Miscellanea Archivistica Studia 225

International Archives Symposium
Internationales Archivsymposion in Namur (2025)

Open Data and AI. New chances for Archives ?
Open Data und KI. Neue Möglichkeiten für Archive ?

PROCEEDINGS/ANNALEN

Editing - Redaktion
Els Herrebout

International Archives Symposium / Internationales Archivsymposion
in Namur (2025)

Open Data and AI. New chances for Archives?
Open Data und KI. Neue Möglichkeiten für Archive?

—

Proceedings / Annalen

Editing – Redaktion
Els HERREBOUT

2026

ALGEMEEN RIJKSARCHIEF
EN
RIJKSARCHIEF IN DE PROVINCIËN

ARCHIVES GÉNÉRALES DU ROYAUME
ET
ARCHIVES DE L'ÉTAT DANS LES PROVINCES

GENERALSTAATSARCHIV
UND
STAATSARCHIVE IN DER PROVINZ

MISCELLANEA ARCHIVISTICA
STUDIA

225

Generalstaatsarchiv – Algemeen Rijksarchief – Archives générales du Royaume
Ruisbroekstraat 2 rue de Ruysbroeck
1000 Brüssel – Brussel – Bruxelles

ISBN : 978 94 6391 661 5
D/2026/531/032
Publ. 6668

Offprint / Sonderdruck

VAN GELDER, Klaas; ROMEIN, Annemieke & GILLARD, Xavier (2026). When archives go digital…! Tools, Practices, Opportunities, and Challenges. In Els HERREBOUT (Ed.), International Archives Symposium in Namur (2025). Open Data and AI. New chances for Archives? Proceedings (Miscellanea Archivistica Studia, 225, pp. 41–69). State Archives of Belgium.

DOI: https://doi.org/10.5281/zenodo.20058876
Handle: https://orfeo.belnet.be/handle/internal/14701

When archives go digital…!
Tools, Practices, Opportunities, and Challenges[1]

Klaas VAN GELDER, C. Annemieke ROMEIN, Xavier GILLARD[2]

Abstract

The digital transformation of archives, driven by Artificial Intelligence and machine learning, represents a fundamental methodological shift in historical research. This paper explores some of the tools, practices, opportunities, and challenges of this transition. It first analyses the role of Automatic Text Recognition (ATR) platforms, exemplified by Transkribus, in converting vast holdings of historical manuscripts into machine-readable, searchable data. It details the practical workflows — from layout analysis to text enrichment — that enable new forms of access and analysis. This paper then addresses the subsequent challenge: processing and enriching these new digital corpora. For that purpose, it introduces AI-rchivist, a prototype tool from the ARKEY project, which uses generative AI to automatically extract metadata, generate multilingual summaries, and identify named entities. Finally, the paper critically assesses the significant limitations of these new AI-based approaches, including high hardware costs, technical constraints, and the unavoidable risk of "hallucinations". It highlights profound methodological risks, such as "automation bias" and the loss of collection-level context, concluding that a "machine-in-the-loop" approach that keeps human archivists central to the process is essential to mitigate the above risks.

Keywords: Automatic Text Recognition; Heritage; Archives; Transkribus; Metadata; Arkey; AI-rchivist

[1] This contribution is an extended version of a joint presentation by the three authors at the International Archives Symposium in Namur, Belgium, on 5 and 6 June 2025, on the topic of open data and AI. The order of authors is based upon the order of presenting the material in both the live presentation and the consequent article here. There is no hierarchical order, the contributions are equal.

[2] Klaas Van Gelder is archivist at the State Archives in Brussels and assistant professor in early modern history at Vrije Universiteit Brussel (SHOC research group); C. Annemieke Romein works on the HAICu project (Digital Humanities Artificial Intelligence Cultural Heritage) at the University of Twente (Enschede); Xavier Gillard is a researcher on the ARKEY project (The Belgian National Archives and UCLouvain).

1. Introduction

The digital transformation of archival practice represents one of the most significant methodological shifts in historical research since the discipline became professionalized. Where previous generations of historians operated within the constraints of physical repositories and manual transcription, contemporary scholars increasingly encounter archives that have been fundamentally reimagined through artificial intelligence and machine learning technologies. This transformation extends far beyond simple digitization; it encompasses a comprehensive reconceptualization of how historical sources are accessed, processed, and analysed.

The emergence of Automatic Text Recognition (ATR) technology, as a technique that can recognize handwritten, printed, and typewritten texts, exemplifies this paradigmatic shift. Platforms such as Transkribus, developed under the coordination of the University of Innsbruck since 2013, and eScriptorium, released in 2018 and currently hosted by the Radboud University Nijmegen, have surpassed the earlier limitations of Optical Character Recognition (OCR) to tackle the complex challenge of converting manuscript images into machine-readable, searchable text.[3] This technological leap, accelerated significantly during the COVID-19 pandemic through unprecedented volunteer mobilisation, has enabled the creation of vast digital corpora comprising tens to hundreds of thousands of transcribed pages of historical documents.

The implications of such developments are profound. Projects like *Chronicling Novelty* (see Figure 1) and *Alle Amsterdamse Akten* have rendered millions of pages of Dutch chronicles and notarial records systematically searchable, transforming research methodologies that previously required

[3] Benjamin Kiessling et al., 'eScriptorium: An Open Source Platform for Historical Document Analysis', *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)* 2 (September 2019): 19–19, https://doi.org/10.1109/ICDARW.2019.10032; *The eScriptorium VRE for Manuscript Cultures – Classics@ Journal*, n.d., accessed 26 September 2025, https://classics-at.chs.harvard.edu/classics18-stokes-kiessling-stokl-ben-ezra-tissot-gargem/; Sebastian Colutto et al., 'Transkribus. A Platform for Automated Text Recognition and Searching of Historical Documents', *2019 15th International Conference on eScience (eScience)*, September 2019, 463–66, https://doi.org/10.1109/eScience.2019.00060; Guenter Muehlberger et al., 'Transforming Scholarship in the Archives through Handwritten Text Recognition', *Journal of Documentation* 75, no. 5 (2019): 954–76.

weeks of manual investigation.[4] Similarly, initiatives such as the FED-tWIN project *ACCESS* (2021-2026) and the BRAIN-be-project *PARDONS* (2021-2025), both funded by the Belgian Science Policy Office (BELSPO) demonstrate how ATR technology can unlock complex multilingual archives, making visible patterns and connections that would otherwise remain buried within the linear kilometres of archival holdings.[5]

Figure 1 Chronicling Novelty. https://kronieken.transkribus.eu

Yet this digital revolution not only presents historians and archivists with a complex array of opportunities, it also raises problems and creates new challenges. The promise of making archives accessible "on a next level" must

[4] 'Chronicling Novelty', Chronicling Novelty, 21 December 2023, https://chronicling-novelty.com/; Amsterdam, 'Alle Amsterdamse Akten', webpagina, Stadsarchief, Gemeente Amsterdam, accessed 26 September 2025, https://www.amsterdam.nl/stadsarchief/alleamsterdamseakten/. For a review of the database: Klaas Van Gelder and Jim van der Meulen, 'Database Lokale kronieken uit de Nederlanden 1500–1850', Tijdschrift voor Geschiedenis 138, no. 2 (2025): 204–209.

[5] 'ACCESS to Court Files and Access to Justice. The Council of Brabant during the Early Modern Era - Rijksarchief in België', accessed 26 September 2025, https://www.arch.be/index.php?l=nl&m=lopend-onderzoek&r=onderzoeksprojecten&pr=access-to-court-files-and-access-to-justice.-the-council-of-brabant-during-the-early-modern-era; 'PARDONS', accessed 26 September 2025, https://pardons.eu/.

be balanced against significant investments in technological infrastructure, personnel training, and volunteer coordination. Furthermore, the transition from traditional archival inventories to ATR-generated datasets raises fundamental questions about research methodology, source criticism, and the nature of historical evidence itself.[6] How do we evaluate the reliability of machine-generated transcriptions? What new forms of analysis become possible when serial documents can be interrogated through computational methods? How do we navigate the tension between the democratising potential of digital archives and the technical expertise required to utilise them effectively?

