AIDAinformazioni
ISSN 1121–0095
ISBN 979-12-5965-292-8
DOI 10.57574/596529286
pag. 115-132 (gennaio-giugno 2023)

Handwritten Text Recognition as a digital perspective of Archival Science

Salvatore Spina*

Abstract: The digital divide in the Humanities scientific field represents a heated debate between traditionalists (analogue scholars) and digital humanists. While some progress has been made towards a dialogue between the two sides, friction persists. Technology, including the development of artificial intelligence tools and algorithms, is not a threat to humanities research but a solution to problems. However, society has changed dramatically, and new generations require new communicative products and systems. The debate becomes increasingly wearying, but "analogues" do not consider the mutation of reality's interpretive patterns and prefer to rely on the "death of traditional Humanities statutes". The closed-mindedness towards Information technology and communication (~~ITC~~) tools, such as Handwritten Text Recognition (HTR), makes no sense. This study aims to demonstrate the potential of automatic transcription as a helpful tool for archival fields and research.

Keywords: Transkribus, Biscari Archive, Italian Administrative model, Filemaker, Digitization.

1. Introduction

If we were to talk about the *digital divide*, indeed, the world of the Humanities is a valid testing ground for the exemplification of the debate that pits, on the one hand, traditionalists (analogue scholars) against digital humanists. The heated and profound debate often shows a position of fear and anxiety about the future of research and does not look for mediation. Of course, some progress has been made in this direction, and a likely dialogue – always calibrated on the absence of trust – could mediate the ideas of both parties. However, there remains constant friction between the parties involved. The underlying assumptions behind the positions can probably be poorly understood, especially by those who look at technology as an inadequate solution to a problem that, actually, has been solved since the invention of the wheel, thanks to a mechanical-technological invention. Just think of Pascaline. If technology has always been the solution to all those problems dictated by

[*] Università degli Studi di Catania, salvatore.spina@unict.it

human resource inadequacy, then the development of artificial intelligence tools, neural systems and algorithms should not be considered a menacing weapon for eliminating the human operator/worker. However, in the face of innovation, the "day of judgment" and the "embodied evil" have been the only weapons available to "traditionalists".

«[Is p]assing, O men, Satan the great»[1] (Carducci 1964) – with these words, for example, trains were greeted as they clattered through Europe, which had recently invented the railway infrastructure. Society has changed dramatically, not in terms of government and political systems –which remain expressions of business interests devoid of concern for the community – but in the demands of new generations for communicative products that have, beyond deliberate consequences, put ancient disciplines to the corner, unable to keep up with humans evolution towards the structuring of a new cognitive entity that, in an epigenetic key, will be perennially connected (it is *Homo-Loggatus*) (Spina 2022a). The Turing Machine and the Internet network will log every intellectual and rational activity. Evolution does not go in a direction different from this just stated. For this reason, the debate among humanists becomes increasingly wearing, especially for "analogues," whose position does not take into account a mutation of reality's interpretive patterns. Thus, rather than promoting the evolution of discipline statutes, trying to make the new language of Sciences –Information Technology – its own, it is preferred to appeal to the "death of humanities research" (Valacchi 2021) or to the impossibility, for instance, of archivists, to move in the direction of building the Big Data complexes of History.

In this one-sided debate perspective, the strong rejection of ~~ITC~~ tools such as HTR makes no sense. For this reason, the study conducted on the part of the "Correspondence" section of the Paternò-Castello family archive aims to demonstrate the potential of the Transkribus technology and, at the same time, be a clear invitation to create an automatic transcription model that could become an essential tool for the correct digitisation of the Italian archives' heritage.

Digitalisation represents a frontier, but that, in many ways, is not viewed from the correct perspective. Despite clear ideas between those who fear (because they do not know) computer technologies and those who see it as a logical, technologically valuable tool for scientific research, the meaning that humanists attribute to the term "digitisation" is still incorrect. Therefore, clarification is mandatory: "To digitise" means translating analogue pieces of information into machine-readable language. To digitalise is to "encode" any statement from one language to another. Remembering, therefore, the events that marked the development of the first computers (in the 1930s), the correct

[1] «Come di turbine / L'alito spande: / Ei passa, o popoli, / Satana il grande. / Passa benefico / Di loco in loco / Su l'infrenabile / Carro del foco».

digitalisation required the creation of punch cards on which information was formalised in binary code or in a “proposition” that the ~~Main Frame~~ could analyse. Fortunately, nowadays, “to digitalise” does not mean writing in binary code, but this does not mean that data must be encoded in digital text to apply various computational tools.

“Digitisation” and “digitalisation” are often used interchangeably but have different meanings[2]. “Digitisation” is often seen as merely “taking a picture” of an archival document and uploading it to a database.

Many projects that involve the digital acquisition of archival documentation beyond metadata do not operate in this direction; that is, complete digital texts are not adequately realised. This deficit, of course, becomes an obstacle to a more profound use of the semantic web.

This viewpoint is limited as it fails to take into account that a picture, in the digital environment, cannot be read by an artificial intelligence platform in the same way historians read the text contained within it. For a computer, an image is simply a data set, such as colour codes, dimensions, and dpi. Consequently, a proper digitisation process of an archival document requires more than photographic acquisition if we aim to create its digital version. The latter necessitates the accurate transcription of every word present in the image to compile the document’s digital edition.

Therefore, starting from the late 1950s with the rise of ~~Dimond~~, the idea of developing technologies for automatic recognition of handwritten text became increasingly (Adamek, O’Connor, and Smeaton 2007; Albertin et al. 2016; Dunley 2018; Spina 2022b) intense:

> In the last five years, much thought and effort have gone into the development of printed character-recognition devices. Varying degrees of success have been achieved. In some cases, ingeniously distorted typefaces have been required. One might wonder why all this interest exists. The answer is simple. Character-recognition devices help reduce the cost of getting information into forms that computers can understand (Dimond 1957).

As we know – of course – when we delve into the field of archival or historical research, the most severe difficulties in using computer technologies are related to the existence of certain limitations: archival documentation mainly consists of manuscripts, a significant challenge for using computer tools to study events in the Modern Age – a situation where challenges of Archival Science and archivists occur.

The debate on the usefulness of computer tools seems not to affect the principles of this discipline. Despite the scientific and technological dynami-

[2] (1) “Digitization,” which refers to the conversion of analog information into a digital format; (2) “Digitalization,” which involves the native creation of information in a digital format; (3) “Digital transformation,” which is the process aimed at the exclusive adoption of technology in production processes.

sm, archivists take refuge behind statutory principles that seem to make this scientific field resistant to the needs of the research world.

