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

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

https://doi.org/10.5281/zenodo.20059676
https://orfeo.belnet.be/handle/internal/14703

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

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

Table of Contents/Inhaltsverzeichnis

FOREWORD

Archives are currently undergoing a period of profound change. As institutions that preserve cultural heritage whilst also providing reliable sources of information for society, public administrations and research, they are faced with new expectations and opportunities. Digitisation has already had a lasting impact on this transformation; with the increased use of artificial intelligence, it is gaining new momentum.

The articles in this volume highlight the diversity of opportunities. AI can help make large volumes of archival material more accessible, make previously hard-to-access sources available, and open up new ways of using them. At the same time, however, the articles also show that this path is not without challenges. Questions regarding the quality of automated results, the handling of uncertainties, and the preservation of historical context are central to the discussion. It is repeatedly emphasised that AI can support archival work – but cannot replace it.

Another important aspect relates to the legal framework. As Laura Drechsler's article highlights, archives today operate within a complex European legal framework in which different regulations governing personal and non-personal data intersect. Careful consideration is required to balance the aim of making data as openly accessible as possible with the obligation to protect sensitive information.

A common thread running through all the contributions is the realisation that dealing with artificial intelligence requires more than just technical adaptation. It challenges us to rethink fundamental aspects of archival practice and to further develop our own mission in the digital age. In this regard, the symposium in Namur provided an important forum for discussion.

The texts collected in the present volume are therefore not intended as definitive answers, but as contributions to an ongoing conversation. They invite you to take up considerations that have already been raised and develop them further.

VORWORT

Archive befinden sich gegenwärtig in einer Phase tiefgreifender Veränderung. Als Institutionen, die kulturelles Erbe bewahren und zugleich verlässliche Informationsgrundlagen für Gesellschaft, Verwaltung und Forschung bereitstellen, sehen sie sich mit neuen Erwartungen und Möglichkeiten konfrontiert. Die Digitalisierung hat diesen Wandel bereits nachhaltig geprägt; mit dem verstärkten Einsatz Künstlicher Intelligenz erhält er eine neue Dynamik.

Die Beiträge dieses Bandes machen deutlich, wie vielfältig die damit verbundenen Chancen sind. KI kann helfen, große Mengen von Archivgut besser zugänglich zu machen, bislang schwer erschließbare Quellen zu erschließen und neue Formen der Nutzung zu eröffnen. Gleichzeitig zeigen die Beiträge aber auch, dass dieser Weg nicht ohne Herausforderungen ist. Fragen nach der Qualität automatisierter Ergebnisse, nach dem Umgang mit Unsicherheiten oder nach dem Erhalt des historischen Kontextes begleiten die Diskussion. Immer wieder wird dabei betont, dass KI archivische Arbeit unterstützen kann – sie jedoch nicht ersetzt.

Ein wichtiger Gesichtspunkt betrifft zudem die rechtlichen Rahmenbedingungen. Wie der Beitrag von Laura Drechsler hervorhebt, bewegen sich Archive heute in einem komplexen europäischen Rechtsraum, in dem unterschiedliche Regelungen zu personenbezogenen und nicht-personenbezogenen Daten ineinandergreifen. Zwischen dem Ziel, Daten möglichst offen zugänglich zu machen, und der Verpflichtung, sensible Informationen zu schützen, sind sorgfältige Abwägungen erforderlich.

Was sich durch alle Beiträge zieht, ist die Einsicht, dass der Umgang mit Künstlicher Intelligenz mehr verlangt als technische Anpassung. Er fordert dazu heraus, grundlegende Fragen archivischer Praxis neu zu bedenken und den eigenen Auftrag im digitalen Zeitalter weiterzuentwickeln. Das Symposium in Namur hat hierfür einen wichtigen Raum des Austauschs geschaffen.

Die hier versammelten Texte verstehen sich daher nicht als abschließende Antworten, sondern als Beiträge zu einem fortlaufenden Gespräch. Sie laden dazu ein, die begonnenen Überlegungen aufzugreifen und weiterzudenken.

Für den Lenkungskreis
Els Herrebout
April 2026

Open Data versus Black Box: Or: How can AI Fulfill Archival Tasks and Professional Requirements? [1]

Bettina JOERGENS[2]

> The digital transformation may challenge us much more as Terry Cook already knew about 30 years ago: „‚He asked: ‚How do we recast our paper minds to deal electronic realities?'"[3]

A colleague of us said recently that Artificial Intelligence (AI) seems to be like sparkling glitter, like a unicorn, everyone wants to have. It is like children urgently want a toy, like children who built and program robots[4]. In fact – this is my argument – it is a game changer, which needs change management – also in archives. Even the UNESCO recommends to use AI technologies "in order to preserve the material, documenting and immaterial cultural heritage, administrate and make accessible".[5] AI is no toy, it is not glittering, but it is crucial also for archives to use AI technologies to hold or expand their positions as accountable institutions of the information infrastructure. But if we want to apply deep learning, machine learning or generative AI professionally and based on archival principles, – this is my thesis – then we have to "think" and to reflect about in many respects like professional, legal issues and issues concerning infrastructure and organisation. Here, I follow Colavizza et al., who "note a still limited reflexive attitude towards AI".[6]

[1] This article is based on my keynote at the International Archival Symposion, 5./6.6.2025 in Namur. I kept the lecture style.

[2] Dr. Bettina Joergens ist Leiterin des Fachbereichs „Grundsätze" beim Landesarchiv NRW in Duisburg.

[3] David Canning / Lise Jaillant, AI to review government records: new work to unlock historically significant digital records, in: AI & Society, 22 February 2025, [D. 1] (https://doi.org/10.1007/s00146-025-02221-0).

[4] Junge Tüftler bauen und programmieren Roboter - Lokalzeit Ruhr - Sendungen A-Z - Video - Mediathek – WDR (Junge Tüftler bauen und programmieren Roboter - Lokalzeit Ruhr - Sendungen A-Z - Video - Mediathek - WDR, 7.11.2025).

[5] Deutsche UNESCO Kommission e.V. (Hg.), Die UNESCO-Empfehlungen zur Ethik der Künstlichen Intelligenz, Wegweiser für die Gestaltung unserer Zukunft, 2023, p. 136, 140.

[6] Giovanni Colavizza / Tobias Blanke / Charles Jeurgens / Julia Noordgraaf, Archives and AI: An Overview of Current Debates and Future Perspectives, in: ACM Journal on Computing and Cultural Heritage, 15 (2021), Issue 1, here p. 10. (https://doi.org/10.1145/3479010, 7.11.2025).

In the first part, I offer some basic information about Artificial Intelligence, also in archives. In the second part "AI meets Archives" I discuss several critical questions which should help to apply AI in archives professionally. Since the development of AI in general, AI-technologies and AI applications in archives are extremely dynamic, consequently, my article can just be a snapshot.

1. AI-technology landed on earth – and in Archives

AI experts (and others) might tell us that there is nothing new about AI: Its history lasts – at least – for 70 years. But in fact, for most of the people, the public discourse and also for archivists AI came up in 2022: Open AI published ChatGPT on the 30th of November 2022 and made AI technology accessible and usable for everyone. Miriam Meckel and Lea Steinacker call this event the "smartphone moment".[7] AI has become part of our daily life and wider critical social discourses. In the following, I will offer some rudimentary, basic aspects like a definition and some historical benchmarks of AI. Since legal issues will be discussed by Laura Drechsler in this issue, I am not going into it.

1.1. What is AI - technologically?

This is a really difficult question, because even experts do not agree with each other. For sure, the problem starts with the unprecise term "Artificial Intelligence". One reason is that there is no firm agreement about the question, what is (human) intelligence. But according to Manuela Lenzen, author of "Künstliche Intelligenz. Fakten, Chancen und Risiken", there is the consensus that "intelligence" has to do with flexibility and learning, also with the ability to cope with changing requirements.[8] Katharina Zweig, professor for informatic, says that there are several meanings of the term "Artifical Intelligence": On the one side, it is a field of research which develops methods to enable computers to make things, for which human beings need intelligence. And on the other side these methods themselves are called "Artificial Intelligence". Further, although it is confusing, software which is

[7] Miriam Meckel / Léa Steinacker, Alles überall auf einmal. Wie künstliche Intelligenz unsere Welt verändert und was wir dabei gewinnen können, Hamburg 2024; s.a. Der iPhone-Moment der Künstlichen Intelligenz - PROMAGAZIN (7.11.2025).
[8] Manuela Lenzen, Künstliche Intelligenz. Fakten, Chancen, Risiken, 3. Aufl., München 2024, p. 14.

based on those methods is also called "Artificial Intelligence".[9] According to Manuela Lenzen, the current discourse uses "Artificial Intelligence" for programmes which work with processes of machine learning, especially with Deep Learning based on neuronal networks.[10] Isto Huvila, a Swedish information scientist emphasizes that AI is much more than just a technological tool, it is a broad and non-definite concept.[11] I will come back to his thesis when I discuss the management challenges of using "AI". All in all: "Artificial intelligence" generally describes technologies from information technology that work with algorithms, automatic processes, recognize patterns and, in some cases, learn in the process. In many respects, this is what computers have always done. It is therefore relevant to name the technologies used and how they work in order to understand the processes in order to design projects professionally and sustainably.

1.2. Some historical benchmarks

As early as 1843, the "world's first computer programmer", Ada Lovelace, predicted: "Technology will one day be able to reproduce everything that can be described with logic. Because these things have recognizable patterns, the machine will be able to compose music and write poetry in the future".[12] Around a century later, in the 1950s, two fundamental approaches were developed within AI research:
The first approach focused on symbolic AI, which is based on explicitly defined rules and logical deductions. The second approach focused on artificial neuronal networks, which are based on data-driven learning and can recognize patterns and correlations directly from the data. According to Meckel and Steinacker, symbolic AI initially seemed to ‘win the race’. Today, we know that the field is dominated by generative AI, which is based on neuronal networks. This uses a variety of algorithms, including models such

[9] Katharina Zweig, Die KI war’s. Von absurd bis tödlich: Die Tücken der künstlichen Intelligenz, 4. Aufl. München 2024, p. 32; see also: Bettina Joergens / Tobias Krafft, Künstliche Intelligenz im Archiv. Annäherungen zwischen Methoden und Praxis in Archiven einerseits und KI-Technologien, in: Archiv. theorie & praxis, 78 (2025), H. 1, p. 6-11; Lenzen 2024, p. 14. Isto Huvila, a Swedish information scientist, emphasizes that AI is much more than just a technological tool, it is a broad and non-definite concept: Isto Huvila, Letting AI Loose in an Archive. Technology to Manage or to Manage with, in: „Archiv theorie & praxis“, 78 (2025), issue 1, p. 12-15, here p. 15.
[10] Lenzen 2024, p. 14.
[11] Huvila 2025. p. 15.
[12] Meckel / Steinacker 2024, p. 43f.

as ChatGPT, machine translation applications, as well as techniques such as machine learning, deep learning and reinforcement learning (RFL). Generative AI enables the creation of new content, such as text, images or music, based on existing data patterns.[13]
But "Generative AI" requires an almost unimaginable amount of data and computers must be able to process these data. Consequently: "The era of big data" – and of generative AI – "began with the World Wide Web, " so Meckel and Steinacker[14], and after two "AI winters". The two authors trace how Fei-Fei Li, a scientist of Chinese descent in the USA, created "the largest database in AI research at the time" in order to make the previously unknown amounts of data usable for AI in the early 2000s: "After years of work, Li's team had compiled 15 million images, which were organized according to 22.000 object classes", by 50.000 people, mainly from the Global South.[15] Later, in 2017, Google introduced the "self-attention mechanism", which makes it possible to weigh up the meaning of different words in a sentence in relation to each other.

This mechanism forms the base of the Transformer Model Architecture, which represents a revolution in the field of natural language processing (NLP).[16] Transformer models can use this method to capture efficiently contextual relationships and thus achieve significantly more precise results in tasks such as machine translation or text generation.[17] Now, Machine Learning and Deep Learning technologies help human beings to process enormous amounts of data to reproduce existing knowledge in order to produce texts, music, images, presentations, classifications etc. And this is attractive for archives.

