo-Zaragoza J. (2025). *Handwritten Text Recognition: a Survey*. <https://doi.org/10.48550/arXiv.2502.08417>

Kahle P., Colutto S., Hackl G. and Mühlberger G. (2017). *Transkribus – A Service Platform for Transcription, Recognition and Retrieval of Historical Documents*. In "2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)", 2017, Kyoto (Japan), pp. 19–24. <https://doi.org/10.1109/ICDAR.2017.307>

Kiessling B., Tissot R., Stokes P. and Stökl Ben Ezra D. (2019). *eScriptorium: an Open Source Platform for Historical Document Analysis*. In "2019 International Conference on Document Analysis and Recognition Workshop (ICDARW)", pp. 19–24. <https://doi.org/10.1109/ICDARW.2019.10032>

Romein C.A., Rabus A., Leifert G. and Ströbel P.B. (2025). *Assessing Advanced Handwritten Text Recognition Engines for Digitizing Historical Documents*. In "International Journal of Digital Humanities", 2025, 7, pp. 115–134. <https://doi.org/10.1007/s42803-025-00100-0>

Semnani S. J., Zhang H., He X., Tekgürler M. and Lam M. S. (2025). *CHURRO: Making History Readable with an Open-Weight Large Vision-Language Model for High-Accuracy, Low-Cost Historical Text Recognition*. <https://doi.org/10.48550/arXiv.2509.19768>