What emerges from monographs and many scientific publications is a desire to defer the confrontation with computer technology, which could improve the activities of archives and archivists, accusing historians and scholars of fragmenting documentary resources when published on some websites (Valacchi 2021) – this is because the multiplication of users through web access requires a reassessment of the paradigms of archival science and a stance that makes the difficulties of archivists in realising digitalisation projects adequate to the demands of the scientific community and, on the other hand, from computer technology, which considers the simple photographic acquisition of documents useless.

While the Italian National Archival System is advertised as one of the best products for digitising the national archival heritage, we face millions of photographs that remain meaningless for the semantic web neural system and the everyday user, who cannot perform an analytical search – this is because no transcription phase follows the photograph, neither human (reasonably impossible) nor automated.

HTR computer tools seem to be little known to the most prepared archivists, whose function, according to the dictates of their discipline, remains to preserve and not to make the archival heritage accessible through the Internet.

The limits are undoubtedly evident. It will not be possible to digitise and transcribe every Italian archival document quickly. However, it is also true that Artificial intelligence like Transkribus and the creation of increasingly performing transcription models could boost the creation of a digital heritage that can guarantee concrete access to the sources of our Past.

So, thanks to the Transkribus artificial intelligence platform (Erwin 2020; Kahle et al. 2017; Massot, Sforzini, and Ventresque 2018; Milioni 2020; Muehlberger et al. 2019), this limit begins to waver, as demonstrated by the test I carried out on the archival documentation of the Paternò Castello family, Princes of Biscari.

2. Biscari Archive and HTR Transkribus

Consider the assumption: a text depicted in a picture – in a technological process that, nowadays, is digitising objects, places, and men – cannot be read by artificial intelligence tools in the same way historians read the exact text. To a computer, a picture is a set of other kinds of data, such as colour code – e.g., #0000 is 'black', #654321 is 'brown', #FF0000 is 'red', and so on –, definition, number of pixels, and location information if the camera has Global Positioning System (GPS). Thus, a proper digitisation process would involve more than just taking pictures of historical documents.

Despite the limitations mentioned, there are still important questions to be addressed when it comes to digitising archival heritage: Is taking a photo the right way to digitise it? Can a computer read handwritten texts? How can computer tools and artificial intelligence enhance the profession of historians and archivists?

Alan Turing already posed the question: "Can machines think?" (Turing 1950) ... a question that has sparked debates and, above all, perspectives that, to this day, place scholars on opposing theoretical fronts based on unscientific and unconvincing foundations. In our case, the question is much simpler: "Can machines read?"

If we considered reading solely from the perspective of mechanising the act of recognising a character, the answer would undoubtedly be affirmative. However, everything changes when computers, hardware, and software do not operate on machine-readable texts.

Nevertheless, starting from the concept of "learning" – which characterises the life of humans who, after proper and adequate schooling, can recognise written signs and reproduce them – even a software or artificial intelligence tool can do the same – it is the "machine learning" (Alpaydin 2020; Bishop 2006; Collobert and Weston 2008; Gori 2018) –, while still maintaining a distinct semantic level, in which the machine remains deficient.

A study of the Biscari Archive in Catania can help address these issues and demonstrate the feasibility of using HTR technology in historical research.

Most archivists believe that the digital methodology consists of scanning or photographing some archival document and uploading it to the official archive website. However, a photo – as previously stated – cannot be read by artificial intelligence tools the same way scholars read the text of the image. If we want to conduct an archival or historical study by analysing some depicted text, no computer can figure out all the information it carries because it is inaccessible even to the Turing Machine. If this were the case, the computer could only reproduce words without relation to reality because language is a code structured on non-explicit elements, inconsistency between terms and actions and other aspects that computers cannot process. Furthermore, even if a computer can analyse a digital text, it cannot read and recognise a handwritten word, not even a single consonant or vowel.

The Biscari Archive has 2,000 folders with hundreds of thousands of sheets documenting legal disputes, political decisions, and business and personal letters. The "Correspondence" section consists of some 42,493 papers, in which, in folder 1.642, we can find 591 sheets, constituting 366 letters and a manuscript by Emile Rousseau. So, to test the capabilities of Transkribus AI, it was decided to use these 367 documents.

The acquisition process was performed using a Nikon D610 camera with an AF-S Nikkor 24-120mm f/4G ED VR lens. The photos were collected in a database created using Claris FileMaker 19 software, as no website was planned for this project stage. After adding metadata, the photos were converted to PDF format to reconstruct the 366 letters, and 53 were uploaded to the Transkribus server for automatic transcription.

Transkribus, powered by Java technology, allows for creating workflows based on deep neural networks that can be trained to recognise specific handwriting, and to achieve the best results, texts must be written by the same hand. For this reason, 53 letters were selected, consisting of 28 letters from Michele Maria Paternò to Princess Anna Maria Morso Bonanno and 25 letters from Marquis Giovanni Fogliani Sforza D'Aragona to various recipients.

With the help of Gephi software (Bastian, Heymann, and Jacomy 2009), 53 letters were pinpointed and selected – consisting of 28 letters from Michele Maria Paternò to Princess Anna Maria Morso Bonanno and 25 letters from Marquis Giovanni Fogliani Sforza D'Aragona to various recipients (Fig. 1) – and sorted into two separate PDF files, which were then uploaded to the Transkribus server.