2. AI meets Archives

It is not exaggerated that there is an AI hype among archives.[18] Since 2022, and more in 2024, many archival conferences focus on AI in archives. I just

[13] Ebda., p. 23f.
[14] Ebda., p. 62. See also Roberto Simanowski, Data Love, Berlin 2014.
[15] Ebda., p. 63-65.
[16] Ebda., p. 78.
[17] Ebda., p. 78.
[18] See also for libraries and other cultural heritage organisations Lise Jaillant / Claire Warwick / Paul Gooding / Katherine Aske / Glen Layne-Worthey / J. Stephen Downie (Hg.), Navigating Artificial Intelligence for Cultural Heritage Organisations, London 2025 (https://doi.org/10.14324/111.9781800088375, 7.11.2025).

mention the AI panel at the German Archival Conference at Suhl in 2024, the Cologne Forum 2024, the internal archival conference of the State Archives of North-Rhine-Westphalia in 2024, the Berliner Archivtag 2024[19], the Symposion of the Archival School at Marburg or the Conference of the Federal Archives and the State Archives of Rheinland-Pfalz in Coblenz, both in 2024.[20] Above that, many archival institutions started AI projects and prove the use of AI for archival jobs. Meanwhile, the international debate about AI and Archives concerning "Theoretical and Professional Considerations", "Automating Recordkeeping Processes and Decisions", "Appraisal", "Handling Sensitive Information", "Metadata", "Organising and Accessing Archives" and "Novel Forms of Digital Archives" reaches back at least until 2018, according to Colavizza, Blanke, Jeurgens and Noordegraaf.[21]

Elena Williams made a survey about AI applications in archives in the year 2023 for her Bachelor thesis at the University for Applied Sciences in Potsdam.[22] According to her survey among 241 (small) archives (with 1-5 employees) 29% already use AI. Most of them (85 %) apply AI for transcriptions and descriptions of images (70%) and texts (66,7%). Only 42% use it for description of films. All in all, the application of AI focuses on archival description in a broader sense. But, as Williams found out, archivists can imagine to apply AI technologies also for many other archival jobs like digital consulting, appraisal, for the online presentation, in the reading room and in the archival stores. William‘s survey makes clear: AI has great potential in and for archives. Colavizza et al. are convinced that "AI is used throughout the record keeping processes that characterise archive".[23] But which potential exactly and what are the hurdles? In the following I will discuss four issues:

[19] Torsten Musial / Yves A. Pillep (Hg.), KI im Archiv - Chance oder Risiko? Referate des 7. Berliner Archivtags am 20. November 2024 (Tagungsdokumentation zum Berliner Archivtag, Bd. 7), Fulda 2025.
[20] See the special focus on AI in archives in the issue Archiv. theorie & praxis, 78 (2025), H. 1, p. 6-53; Joergens / Krafft 2025. See also Robert Klugseder, Der Einsatz von KI in Archiv und historischer Forschung Eine Betrachtung aktueller Trends und Zukunftsperspektiven Keynote beim "13. Bayerischen Archivtag" im Kongress am Park Augsburg am 25.03.2025 (PowerPoint-Präsentation, 7.11.2025).
[21] Colavizza et al. 2021, p. 2. (https://doi.org/10.1145/3479010, 7.11.2025); see also: Lise Jaillant (Hg.), Archives, Access and Artificial Intelligence. Working with Born-Digital and Digitized Archival Collections, Bielefeld 2022.
[22] Elena Williams, Alexa trifft Archiv. Der aktuelle Stand des Einsatzes von KI in Archiven, in: Archiv. theorie & praxis, 78 (2025), H. 1, p. 16-19.
[23] Colavizza et al. 2021, p. 2.

1) In which way can AI be useful for archives?
2) Who is the boss? Who takes control? Who is responsible?
3) AI needs Data – Archival Data need AI
4) AI – a challenge for archival management

2.1. In which Way can AI be Useful for Archives?

Archives are institutions of cultural heritage and trustworthy information infrastructure. Their job is to hold, preserve and give access to information. Archival material is – hopefully completely – described by metadata, is partly digitised, hence transformed into data, or originally digital (meaning data). The main and almost first access to archives and archival material is done via online information and search tools, hence via data processing. Coming from – almost – written cultural heritage and administrative documents, our job is to a large extent data driven. Colavizza et al. even state that "the digital transformation is reconfiguring the archive from a collection of administrative records into a collection of data", meaning datafication.[24] Above that, we "give" our data and metadata in the world wide web, in archival and other cultural portals and data rooms – not only to improve the access to archival material, but to be part in the big realm of information infrastructure, which is crucial for scientific, social and economic progress. By the way: That is why it is so important to make our data fit for data linkage and the semantic web.

One crucial precondition of playing with data delivering players is the transformation of archival material into machine readable data, meaning: unlock records[25] via description, indexing and transcription of archival material and archival information. But to be honest, I think, that no archival institution can state that 100% of their holdings are machine readably described or digitised. For example, that State Archives of North-Rhine-Westphalia has – roundabout – 10-15% of its holdings digitised. We all would need much more resources and – first of all – employees to approach the 100%. Probably, that won't happen. But now, AI may help to transform archival material (analogue or digitised archival objects and representatives) into useful, machine-readable data.

[24] Ebda., p. 2, 9.
[25] Lise Jaillant / Lingjia Zhao, Introduction: When data turns into archives: making digital records more accessible with AI, in: AI and Society. Journal of Knowledge, Culture and Communication (April 2025) (https://doi.org/10.1007/s00146-025-02374-y, 7.11.2025).

Florian Detjens wrote his exam thesis during his archival traineeship about AI in archives in this spring. He gives a precise summary how AI may be usefully applied in archives.[26] In the following, I will refer to some of his – not yet published – results, which focus on archival description, as well as on a spontaneous non-representative poll among state and federal archives in 2025: No matter how one will use AI for improving the access to the contents of archival material, first of all one has to transform the analogue writing into machine readable text by Optical Character Recognition (OCR) or by Handwritten Text Recognition (HTR)-tools. HTR is based on *deep learning*, which has "learned" with training data and is able to interpret characters as letters. *Transformer Models*, which are *deep learning models*, help very efficiently to cope with different handwritings. Further, one can extract entities (like names of persons or places) from unstructured texts with Named Entity Recognition (NER); it is based on linguistic rules and statistic processes.[27] NER can be combined with data- or text-mining to structure data. Data-mining as an algorithm-based analysis process is able to extract information from weakly structured texts. In this way, it is possible to generate indexes automatically.[28] For example, Florian Probst and Thomas Reich introduce their project to describe 156 Reichstags reports from the time between 1681 and 1803. In three steps (text and handwriting recognition, text processing via Natural Language Processing (NLP), and evaluation with generative AI), geographical entities and suggestions of archival metadata should be extracted.[29]

[26] Florian Detjens, Das automatisierte Archiv? Chancen und Risiken beim Einsatz von KI, Transferarbeit, vorgelegt am 1.4.2025 (noch unveröffentlicht).
[27] Detjens 2025, p. 5. See also Andreas Neuburger, Die Zähmung des Chaos. Perspektiven zur KI-gestützten Erschließung von Entschädigungsakten im Landesarchiv Baden-Württemberg, in: Archiv. theorie & praxis, 78 (2025), H. 1, p. 34-37; Benjamin Rosemann: Deutscher Archivtag 2024 | FDMLab@LABW (29.5.2025).
[28] Detjens 2025, p. 11f.
[29] Florian Probst / Thomas Reich, Das Digitalisat als Datenquelle. Ein Workflow für die Nutzung bereits digitalisierter Archivbestände, in: Archiv. theorie & praxis, 78 (2025), H.1, p. 38-42.
Further examples: Also the coordination office for provenance research of the State North-Rhine-Westphalia works with a combination of OCR and Language Models in order to describe and unlock archival holdings for the provenance research. (Ruth von dem Bussche / Jasmin Hartmann, Die KI-basierte Erschließung von Archivbeständen für die Provenienzforschung. Best Practice – Next Practice, in: Archiv. theorie & praxis, 78 (2025), H.1, p. 43-48.)
The Federal Archives of Germany uses as well OCR and HTR to describe and transcribe handwritten documents of the Reichskolonialamt, index cards (Ministry of Statesecurity, GDR, and NSDAP-members) and the Reichskanzlei (19th/20th century). (Esther Lemmerz, Welche

Generative AI tools like Large Language Models are able to produce new content and are – somehow – creative: Generative AI also has to be trained with data. But the data does not always or necessarily have to be annotated like for the non-generative AI. This is called self-supervised learning (selbstüberwachtes Lernen). However, the training data have to be somehow structured and the training-job must be clearly defined with prompt engineering, like "make a summary of a text" or "translate". But additionally, archives need tools which recognize layout like forms or tables. Multimodal Large Language Models are able to structure the labelled text according to the fields in the form or table, because MLLMs understand texts semantically with its visual context. The State Archives of North-Rhine-Westphalia and the Technical University of Dortmund are preparing a project to develop a Large Vision Language Model (LVLM), which is a specialised Multimodal Large Language Model. This project has several goals, mainly: archival description of about 130.000 personal index cards of prisoners from about 1925 to 1960, the mapping of the recognized information to the defined labels and the definition of the term of data protection. Above that, the developed MLLM should be usable for further projects for archival material from the first half of the 20th century which shows a combination of texts and forms.

Archives hold also audiovisual material, but it is often enough not described. The city archives of Heilbronn use Software of the platform DeepVA (by Aiconix) for tools to recognize persons, buildings and places on photos. The offered AI tools to analyse and administrate e.g. audiovisual media are analytic as well as generative.[30] In a study, led by Emmanuele Frontoni, a Deep Learning tool, is been used to "derive meaningful information from digital images, videos and other visual inputs", especially "the signum, a specific and personally drawn mark used by a single notary".[31] Other archives, like the one of the Fritz-Bauer-Institut in Frankfurt (Main), use Whisper,

Möglichkeiten bietet uns KI? Erste Erfahrungen aus dem Bundesarchiv, talk at the Deutscher Archivtag in Suhl, 2024 (will be published).

[30] Miriam Eberlein, Bild-Erschließung mit KI. Sechs Vorschläge für die Nutzung KI-generierter Metadaten, in: Archiv. theorie & praxis, 78 (2025), H.1, S. 49-53.

[31] Lucia Duranti, I Trust AI. The Fifth Phase of the Interpares Project, in: Archiv. theorie & praxis, 78 (2025), H.1, p. 20-23, here p. 21. See also Pier Luigi Mazzeo / Emmanuele Frontoni / Stan Scarloff / Cosmimo Distante (Hg.), Image Analysis and Processing. ICIAP 2022 Workshops. ICIAP Internaional Workshops, Lecce, Italy, 2022, also: Image Analysis and Processing. ICIAP 2022 Workshops: ICIAP International Workshops, Lecce, Italy, May 23–27, 2022, Revised Selected Papers, Part II | SpringerLink (29.5.2025).

which is a non-generative tool for speech recognition.[32] The State Archives of North-Rhine-Westphalia also plans to test Whisper to transcribe the spoken text in video films which show interviews with victims of the national socialist regime concerning their pension payments (called "Ghetto-Renten").[33]

AI tools are also interesting for other archival tasks besides description: David Canning and Lise Jaillant introduce the application of AI in order to cope with the great amount of not yet archived digital records of the British government. They say, since this "process of appraisal cannot be done manually", they developed a system which helps to classify with at least three levels of filtration partly done by a machine, partly by human beings.[34]

Another field of AI application in archives is the service for users: The Federal Archives of Switzerland uses a chatbot to help users. It answers questions concerning the research, the holdings, requests, the online-access or the transfer of documents from agencies to the archives.[35]

Reports about using AI for conservation or preservation are still seldom, but existent: Fraunhofer IOSB-INA (Lemgo) works on an automated climate

[32] Johannes Beermann-Schön, Fritz-Bauer-Institut (Frankfurt/M.) presented this project at the Deutscher Archivtag at Suhl 2024. His paper is to be published.