This contribution examines these questions through a comprehensive analysis of current digital archival practices, with a particular focus on ATR technology, which serves as both a representative of broader technological trends and a case study in the practical implementation of AI-driven archival solutions. Through an examination of specific projects in the Low Countries and beyond, we explore how digital tools are reshaping not only access to historical sources but also the fundamental practices of historical research itself.

2. Tool example – Transkribus in Practice: From Document to Data[7]

The AI-based platform Transkribus (see Figure 2) serves as a central tool for digitizing and cataloguing historical documents.[8] The software enables the

[6] C. Annemieke Romein et al., 'Exploring Data Provenance in Handwritten Text Recognition Infrastructure: Sharing and Reusing Ground Truth Data, Referencing Models, and Acknowledging Contributions. Starting the Conversation on How We Could Get It Done', *Journal of Data Mining and Digital Humanities*, ahead of print, 24 March 2023, https://doi.org/10.5281/zenodo.8116009; C. Annemieke Romein et al., 'From Research Proposal to Project Management. A Guide from the Transkribus Community on Planning and Executing Workflows for Researchers and GLAM-Professionals', *International Journal of Digital Humanities*, ahead of print, 1 September 2025, https://doi.org/10.1007/s42803-025-00107-7.

[7] This section on Transkribus is based upon several other publications and workshops of C.A. Romein on Transkribus, but predominantly on: Helene Prokop and Christel Annemieke Romein, 'Einsatz von künstlicher Intelligenz bei der automatischen Handschriftenerkennung. Das Beispiel Transkribus', in *Lauter weiße Flecken? Aktuelle Brennpunkte der Archivarbeit. Referate des Landesarchivtags in Dessau-Roßlau am 12. und 13. Juni 2024.*, Landesarchivtag Sachsen-Anhalt 2024 (VdA - Verband deutscher Archivarinnen und Archivare e.V, 2025). This English text, which include translated parts from the previously mentioned German article has updated references, textual revision, updates on newly incorporated features and region specific examples.

precise and efficient capture of both printed and handwritten texts, which is particularly significant for work with extensive historical text collections. The integration of cutting-edge technologies, particularly ATR, which combines functions of OCR and Handwritten Text Recognition (HTR), enables the efficient conversion of texts into machine-readable data.[9] This conversion unlocks numerous advantages for researchers and archives, including the ability to conduct full-text searches within archival holdings and a significant enhancement in access to relevant information.

The Transkribus platform supports the entire process, from digitization through text recognition to publication, thereby opening up diverse application possibilities.[10] Furthermore, the software enables automated analysis of layout structures, whereby elements such as columns, images, and other document components can be recognised.[11] The precision of text recognition improves with the amount of training data provided, ensuring continuous improvements in efficiency and accuracy when processing historical materials.

The use of ATR technologies offers numerous advantages. The automated capture of archival sources not only facilitates the digitisation of holdings but also enables their broad public accessibility. Simultaneously, the digital processing of large quantities of manuscripts enables efficient cataloguing and searchability. In this context, Transkribus makes a substantial contribution to the sustainable preservation and utilisation of cultural resources.

Figure 2 The Transkribus Logo since 2023.

[8] Transkribus uses Machine Learning, which is a subclass of Artificial Intelligence, to be more precise. 'Transkribus', accessed 16 September 2024, https://app.transkribus.org/nl.
[9] Romein et al., 'Exploring Data Provenance in Handwritten Text Recognition Infrastructure'.
[10] Muehlberger et al., 'Transforming Scholarship in the Archives through Handwritten Text Recognition'.
[11] Colutto et al., 'Transkribus. A Platform for Automated Text Recognition and Searching of Historical Documents'.

2.1. The Origins and Development of Transkribus

The origin of Transkribus and the founding of the cooperative READ-COOP SCE reflect a pioneering development in the field of historical sciences and archival work. Originally developed within the "READ-Project" (Recognition and Enrichment of Archival Documents) at the University of Innsbruck, the idea for Transkribus emerged to solve the challenge of handwriting recognition in historical documents. The foundation for this was laid by two projects funded by the European Union: TranScriptorium (2013–2015) and READ (2016–2019).[12] Whilst TranScriptorium developed technologies for HTR and demonstrated their potential, the READ project built upon this and introduced Transkribus as a central platform.

In 2019, the question of long-term financing and sustainable development of the platform arose. To address this challenge, the European cooperative READ-COOP SCE was established on 1 July 2019, comprising both private individuals and international archives, libraries, and research institutions.[13] This model promotes collaboration between the actors and the public. With a democratic administrative structure and the principle of reinvesting all incomes into the further development of the platform, the cooperative ensures the sustainable continuation and continuous improvement of Transkribus.

Today, READ-COOP SCE counts over 250 cooperative members, including institutions such as the University of Cambridge, the State Archives of Belgium, and the KB National Library of the Netherlands. It is supported by a global community of more than 300,000 users, including scholars, archivists, and volunteers. These have already processed over 100 million pages of historical documents, thus making a considerable contribution to the digitisation and searchability of archival material. The platform combines advanced machine learning technologies with digitisation, enabling the transcription and analysis of texts. This opens up new possibilities for the processing of historical documents and sustainably changes research practice.

The cooperative model, which places *purpose before profit*, demonstrates the high acceptance and effectiveness of the technology within the international research community. The sustainable development and success of Transkribus

[12] 'Recognition and Enrichment of Archival Documents | READ Project | H2020 | CORDIS | European Commission', accessed 29 July 2021, https://cordis.europa.eu/project/id/674943.
[13] uibk, '+ READ-COOP SCE Formally Established!', READ-COOP, 15 November 2019, https://readcoop.eu/read-coop-sce-formally-established/.

underscore the importance of cooperation and innovation in digital historical science.[14]

2.2. From Handwriting to Digital Edition: Technical Processes

The digital transformation of historical manuscripts is accompanied by complex challenges that can be addressed through the use of modern technologies and systematic approaches. With the development of specialized platforms such as Transkribus, new possibilities for automatic text recognition and structural analysis have emerged, making the digitisation process substantially more efficient.

The comprehensive approach of these tools encompasses initial text capture, layout recognition, and semantic markup. It is suitable for both individual research projects and larger digitisation undertakings. The following sections provide a detailed discussion of the methodological foundations and practical application possibilities of automated manuscript processing.

2.3. Methods of Text Capture and Processing

ATR and Layout Analysis (LA) are essential methods for capturing and processing text. In recent years, ATR has developed into an indispensable tool of the digital humanities. With the introduction of Transkribus, the research community now has access to a mature platform that supports the systematic process of handwriting recognition and digitization. The processing of historical documents occurs through two distinct yet complementary workflows, which are tailored to the specific requirements of the respective research projects.

Standard Workflow

Figure 3 The Standard Workflow.

[14] 'Our Members', accessed 17 November 2024, https://readcoop.org/members.