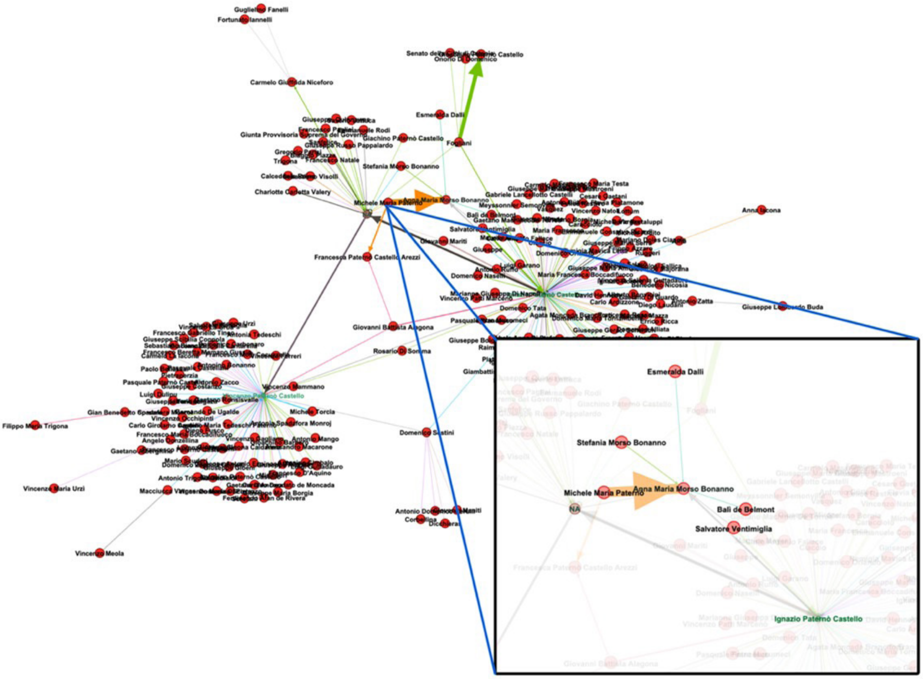


Figure 1: Correspondence between Michele Maria Paternò and Anna Maria Morso Bonanno.

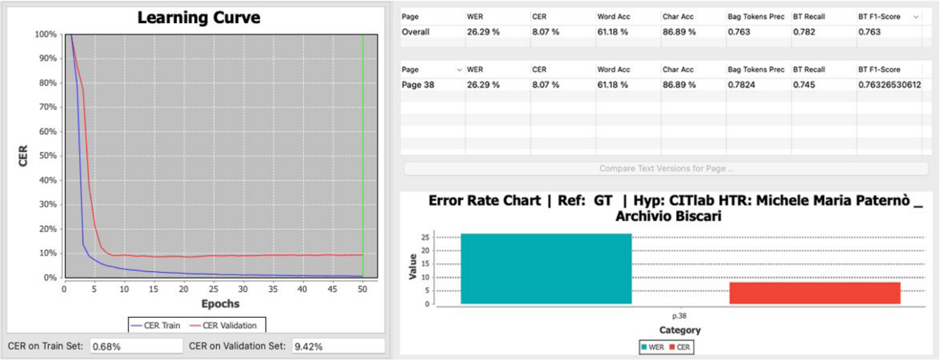
Thanks to the small corpus size, it allowed for quick and accurate automatic transcription. Transkribus aim to support historical research by transcribing large volumes of handwritten texts. However, the Italian scientific community is still reluctant to use state-of-the-art IT tools for historical research, leading to a delay in digitising historical heritage. In contrast, Finland has embraced AI for historical research by involving students and scholars in training models to minimise errors in transcriptions. One of the core features of Transkribus is the ability to involve the broader community in digital transcription projects, making the goal of Public History achievable.

3. Outcome

The automatic transcription of the letters was done without any prior training to evaluate the accuracy of the public Transkribus AI model. The "Italian Administrative Hands 1550-1700" model (Midura 2020)[3], created by Jake Dyble, Antonio Iodice, Sara Mansutti and Rachel Midura, with a Character Error Rate (CER) of 9.15%, was used for the transcription. This result is considered excellent for a public model in the automatic transcription of manuscripts in Italian archives, which hold the most diverse historical heritage in Europe due to various writing styles, languages, and institutions – *e.g.*, Latin in the Vatican, Spanish in the Kingdom of Sicily and the Duchy of Milan, Italian in the Grand Duchy of Tuscany.

The model was applied when the two PDF files were uploaded to the Transkribus server. The error rate for Marquis Fogliani's letters was 7% (e.g., 6 out of 82 words were not transcribed correctly in the first letter) and remained constant throughout. However, the error rate was much higher, over 10%, for the 28 letters written by Michele Maria Paternò, due to incorrect Italian words and words unknown to the model. The graphemes "V.S." and "S.M." were also not transcribed, even though they were present in the text image segmentation, resulting in a higher error rate and an inability for the user to correct it.

[3] The "Italian Administrative Hands" model features a variety of Italian-language documents only from State Archives in Milan (from the *Carteggio delle cancellerie dello Stato*, *Atti di Governo*, and *Registri delle cancellerie dello Stato* collections), Venice (from the *Compagnia dei corrieri*, *Senato*, and *Inquisitori di Stato* collections), Florence (fond *Mediceo del Principato*, series *Relazioni con stati italiani ed esteri*), Pisa (from the *atti civili* sub-series of the archive of the *Consoli del Mare di Pisa*), and Genoa (from the *Notai Giudiziari* and the *Conservatori del Mare* collections).



CER on Train Set: 0.68%   CER on Validation Set: 9.42%

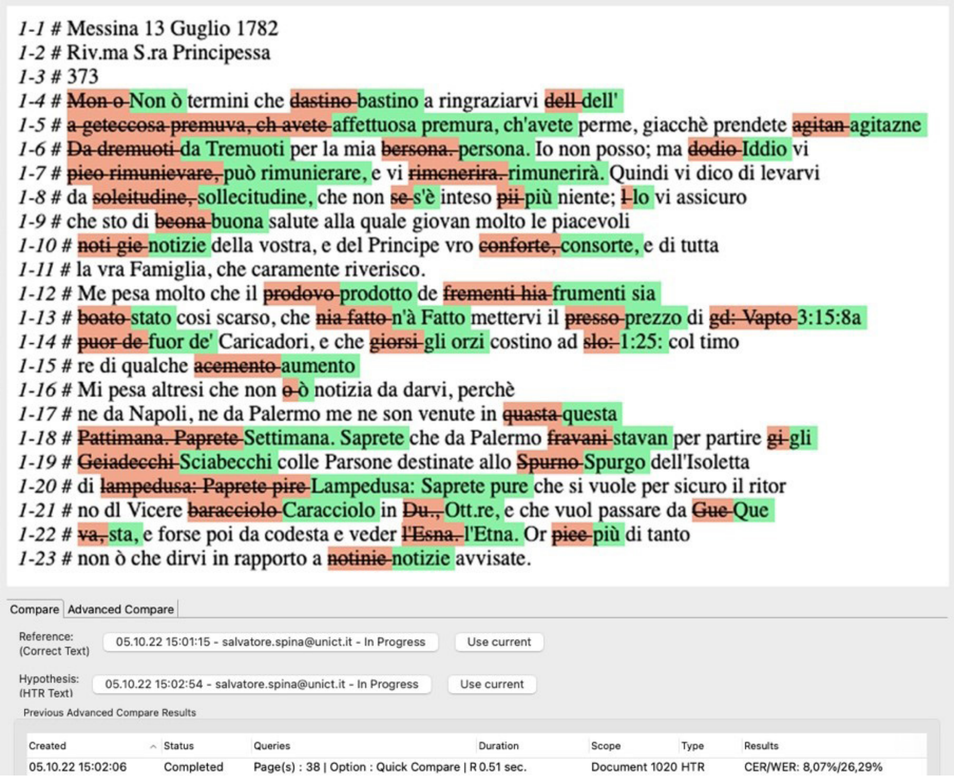
Page	WER	CER	Word Acc	Char Acc	Bag Tokens Prec	BT Recall	BT F1-Score
Overall	26.29 %	8.07 %	61.18 %	86.89 %	0.763	0.782	0.763