[33] Transcriptions of spoken testimonies in video records of the context of proceedings concerning pension payments for former ghetto inmates: At the beginning of the 2000er years, a judge of the Landessozialgericht Düsseldorf travelled to Israel to take videos of NS-victims. These about 150 audio videos are archival objects of the State Archive of North-Rhine-Westphalia (LAV NRW Abt. R Gerichte Rep. 871). Whisper, an open source tool which can be used on premis, should help to transform spoken language into text to support the archival description. See also: Benedikt Nientied, Ghettorentenakten - Das letzte Kapitel der sozialversicherungsrechtlichen Aufarbeitung des NS-Unrecht, in: Jens Heckl (Hg.), Unbekannte Quellen: „Massenakten des 20. Jahrhunderts. Untersuchungen seriellen Schriftguts aus normierten Verwaltungsverfahren, Bd. 5, hg. i. A. des Landesarchivs NRW (Veröffentlichungen des Landesarchivs Nordrhein-Westfalen 99). Duisburg 2025, p. 96-111. DeepSpeech, also a non-generative speech recognition or audio-mining tool is been used for language and speech analysis like in "Oachkatzl". "Oachkatzl" is the Bavarian word for "Eichkätzchen" or "Eichhörnchen", in English "squirrel". It is also the name of a project to transcribe audiomedia sources with spoken Bavarian dialects: Johannes Lederle, Oachkatzl. Training and Benchmarking von KI-basierten Audio-Mining-Systemen auf bayerische Dialekte, in: Info 7, 25 (2020), H. 2, p. 33-38) (https://www.vfm-online.de/newcomerforum/preistraeger/info7_2020-2_S33-38.pdf, aufgerufen am 29.5.2025).

[34] Canning / Jaillant 2025 (https://doi.org/10.1007/s00146-025-02221-0).

[35] See Chatbot (aufgerufen am 29.5.2025).

monitoring in the magazine. They plan to cooperate with the regional departement Ostwestfalen-Lippe of the State Archives of North-Rhine-Westphalia in order to participate in a funding competition on the topic of "mold growth" in archives. The intention is to use intelligent sensor technology to monitor the climate in archives and to collect and analyze these data. In a second step, these data will be used with AI as the basis for intelligent ventilation and/or heating control in archives.

The future range of possible AI applications in archives might be much broader than we can imagine now. In the end, it is a management issue to define a professional goal by using AI and to consider the profits, the expected results and the costs of applying AI technologies, before starting a new project.[36] Above that, using AI technology is not only adding a tool to well-known processes like using text programmes like "Word" instead of a typewriter. It is not least also a management question and an issue of responsibility, accountability and transparency.

2.2 Who is the Boss? Who does Control? And who is Responsible?

Is it the AI-tool or the archivist? AI can make our work and lives more comfortable, because AI is much faster than human beings in processing data. But we should not be lazy and lean back. We are still responsible and, therefore, have to know the range of the applied AI technology. Here is an example for it: The German "Tatort" on the 21st of April 2025, a crime story series, showed a story about a murder by knife at the station of Hannover, in a very crowded place. Because the investigator team has no clues, indications, motives etc. one part of the team relied on a new AI application. This new tool calculated on the basis of some given aspects of personality and the login information of smartphones at the time of the crime at the station. Finally, they found a man who fitted perfectly in this pattern, but he was not the murderer. Tragically, he was mentally ill and committed suicide under the pressure by the police. The investigation team makes two crucial mistakes: 1) It mixes up correlations (e.g. a data pattern matches the profile of an individuum) with causality, 2) It doesn't understand that AI is often based on probability calculations. Probability calculations do not give specific information about an individual person, case or object. But: In the end, they

[36] See also Martin Vogel, Künstliche Intelligenz im Archiv. Herausforderungen und Chancen, in: Archiv theorie & praxis, 78 (2025), H.1, p. 24-28, here p. 25f.

found the murderer by combining AI technology and human thinking and experienced knowledge.
Sharing the jobs with the machine, it is crucial to understand the machine. But deep learning models work with so called black boxes. Black boxes stand for a complex architecture with many levels and parameters, and, additionally, with many non-linear processes. Consequently, the results are often enough not interpretable for human beings. Junny Bunn states: "The increasing use of more opaque AI techniques is generally framed as disruptive for recordkeeping."[37] Therefore, colleagues like Bunn (UK) and Huvila (Sweden), discuss about XAI in archives: Explainable AI. Further, there is a risk of hallucinating by AI. Hallucinating means that the machine invents results, e.g., in the case of too less (training) data. Therefore, it is necessary to take those failures into account of the project design.[38]

Jean-Claude Mbassi Ndzengue emphasizes in his talk "Quand l'intelligence artificielle sème le doute: la prolifération des faux documents face aux enjeux de gestion électronique des documents": "AI, a powerful but risky tool, is revolutionizing document management while promoting the proliferation of 'fake documents'. These threaten the probative value of digital archives. To address this, advanced technologies, regulatory frameworks, and professional training must be combined. Archivists and managers, on the front lines, must adopt a proactive stance, integrate technological skills, and strengthen trust in digital systems."[39] Dominique Luster "explores [in her paper] culturally competent and racial-conscious archival practices in the African diaspora, focusing on ethical metadata creation and equitable representation. To complement archivists' efforts, it incorporates experiments with AI tools like CustomGPT, which are trained to revise metadata using inclusive guidelines.

[37] Jenny Bunn, Working in Contexts for which Transparency is Important. A Recordkeeping View of Explainable Artificial Intelligence (XAI), in: Records Management Journal, 30 (2020), 2, p. 143-153, here: p. 143.
[38] "The increasingly apparent risks relating to hallucinating generatives AI systems", so Huvila, "… led to mounting demands of transparency, explanations and information on how AI techniques function" (Huvila 2025 p. 13.) See also Duranti 2025, p. 20.
See also Gabor Mihaly Toth / Richard Albrecht / Cedric Pruski, Explainable AI, LLM, and digitized archival cultural heritage: a case study of the Grand Ducal Archive of the Medici, in: AI & Society (March 2025) (https://doi.org/10.1007/s00146-025-02238-5, 7.11.2025).
[39]Jean-Claude Mbassi Ndzengue, Quand l'intelligence artificielle sème le doute: la prolifération des faux documents face aux enjeux de gestion électronique des documents, talk at the Congress of the International Council on Archives at Barcelona 2025 (Programme - ICA Barcelona 2025, abstract, 7.11.2025).

The paper ties these approaches to digital access, colonial records, and fostering a global dialogue on creating more inclusive archives." It shows that the output of the AI tool must be thoroughly controlled. Consequently, the quality management is a big workload as part of the project, as Luster explaines.[40]

Archives as public institutions stand for accountability and transparency. But what about archival decisions about access and appraisal? These are decisions within the scope of discretion and decisions with long-term and legal impact. The State Archives of NRW stated (so far) that the machine may define which documents can be given online access by legal terms, but the quality management must exclude any mistakes (zero tolerance). Further, archivists have to decide, if they want to use AI for appraisal processes. The argument against it is that appraisal is the result of a highly responsible and archival professional consideration. David Canning and Lise Jaillant say that we need AI for appraisal, otherwise big amounts of documents were not transferred to archives and, hence, are not accessible.[41] Concludingly, the responsible archivists have to know the technologies and consider how to use AI and how to establish a good quality management.[42]

Luciana Duranti, head of the InterPARES project "I trust AI" (2021-2025), knows that explainable AI is still difficult to establish. But: "It is important to understand", she writes, "that Accountable AI is different from Explainable AI as the latter focuses in why a given tool produced a given output from a given set of inputs. Building accountable AI must also consider the individuals, organization, and environment in which the AI tool operates, and paradata is necessary to explain why, how, by whom, and to what effect a given tool was used in a particular context. "[43] Further: "paradata is necessary to document the AI process and promote transparency and accountabilty... ".[44] This includes being transparent of possible reproduction of discriminating terms and

[40] Dominique Luster, Inclusive Metadata: Can AI Think Like an Archivist?, talk at the Congress oft he International Council on Archives at Barcelona 2025 (Programme - ICA Barcelona 2025, 7.11.2025).
[41] Canning / Jaillant 2025, [D. 1] (https://doi.org/10.1007/s00146-025-02221-0).
[42] See also Vogel 2025, p. 26.
[43] Duranti 2025. p. 22.
[44] Duranti 2025, p. 23. See also Annette Zimmermann / Zoe Porter / Phillip Morgen / John McDermid / Tom Lawton / Ibrahim Habli, Distinguishing two features of accountability for AI technologies, in: Nature Machine Intelligence 4, (2022), p. 734–736 (https://doi.org/10.1038/s42256-022-00533-0).

contents: If the used data include discriminating terms or standpoints – as we are familiar with in historical documents –, discrimination is reproduced.[45]

2.4. AI needs Data – Archival Data need AI

It is not enough to have enough data. Furthermore, the data quality is also an essential issue. Crucial questions about data quality concern the kind of data which should be processed (structured data (tables), texts, spoken data in audio/video-files, pictures, videos etc). And: Are the data and metadata machine readable and based on the FAIR principals, meaning findable, accessible, interoperable and re-usable?[46] Further, AI needs enough data to learn and not to hallucinate. Some data need to be annotated – which is an additional benchmark in the wanted AI-project.

Data protection is another important issue: As we know, generative AI technologies need big amounts of data to learn and so to improve the results. But if an AI tool should process protected data, tools in public clouds are not suitable for the project. I would include sensible personal data concerning the intimate sphere of persons which might be open for access by legal terms, but are too sensible for online publication. Further, AI tools which learned by protected data 'learned protected data'. Consequently, these tools cannot be used anymore for open data projects. So, depending on the tool which is to be used, the project management needs to be clear about data protection in the context of AI technology and IT infrastructure.[47]

[45] I will give you an example: I asked ChatGPT to give me good quotations about AI by AI experts. Although I still haven't proved the authenticity of these quotes, it was clear: The authors were all men. Then I asked about quotations by women. Then I got some quotes with the explanation that women often emphasize more on social and emotional aspects of AI. This was a clear gender-based connotation. This harmless example may give an idea of what we have to be aware of when we use AI technologies.

[46] See e.g. FAIR data principles – Forschungsdaten.org (7.11.2025). See also Merc Crosas, AI-Ready Archives: Understanding the past with the tools of the future, Keynote at the Congress of the International Council on Archives, Barcelona 2025 (Unlocking the Future of Archives: Keynote Themes Revealed for the International Archives Congress - ICA Barcelona 2025 - ICA Barcelona 2025, 7.11.2025).

[47] Luckily, some nations and states are in the process of developing their own closed up administration platforms, which might be used for protected data. For example, North-Rhine-Westphalia is going to deliver "NRW.Genius". This is part of an AI infrastructure inside the protected sphere of the administration network of the state NRW (nrw.genius_-die_ki-verwaltungsassistenz_fuer_nrw.pdf and PowerPoint-Präsentation, 21.11.2025).

Consequently, if an AI project deals with protected data, the AI technology must be operated on premise. This is much more expensive and needs more training data of our own holdings to build a good enough "ground truth". Consequently, legal preconditions and technological decisions entail economic questions, also concerning the costs for storage and computer power, new expert knowledge and the development of archival information systems. This leads me to my last point concerning the AI ecology.

2.5 AI – a Challenge for Archival Management

AI technology does not just help us doing a lot of work. It also helps us to take the next step of the digital transformation. But it is also challenging to apply AI professionally and sustainably. I will take a closer look at the technical infrastructure, change of processes and knowledge enhancement.[48]

"AI needs to be considered as part of an entire system"[49], say Canning and Jaillant. Hence, it is crucial to consider, which interfaces are needed for the AI application, like a user interface, the interface to relevant data or the interface to observe the AI and to collect further training data. For example, the integration of the data generated by AI into the existing archival information systems and the IT-infrastructure may cause a change of the archival system. Consequently, AI applications on premise become part of the operated IT infrastructure.[50]

As mentioned, AI applications do also have an impact on the working processes. Just to give an example: The State Archives of North-Rhine-Westphalia has defined that archival material must well enough be described, before it can be digitised. The describing metadata are basic for the "link" between the image and the finding aids. In the planned AI project, the process will be changed: Firstly, the archival material will be digitised, although with rudimentary metadata. Secondly, AI will then extract the necessary metadata which are to be linked with the images.