The *Standard Workflow* (see Figure 3) provides a straightforward introduction to ATR. Following the upload of documents, processing occurs through pre-trained AI models (see Figure 4) that already cover a broad spectrum of common script types. This approach enables time-efficient and effective initial transcription, particularly with frequently encountered script types, yielding convincing results. The pre-trained models are based on extensive datasets of historical documents and are thus capable of delivering reliable results even with various handwriting variants.[15]

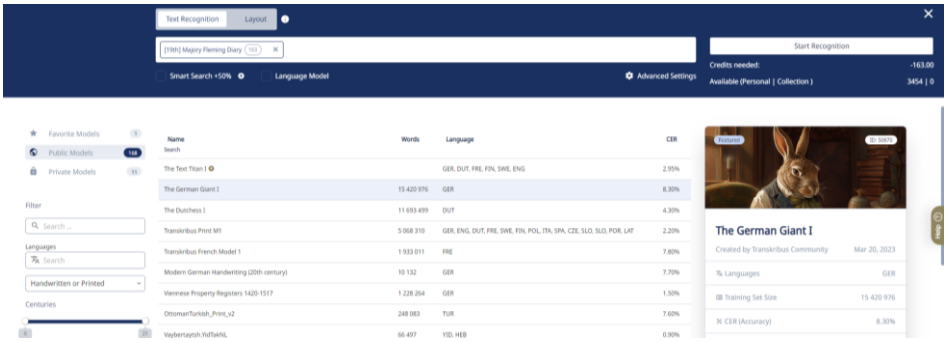

Figure 4 Public models, available to all users of Transkribus.

Advanced Workflow

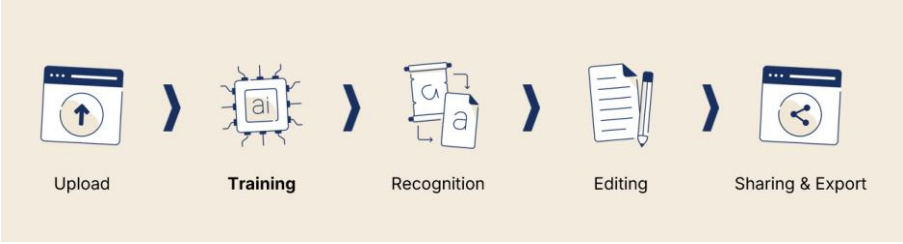

Figure 5 Advanced Workflow.

For advanced requirements or particularly demanding handwriting, the *advanced workflow* is available (see Figure 5). The advanced workflow enables the development of individual AI models that are precisely tailored to

[15] Melissa Terras et al., 'READ-COOP and Transkribus: A Cooperative Model for Responsible Technology', 24 May 2025, https://doi.org/10.5281/zenodo.15503325.

the particularities of the respective source material. The process begins with careful selection of representative sample pages, which are manually transcribed and serve as the foundation for the training phase of the AI model (see Figure 6). Although these initial work steps require a higher workload, this is amortised through the significantly more precise recognition results as well as the considerably reduced correction effort in the further processing of larger documents.[16]

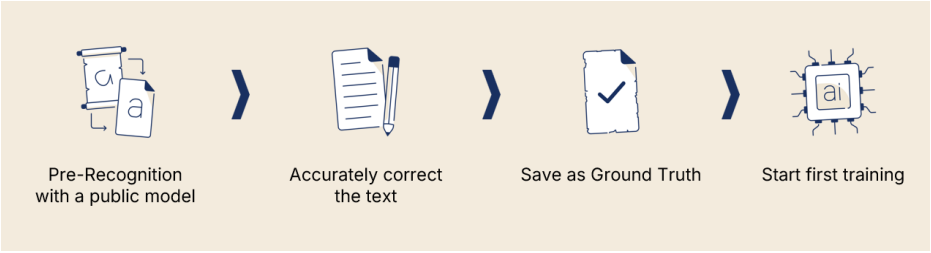

Figure 6 The training workflow, itself.

Further Processing and Refinement of Transcriptions

After automatic text recognition, users have at their disposal, through Transkribus' document editor, a comprehensive set of tools for further processing and refinement of the created transcriptions (see Figure 7). In the layout editor, you can review the captured structure of the document, particularly regarding page division and columns.[17] The integrated text editor enables users to correct erroneous text passages immediately and utilize advanced functions for the semantic enhancement of transcriptions.

Quality assurance is facilitated by various integrated tools. The layout editor enables a visual review of the recognized page structure, while validation functions ensure the consistency of the transcription and the markups used. The multi-layered control mechanisms ensure that the resulting digital editions correspond to the highest scholarly standards.

[16] '1. Automatically Transcribing Your Documents', accessed 26 September 2025, https://help.transkribus.org/automatically-transcribing-your-documents.
[17] '2. Training Text Recognition Models', accessed 26 September 2025, https://help.transkribus.org/training-text-recognition-models.

Due to the comprehensive methodological approach, Transkribus establishes itself as a central platform for the systematic cataloguing of historical manuscripts. The combination of automated text recognition, flexible post-processing, and standardised output creates the prerequisites for a sustainable digital transformation of archival sources.

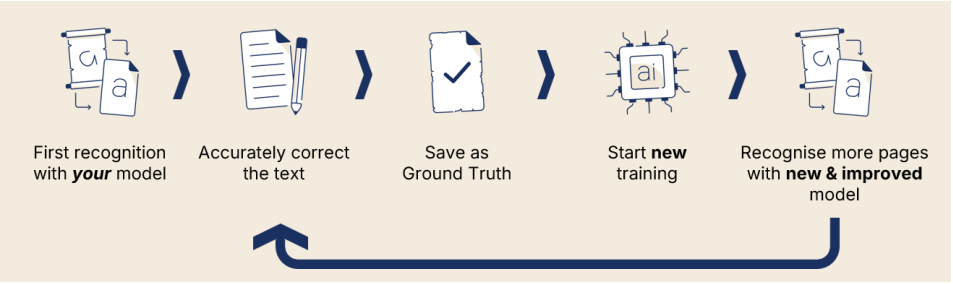

Figure 7 Transkribus includes functions to refine and improve texts.

Text Enrichment

The use of "tags" enables users to systematically identify and highlight essential sections such as names, place names, or date entries (see Figure 8). The markups adhere to established standards within the Digital Humanities, ensuring the interoperability of the created transcriptions. Additionally, Transkribus offers advanced metadata functions that enable the storage of additional information about the processed documents, such as source, script type, backlink, or external ID, which supports more precise contextualization of the transcripts.[18] The diverse editing options in the document editor not only improve the accuracy of the transcriptions but also facilitate users' further analysis and use of the digitized content.

[18] '3. Textual Tags', accessed 26 September 2025, https://help.transkribus.org/textual-tags.

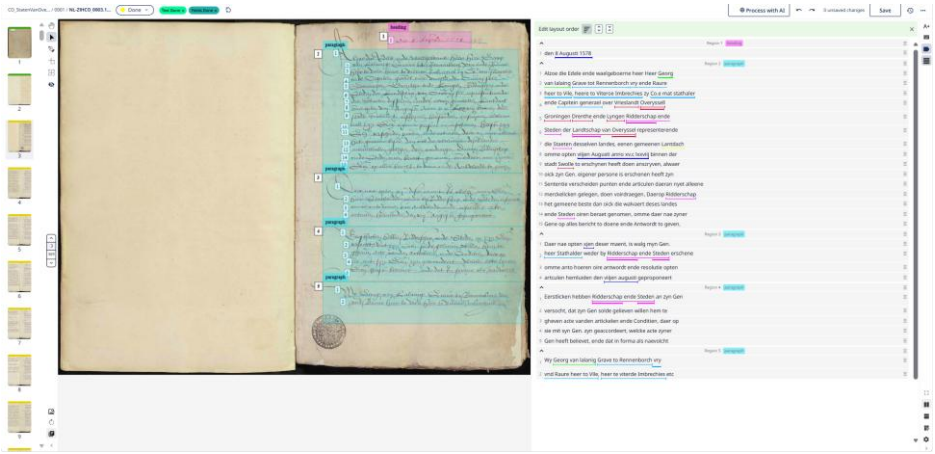

Figure 8 Lay-Out Analysis (Structure) and Tags are visible in this image and the transcription on the right. Source: Collectie Overijssel Resolutions of the Provincial States (0003.1 0001 page 3).