Page	WER	CER	Word Acc	Char Acc	Bag Tokens Prec	BT Recall	BT F1-Score
Page 38	26.29 %	8.07 %	61.18 %	86.89 %	0.7824	0.745	0.76326530612

Compare Text Versions for Page ..
Error Rate Chart | Ref: GT | Hyp: CITlab HTR: Michele Maria Paternò _ Archivio Biscari
Value
25
20
15
10
5
0
p.38
Category
WER
CER

Figure 2: Correspondence between Michele Maria Paternò and Anna Maria Morso Bonanno.

Therefore, a new model was trained based on the "Italian Administrative" one, using the 24-page "ground truth" as the starting point (pages 1-5, 7-9, 11-18, 20-24, 26, 27, 35). The CER of the Training Set dropped to 0.68% and that of the Validation Set to 9.42% after the training and formation of the "Michele Maria Paternò_Archivio Biscari_Model" (Fig. 2). Comparing the "ground truth" and automatic transcription using the latter model showed a CER of 9.88% (Fig. 3). The validation set had a relatively high percentage of 35.57%, indicating the need for training on a more extensive "ground truth" of at least 50 pages. Nevertheless, the "Michele Maria" model resulted in a transcript with a CER of 8.07%/26.29 and minor errors that do not impede scholarly inference but make computational analysis unreliable.



1-1 # Messina 13 Guglio 1782
1-2 # Riv.ma S.ra Principessa
1-3 # 373
1-4 # ~~Mon o~~ Non ò termini che ~~dastino~~ bastino a ringraziarvi ~~dell~~ dell'
1-5 # ~~a geteccosa premuva, ch avete~~ affettuosa premura, ch'avete perme, giacchè prendete ~~agitan~~ agitazne
1-6 # ~~Da dremuoti~~ da Tremuoti per la mia ~~bersona.~~ persona. Io non posso; ma ~~dodio~~ Iddio vi
1-7 # ~~pieo rimunievare,~~ può rimunierare, e vi ~~rimenerira.~~ rimunerirà. Quindi vi dico di levarvi
1-8 # da ~~soleitudine,~~ sollecitudine, che non ~~se~~ s'è inteso ~~pii~~ più niente; ~~l~~ lo vi assicuro
1-9 # che sto di ~~beona~~ buona salute alla quale giovan molto le piacevoli
1-10 # ~~noti gie~~ notizie della vostra, e del Principe vro ~~conforte,~~ consorte, e di tutta
1-11 # la vra Famiglia, che caramente riverisco.
1-12 # Me pesa molto che il ~~prodovo~~ prodotto de ~~frementi hia~~ frumenti sia
1-13 # ~~boato~~ stato cosi scarso, che ~~nia fatto~~ n'à Fatto mettervi il ~~presso~~ prezzo di ~~gd: Vapto~~ 3:15:8a
1-14 # ~~puor de~~ fuor de' Caricadori, e che ~~giorsi~~ gli orzi costino ad ~~slo:~~ 1:25: col timo
1-15 # re di qualche ~~acemento~~ aumento
1-16 # Mi pesa altresi che non ~~o~~ ò notizia da darvi, perchè
1-17 # ne da Napoli, ne da Palermo me ne son venute in ~~quasta~~ questa
1-18 # ~~Pattimana. Paprete~~ Settimana. Saprete che da Palermo ~~fravani~~ stavan per partire ~~gi~~ gli
1-19 # ~~Geiadecchi~~ Sciabecchi colle Parsone destinate allo ~~Spurno~~ Spurgo dell'Isoletta
1-20 # di ~~lampedusa: Paprete pire~~ Lampedusa: Saprete pure che si vuole per sicuro il ritor
1-21 # no dl Vicere ~~baracciolo~~ Caracciolo in ~~Du.,~~ Ott.re, e che vuol passare da ~~Gue~~ Que
1-22 # ~~va,~~ sta, e forse poi da codesta e veder ~~l'Esna.~~ l'Etna. Or ~~piee~~ più di tanto
1-23 # non ò che dirvi in rapporto a ~~notinie~~ notizie avvisate.

Reference: (Correct Text) 05.10.22 15:01:15 - salvatore.spina@unict.it - In Progress

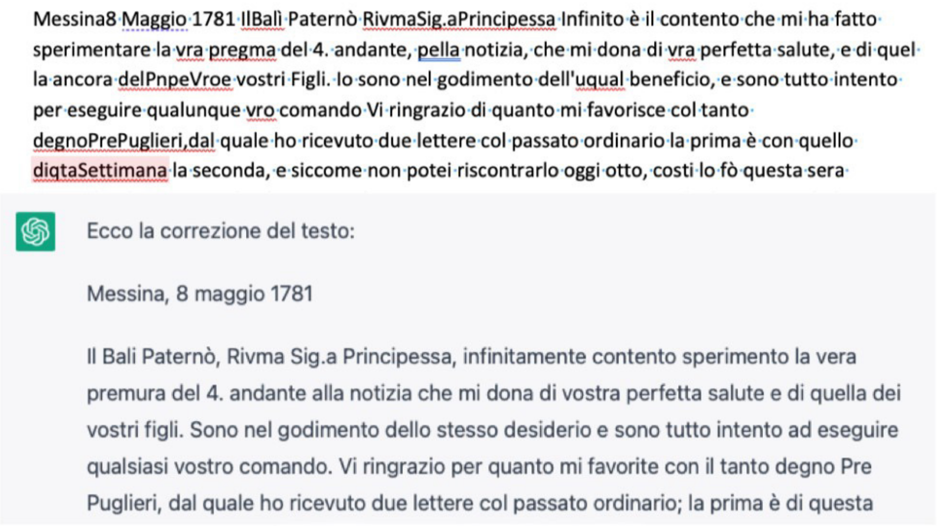
Hypothesis: (HTR Text) 05.10.22 15:02:54 - salvatore.spina@unict.it - In Progress

Previous Advanced Compare Results

Created	Status	Queries	Duration	Scope	Type	Results
05.10.22 15:02:06	Completed	Page(s) : 38 | Option : Quick Compare | R	0.51 sec.	Document 1020	HTR	CER/WER: 8,07%/26,29%

Figure 3: Comparison set.