[48] See also: Abdulhalik Pinar / Andrew Cox, An Analysis of Artificial Intelligence (AI) Capability, in: Cataloging & Classification Quarterly, 63 (2025), 6-7, p. 566-599.
[49] Canning / Jaillant 2025, [last page].
[50] For example, the State Archives of Niedersachsen established a testing ecology to find out with which technological infrastructure and with which tool the best and most efficiently results can be achieved (Vogel 2025, p. 27f.). See also Colavizza et al. 2021, p. 10.

Furthermore, the change of technology and of archival processes require enhanced knowledge and change management. One can say that archivists and archival institutions are now in the middle of a "deep learning" process in order to understand how the different AI technologies work. However, it is crucial to offer AI trainings for archival employees (archivists, IT-technicians and administrative staff) and start an open minded exchange about experiences with other archival and other professionals.[51] Applying AI hopefully will make processes more efficient, but it is also a new task which demands resources.[52] One challenge is that archival principles are under pressure in the context of datafication and using AI: Colavizza et al. see "a general awareness that the digital transformation has put pressure on archival concepts such as provenance and original order… (…). …technical advances require a significant shift in archival thinking, which brings usability, trust, and context in the center of archival work".[53] Or as Tom Scheinfeldt states: "The practical shift is moving our core function from the relatively passive act of *curation* to the active, authoritative act of *certification*."[54]

Katharina Zweig, an AI professional, confirms that applying AI is not only about technology, but it is designing socio-technical systems consisting of human beings and machines.[55] And above that: The consequences for archival science are almost not at yet enough an issue of the archival community. We are at the eve of a change, I would state.

3. Conclusion

Applying AI technology is – concludingly – a great chance for archives as big data holding institutions which are important for the democratic information-based societies. Since archives stand for accountability and transparency, for

[51] See also Vogel 2025, p. 25.
[52] A survey of the Renmin University of China among Chines 34 provincial archive comes also to the conclusion that archivists expect resp. experienced that the application of AI can make archival processes much more efficient, but meanwhile new tasks ("burdens") like training, review or supervision: Yuenan Liu / Xiya Zhang / Jilanliang Yang / Sishi Huang, AI Adoption in provincial Archives in China: Effects, Challenges and Prospects, talk at the Congress of the International Council on Archives at Barcelona 2025.
[53] Colavizza et al. 2021, p. 5.
[54] Tom Scheinfeldt, Generative Artificial Intelligence and Archives: Two Years On, 2025 (Generative Artificial Intelligence and Archives: Two Years On – Found History, 7.11.2025).
[55] Katharina Zweig / Tobias D. Krafft / Anita Klingel / Enno Park, Sozioinformatik. Ein neuer Blick auf Informatik und Gesellschaft, München 2021.

access and reliability, it is crucial to acknowledge that applying AI is much more than using a new tool.[56] A responsible archival management takes the necessary change management, the employee training, the integration in the IT infrastructure, the archival parameter, the economic costs and also (new) legal conditions into account. This leads to a statement by Isto Huvila: "Even if it spontaneously might feel reasonable to think that the most significant challenges posed by AI for archives and records management … are technical by their nature, they are to a large extent elsewhere."[57]

[56] Sarah Rachut, Digital souverän oder so souverän wie gerade möglich?, in Tagesspiegel am 20.11.2025 (https://background.tagesspiegel.de/digitalisierung-und-ki/briefing/digital-souveraen-oder-so-souveraen-wie-gerade-moeglich, 21.11.2025).
[57] Huvila 2025, p. 15.

The EU's legal framework for public archives: from personal data to artificial intelligence

Laura DRECHSLER[1]

1. Introduction

Public archives are evolving in their societal role as defined by EU legislation.[2] In the past, public archives hosted mainly paper-based documents, stored in specially designed rooms in various locations. From a legal perspective, this was regulated in national archival laws with little to no direct involvement of the EU legislator. Anno 2026 this has fundamentally changed as more and more artefacts stored by archives fall under the legal definition of (personal) data, which is experiencing a 'tsunami'[3] of regulation at EU level since the EU's data strategy of 2020[4] (updated in 2025 by a Data Union Strategy).[5] Within a short time-frame, the EU legislator has adopted a variety of rules, often summarized as the EU's 'digital rulebook' which include *inter alia* rules on making data from the public sector more accessible (Open Data Directive[6] and Data Governance Act),[7] but also how to develop trustworthy AI systems for the market (AI Act).[8] These rules come on top of

[1] Dr. Laura Drechsler is affiliated with the State Archives of Belgium, the KU Leuven Centre for IT and IP Law, and the Open University (Heerlen).
[2] See further Mikulás Ctvrtník, *Archives and Records: Privacy, Personality Rights, and Access* (Palgrave Maximillian, e-book 2023).
[3] The term was coined by Prof. Jan De Bruyne in his inaugural lecture at KU Leuven on 10 October 2024 on 'In the Aftermath of the 'Digital Law Tsunami' – Some Remaining Challenges and the Way Forward', available at https://www.linkedin.com/posts/jan-de-bruyne-81931a159_digital-law-tsunami-activity-7250423929188626432-DksA?utm_source=share&utm_medium=member_desktop&rcm=ACoAABQcfuoBREH5oC1vsjbaVVNeHMBfjgBxRc4.
[4] European Commission, *A European Strategy for Data*, COM (2020) 66 final.
[5] European Commission, Data Union Strategy: Unlocking Data for AI, COM (2025) 835 final.
[6] Directive (EU) 2019/1024 of the European Parliament and of the Council of 20 June 2019 on open data and the re-use of public sector information, OJ 2019 L 172/56.
[7] Regulation (EU) 2022/868 of the European Parliament and of the Council of 30 May 2022 on European data governance and amending Regulation (EU) 2018/1724 (Data Governance Act), OJ 2022 L 152/1.
[8] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act), OJ L 2024/1689.

another set of data rules at EU level – EU data protection law, in particular, in the form of the General Data Protection Regulation (GDPR), which regulate all uses of personal data from their creation to their deletion.[9]

Navigating the EU legal landscape for data is no easy task for public archives, since the different regimes are not well aligned at EU level; the newer pieces of legislation lack guidance and case law, and existing archival legislation is often developed without any consideration of the EU's data legislation. With this contribution, I offer a research agenda in order to better detail the route that public archives will have to take in the coming years in order to be compliant with the most relevant legal frameworks, namely the Open Data Directive, the Data Governance Act, the AI Act, but also the GDPR. To do so, I first detail the legal definitions of data that have transformed public archives into so-called 'data spaces', and therefore require compliance with the EU's data acquis. To better outline this acquis, I map the main legal challenges in both strands of EU data legislation: EU data protection law and the EU's digital rulebook. The objective is thereby to provide a legal research agenda for the coming years. Finally, I highlight the legal questions raised by potential uses of Artificial Intelligence (AI) by and through archives as a case study to further detail some of the challenges mapped.

1. Of documents and data in EU law

The classification of archival documents and artefacts as data often comes as a surprise for archival experts, who have a more nuanced understanding of the material they are handling. However, in EU law the question of data is typically independent from both the medium the information is stored on and its content, leading to a large number of materials being considered as falling under the legal definition of data.

The evolution towards a datafication of everything archival starts with the Open Data Directive of 2019, which regulates the access to information held by public authorities, including archives.[10] The Directive, despite its name, uses the concept of 'document' to describe its scope of application.[11] Yet,

[9] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), OJ 2016 L 88/1.
[10] Open Data Directive (n 6) Art. 1.
[11] Ibid., Arts. 1(1) and 2(6).

'document' is defined in a manner that disregards the medium the 'content' is stored on, and does not require its completeness either.[12] A document can thus be anything from the digital to the analogue, according to the EU legislator. This can also be seen in the title of the Directive itself, which, unlike its predecessor who referred to access to public sector information, names 'Open Data' as its central tenet.[13]

The Data Governance Act (DGA) and the Data Act, which were both adopted as a direct consequence of the EU's data strategy of 2020, include a definition of data with a similarly broad scope, although curiously excluding non-digital content. Both define data as 'any digital representation of acts, facts or information and any compilation of such acts, facts or information, including in the form of sound, visual or audiovisual recording'.[14] Yet, the from a legal perspective most important sub-set of such data – 'personal data' – does not exclude the non-digital from its definition.[15] In fact, the Court of Justice of the EU (CJEU) – the highest court for all questions of EU law – has explicitly confirmed that physical information, such as that contained in paper documents, can fall under the definition of personal data and thereby under the scope of EU data protection law.[16]

The diverging definitions are not easy to navigate. However, given the more stringent nature of both the Open Data Directive and the GDPR when

[12] Ibid., Art. 2(6) ''document' means: (a) any content whatever its medium (paper or electronic form or as a sound, visual or audiovisual recording); or (b) any part of such content'.
[13] Compare Directive 2003/98/EC of the European Parliament and of the Council of 17 November 2003 on the re-use of public sector information, OJ 2003 L 345/90 (repealed); to Open Data Directive (n 6).
[14] See DGA (n 7) Art. 2(1); and Regulation (EU) 2023/2854 of the European Parliament and of the Council of 13 December 2023 on harmonised rules on fair access to and use of data and amending Regulation (EU) 2017/2394 and Directive (EU) 2020/1828 (Data Act), OJ 2023/2854, Art. 2(1) DA.
[15] GDPR (n 9), Art. 4(1). Compare to Directive (EU) 2016/680 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and on the free movement of such data, and repealing Council Framework Decision 2008/977/JHA, OJ 2016 L 119/89, Art. 3(1); and Regulation (EU) 2018/1725 of the European Parliament and of the Council of 23 October 2018 on the protection of natural persons with regard to the processing of personal data by the Union institutions, bodies, offices and agencies and on the free movement of such data, and repealing Regulation (EC) No 45/2001 and Decision No 1247/2002/EC, OJ 2018 L 295/39, Art. 3(1).
[16] Case C-25/17, *Proceedings brought by Tietosuojavaltuutettu (Jehovan todistajat)*, judgment of 10 July 2018 (Grand Chamber) (ECLI:EU:C:2018:551).

compared to the DGA and the Data Act, it seems prudent to consider most material stored in public archives, be it digital or analogue, as (personal) data or at least as 'document' for the purposes of EU law. This, in turn requires a thorough assessment on whether, in a concrete situation, a piece of EU data legislation applies and what consequences such an application has. It also turns public archives into institutions handling data, which are further defined in the different legislations. For the Open Data Directive and the DGA, archives essentially transform into 'public sector bodies'[17] and 'data holders',[18] whereas for the GDPR they (often) become 'controllers'.[19] Each legal classification comes with its own set of obligations, some of which will be further explored in sections 3 and 4.

From an EU policy perspective, this also makes public archives into potential 'data spaces' – a concept used by the European Commission since 2020 to describe all types of structures where a large amount of data is shared with multiple parties.[20] While data spaces are not itself fully legally defined,[21] some parts of the new data legislation aims at regulating its activities.[22] The DGA, for example, includes rules for 'data intermediaries' as institutions mediating data between those who have them and those who want to use them,[23] whereas the Data Act puts in place a framework to make different data spaces interoperable with one another.[24] Despite various efforts of the European Commission to foster the set-up and functioning of data spaces, practice has so far yielded few success stories, a continuous obstacle arguably being compliance with the various legal requirements concerning (personal) data.[25] This obstacle would also have to be overcome by public archives as data spaces.

[17] Open Data Directive (n 6) Art. 2(1).
[18] DGA (n 7) Art. 2(8).
[19] GDPR (n 9) Art. 4(7).
[20] The concept was especially developed in two working papers of the European Commission. See European Commission, Commission Staff Working Document on Common European Data Spaces, SWD (2022) 45 final; Commission Staff Working Document on Common European Data Spaces, SWD (2024) 21 final.
[21] The Data Act introduces some specific requirements for data spaces and thereby conceptualizes data spaces as 'purpose- or sector-specific or cross-sectoral interoperable frameworks for common standards and practices to share or jointly process data for, inter alia, the development of new products and services, scientific research or civil society initiatives'. See Data Act (n 14) Art. 33(1).
[22] See further Data Strategy (n 4); Data Union Strategy (n 5).
[23] DGA (n 7), Chapter III.
[24] Data Act (n 4) Chapter VIII.
[25] See further Data Union Strategy (n 5).