Structural Analysis and Layout Recognition

Layout recognition represents a fundamental component of modern digitisation, which is why a differentiated examination of both its technical implementation and its practical application is required. The technical foundation is formed by the advanced analysis method implemented in Transkribus, which enables systematic processing of complex document structures. This functionality proves particularly valuable in processing historical documents with complex layout elements, such as tables, forms, marginalia, or multi-column text arrangements.

The methodological approach is based on the integration of various analysis procedures. Layout recognition complements ATR, enabling a multi-layered analysis of documents. The resulting digital representation captures both the content and structural dimensions of the sources, thereby generating synergy. The practical implementation occurs through a user-friendly interface that ensures precise control of the automatically recognised structures. Users have the possibility not only to review the generated layout elements and, if necessary, to correct them, but also to extend them through additional markups. This ensures accurate capture even with complex document structures.

Baselines

The implementation of baseline models represents an advanced method for precise text line recognition, which proves extremely valuable, particularly in

the processing of complex document structures. While well-preserved documents with clearly structured scripts are frequently sufficient for standard text recognition, the particular benefit of these specialised models is especially evident with demanding materials. This particularly concerns documents that exhibit diagonally running or strongly distorted text lines, unusually long or short line formats, as well as materials with significant quality impairments, such as those caused by damage or aging processes (see Figure 9).

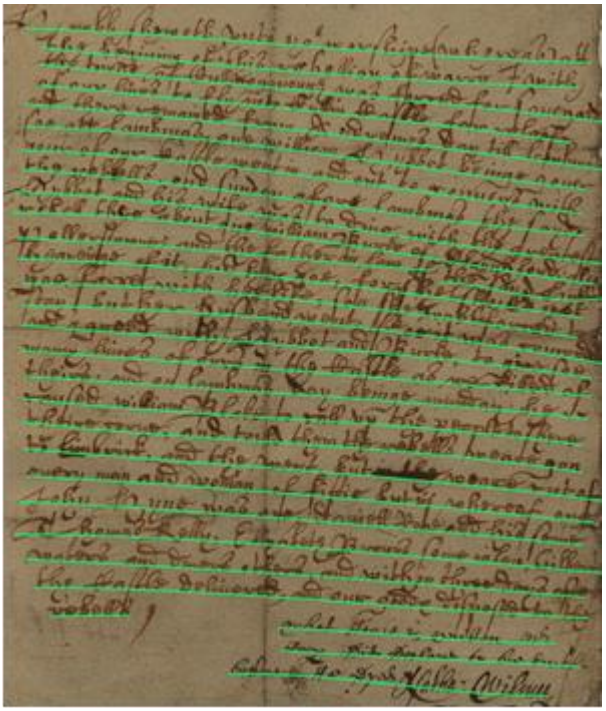

Figure 9 Baselines. Example with screwed baselines.

Tables

The specialized Table Models represent a central element of these extended analysis capabilities—the development of these ML/AI-supported models aimed to recognize and process tabular structures in historical documents automatically. The relevance of this function becomes particularly evident in work with archival sources, like, amongst others, tax lists, trade registers, church books, and academic matriculation records (see Figure 10). The Table Models are also capable of not only identifying the basic table structure but

also precisely assigning the text to the respective cells. The automated structure recognition leads to significant time savings in processing and enables systematic searchability and analysability of the contained information in digital format.[19]

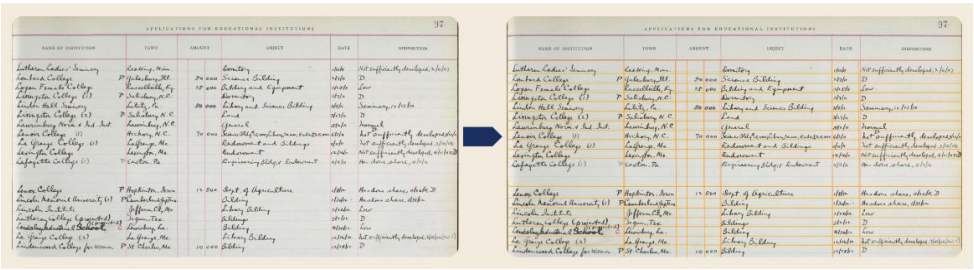

Figure 10 Tables with Table Recognition.

Field models

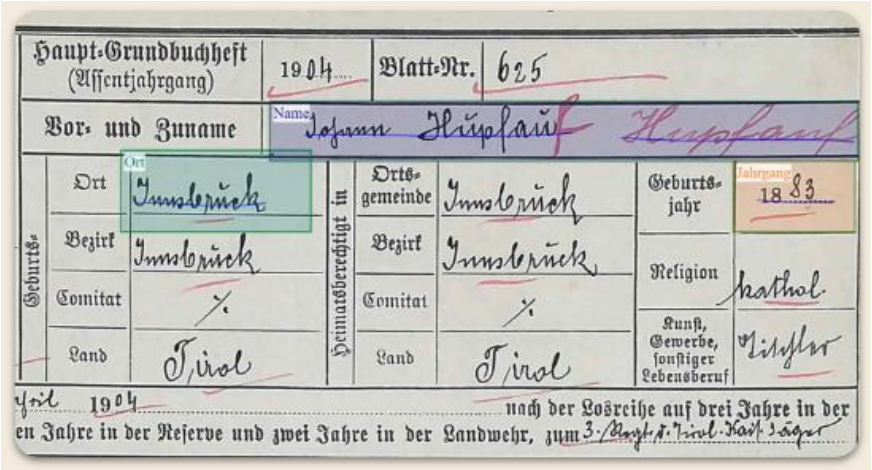

Figure 11 Field-model example

Additionally, the Field Models offer a specialized solution for processing form-like documents. The models have been trained to automatically recognise and semantically mark recurring structural elements and specific

[19] '2. Managing Documents and Pages', accessed 26 September 2025, https://help.transkribus.org/managing-documents.

information fields. This functionality proves particularly valuable in the processing of standardised historical documents, which include, for example, civil status certificates, registration forms, or official forms (see Figure 11). The Field Models are also capable of identifying the position of relevant fields and assigning them corresponding tags, which considerably simplifies the extraction of structured data.[20]

End-to-End models

Traditional ATR systems rely on sequential processing pipelines that begin with document layout analysis and proceed through multiple separate stages, including line detection, character classification, and text recognition. Nevertheless, this sequential approach creates several significant problems: errors from early processing steps propagate through the entire pipeline, the system becomes complex and challenging to maintain, computational inefficiencies arise from chained processes, and each component requires separate training data. Most critically, models that process only individual text lines or visual pixel information cannot incorporate broader page context, meaning that clearly readable text at the beginning of a page cannot help improve recognition of uncertain passages elsewhere on the same page.

Upcoming End-to-End (E2E) architectures fundamentally address these limitations by processing entire document pages as unified visual units, rather than breaking them down into sequential components. The ongoing development with technologies like the Document Attention Network (DAN) introduces a novel two-dimensional attention mechanism specifically optimized for text recognition.[21] These systems combine a fully convolutional network encoder that extracts spatial features from the entire page with a transformer decoder that directly accesses these two-dimensional feature maps. This enables the model to learn latent representations and flexibly navigate between columns, marginal notes, or tables without requiring explicit segmentation or manual reading-order annotations. E2E models simultaneously answer three essential questions: what text appears in the document, where it is located spatially, and what it means in terms of semantic and functional structures such as titles, form fields, or dates.

[20] '2. Field Models', accessed 26 September 2025, https://help.transkribus.org/field-models.
[21] For example: Denis Coquenet et al., 'DAN: A Segmentation-Free Document Attention Network for Handwritten Document Recognition', *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, no. 7 (2023): 8227–43, https://doi.org/10.1109/TPAMI.2023.3235826.