After testing the new model, it was HTR the remaining part of the corpus, testing both the Michele Maria Paternò model and the Italian administrative, obtaining, with the latter, a transcription with minimal errors. Once exported, ChatGPT – a Large Language Model (LLM) technology by OpenAi (a digital project by Elon Musk and Sam Altman in 2015) (Fig. 4) – corrected the transcription managing to obtain a digital corpus on which it was possible to apply additional text analysis tools, such as Keyphrase Digger (Moretti, Sprugnoli, and Tonelli 2015), which allowed to dig new historical information.



Messina8 Maggio 1781 IlBalì Paternò RivmaSig.aPrincipessa Infinito è il contento che mi ha fatto
sperimentare la vra pregma del 4. andante, pella notizia, che mi dona di vra perfetta salute, e di quel
la ancora delPnpeVroe vostri Figli. Io sono nel godimento dell'uqual beneficio, e sono tutto intento
per eseguire qualunque vro comando Vi ringrazio di quanto mi favorisce col tanto
degnoPrePuglieri,dal quale ho ricevuto due lettere col passato ordinario la prima è con quello
diqtaSettimana la seconda, e siccome non potei riscontrarlo oggi otto, costi lo fò questa sera

Ecco la correzione del testo:

Messina, 8 maggio 1781

Il Bali Paternò, Rivma Sig.a Principessa, infinitamente contento sperimento la vera premura del 4. andante alla notizia che mi dona di vostra perfetta salute e di quella dei vostri figli. Sono nel godimento dello stesso desiderio e sono tutto intento ad eseguire qualsiasi vostro comando. Vi ringrazio per quanto mi favorite con il tanto degno Pre Puglieri, dal quale ho ricevuto due lettere col passato ordinario; la prima è di questa

Figure 4: Correction in ChatGPT.

4. Transkribus and Filemaker

Beyond conflicting positions, the digital divide brings out the need for an innovative approach to humanities studies. Of course, information technology can boost digital asset organisation workflows and those to be digitised. Technology allows the development of heterogeneous approaches and produces digital documentation that can be analysed, organised, and structured differently. Digitalisation is total control of the archival and historical sources.

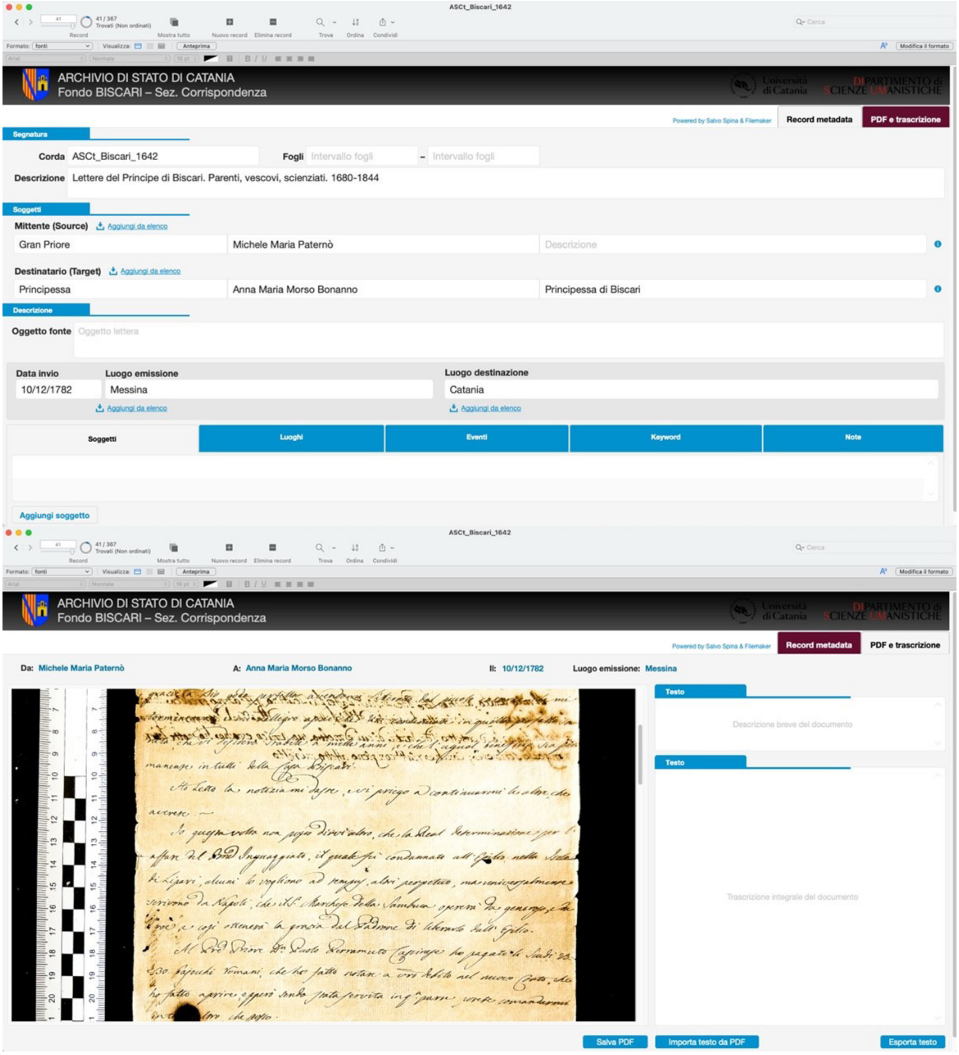
The study conducted at the Biscari Archive demonstrates how these technologies can interact and assist historians in reconstructing past events with more comprehensive and accurate information. Transkribus allows for the export of transcriptions in PDF files, which can be analysed, processed, and organised into databases, enabling searches through a user interface and relational system that highlights new information.

