UPPSALA
UNIVERSITET

Archives in the Digital Age

The use of AI and machine learning in the Swedish archival sector

Gijs Aangenendt

Department of ALM
Theses within Digital Humanities
Master’s thesis (two years), 30 credits, 2022, no. 17.

Author
Gijs Aangenendt

Title
Archives in the Digital Age. The use of AI and machine learning in the Swedish archival sector

Supervisor
Matti La Mela

Abstract

The arrival of AI and machine learning in the archiving world implies a reconceptualization of archival institutions from consisting of archives as records to be read towards archives as data to be mined. This thesis aims to survey and analyse the arrival of AI and machine learning within the Swedish archival sector.

Through the analyses of four archival institutions and semi-structured interviews with archive professionals, this study illuminates the current state of AI in the Swedish archival sector, opportunities and obstacles of AI implementations, and the impact of AI on the Swedish archival profession. Swedish archival institutions have interest in and expectations of AI implementation. Yet, the current level of AI implementation is generally on experimental instead of an operational level. The findings show that collaboration and knowledge exchange are key drivers to accelerate the implementation of AI in the archival sector. The study concludes with practical recommendations for archival institutions and archive professionals.

Key words
Archives, Archival profession, Sweden, Artificial Intelligence, Machine Learning.

Table of contents

Abbreviations

AI	Artificial Intelligence
CAS	Computational Archival Sciences
HTR	Handwritten Text Recognition
NAD	National archive database
NLP	Natural Language Processing
OCR	Optical Character Recognition
XAI	Explainable Artificial Intelligence

Introduction

The digital age presents archives continuously with new challenges. The arrival of new information and communication technologies radically alters the ways in which people create, access, and engage with information and archival records. (Goudarouli et al. 2019) As a result, archives are tasked to find new ways to collect and preserve the wide range of born-digital media formats and content types in which information about society and public life is nowadays stored and created. Besides the incoming born-digital material, archives have been digitizing archival collections on ever larger scales the past decades. Combined, the nature, diversity, and size of this digital archival material puts pressure on the traditional archival concepts, practices, and workflows that were originally developed for analogue paper records. (Moss, Thomas & Gollins 2018) In order to handle the influx of the digitized and born-digital records, scholars have argued that archives and archivists are to develop and adopt new practices, methodologies, and tools. (Esteva et al. 2013; Ranade 2016)

In recent years, the number of available computational tools and methods for archives to deal with the large quantities of digital material in archival collections has grown exponentially. (Colavizza et al. 2021) Taking advantage of the newest advancements in the fields of artificial intelligence (AI) and machine learning, these tools give archives the possibility to tackle the archival challenges in the digital age. The widespread availability of open-source software libraries for AI and machine learning applications and the lowered threshold to sufficient computational power have made the use of AI-based tools and methods a reality for archives. (Lee 2019) For textual documents for instance, developments in the field of computer vision for digitalization of records through optical character recognition (OCR) and/or handwritten character recognition (HTR), and the field of natural language processing (NLP) for text analysis, offer archives great opportunities. (Hutchinson 2020) It enables archives to process, analyze and present documents in new ways, uncovering unknown perspectives and links in and between documents. (Ranade 2016)

Introducing and applying AI and machine learning-based techniques in archival processes implies a transformation of the archive and the archival profession. It transforms the outlook or perspective on archives, from archives as collections of records to be read to archives as sources of data to be made sense of. (Moss, Thomas & Gollins 2018) The use of AI-technology within archives also poses methodological and ethical issues that archivists need to be wary of, such as bias in the datasets. (Mordell 2019) Archivists will need to develop new skills and acquire new knowledge in order to use the techniques, analyze the output, and collaborate successfully with computer and data scientists. (Marciano et al. 2018)

This thesis focuses on the introduction of computational tools and methods in the Swedish archival sector. The scope of thesis covers how archival institutions currently make use of AI-technology and identify the opportunities and obstacles for (further) implementation in the future. Additionally, it dives into the question how the implementation of AI changes the archival profession and workflows. The topic is approached from the perspective of archivists at Swedish archival institutions, whose experience with and outlook on AI are uncovered through interviews. In the end, the study hopes to be a useful resource for any archival institution or archivist interested in exploring the potential and limits of AI technology.