While EU law knows various definitions of data, it does not define 'archive' or 'archiving'. This is unfortunate as the institution of an archive is explicitly included in the scope of the Open Data Directive[26] (and not excluded from the scope of the GDPR), whereas the activity of 'archiving in the public interest' has a specific legal regime for personal data.[27] Without definitions, such references are not always easy to decipher, particularly opening questions on the role of private archives and archiving as an activity done outside of public state archives.

The legal definition of (personal) data and documents also further complicates the distinction between data as a regulatory object of much EU legislation, and information, which, following the idea of freedom of information as part of the fundamental right of freedom of expression as, for example, safeguarded in Article 11 of the Charter of Fundamental Rights of the EU (CFR), ought to be easy to share and thus not heavily regulated. Data used to be understood as a particular form of information, which is regulated, whereas the information itself is not.[28] This understanding is, for example, still at the core of copyright law, where protection is awarded to the concrete expression (data) but not to the idea (information). Yet, the definitions of data and personal data muddle this distinction by making information part of the definition. If all information is data, is there still anything left to freely share under freedom of information? This question is especially relevant for public archives, whose ultimate objective is the broad sharing of information with the public at large.

3. Archives and EU data protection law

Perhaps the most established form of regulating 'data' and those who use them can be found in EU data protection legislation. While the General Data Protection Regulation, as its most prominent current representative, is directly applicable since 2018, many of its core principles and concepts have been around much longer, as most European countries introduced some form of

[26] The inclusion is achieved in a negative manner, as archives are explicitly removed from the scope of one of the exceptions of the Directive. See Open Data Directive (n 6) Art. 1(2)(j).
[27] See in particular GDPR (n 9) Art. 89.
[28] See further Lee A. Bygrave, '*Information Concepts in Law: Generic Dreams and Definitional Daylight'*, 35(1) Oxford Journal of Legal Studies (2015), pp. 91-120.

them already in national data protection legislation in the 1970s.[29] This, in turn, inspired first international[30] and then EU legislation.[31]
With the GDPR, it is clear that for the whole of the EU, public archives have to follow its rules when handling personal data.[32] The GDPR is thereby, to a certain extent, mindful of the specific societal value of archiving as it grants some privileges to 'archiving in the public interest' that are either directly applicable or provide options to the Member States to provide for exceptions in national law.[33] No definition for such archiving is provided. Recital 159 suggests that this only concerns archiving which is done by private or public actors based on legal obligations. However, recitals are not binding and the CJEU has yet to discuss who can rely on the exceptions for archiving in the public interest.

Based on existing case law of the CJEU and other European courts, there are two main challenges stemming from EU data protection law for public archives. The first challenge concerns the definition of personal data and the difficulties in assessing whether a particular piece of information qualifies as such. This challenge is not unique to public archives – every entity handling data is confronted with it. In an archiving context, the assessment is made more difficult due to the amount and the age of the data in question.

The GDPR (and all other pieces of EU data protection law) define personal data broadly. It refers to any information relating to a natural person, with

[29] See in detail Frits W. Hondius, *Emerging data protection in Europe* (North Holland Publishing Company 1975); Gloria González Fuster, *The Emergence of Personal Data Protection as a Fundamental Right of the EU* (Springer 2014).
[30] Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data, ETS No. 108, 28 January 1980; Organisation for Economic Cooperation and Development, 'Annex to the recommendation of the Council of 23 September 1980: Guidelines governing the protection of privacy and transborder flows of personal data' (23 September 1980).
[31] Starting with Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data, OJ 1995 L 281/31 (no longer in force).
[32] See also GDPR (n 9) rec. 158 'Where personal data are processed for archiving purposes, this Regulation should also apply to that processing'.
[33] See for a full overview Laura Drechsler and Charlotte Somers, 'Article 89. Safeguards and derogations relating to processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes' in Kuner, Bygrave and Docksey (eds.), The General Data Protection Regulation: A commentary (OUP 2026, forthcoming). A pre-print is available at https://www.law.kuleuven.be/citip/en/news/item/archived_news/compilation-from-the-eu-general-data-protection-regulation-a-commentary-second-edition.

which said person is identifiable.[34] Identifiability requires a test based on the means an entity or natural person holding the data is reasonably likely to use to achieve identification.[35] This test does not only consider information that that entity or natural person has already at their disposal but also information they could access and combine with what they have.[36] Current case law of the CJEU also confirms that the assessment is heavily context-dependent and even fluctuates depending on the provision or principle of the GDPR that is being investigated.[37]

There are only two limits to the GDPR's broad conceptualization of personal data: anonymous data (understood as data where it is impossible to identify a natural person with) and data on deceased persons.[38] The latter, however, only excludes the data from being personal for the person that is deceased, it leaves open the possibility that this data is personal for their family members.
For public archives, the classification of data as personal is made more complex by the fact that a lot of the data potentially qualifying as personal is of a certain age, not always easy to determine exactly, which makes it possible that data relates to a deceased person. Yet, for the exclusion to work, public archives would have to be sure that this is indeed the case. Moreover, given that such data could still relate to a living relative, the scope of the GDPR might still be opened. While the issue of post-mortem privacy and the application of the GDPR to the deceased has been considered in some of the academic literature,[39] a thorough understanding from the perspective of the archives is missing and remains to be addressed by future research.

With increasing digitization of archival records, public archives are also playing an important role in the identifiability assessment for potential personal data. If the assessment is done from their own perspective, the amount of data a public archive holds can make it more likely that data can be linked to an individual and thus qualifies as personal data. In particular,

[34] GDPR (n 9) Art. 4(1).
[35] Ibid., rec. 26.
[36] See Article 29 Working Party, 'Opinion 4/2007 on the concept of personal data' (WP 136, 20 June 2007).
[37] See for example Case C-413/23 P, *European Data Protection Supervisor v Single Resolution Board*, judgment of 4 September 2025 (ECLI:EU:C:2025:645).
[38] GDPR (n 9) rec. 26 and 27.
[39] See for example Edina Harbinja, *Digital Death, Digital Assets and Post-Mortem Privacy* (Edinburgh University Press, e-book, 2023); David Erdos, 'Dead ringers? Legal persons and the deceased in European Data Protection law', (40) Computer Law and Security Review (2021) 105495.

considering that identity of a natural person is not further defined in itself, the text of the definition notes that it is not necessarily civil identity but that it can be 'physical, physiological, genetic, mental, cultural or social identity' instead.[40]

Additionally, public archives play a role in the identifiability assessment of other entities. Archival data is of high quality as it is verified and described with meta-data. This can make it an excellent means for cross-checking information and thus making it identifiable. The more accessible and 'younger' information in an archive is, the more likely it can be used for transforming information into personal data. This development is further amplified by the emergence of AI, which makes it even easier to connect different data points across the internet and thus transforms more data into personal data. More research is needed to understand the exact legal role of archives for making data personal.

It has to be remembered that the qualification of data as personal does not necessarily mean that such data cannot be used or made publicly available by public archives. Qualification as personal data just leads to the application of the GDPR, if the other conditions of the material scope are met and there is no exception. The principles of the GDPR, originally derived from information management practices, offer a lot of leeway for individuation in particular contexts. Future research must explore how to optimize this flexibility for the task of archiving.

The second challenge with the GDPR for public archives is the so-called 'right to be forgotten' and its implications for archiving. The right to be forgotten is a sub-title given by the GDPR to the right to erasure.[41] Both names are misleading.[42] EU data protection law, including the GDPR, only grant individuals the right to erase personal data, when such data was not meant to be stored or used in the first place.[43] In other words, when the principles of the GDPR for using personal data are not complied with, individuals are granted a right to delete what has no justification to exist in the first place. This right does not allow individuals to delete data they themselves no longer deem relevant, except arguably in the context of data

[40] GDPR (n 9) Art. 4(1).
[41] Ibid., Art. 17.
[42] See further Jef Ausloos, The right to erasure in EU data protection law: From individual rights to effective protection (Oxford University Press 2020).
[43] GDPR (n 9) Art. 17(1).

collected by information society services, in particular social media companies, when the individual in question was a child.[44] The right to erasure also has many exceptions, including situations where processing of personal data is necessary for archiving in the public interest.[45]

The right to erasure has to be distinguished from the right to delist, which is a right developed by the CJEU in the context of search engines.[46] The right to delist allows individuals to ask search engines to take down certain results that appear when their own name is being searched, provided that they are outdated and no longer relevant.[47] Search engines have to balance delisting requests with the value of the information for the general public based on criteria developed by the European Court of Human Rights (ECtHR) when balancing the right to privacy with freedom of the press.[48]

The CJEU has yet to decide a case of erasure in the context of archiving, though a reference from Belgium concerning erasure from the baptism registry is pending.[49] The ECtHR has issued a number of judgments on the right to erasure in the context of newspaper archives.[50] The most prominent judgment *Hurbain v Belgium* of 2023 confirms a nuanced approach taken by the ECtHR.[51] In this case, the applicant tried to erase information from a newspaper archive relating to a traffic accident he was involved in. The archive was directly connected to the search engines and easily accessible without an account to get specific results, so name-based searches would already return the information. The ECtHR found that, while the newspaper

[44] Ibid., Art. 17(1)(f).
[45] Ibid., Art. 17(3)(d).
[46] See in particular Case C-131/12, *Google Spain SL and Google Inc. v Agencia Española de Protección de Datos (AEPD) and Mario Costeja González*, judgment of 13 May 2014 (Grand Chamber) (ECLI:EU:C:2014:317).
[47] See further European Data Protection Board, 'Guidelines 5/2019 on the criteria of the Right to be Forgotten in the search engines cases under the GDPR (part 1)' (Version 2.0, 7 July 2020).
[48] As developed by the CJEU inter alia in Case C-131/12, *Google Spain*; Case C-507/17, *Google LLC v Commission nationale de l'informatique et des libertés (CNIL)*, judgment of 24 September 2019 (Grand Chamber) (ECLI:EU:C:2019:772); Case C-136/17, *GC, AF, BH, ED v Commission nationale de l'informatique et des libertés (CNIL),* judgment of 24 September 2019 (ECLI:EU:C:2019:773).
[49] Case C-12/25, *Bisdom Gent v GBA* (pending).
[50] See for example ECtHR, *M.L. and W.W v Germany*, Appl. Nos. 60798/10 and 65599/10, judgment of 28 June 2018; ECtHR, *Biancardi v Italy*, Appl. No. 77419/16, judgment of 25 November 2021.
[51] ECtHR, *Hurbain v Belgium*, Appl. No. 57292/16, judgment of 4 July 2023 (Grand Chamber).

could continue archiving the information in question, in this case it was justified , to at least anonymize the article in the publicly available part of the archive considering the ease with which the information could be accessed.

The *Hurbain* judgment clarifies an important legal distinction when it comes to the activities of public archives, namely between the processing of personal data for archiving as such (preserving material for eventual publication) and the publication of material. Erasure does not touch the archiving as such (also due to the exception), although some form of it can be applied to the publication process depending on the details of such a publication. Future research should further investigate the factors which determine the possibilities for individuals to temporarily suppress access to their own information and how this plays out if the archiving is not done by a newspaper but by the public archive of a country.

4. Archives and the EU's digital rulebook

Since the EU's data strategy of 2020, there has been much legislative activity focused on data in a broad sense (personal and non-personal data) as already noted in the Introduction above. The objective of this activity is to create a European market for data, which in turn is meant to make data more accessible and easier to re-use for commercial purposes.[52] According to the European Commission, this is necessary to keep the EU competitive, in particular with regard to the development of AI, which is a very data-intensive activity. The European Commission thereby assumes that there is principally enough high quality data within the EU to make progress with AI. However, that data is insufficiently shared and the economic potential of such data missed.