Thus, Transkribus is a critical tool for building the Big Data of History (Kaplan and di Lenardo 2017) and aligns with Gardin's belief that the creation of databases and ontologies must be accompanied by the proper encoding and formalisation phase, which converts historical texts into a computable format.

The 367 transcriptions from the Biscari Archive were entered into a database created with FileMaker 19, a software that facilitates the creation of a relational structure between the data. FileMaker 19.5 has also added a new JSON function, allowing for the differentiation of numbers and text and using system libraries, Optical Character Recognition (OCR), and HTR images in a “container field” (Fig. 5). This addresses the need to digitise photographic sources and prepares the databases for creating machine-readable sources.

Transkribus highlights the importance of upgrading historians’ skills in a digital environment, focusing on the most crucial aspects of source analysis: organisation and transcription.

The collections created by Transkribus users represent a valuable archival heritage beyond automated transcription capabilities. Despite being often wrongly considered disconnected fragments from the primary documentation, these collections demonstrate the perspective that Archival Science should follow to meet the needs of the highly technological society. Additionally, by opening up to society (crowdsourcing), there will be opportunities for more profound training, possibly creating a model that can transcribe the Italian archival corpus. Furthermore, the transcriptions can be exported in formats that text analysis tools can use, and, as demonstrated in our case, PDFs can be obtained that can be inserted into databases, platforms, and websites for dynamic indexing and searching. The 367 documents will be added to the website “Biscari Epistolography – Archivio Biscari Archivio di Stato di Catania” (Spina 2023), providing a means of disseminating knowledge to all connected users. The dynamism of the PDFs will also provide a text search and export function (Fig. 6), taking advantage of the full potential of computer tools for data analysis.



Segnatura

Corda: ASCt_Biscari_1642
Fogli: Intervallo fogli – Intervallo fogli
Descrizione: Lettere del Principe di Biscari. Parenti, vescovi, scienziati. 1680-1844

Soggetti

Mittente (Source): Gran Priore | Michele Maria Paternò | Descrizione
Destinatario (Target): Principessa | Anna Maria Morso Bonanno | Principessa di Biscari

Descrizione

Oggetto fonte: Oggetto lettera
Data invio: 10/12/1782
Luogo emissione: Messina
Luogo destinazione: Catania

Da: Michele Maria Paternò
A: Anna Maria Morso Bonanno
Il: 10/12/1782
Luogo emissione: Messina

[illegible] mi
terminerann[o] [illegible] allegro [illegible] in questo [illegible]
vota che vi desidero stabile a mille anni, e che l'ugual beneficio sia per
mancarsi in tutti della Casa Biscari.
Ho letto la notizia mi date, e vi priego a continuarmi le altre, che
averete —
Io questa volta non posso dirvi altro, che la Real determinazione per l'
affare del Rev.do Ingnaggiato, il quale fu condannato all'Esilio nella Isola
di Lipari; alcuni lo vogliono ad tempus, altri perpetuo, ma universalmente
scrivono da Napoli, che il S.r Marchese della Sambuca opererà da generoso, e da
[illegible]poi; e così otterrà la grazia dal Padrone di liberarlo dall'Esilio.
Al Rev.do Priore D.n Paolo Dormmuto Cassinese ho pagato li Scudi 23
e 30. bajocchi Romani, che ho fatto notare a vro debito nel nuovo Conto, che
ho fatto aprire, essendo stata servita ing.ta parte; ove te comandarmi
[illegible] altro che posso.

Testo: Descrizione breve del documento
Testo: Trascrizione integrale del documento

Figure 5: The database, Transkribus and Filemaker.

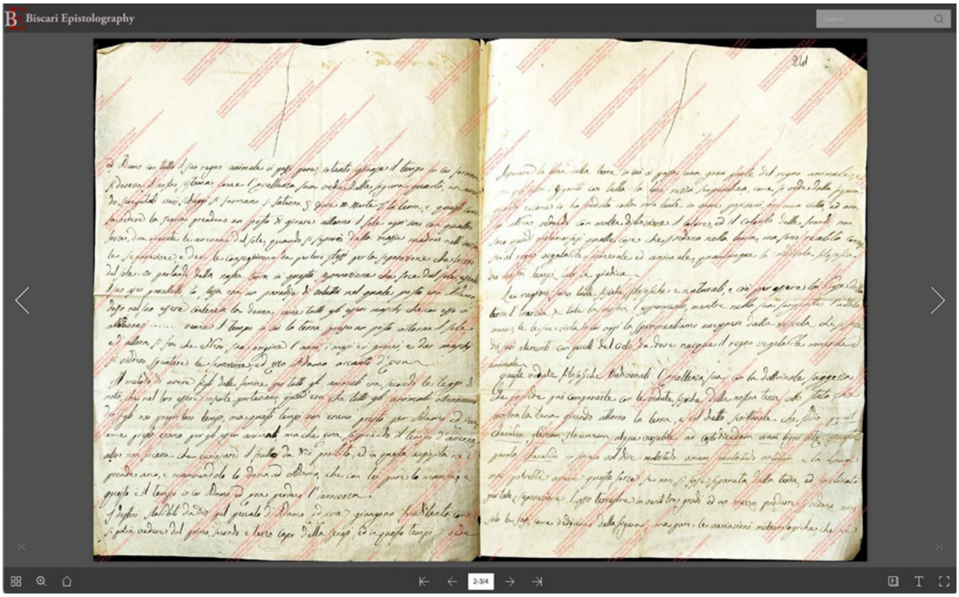


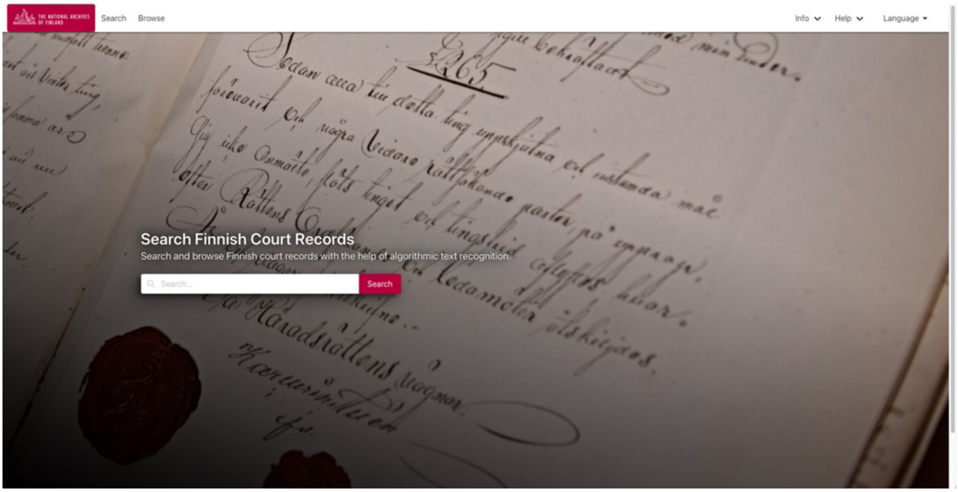
Figure 6: Pietro Musumeci flipbook on "www.biscariepistolography.it" website.