The practical advantages of E2E processing for archival digitization are expected to be substantial. These systems eliminate error propagation between processing steps, reduce annotation requirements to simple document transcripts, and enable the sharing of mutual information between text recognition and layout understanding. This integrated approach not only simplifies the entire processing chain but also provides significantly more robust handling of complex and variant-rich layouts, making them particularly valuable for historical documents with challenging features, such as changing text direction, marginalia, tables, or insertions between lines.

Export and Digital Edition

A particular advantage of the platform is its flexible export functionality. Depending on the intended purpose of use, the processed documents can be converted into various output formats (Figure 12).

Figure 11 Various export options (standard availability).

For scholarly further processing, structured formats such as TEI-XML are available, whilst for presentation purposes, PDF or Word documents can be generated.[22] The properties above ensure seamless integration of the transcriptions into different workflows, whether for online publications, print editions, or further computer-assisted analyses.

[22] More and additional (new) information can be find here: '2. Downloading', accessed 26 September 2025, https://help.transkribus.org/downloading The Subscriptionmodel(s) of Transkribus provide additional Export options to the users.

Transkribus Sites

A particular added value results from the possibility of making the structured data accessible to the community through Transkribus Sites. The platform thereby offers not only functions for presenting digitized documents but also the possibility of targeted searches within recognized structures. Public accessibility fosters scholarly collaboration and enables innovative research approaches through the systematic analysis of extensive document holdings (see Figure 13).

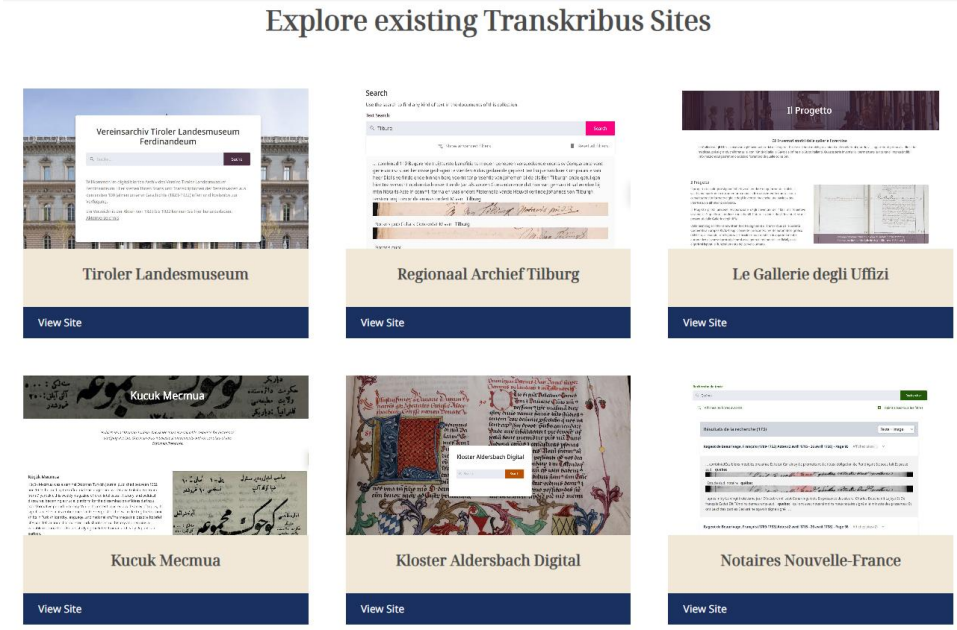

Figure 12 Transkribus Sites with several publicly available websites: https://www.transkribus.org/sites.

2.4. Application Examples from Practice

The performance and versatility of Transkribus are demonstrated in numerous documented use cases from various research areas. A particularly significant area of application is genealogical research, where the platform has established itself as an indispensable tool.[23] The digitisation and transcription of personal documents, from handwritten letters to ecclesiastical archival materials such as birth and death registers, enable both private and institutional users to efficiently and sustainably document historical family documents.

The analysis of politically and culturally significant documents demonstrates the platform's performance. An example of this is the systematic investigation of the forged Hitler diaries, in which AI-supported text recognition enabled detailed analysis of the difficult-to-read handwriting.[24] In the institutional context, Transkribus provides archives and libraries with resources for managing extensive digitisation projects. In the National Archives of the Netherlands, over three million scans have been successfully transcribed and made accessible for research. In the HAICu Project, the Collectie Overijssel has made its corpus of the Resolutions of the Provinciale Staten (0003.1) available.[25] The collection consists of approximately 60,000 pages and the accessibility will be improved by applying layout analysis, text recognition, entity tagging and consequently, adding additional metadata to the text. This is done together with citizen scientists.[26]

The cited examples illustrate the impressive performance of AI-supported text recognition as well as the comprehensive infrastructure of Transkribus. The platform enables not only the large-scale digitization and transcription of

[23] See for a wide array of projects and applications of ATR: Christel Annemieke Romein et al., eds, *Praeteritum Transcriptum. A Transkribus Tribute: Celebrating Our First Five Years as a Cooperative (2019-2024)* (Zenodo, 2025), https://doi.org/10.5281/zenodo.15308678.
[24] *Böse Fälschung: Was Steht in Den 'Hitler-Tagebüchern'? | STRG_F*, directed by STRG_F, 2023, 27:22, https://www.youtube.com/watch?v=NNspVJCdaQw; NDR, 'Datenbank: Die gefälschten "Hitler-Tagebücher" zum Durchsuchen', accessed 17 November 2024, https://www.ndr.de/geschichte/tagebuecher/Datenbank-Die-gefaelschten-Hitler-Tagebuecher-zum-Durchsuchen,hitlertagebuecherdatenbank102.html.
[25] 'HAICu - Digital Humanities Artificial Intelligence Cultural Heritage', HAICu, accessed 26 September 2025, https://www.haicu.science/.
[26] C. A. (Annemieke) Romein, *Handleiding Citizen Scientists Collectie Overijssel i.s.m. HAICu/ UTwente Archiefdeel: Resoluties van de Staten van Overijssel*, Zenodo, 14 May 2025, https://zenodo.org/records/15401991.

historical manuscripts but also their scholarly cataloguing and public accessibility. The combination of technical innovation and practical applicability makes Transkribus a central instrument in modern historical research and archiving, opening up new perspectives for scholarly work.

3. Arkey and AI-rchivist

While platforms like Transkribus are essential for the first step of ATR, the subsequent challenge lies in processing and enriching these new digital transcriptions. The ARKEY project (2023–2028) provides an example. It is a FED-tWIN project funded by BELSPO[27]. In essence, it aims at making the archives more broadly and easily accessible: for the general public, for researchers, and for archives practitioners. It stems from the observation that many treasures in the Belgian State Archives (BSA) collection have remained relatively unknown until now, and hence are not exploited to the fullest of their potential. The role of the ARKEY project is then to bridge this gap and apply computational methods to *extract* information*, interpret* it, and help users *navigate* the archival holdings.

Getting a nice transcription of historical documents – as is the case with Transkribus explained above – is a natural first step to achieve those objectives. However, it only gets the archive users so far. Indeed, ATR "only" takes care of the paleographic deciphering aspects of the problem. This dramatically reduces the expertise needed for one to be able to read the text, but it does very little to help users find the text in our collections, nor does it help them make sense of the document. This is where the value added by archivists remains crucial.

In the context of budget cuts that are impacting all culture and heritage services across Europe, Belgium is no exception. Therefore, the availability of this added value is scarcer and scarcer. Still, the amount of documents that should be processed is immense and growing. Which is why a prototype tool called AI-rchivist has been developed in the context of ARKEY. This tool is designed not as a replacement for human expertise, but as an automated "assistant" built upon a "*machine-in-the-loop*" approach[28]. This philosophy

[27] *'ARKEY - AI meets archives'*, Archives de l'État en Belgique, accessed 22 may 2025, https://arch.arch.be/index.php?l=fr&m=nos-projets&r=projets-de-recherche&pr=arkey-ai-meets-archives.