Within the 'tsunami' of legislation since adopted (at least 8 major regulations since 2020), two sub-categories of the EU's digital rulebook emerge. On the one hand, legislation with data as the regulatory object, which is directly focused on better harnessing the economic potential of data. The Data Governance Act, the Data Act, and the European Health Data Space Regulation[53] are the prime examples of this type of data legislation. On the

[52] Data Strategy (n 4); Data Union Strategy (n 5).
[53] Regulation (EU) 2025/327 of the European Parliament and of the Council of 11 February 2025 on the European Health Data Space and amending Directive 2011/24/EU and Regulation (EU) 2024/2847, OJ L 2025/327.

other hand, legislation that wants to ensure a fair digital market, where data is regulated more as a side question. The Digital Services Act,[54] the Digital Markets Act,[55] and the AI Act are examples of this type. Neither of them focuses on data directly, but its provisions indirectly affect how data can be used by whom within the EU. All legislation noted is comparably young and most of the provisions of the EU's digital rulebook are completely new, unlike EU data protection law, which was based on principles and concepts pre-existing in the Member States and at international level,. This lack of history, combined with relatively sparse case law and guidance, makes it extremely difficult to understand how to apply its provisions in an archiving context.

Compliance with the EU's digital rulebook is further complicated by the fact that the different legislations are not well aligned amongst themselves (and in some cases within themselves)[56] and that they seem to pull the regulation of personal data in the opposite direction of what the GDPR is trying to do. The digital rulebook tries to encourage more data sharing, whereas the GDPR endorses a less-is-more approach expressed with the principle of data minimization, which requires to only collect and use the personal data truly necessary for the purpose the entity has set for the operation in question.[57] This purpose that organizations need to determine the moment they collect the data, also defines what the data can be used for and ensures that it cannot be just re-used for new ventures, including sharing of data with new partners.[58] In essence, the EU legislator is pulling in two directions at the same time.

For the moment, most legislation of the EU's digital rulebook specifies that its rules are 'without prejudice' to EU data protection law, giving clear priority to the protection of personal data over the sharing of data as such.[59]

[54] Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act), OJ 2022 L 277/1.
[55] Regulation (EU) 2022/1925 of the European Parliament and of the Council of 14 September 2022 on contestable and fair markets in the digital sector and amending Directives (EU) 2019/1937 and (EU) 2020/1828 (Digital Markets Act), OJ 2022 L 265/1.
[56] See illustrated for the concept of fairness between and within the DGA and the DA in Laura Drechsler and Charlotte Ducuing, 'The notion of fairness in the Data Governance Act and the Data Act', in Van Cleynenbreugel and Grozdanovski, The Standards of Fairness in Digital Law (Edward Elgar Publishing, 2025), pp. 91-128.
[57] GDPR (n 9) Art. 5(1)(c).
[58] Ibid., Art. 5(1)(b).
[59] See for example DGA (n 7) Art. 1(3).

Nevertheless, applying this priority depends on correctly classifying data as personal, which, as described in the previous section, comes with its own challenges. Future research needs to further untangle the interplay of the EU's digital rulebook with EU data protection law in an archiving context, as public archives find themselves in a situation where some laws tell them to make all data freely available for commercial re-use, whereas the GDPR mandates caution when publishing personal data.

The EU's digital rulebook is also marked by a high degree of volatility. Despite its short terms of existence and even shorter term of application, the European Commission has already proposed a radical reform. In November 2025, it proposed to merge the Open Data Directive and the Implementing Regulation on High Value Datasets, with the Data Governance Act and the Data Act into one comprehensive Data Act.[60] The implications of this merge on public archives are not yet clear. Nevertheless, the proposal definitely confirms that commercial re-use of data is becoming the core objective of the European Commission's data strategy, pushing aside protection such as the one provided with the GDPR (for which the proposal also includes amendments).

5. Archives and artificial intelligence

The EU's data strategies of 2020 and 2025 are clearly motivated by one key driver, namely the development of AI, as it is feared that the EU is falling behind in terms of innovation due to its complex regulatory framework surrounding data.[61] Data is the fuel of AI, which needs particularly high quality data to be properly developed and trained. From a legal perspective, this development and training of AI raises a number of legal questions during each stage of the AI lifecycle. While the EU adopted the AI Act to regulate the use of AI systems (which are trained AI models), the AI lifecycle is not only regulated therein. Especially for training and development of AI, both EU data protection legislation and the EU's digital rulebook are of relevance and have to be navigated.

[60] Proposal for a Regulation of the European Parliament and of the Council amending Regulations (EU) 2016/1679, (EU) 2018/1724, (EU) 2018/1725, (EU) 2023/2854, (EU) 2024/1689 and Directives 2002/58/EC, (EU) 2022/2555 and (EU) 2022/2557 as regards the simplification of the digital legislative framework, and repealing Regulations (EU) 2018/1807, (EU) 2019/1150, (EU) 2022/868, and Directive (EU) 2019/1024 (Digital Omnibus), COM(2025) 837 final, 19 November 2025.
[61] Data Union Strategy (n 5).

From existing literature, three legal challenges emerge for public archives. The first challenge concerns the use of archival data for training of AI systems either within the archives themselves or by external developers for their own purposes. The EU legislation mapped above has various, at times seemingly contradicting requirements for the use of data for AI development, which would have to be complied with. Re-using data for AI development by external actors, often located outside of the EU, also raises ethical and policy questions. Archival data is of high quality in order to ultimately inform a society's understanding of history and hence of great interest for AI development. Yet, the investment in quality was not necessarily motivated by feeding further Big Tech companies. For the moment, the EU legal regime does not distinguish between the types of re-use when obliging archives to make data available for it.[62]

The second challenge is linked to the use of AI systems by public archives and archivists themselves. The AI Act regulates the use of AI systems based on the level of risk a concrete use poses. It thereby defines a number of uses that are prohibited, and a number of uses that are considered high risk.[63] It is only to the latter that most of the obligations of the AI Act apply to. For the moment, none of the prohibited or high-risk uses directly concern archiving, although the specific rules for all-purpose AI systems, e.g., Large-Language-Models (LLMs) might be of relevance.[64] The classification as prohibited or high risk might also change as the European Commission plans to regularly update both categories depending on technological development. The potential application of the AI Act to different stages in the archival lifecycle of material remains unexplored in current literature and case law. Future research should, in particular, classify the risks such use of AI brings to the fundamental rights and freedoms of individuals, as they can make a use prohibited or high risk.[65]

The final challenge concerns the archiving of the use of AI systems by public authorities. The AI Act comes with extensive transparency and documentation requirements for systems that qualify as high risk, including logging obligations for certain actions of the AI system itself.[66] Archiving such

[62] See for example Open Data Directive (n 6).
[63] AI Act (n 8) Arts. 5 and 6.
[64] Ibid., Chapter V.
[65] Ibid., rec. 1.
[66] Ibid., Arts. 11, 12 and 13.

documentation for future generations might be crucial to understand how states operationalized AI and how individual decisions were made. Even for AI systems currently not covered by these obligations of the AI Act, or those that would no longer be covered if the proposal of the European Commission to alleviate some of the transparency obligations of the AI Act is followed,[67] preserving some information on their working might be highly relevant from a societal perspective. National archival laws might have to further specify how to document AI systems for eventual publication via the archives.

6. Concluding remarks

Public archives are no longer exclusively the domain of national legislation. The present contribution highlighted that EU data legislation consisting of EU data protection law and the EU's digital rulebook is highly relevant for understanding how public archives can and must handle (personal) data. For the moment, each of these areas of EU law comes with a number of legal challenges that remain unresolved for the archiving context and require future legal research as further outlined above. The emergence of AI makes such research and the resolution of the underlying questions even more urgent, due to the fact that AI heavily depends on data.

[67] Proposal for a Regulation of the European Parliament and of the Council amending Regulations (EU) 2024/1689 and (EU) 2018/1139 as regards the simplification of the implementation of harmonized rules on artificial intelligence (Digital Omnibus on AI), COM (2025) 836 final, 19 November 2025.

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]

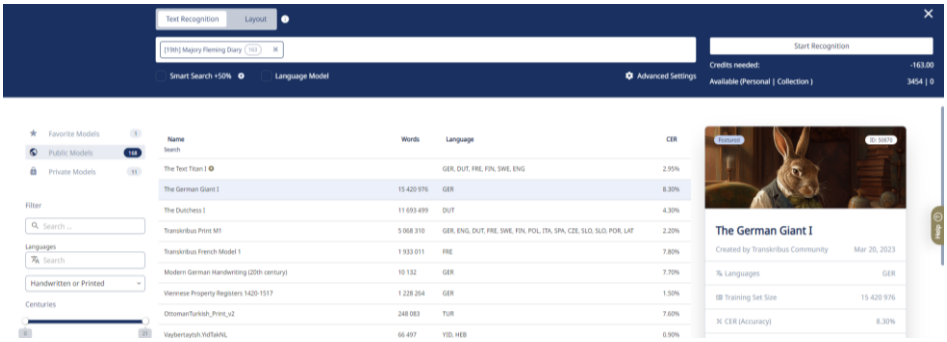

Figure 4 Public models, available to all users of Transkribus.

Advanced Workflow

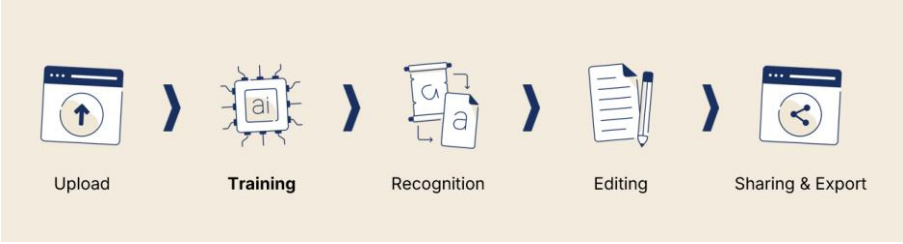

Figure 5 Advanced Workflow.

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]

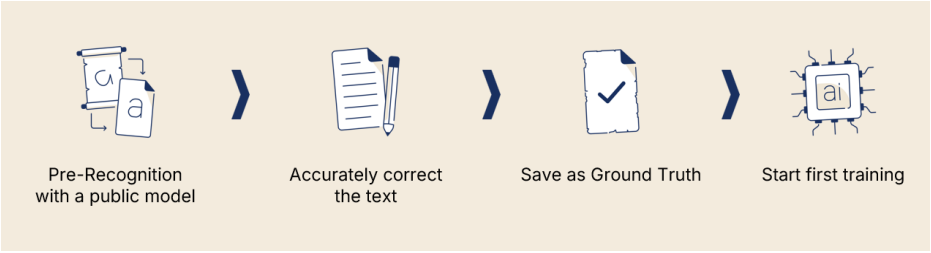

Figure 6 The training workflow, itself.

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.

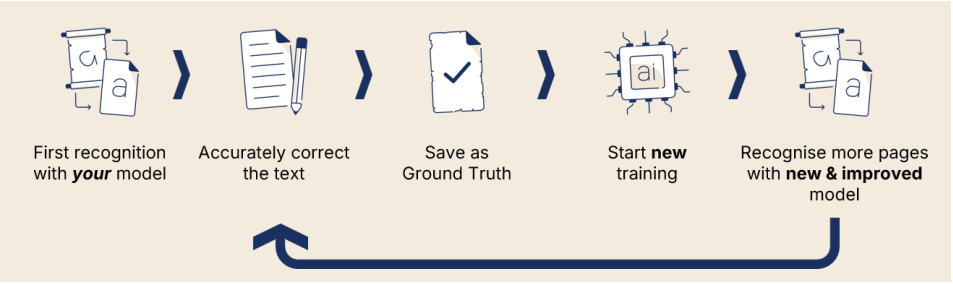

Figure 7 Transkribus includes functions to refine and improve texts.

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.