5. Achievement

The concept of "digitisation" has become widespread in our society. However, if digitisation can be considered a space-time acceleration towards a dimension of participation, communication, versatility, and usability, surely Archival Science can redefine itself within this assumption. Ordering is the most significant activity of the archivists' craft, and this requires them to give a controlled and scientific sense to a community that, as a whole, is participating in the construction of its heritage to be entrusted to the care of the archives. Indeed, anxiety arises when referring to contemporaneity. Today, every single document finds space in sedimentation that is based on digital archives, which, in addition to their traditional function, acquires the function of an open-source structure and rapid access. The discourse, although similar, takes on an even more pressing tone when it comes to historical archives, that is, documentary complexes on which, through the web, everyone wants to organise research. The paper, indeed, is there and remains there, at complete disposal. However, if the archivist is also the one who holds the keys to reconstructing the time of events, he must reasonably accept that everything that is not on the web does not exist. The post-modern, digital, digitised cultural structure is based on a clear epigenetic principle: the world changes our being in relationships and the vision of spaces, and this became an appropriate request from a community that constantly lives on the web and interpersonal communication platforms – this requires a stance that is not destructive but wants to enhance the discipline of archival science and the profession of the archivist to meet the will of web users to access their documentary heritage (current and historical). Therefore, the possibility of using tools such as Transkribus must become a "necessity". Not only to allow the scientific community to deepen its research but to decode complex documentary texts into a machine-readable digital structure.

Moreover, the task – which is currently in the hands of researchers who put some small documentary series on the network, extracted from certainly larger corpora – of this digitisation must be foundational to the vocation of archivists, who, like everyone else nowadays, can no longer exempt himself from computer training that enhances his traditional function. To this, there is a need to create automatic transcription models or boost the "Italian Administrative" with new training that features Italian-language documents from archives in the South of Italy, which can be an open-source tool for the Italian archival system.

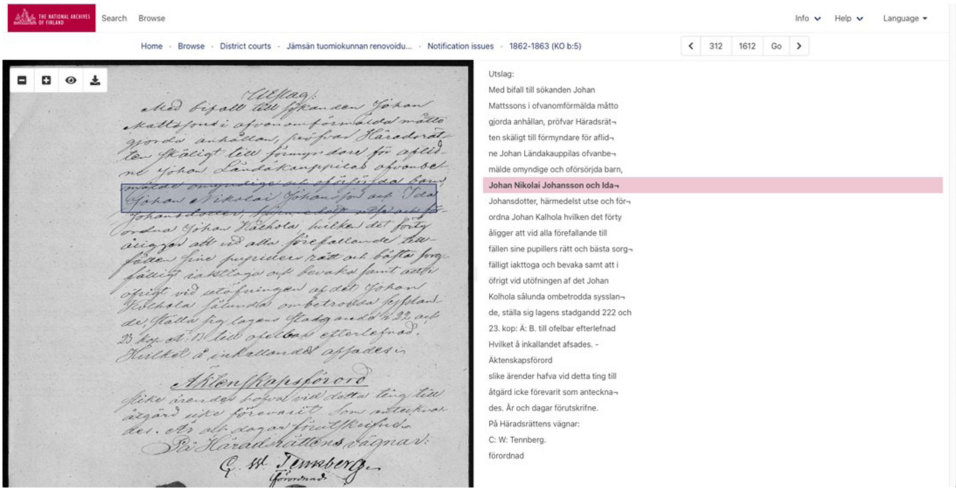
Furthermore, despite the reluctance of the Italian archivist community, Transkribus is one of the best tools to open up to society. Whilst the international archivists' community aspires to build enough digitised historical heritage, the Italians – apart from the case of the (Fondazione Banco di Napoli, n.d.) and the (Archivi Storici e Biblioteche Istituto Suor Orsola Benincasa, n.d.) – are failing to break out of the constraints of the legal protection of archival heritage, leading to a delay in digitising historical heritage. In addition, Italy lacks a scientific mindset to ensure the dissemination of its national archival heritage, unlike, for example, Finland, where, thanks to the "National Archives project" (The National Archives of Finland 2023), the Central Government wanted to restore to its community the archival heritage related to the Second World War, the judicial registers and property inventories of the Finnish nobility (Fig. 7 and Fig. 8).



Search Finnish Court Records

Search and browse Finnish court records with the help of algorithmic text recognition.

Figure 7: The National Archives of Finland website.


THE NATIONAL ARCHIVES OF FINLAND
Search
Browse
Info
Help
Language
Home · Browse · District courts · Jämsän tuomiokunnan renovoidu... · Notification issues · 1862-1863 (KO b:5)
312
1612
Go

Utslag:
Med bifall till sökanden Johan
Mattssons i ofvanomförmälda måtto
gjorda anhållan, pröfvar Häradsrät¬
ten skäligt till förmyndare för aflid¬
ne Johan Ländakauppilas ofvanbe¬
mälde omyndige och oförsörjda barn,
Johan Nikolai Johansson och Ida¬
Johansdotter, härmedelst utse och för¬
ordna Johan Kalhola hvilken det förty
åligger att vid alla förefallande till
fällen sine pupillers rätt och bästa sorg¬
fälligt iakttoga och bevaka samt att i
öfrigt vid utöfningen af det Johan
Kolhola sålunda ombetrodda sysslan¬
de, ställa sig lagens stadgandd 222 och
23. kop: Ä: B. till ofelbar efterlefnad
Hvilket å inkallandet afsades. -
Äktenskapsförord
slike ärender hafva vid detta ting till
åtgärd icke förevarit som anteckna¬
des. År och dagar förutskrifne.
På Häradsrättens vägnar:
C: W: Tennberg.
förordnad

Figure 8: An example of a digital edition of the National Archives of Finland's document.