[28] CLARK, E. et al., '*Creative Writing with a Machine in the Loop: Case Studies on Slogans and Stories'* in 23rd International Conference on Intelligent User Interfaces, 2018, https://doi.org/10.1145/3172944.3172983; ACM (Tokyo Japan), https://dl.acm.org/doi/10.1145/3172944.3172983.

underscores that the archivist remains central to the process, supported by an AI that facilitates the most time-consuming tasks.

As shown per Figure 14 the tool uses a generative AI model (Large Language Model or LLM) to process raw transcriptions. During the symposium, several examples of historical documents have been used: a 1651 judgment from the Council of Brabant (partially reproduced in Figure 14), a 1791 tax document from Overijsel, and a 1521 letter of remission issued by the Privy Council, the tool:

- **Extracts key metadata**: This includes "Type of Document," "Act Date," and "Facts Date" as shown in Figure 15.
- **Generates succinct summaries:** The content of the original text is summarized in multiple languages (English, French, Dutch, and German) to make it easier to find and grasp by a non-specialist (see Figure 15).
- **Identifies entities:** The tool automatically lists persons, locations, and their corresponding roles or types as shown in Figure 16.

Machine-in-the-loop

The "machine-in-the-loop" design is operationalized via an interactive interface. Every field of information generated by the AI is fully editable by the archivist. This allows one to complete missing data, fix errors and refine the automated suggestions. Furthermore, a chat bot feature has been integrated to AI-rchivist. This allows the user to ask for transversal modifications /adaptations to correct and refine the AI's output in natural language. For instance, the archivist can instruct the AI to correct omissions (e.g., "You forgot to mention the king...") or refine details (e.g., "Actually, the king is Charles Quint. Can you fix that... and mention it in the summaries?").

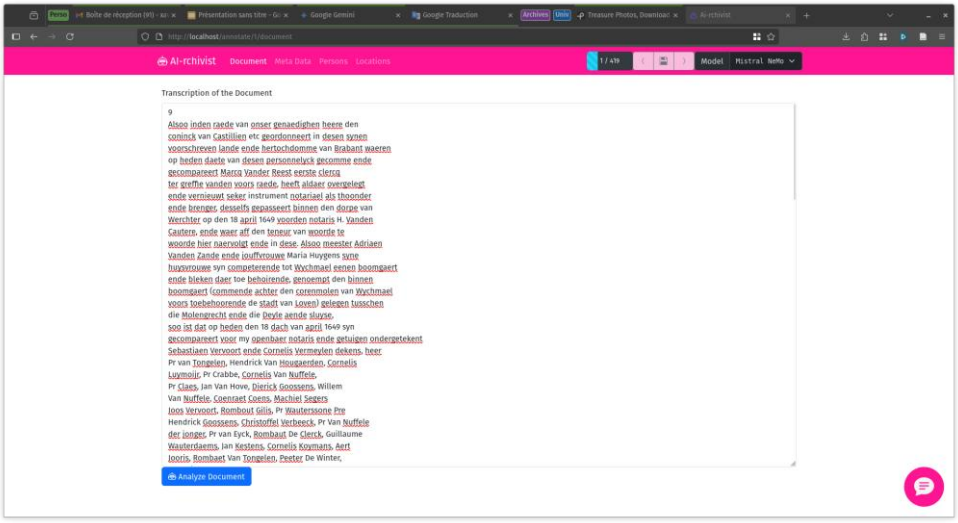

Figure 13: A legal document dated April 18, 1649 (formalized in 1651) ended a dispute over the use of the banks of an orchard in Wychmael. The owners, Adriaen Vanden Zande and Maria Huygens, had been harassed by boatmen from the Deyle River, who dragged their boats onto their land. (Source of the transcription: ACCESS project)

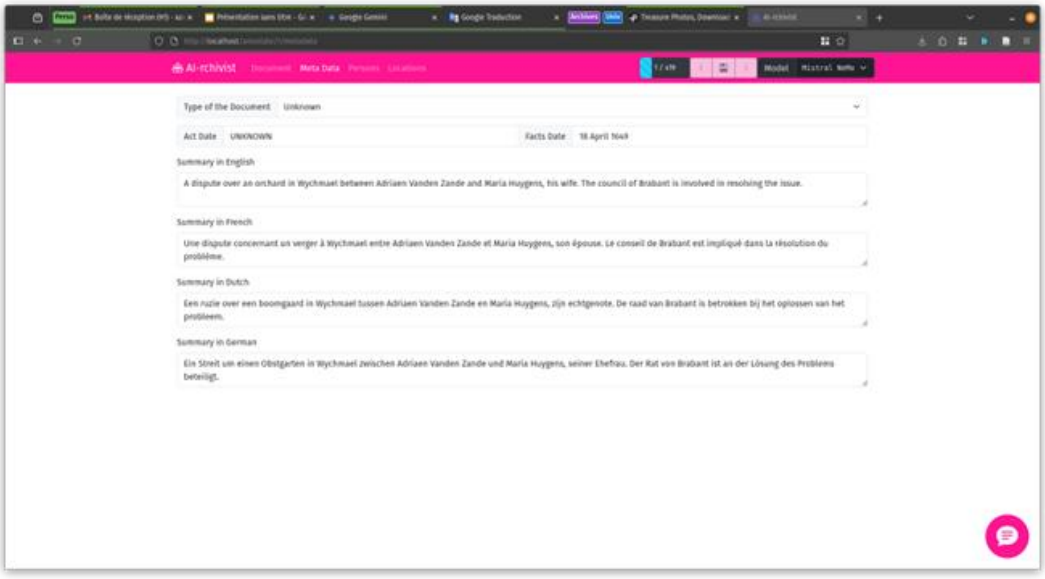

Figure 15: Key meta-data (dates, and multi-lingual summaries), extracted from the example transcribed document in figure 14.

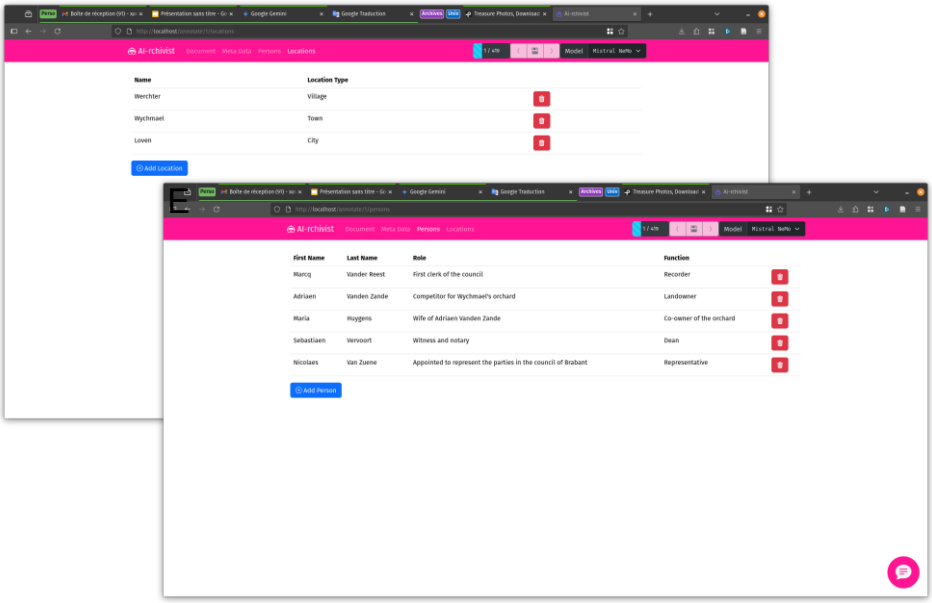

Figure 14: Key meta-data (dates, and multi-lingual summaries), extracted from the example transcribed document.