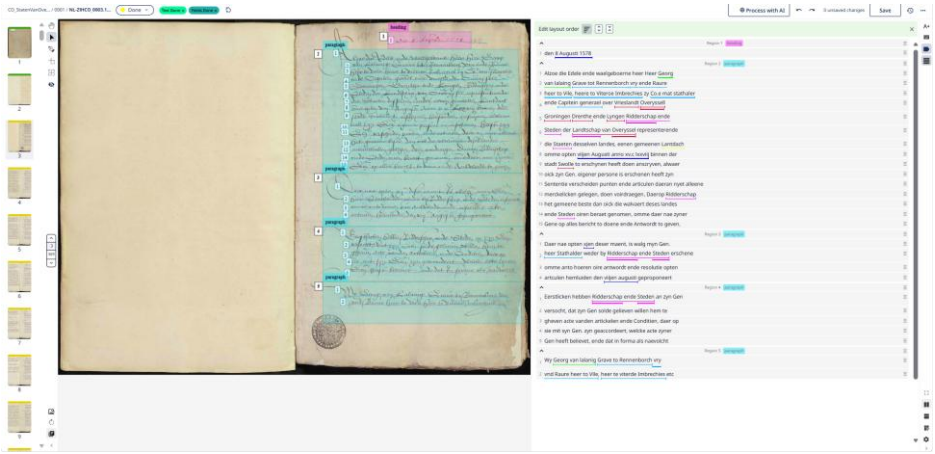

Figure 8 Lay-Out Analysis (Structure) and Tags are visible in this image and the transcription on the right. Source: Collectie Overijssel Resolutions of the Provincial States (0003.1 0001 page 3).

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).

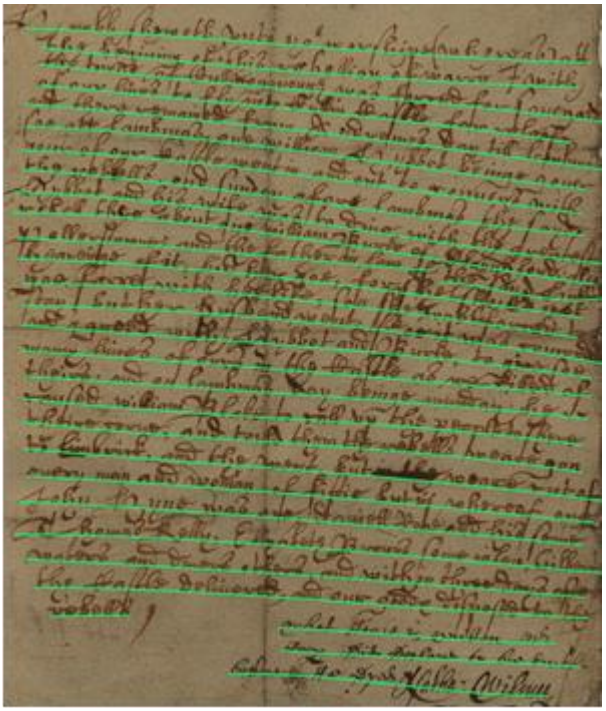

Figure 9 Baselines. Example with screwed baselines.

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]

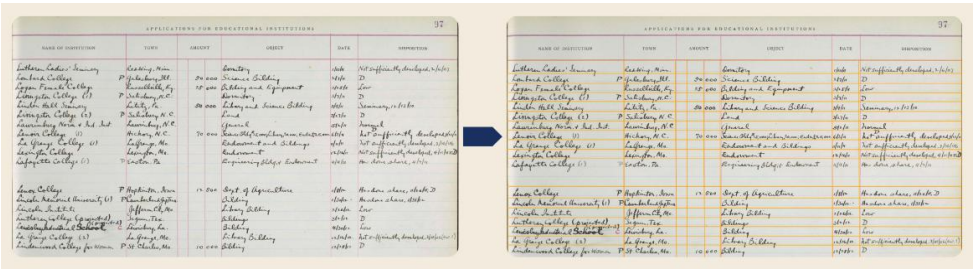

Figure 10 Tables with Table Recognition.

Field models

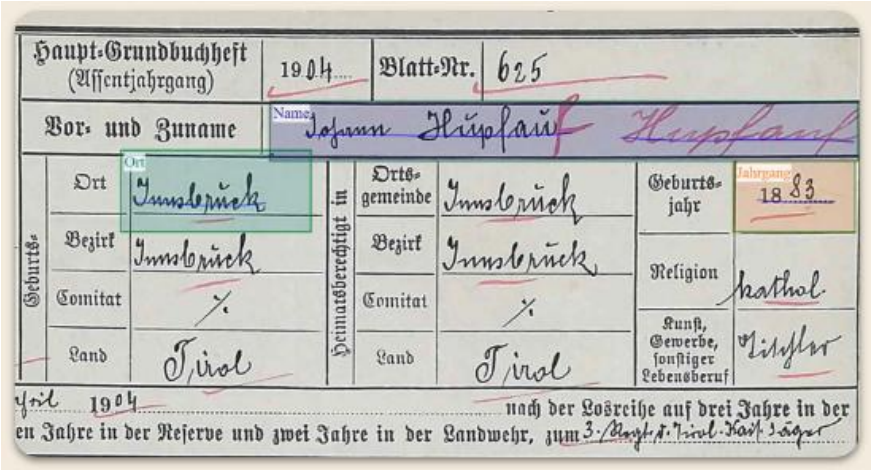

Figure 11 Field-model example

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).

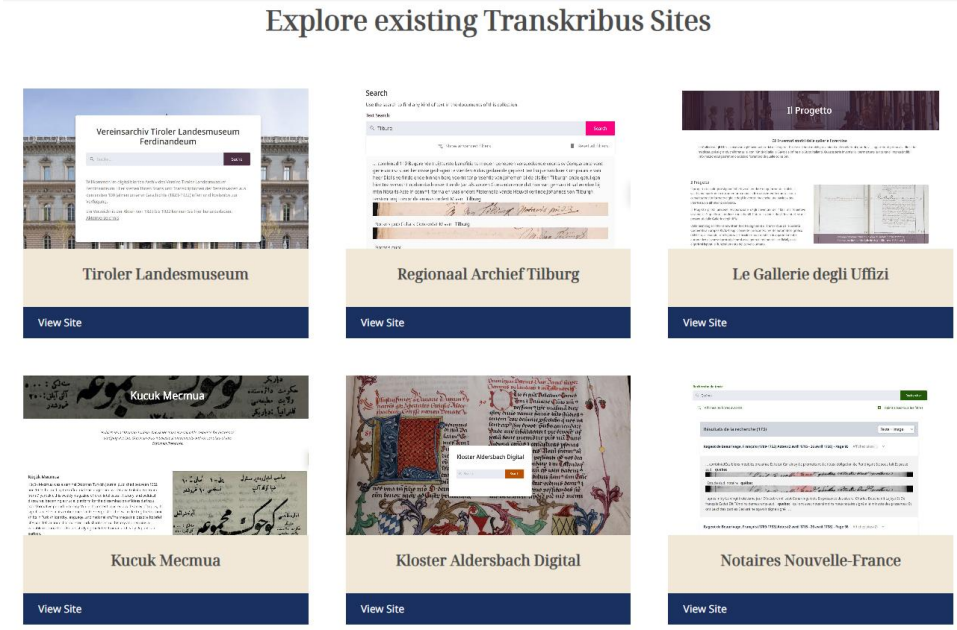

Figure 12 Transkribus Sites with several publicly available websites: https://www.transkribus.org/sites.

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?").

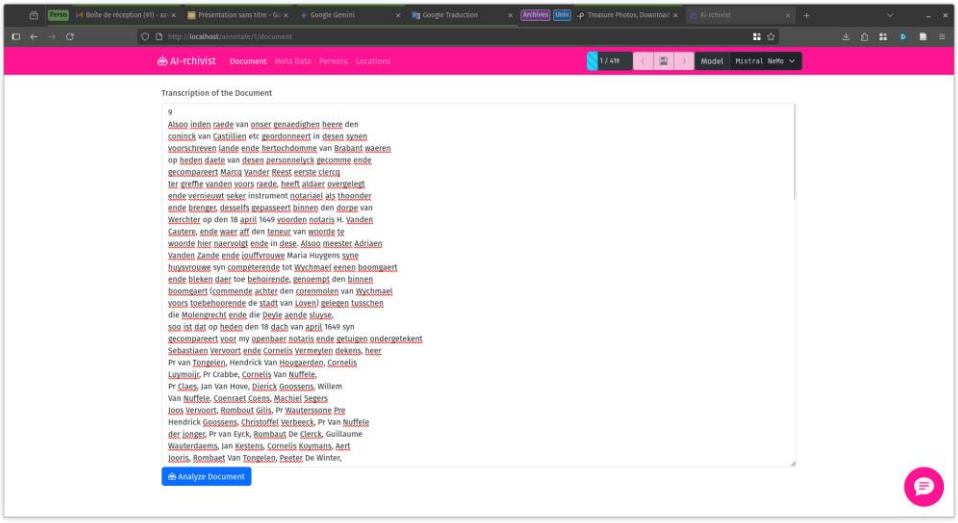

Figure 13: A legal document dated April 18, 1649 (formalized in 1651) ended a dispute over the use of the banks of an orchard in Wychmael. The owners, Adriaen Vanden Zande and Maria Huygens, had been harassed by boatmen from the Deyle River, who dragged their boats onto their land. (Source of the transcription: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

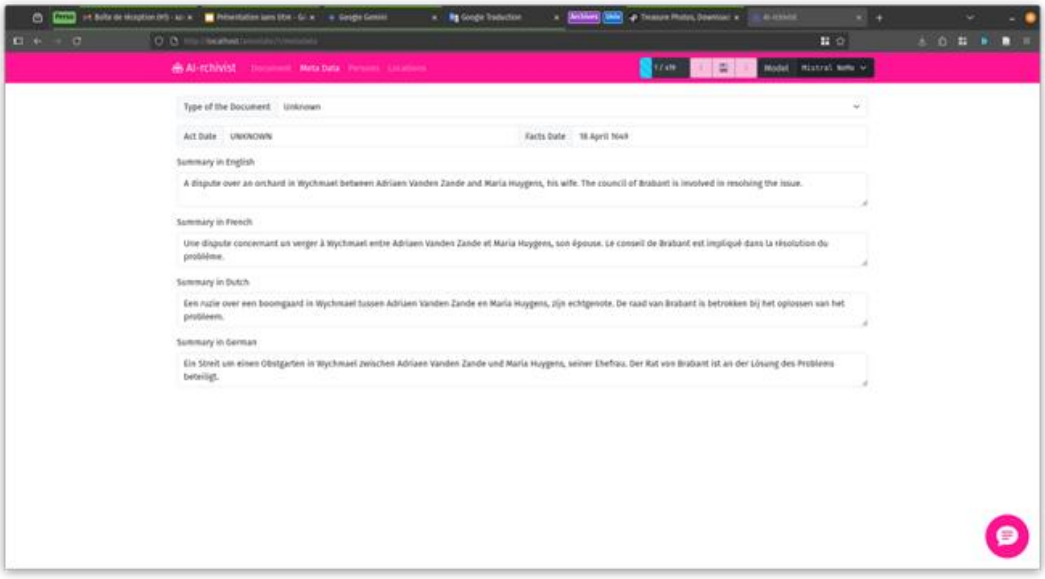

Figure 15: Key meta-data (dates, and multi-lingual summaries), extracted from the example transcribed document in figure 14.

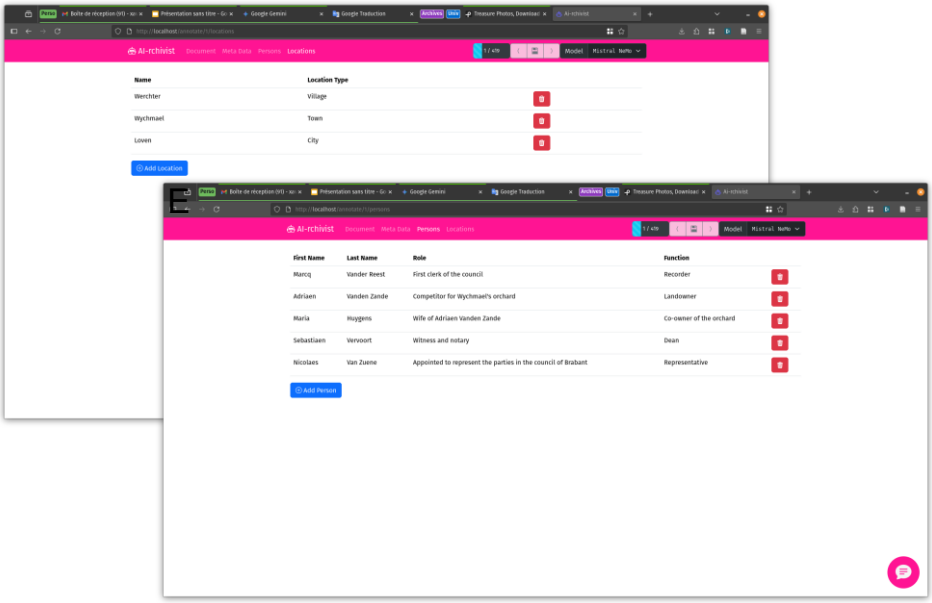

Figure 14: Key meta-data (dates, and multi-lingual summaries), extracted from the example transcribed document.