To do this, it involved students and scholars in training a model to obtain transcriptions with minimal margins of error. Detailed and in-depth training will enhance artificial intelligence tools, enabling Transkribus to automatically transcribe the entire human heritage in all languages and eras of the Past.

A further core feature of Transkribus is that digital transcription projects of archival heritages can be open to the entire community, who thus participate in the training phases, thus realising the grand goal of Public History.

Bibliographic References

Adamek, Tomasz, Noel E. O'Connor, and Alan F. Smeaton. 2007. "Word Matching Using Single Closed Contours for Indexing Handwritten Historical Documents." *International Journal of Document Analysis and Recognition* 9 (2): 153–65. https://doi.org/10.1007/s10032-006-0024-y.

Albertin, Fauzia, Alessandra Patera, Iwan Jerjen, S. Hartmann, Eva Peccenini, Frédéric Kaplan, Marco F.M. Stampanoni, Rolf Kaufmann, and G. Margaritondo. 2016. "Virtual Reading of a Large Ancient Handwritten Science Book." *Microchemical Journal* 125 (March): 185–89. https://doi.org/10.1016/j.microc.2015.11.024.

Alpaydin, Ethem. 2020. *Introduction to Machine Learning*. Massachusetts Institute of Tecnology.

Archivi Storici e Biblioteche Istituto Suor Orsola Benincasa. n.d. Consultato il 10 febbraio 2023. https://archivistoriciisob.transkribus.eu/.

Bastian, Mathieu, Sebastien Heymann, and Mathieu Jacomy. 2009. "Gephi - The Open Graph Viz Platform." https://gephi.org/.

Bishop, Christopher M. 2006. *Pattern Recognition and Machine Learning*. New York, NY: Springer International Edition.

Carducci, Giosuè. 1964. *Tutte Le Poesie, Juvenilia, Levia Gravia, A Satana, Giambi Ed Epodi, Intermezzo, Rime Nuove, Odi Barbare, Rime E Ritmi, Canzone Di Legnano*. Segrate: Rizzoli. https://www.ibs.it/tutte-poesie-ijuvenilia-levia-gravia-libri-vintage-giosue-carducci/e/2560035385391.

Collobert, Ronan, and Jason Weston. 2008. "A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning." In *Proceedings of the 25th International Conference on Machine Learning*, 160–67. ICML '08. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/1390156.1390177.

Dimond, Tom L. 1957. "Devices for Reading Handwritten Characters." In *Papers and Discussions Presented at the December 9-13, 1957, Eastern Joint Computer Conference: Computers with Deadlines to Meet*, 232–37. *IRE-ACM-AIEE '57 (Eastern)*. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/1457720.1457765.

Dunley, Richard. 2018. "Machines Reading the Archive: Handwritten Text Recognition Software." The National Archives Blog. The National Archives. March 19, 2018. https://blog.nationalarchives.gov.uk/machines-reading-the-archive-handwritten-text-recognition-software/.

Erwin, Brittany. 2020. "Digital Tools for Studying Empire: Transcription and Text Analysis with Transkribus." *Not Even Past*. November 6, 2020. https://notevenpast.org/digital-tools-for-studying-empire-transcription-and-text-analysis-with-transkribus/.

Fondazione Banco di Napoli. n.d. Consultato il 10 febbraio 2023. http://www.fondazionebanconapoli.it/en/.

Gori, Marco. 2018. *Machine Learning. A Constraint-Based Approach*. Burlington, Massachusetts: Morgan Kaufmann Publishers. https://doi.org/10.1016/C2015-0-00237-4.

Kahle, Philip, Sebastian Colutto, Günter Hackl, and Günter Mühlberger. 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)*, 19–24. https://doi.org/10.1109/ICDAR.2017.307.

Kaplan, Frédéric, and Isabella di Lenardo. 2017. "Big Data of the Past." *Frontiers in Digital Humanities* 4. https://doi.org/10.3389/fdigh.2017.00012.

Massot, Marie-Laure, Arianna Sforzini, and Vincent Ventresque. 2018. "Transcribing Foucault's Handwriting with Transkribus." https://hal.archives-ouvertes.fr/hal-01913435.

Midura, Rachel. 2020. "Italian Administrative Hands." Early Modern Digital Itineraries. July 21, 2020. https://emdigit.org/tool/2020/07/21/italian-administrative-hands.html.

Milioni, Nikolina. 2020. "Automatic Transcription of Historical Documents. Transkribus as a Tool for Libraries, Archives and Scholars." PhD Diss., Uppsala Universitet. http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-412565.

Moretti, Giovanni, Rachele Sprugnoli, and Sara Tonelli. 2015. "Digging in the Dirt: Extracting Keyphrases from Texts with KD." In *Proceedings of the Second Italian Conference on Computational Linguistics CLiC-It 2015*. https://doi.org/10.4000/books.aaccademia.1518.

Muehlberger, Guenter, Louise Seaward, Melissa Terras, Sofia Ares Oliveira, Vicente Bosch, Maximilian Bryan, Sebastian Colutto, et al. 2019. "Transforming Scholarship in the Archives through Handwritten Text Recognition. Transkribus as a Case Study." *Journal of Documentation* 75 (5): 954–76. https://doi.org/10.1108/JD-07-2018-0114.

Spina, Salvatore. 2022a. *Digital History. Metodologie Informatiche per La Ricerca Storica*. Napoli: Edizioni Scientifiche Italiane.

Spina, Salvatore. 2022b. "Historical Network Analysis & Htr Tool. Per un approccio storico metodologico digitale all'archivio Biscari di Catania." *Umanistica Digitale* 14: 163–81. https://doi.org/10.6092/issn.2532-8816/15159.

Spina, Salvatore. 2023. *Biscari Epistolography.* Ministero della Cultura, MIC|MIC_AS-CT|03/02/2023|0000173-P. https://www.biscariepistolography.it/.

Turing, Alan Mathison. 1950. "Computing Machinery and Intelligence." *Mind* LIX (236): 433–60. https://doi.org/10.1093/mind/LIX.236.433.

Valacchi, Federico. 2021. *Gli archivi tra storia uso e futuro*. Milano: Editrice Bibliografica.