Practical and technical limitations

Despite the "magic" of the demonstration, applying this technology to archival materials reveals significant limitations, which we briefly address in the following sections.

Hardware and Cost

The practical challenges are significant. Given the sheer amount of computation to be done just to output one single “word”[29], AI model operators have come to rely heavily on specific hardware. Those chips are called *Graphical Processing Units* – GPU in short -- because these chips were initially developed for gaming. GPU[30] have become so central to AI that one cannot think of running more than a very small-scale language model on a Central Processing Unit – CPU for short; the kind of chip that equip all of our

[29]Token would be more correct.
[30]TPU : Tensor Processing Units are an other option albeit it is not commonly available. Both GPU and TPU serve the same purpose and come with mostly the same set of limitations.

desktop, laptop and cell phones… The requisite hardware for self-hosting large models is substantial; analysis shows that models like llama3.1-405b require more than nine A100 80G GPUs which amounts to a price tag of about € 200, 000 in total, just for the GPUs. This means that the actual price tag to run and operate such models can be about double that sum. While cloud APIs remove this barrier, they introduce a high operational cost and a dependency to an external provider, which might be problematic when dealing with sensitive personal data.

Tokenization and Context Window

Beyond cost, technical hurdles are substantial. Indeed, the very first step in the processing performed by an LLM is to “tokenize” the text it is given (the so-called prompt). That is, the very first step performed when interacting with an LLM is to split the text in smaller entities called “tokens”. These tokens correspond to either a phrase, a word, or a few letters. The action of splitting the text to tokens is called tokenization. It is an essential step, because models cannot act with words immediately. All they know are numbers, which is why the model creators have devised a large vocabulary where each token is assigned to a numeric identifier. The tokenized text is then converted to a sequence of numbers which can be fed through the actual model.

Because modern LLMs have essentially been trained using all the text from the Internet[31], they did not have a chance to often encounter historical languages with their huge lexical and orthographic variety. This is one of the reasons why LLMs are less performant on historical languages than their current counterparts. Data shows historical texts are tokenized quite inefficiently, requiring 1.8 to 2.2 times more tokens than an ideal tokenizer. This inefficiency directly impacts the "context window" (the model's working memory). Analysis of available token headroom shows that many models, have almost no available token space left after ingesting a single historical document. This creates a hard technical limit on the length of documents that can be processed because that same working memory is required both to “read” text and to write some.

[31] *Common Crawl* - Open Repository of Web Crawl Data, accessed June 26th 2025, https://commoncrawl.org

Output Quality and Hallucinations

Hallucination is the phenomenon that occurs when an LLM outputs text that is factually incorrect while being plausible[32]. This phenomenon is known to impact smaller-sized LLMs more severely. And while techniques have been devised that lower the probability of a model hallucinating, it has recently been shown that hallucinations are unavoidable when working with LLMs[33].

This technical strain contributes to poor output quality, especially in cases like historical documents where the text is long, and tokenization is subpar. Our experiments revealed that while generative models can create summaries, they perform poorly on this specific extraction task when compared to smaller size models trained for that particular purpose. **A posteriori** evaluation by human archivists, who ranked the quality of extracted data on a 5-point scale, confirmed this. The generative models (Mistral, Llama, gpt4o-mini) were all ranked relatively low, with median scores between 1 and 2.

The Human Factor

A significant risk is the introduction of "automation bias". Horowitz et al. have shown that this bias often occurs when AI systems are deployed[34]. Because of it, professionals tend to over-trust the AI's suggestions, even when they conflict with their own expertise. This can lead to a long-term degradation of professional practices and data quality. It could also lead to catastrophic data loss if AI were to be used in the appraisal – the procedure during which a selection is operated regarding the archives that may be destroyed, and those which must be kept for the long term. The machine-in-the-loop approach, which we selected in AI-rchivist is a direct countermeasure to mitigate this risk since it keeps the archivist in charge of the whole process.

A second risk factor stemming from the introduction of AI systems in the archival practice is the loss of appropriation by the archivists. Indeed, by working closely with their funds, archivists build an in-depth knowledge of the material in their collections. This knowledge is then used to build search heuristics that serve the public when it comes to answering archives questions. Altogether forfeiting the manual – and labor-intensive – task of

[32]XU, Z., S. JAIN, & M. KANKANHALLI, '*Hallucination is Inevitable: An Innate Limitation of Large Language Models'*, 2025; arXiv, https://doi.org/10.48550/arXiv.2401.11817.
[33]Idem.
[34]HOROWITZ, M. C. & L. KAHN, 'Bending the Automation Bias Curve: A Study of Human and AI-based Decision Making in National Security Contexts', International Studies Quarterly, Volume 68 Issue 2, 2024, Oxford Academic, https://doi.org/10.1093/isq/sqae020

reading and analyzing the funds would mean this in-depth knowledge would be lost. And hence, the searchability of the archives might be hampered by the introduction of AI rather than boosted by it. Again, the machine-in-the-loop approach, which we opted for, is directly meant to mitigate that risk.

Methodological Context Loss

Finally, a core methodological challenge remains. AI tools typically operate at the document level (the *item)*. Modern archival practice, however, emphasizes the collection level, where the context of creation — and not just the text itself — provides essential meaning. An overemphasis on automated, document-level extraction risks creating vast amounts of disassociated data rather than contextualized, meaningful information.

4. Concluding Remarks

The digital transformation of archival practice, driven by artificial intelligence and machine learning, represents one of the most significant methodological shifts in modern historical and archival practices. This paper has explored some of the possibilities offered by these new technologies through the lens of three projects investigated at UTwente, the Belgian State Archives, VUB, and UCLouvain.

First, platforms like Transkribus have powerfully addressed the foundational challenge of access, converting vast quantities of complex handwritten and printed documents into machine-readable, searchable text. Through a cooperative model and a sophisticated suite of tools — from custom-trained ATR models to advanced layout analysis — Transkribus has unlocked new research possibilities and made millions of pages[35] of historical documents accessible on an unprecedented scale.

However, Automatic Text Recognition is only the first step. The resulting abundance of digital text presents a new, complex challenge: how to process, enrich, and interpret this information in a meaningful way. The ARKEY project's AI-rchivist prototype illustrates a potential path forward, using generative AI to assist archivists in summarizing content, extracting key metadata, and identifying entities.

Yet, this next technological phase introduces its own formidable limitations. These include practical barriers, such as the prohibitive hardware costs and processing dependencies of large language models, as well as critical

[35] Over 150 million of pages, to the best of the author’s knowledge.

technical roadblocks, like the inefficient tokenization of historical languages and the unavoidable risk of factual "hallucinations".

More profoundly, this analysis highlights crucial human and methodological risks. A significant danger lies in "automation bias" where professionals may over-trust flawed AI suggestions, potentially degrading data quality. Furthermore, automating the analysis of archives risks the loss of "appropriation" — the deep, contextual knowledge that archivists build by working manually with their funds. An overemphasis on AI-driven, document-level extraction may create vast amounts of disassociated data, severed from the collection-level context that provides its essential meaning.

This paper demonstrates that AI-driven tools present unprecedented opportunities. It also revealed that such tools cannot replace the archivist. The "machine-in-the-loop" philosophy, which positions AI as an assistant rather than an autonomous replacement, is therefore essential. This approach provides a necessary countermeasure to the weaknesses of current AI, keeping human expertise and critical judgment central to the archival process. It ensures that the digital transformation enriches, rather than flattens, our understanding of historical sources in context.

Acknowledgements

The NWO NWA 1518.22.105 grant funds Annemieke Romein's research on the Collectie Overijssel. Additionally, she is honorary Community Director at the READ-COOP SCE and Chair of the Board of Directors.