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

The legacy of a Nazi photographer in Gau Moselland - workshop report on AI supported indexing in collaboration with the Fraunhofer IAO[1]

Daniel Heimes[2]

Herbert Ahrens was a photographer integrated into the structures of the Nazi regime. He was most likely employed by the *Nationalblatt* in Koblenz or had a freelance contractual relationship with it. The *Nationalblatt* was the propaganda newspaper of the NSDAP in the *Regierungsbezirk Koblenz* (administrative district of Koblenz). Ahrens produced around 66,000 photographs between 1933 and 1945. Most of these were taken in the Gau Moselland. Before the annexation of Luxembourg the designation was Gau Koblenz-Trier. The photographer's legacy from the Nazi era also includes images from other parts of the German Reich as well as the Western Front and the Eastern Front. The result is a photographic legacy of unique significance.

Years of negotiations by various organisations to acquire this legacy failed again and again. The Rhineland-Palatinate State Archives Administration seized the fortunate opportunity to purchase this legacy in 2022. It received financial support from the *Kulturstiftung der Länder* (Cultural Foundation of the Federal States). Not only was ownership acquired, but also the rights of use and exploitation arising from the copyright. This is far more important from an archive's point of view.

The largest part of the legacy consists of negatives. The indexing will have to show whether a relevant part of the prints on paper are not available as negatives. The negatives proved to be a particular challenge. They are nitrate films, also known as cellulose films. In Germany, these are subject to the *Gesetz über explosionsgefährliche Stoffe*, *das Sprengstoffgesetz* (law on explosive substances, the Explosives Act. The nitrofilms could not be stored in the *Landeshauptarchiv Koblenz* (State Archives Koblenz).

[1] The translation from the German original was done with DeepL. The post-editing was done by humans.

[2] Daniel Heimes is archivist at the Landesarchivverwaltung Rheinland-Pfalz / Landeshauptarchiv Koblenz (Rhineland-Palatinate State Archives Administration / State Archives Koblenz), Head of the Archival Services and official data protection officer.

The *Bundesarchiv* (Federal Archives) provided support with its special storage centre in Berlin. The nitrate films were digitised by a specialist company. When this was achieved in May 2024, the enormous challenge was to make these digitised files accessible. As part of a training course on artificial intelligence in September 2024, the *Landesarchivverwaltung Rheinland-Pfalz* (Rhineland-Palatinate State Archives Administration) was suddenly presented with unprecedented opportunities for rapid indexing. This would have taken many staff years under traditional circumstances. The networking made possible by this training led to contact with the Fraunhofer IAO[3]. It has particular expertise in supporting the introduction of AI into work processes. A contract was signed between the Fraunhofer IAO and the Rhineland-Palatinate State Archives Administration in 2024. The content is to support the AI-based indexing of the Herbert Ahrens photographic legacy and, as a sustainable product, the simultaneous development and creation of a tool that can make such photographic legacies and collections AI-supported accessible in the future.

The project is divided into various phases, which are described below. Because this project is developing something new and the two project partners come from different, in some cases completely unfamiliar areas, it was clear from the beginning that it would have to be a project with evaluation cycles. Mistakes and wrong turns must also be justified to find the better way.

First of all, a contract for order processing in accordance with Art. 28 General Data Protection Regulation had to be concluded so that the photographs, some of which are subject to legal restrictions, for example personal retention periods from the *Landesarchivgesetz Rheinland-Pfalz* (Rhineland-Palatinate State Archives Act), could be processed by the Fraunhofer IAO. The entire files were then transferred to a protected area at the Fraunhofer IAO.

At the same time, common requirements for the result of the indexing were developed. The Rhineland-Palatinate State Archives Administration works with indexing guidelines, which also contain a special section on the indexing of photographs. Their specifications became part of the joint requirements.

The Fraunhofer IAO also researched suitable multimodal language models. Initial tests were also carried out with various providers. The basic principle for all applications was that they were carried out in a protected area from which no data was or is fed back into internet-based AI models. The image

[3] The official English designation is Fraunhofer Institute for Industrial Engineering. The official German Designation is Fraunhofer-Institut für Arbeitswirtschaft und Organisation.

indexing models used for the analysis mirror the entered data back to the internet. The assurances sometimes given by the operators of AI models for paid services that the data would not be used for training purposes and would not be mirrored back do not offer the necessary reliability for such data, some of which is very sensitive. For this reason, the Fraunhofer IAO has proposed working with a local version of Gemma 3, among other things, with all the limitations that this entails - both positive and negative. At this stage, we assume that not all the requirements for indexing the legacy can be met by a single AI model, which is why other models will have to be used for different tasks, as will be shown later on. In the end, however, there should be an overall model that integrates these individual tasks.

The next step was to enable facial recognition by means of a special AI model, which first had to be selected from a range of different models, prepared with regard to the existing requirements and is still being prepared so that it becomes a sub-tool for the entire workflow of AI-supported indexing of image data. In simple terms, this works without any problems using the basic configuration of the AI models for well-known personalities of the Nazi state - Adolf Hitler, Rudolf Hess, Robert Ley, Hjalmar Schacht, etc. come to mind here. Persons in the second and third ranks, as well as unknown persons, had and still have to undergo so-called labelling. In principle, this is a manual process, which requires the intervention of archival colleagues. In the case of the Herbert Ahrens legacy, however, there are also documents that are part of the legacy. One of them is a list of photos compiled mainly with a typewriter. This list was only compiled by Ahrens (or complete or in parts perhaps his daughter – it is not sure) after 1945. There is also a list that is to be regarded as a primary source, which was created at the same time as the photographs. Therefore it reflects the unfiltered view of the Nazi photographer. This is Herbert Ahrens' photo diary, which is only available for the period from 1940 to 1945. In it, he himself recorded all photographs or photo series with motifs, often including personal names and locations, picture numbers and dates. In this context, it was an essential aid for labelling. This data should be able to be assigned to the individual images or image series using AI and consequently be included in the image indexing.

Another aspect of facial recognition is determining the age of the people depicted. These are to be used for the automated allocation of blocking periods, which - according to the objective - should be correct with a very high probability. At this stage of the project, it is not yet possible to say whether this will be achieved with the necessary high probability and therefore legally justifiable.

Further AI models will be used to obtain additional information. So-called CLIP models (Contrastive Language-Image Pretraining) are to be used to identify similar images, image motifs and image series. The benefit lies in the recognition of duplicates, similar locations and the linking of series that are close in time. In addition, the Fraunhofer IAO suggested the possibility of colourising the images, for which the necessary tools would also be available. However, this was rejected by the archive as an intervention going beyond a restoration measure. In the recent past, two colourised films about the so-called old Rhineland were shown in cinemas and also on a streaming provider. Due to their great success, the possibility of AI-supported colourisation is at least mentioned here with regard to public relations work.

As part of the evaluation cycles already mentioned, regular joint discussions are held on the results obtained, at which clear progress can be seen, but also open problems and challenges are discussed. A few translated examples are presented here to illustrate the development:

1. KAISER WILHELM MONUMENT AT THE DEUTSCHES ECK

LHA Ko archives group 700,391 No. 4987 Sub-No. 3 [4]

[4] The official german designation is LHA Ko Best. 700,391 Nr. 4987 UNr. 3.

Status 02.04.2025:

A detailed, black and white image of an equestrian statue on a high pedestal. The statue shows a person on a horse, both figures are equipped with wings. The pedestal is decorated with architectural details and extends high into the sky.

Status 30.04.2025:

Description: Deutsches Eck - Imperial Monument

Details: An equestrian statue is enthroned on an imposing column.

Criticism: The improvement of the indexing performance can be seen in the transfer of the title (here still - in deviation from the indexing guidelines - named as description) from information compiled by Herbert Ahrens. However, the AI model has still failed to adopt the dating ‘vor 1934’ (before 1934) as well. It can be assumed that this was due to the formulation deviating from standardised dates. The more precise indexing under Details is not correct in terms of content. Firstly, the definite article ‘the’ and not the indefinite article ‘an’ should be used. The Deutsches Eck does not include several equestrian statues. Similarly, the statue does not stand on a column, but on a pedestal. The perspective and the visitors to be seen are not mentioned.

2. German soldiers in front of the Arc de Triomphe, Paris 1940

LHA Ko archives group 700,391 No. 2431 Sub-No. 3[5]

[5] The official german designation is LHA Ko Best. 700,391 Nr. 2431 UNr. 3.

Status: 02.04.2025:
Soldiers in uniform stand in front of the Arc de Triomphe in Paris. In the foreground are several military vehicles.

Status: 30.04.2025:
Date: 28.07.1940
Description: Cityscape in Paris - Arc de Triomphe
Details: In front of the Arc de Triomphe are soldiers and vehicles.

Criticism: The Arc de Triomphe was already known at the first attempt due to the data known in advance to the AI model, the local version of Gemma 3, that is, without training by the project participants. The improvement in the indexing performance can be seen in the transfer of the recording date from information compiled by Herbert Ahrens. However, it is problematic that it changed the general series name ‘Städtebilder Paris’ (Cityscapes Paris) to ‘Städtebild in Paris’ (Cityscape in Paris), which represents a qualitative deterioration.

As part of the evaluation on 30th of April 2025, which included a total of 20 photographs with the respective indexings, it was discussed to provide the AI model with further training data to increase quality. To this end, it was firstly agreed to index a package of images from the Ahrens legacy exactly according to the specifications of our indexing guidelines. These images were to be widely distributed, even though it was of course not possible to provide a cross-section of all types of motifs found in the approximately 66,000 photographs. Within a short space of time, 80 images were catalogued as examples and the indexing data was sent to the Fraunhofer IAO for training purposes. Secondly, it was decided that the data from the archive database Dr.Doc to the actual image collection of the State Archives Koblenz, namely those for which digitised images and indexing data – both – are available, are exported and made available to the local AI model at the Fraunhofer IAO for training purposes. Thirdly, a digitised copy of the NSDAP handbook, which is now in the public domain, was also included for training purposes, as it contains, among other things, precise illustrations and classifications of the uniforms and insignia, which should make it easier for the AI model to classify them.

Repeated evaluation cycles are foreseeable for the future in order to achieve a result that fulfils archival requirements. Further possibilities will be offered by increasing computing power, which has already been addressed by the Fraunhofer IAO. The broadening and improvement of the training data also promises qualitative progress for the project.

As soon as the desired quality levels have been achieved from an archival point of view, the next task will be to create an overall processing workflow for all the images in the Ahrens legacy on Fraunhofer IAO servers. All additional information, such as the list that Herbert Ahrens created after 1945, his photo diary from 1940 to 1945, the exemplary indexing information, which means the information created using images from the legacy and the additional information imported from the archive database of the Rhineland-Palatinate State Archives Administration and obtained during the evaluations, etc., must be integrated into the selected overall AI model. In addition, this overall model must be able to incorporate other context data identified as relevant that originates from other AI models, such as the recognition of duplicates, image series, identical motifs and facial recognition. It must also be taken into account that this model must be able to function for other cases, that is other, future image collections, donations with extensive image material, pure photographic legacies, etc.

At the end of the project, Fraunhofer IOA will provide the Rhineland-Palatinate State Archives administration with knowledge and skills. This should enable the Rhineland-Palatinate State Archives Administration to carry out previously implemented processing procedures independently and within the framework of its own IT infrastructure. This can not only be done for archivists, but must also fully involve the archive's IT staff. The Fraunhofer IAO will provide a comprehensive and detailed explanation of the procedures for processing the Herbert Ahrens legacy as well as for future photo collections or partial photo collections. This will also include the implemented codes for image processing, including support through discussion and documentation as well as further information material for the Rhineland-Palatinate State Archives Administration.

6668

ISBN 978-94-6391-661-5

9789463916615

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

belspo