Miscellanea Archivistica Studia 225

International Archives Symposium
Internationales Archivsymposion in Namur (2025)

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

PROCEEDINGS/ANNALEN

Editing - Redaktion
Els Herrebout

International Archives Symposium / Internationales Archivsymposion
in Namur (2025)

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

—

Proceedings / Annalen

Editing – Redaktion
Els HERREBOUT

2026

ALGEMEEN RIJKSARCHIEF
EN
RIJKSARCHIEF IN DE PROVINCIËN

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

GENERALSTAATSARCHIV
UND
STAATSARCHIVE IN DER PROVINZ

MISCELLANEA ARCHIVISTICA
STUDIA

225

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

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

Offprint / Sonderdruck

JOERGENS, Bettina (2026). Impulsvortrag / Keynote speech. Open Data versus Black Box, or: How can AI Fulfill Archival Tasks and Professional Requirements? In Els HERREBOUT (Ed.), International Archives Symposium in Namur (2025). Open Data and AI. New chances for Archives? Proceedings (Miscellanea Archivistica Studia, 225, pp. 11–26). State Archives of Belgium.

DOI: https://doi.org/10.5281/zenodo.20058394
Handle: https://orfeo.belnet.be/handle/internal/14700

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.

Table of Contents/Inhaltsverzeichnis

ISBN 978-94-6391-661-5

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

belspo