Bibliography

'1. Automatically Transcribing Your Documents'. Accessed 26 September 2025. https://help.transkribus.org/automatically-transcribing-your-documents.

'2. Downloading'. Accessed 26 September 2025. https://help.transkribus.org/downloading.

'2. Field Models'. Accessed 26 September 2025. https://help.transkribus.org/field-models.

‘2. Managing Documents and Pages’. Accessed 26 September 2025. https://help.transkribus.org/managing-documents.

‘2. Training Text Recognition Models’. Accessed 26 September 2025. https://help.transkribus.org/training-text-recognition-models.

‘3. Textual Tags’. Accessed 26 September 2025. https://help.transkribus.org/textual-tags.

‘ACCESS to Court Files and Access to Justice. The Council of Brabant during the Early Modern Era - Rijksarchief in België’. Accessed 26 September 2025. https://www.arch.be/index.php?l=nl&m=lopend-onderzoek&r=onderzoeksprojecten&pr=access-to-court-files-and-access-to-justice.-the-council-of-brabant-during-the-early-modern-era.

Amsterdam. ‘Alle Amsterdamse Akten’. Webpagina. Stadsarchief, Gemeente Amsterdam. Accessed 26 September 2025. https://www.amsterdam.nl/stadsarchief/alleamsterdamseakten/.

Chronicling Novelty. ‘Chronicling Novelty’. 21 December 2023. https://chronicling-novelty.com/.

Colutto, Sebastian, Philip Kahle, Günter Hackl, and Günter Mühlberger. ‘Transkribus. A Platform for Automated Text Recognition and Searching of Historical Documents’. 2019 15th International Conference on eScience (eScience), September 2019, 463–66. https://doi.org/10.1109/eScience.2019.00060.

Coquenet, Denis, Clément Chatelain, and Thierry Paquet. ‘DAN: A Segmentation-Free Document Attention Network for Handwritten Document Recognition’. IEEE Transactions on Pattern Analysis and Machine Intelligence 45, no. 7 (2023): 8227–43. https://doi.org/10.1109/TPAMI.2023.3235826.

HAICu. ‘HAICu - Digital Humanities Artificial Intelligence Cultural Heritage’. Accessed 26 September 2025. https://www.haicu.science/.

Kiessling, Benjamin, Robin Tissot, Peter Stokes, and Daniel Stökl Ben Ezra. ‘eScriptorium: An Open Source Platform for Historical Document Analysis’. 2019 International Conference on Document Analysis and

Recognition Workshops (ICDARW) 2 (September 2019): 19–19. https://doi.org/10.1109/ICDARW.2019.10032.

Muehlberger, Guenter, Louise Seaward, Melissa Terras, et al. 'Transforming Scholarship in the Archives through Handwritten Text Recognition'. Journal of Documentation 75, no. 5 (2019): 954–76.

NDR. 'Datenbank: Die gefälschten "Hitler-Tagebücher" zum Durchsuchen'. Accessed 17 November 2024. https://www.ndr.de/geschichte/tagebuecher/Datenbank-Die-gefaelschten-Hitler-Tagebuecher-zum-Durchsuchen,hitlertagebuecherdatenbank102.html.

'Our Members'. Accessed 17 November 2024. https://readcoop.org/members.

'PARDONS'. Accessed 26 September 2025. https://pardons.eu/.

Prokop, Helene, and Christel Annemieke Romein. 'Einsatz von künstlicher Intelligenz bei der automatischen Handschriftenerkennung. Das Beispiel Transkribus'. In Lauter weiße Flechen? Aktuelle Brennpunkte der Archivarbeit. Referate des Landesarchivtags in Dessau-Roßlau am 12. und 13. Juni 2024. Landesarchivtag Sachsen-Anhalt 2024. VdA - Verband deutscher Archivarinnen und Archivare e.V, 2025.

'Recognition and Enrichment of Archival Documents | READ Project | H2020 | CORDIS | European Commission'. Accessed 29 July 2021. https://cordis.europa.eu/project/id/674943.

Romein, C. A. (Annemieke). Handleiding Citizen Scientists Collectie Overijssel i.s.m. HAICu/ UTwente Archiefdeel: Resoluties van de Staten van Overijssel. Zenodo, 14 May 2025. https://zenodo.org/records/15401991.

Romein, C. Annemieke, Tobias Hodel, Femke Gordijn, et al. 'Exploring Data Provenance in Handwritten Text Recognition Infrastructure: Sharing and Reusing Ground Truth Data, Referencing Models, and Acknowledging Contributions. Starting the Conversation on How We Could Get It Done'. Journal of Data Mining and Digital Humanities, ahead of print, 24 March 2023. https://doi.org/10.5281/zenodo.8116009.

Romein, C. Annemieke, Süphan Kırmızıaltın, Ronny Reshef, et al. ‘From Research Proposal to Project Management. A Guide from the Transkribus Community on Planning and Executing Workflows for Researchers and GLAM-Professionals’. International Journal of Digital Humanities, ahead of print, 1 September 2025. https://doi.org/10.1007/s42803-025-00107-7.

Romein, Christel Annemieke, Melissa Terras, Andy Stauder, Bettina Anzinger, and Florian Stauder, eds. Praeteritum Transcriptum. A Transkribus Tribute: Celebrating Our First Five Years as a Cooperative (2019-2024). Zenodo, 2025. https://doi.org/10.5281/zenodo.15308678.

STRG_F, dir. Böse Fälschung: Was Steht in Den ‘Hitler-Tagebüchern’? | STRG_F. 2023. 27:22. https://www.youtube.com/watch?v=NNspVJCdaQw.

Terras, Melissa, Bettina Anzinger, Günter Mühlberger, C. A. (Annemieke) Romein, Andy Stauder, and Florian Stauder. ‘READ-COOP and Transkribus: A Cooperative Model for Responsible Technology’. 24 May 2025. https://doi.org/10.5281/zenodo.15503325.

The eScriptorium VRE for Manuscript Cultures – Classics@ Journal. n.d. Accessed 26 September 2025. https://classics-at.chs.harvard.edu/classics18-stokes-kiessling-stokl-ben-ezra-tissot-gargem/.

‘Transkribus’. Accessed 16 September 2024. https://app.transkribus.org/nl.

uibk. ‘+ READ-COOP SCE Formally Established!’ READ-COOP, 15 November 2019. https://readcoop.eu/read-coop-sce-formally-established/.

'ARKEY - AI meets archives - Archives de l’État en Belgique', accessed 22 may 2025, https://arch.arch.be/index.php?l=fr&m=nos-projets&r=projets-de-recherche&pr=arkey-ai-meets-archives

CLARK, E. et al., ‘Creative Writing with a Machine in the Loop: Case Studies on Slogans and Stories’; 23rd International Conference on Intelligent User Interfaces, 2018, ACM (Tokyo Japan), https://doi.org/10.1145/3172944.3172983

‘Common Crawl - Open Repository of Web Crawl Data’, accessed june 26th 2025, https://commoncrawl.org

XU, Z., S. JAIN, & M. KANKANHALLI, ‘Hallucination is Inevitable: An Innate Limitation of Large Language Models’, 2025, arXiv, https://doi.org/10.48550/arXiv.2401.11817.

HOROWITZ, M. C. & L. KAHN, ‘Bending the Automation Bias Curve: A Study of Human and AI-based Decision Making in National Security Contexts’, International Studies Quarterly, Volume 68 Issue 2, April 2024, Oxford Academic, https://doi.org/10.1093/isq/sqae020

Table of Contents/Inhaltsverzeichnis

6668

ISBN 978-94-6391-661-5

9789463916615

Umschlagbild : Archives de l'Etat à Namur ©AENamur.

belspo