I encountered the topic of AI and archives for the first time during my internship at the City Archive in Amsterdam. The main objective of the internship was to find out whether Topic Modeling could be an appropriate tool for archivists to structure and organize born-digital archives. The City Archive was keen on finding an alternative to the manual processing of born-digital archives of private actors. Opening each file manually to discover its content was considered unfeasible and inefficient, due to the number of files and lack of informative filenames within these collections. Diving into Topic Modeling meant that I had to explore machine learning and NLP concepts, use programming languages, and study the implications of approaching archival records as sources of data. It became clear that using computational methods like Topic Modeling would bring about a transformation of archives and the archival profession. The thematic structuring of born-digital archives could for example results in alternative ways for users to access archival material and discover connections between individual records. The internship sparked my interest to explore the transformation of archives in the digital age further in my thesis, with the archival sector in Sweden being an attractive focus of study.

Research field & theoretical concepts

In this section, the topic of AI and archives is centered in its larger research context. First, research on the transformations of archives in the digital age is introduced since the arrival of AI within the archival sector has strong connections with the digitalization of society and archival institutions. After that, the concept of AI and interrelated fields and terms are defined. Then, research on the use of AI in the broader cultural heritage sector is briefly introduced, in order to be able to situate the research discussed on the development and implementation of AI specifically within the archival sector. Lastly, the archive-as-data paradigm is put forward as a theoretical lens to investigate the impact of the arrival of AI in the archival sector and the archival profession.

Transforming archives in the digital age

The nature of archival institutions has transformed significantly in past decades, largely due to the digitalization and the growing use and capabilities of technologies within the archival sector and society at large. The transformation of archives in the digital age has led to debates on the differences between analogue versus digital records, the future of traditional archival concepts, and the responsibilities and tasks of the archive and its archivists. As the transformations of archives in the digital age form the backdrop of the recent arrival and use of AI within the archival sector, these debates will be briefly discussed here.

Meehan (2014) describes how the increasing use of computer technology for information seeking and information use has greatly impacted the expectations and demands of archival users on how information within archival collections can be accessed. Access is increasingly interpreted as online access to digitized archival records and collections. A second change discussed by Meehan (2014) relates to where the archival user finds and accesses archival information. This is increasingly done on a network level, through for example Google, instead of on the local/institutional level. These two examples illustrate the new challenges archival institutions have had to face in the digital age, namely how to accurately represent the content, context, and structure of digitized archival material online as well as the technical requirements of structured metadata that can be exchanged across systems.

Besides making records available online, research points out that the character of digitalized and born-digital records lends itself to providing access on another, more granular level, that of the individual record. Ranade (2016) describes how digital access is currently designed based on how physical archives are explored, with visitors coming to the reading room of the archive to study one record at a time. According to Ranade (2016), this view constrains the potential that digital collections offer in terms of access and research. The fluid nature of digital records offers an opportunity for collections to be arranged and explored in multiple flexible ways, based on the content (themes), relationships (links), or content form.

The impact of the digital age is also discussed in relation to the transforming content of archival collections. Where analogue archives tend to consist of mainly paper-based material with the occasional photograph, digital archival holdings can consist of a wide variety of materials. The latter can include digital photographs, video and audio recordings, e-mails, but depending on its origin and purpose, also websites, web pages, blog posts, and social media messages. (Miller 2017) The influx of the digitized and born-digital material puts pressure on traditional archival concepts developed for physical textual collections. Concepts related to provenance, hierarchy, and original order do not translate easily to the digital material. Thibodeau (2016, p. 3277), for example, argues that these traditional concepts form an “invisible wall”, that constrains the ability of archives to adequately handle the diverse types of digital information and profit from emerging technological opportunities. In a similar vein, Moss, Thomas & Gollins (2018) argue that archivists need to alter the labor-intensive manual methods and processes in order to handle the quantity and diversity of newly incoming digital materials. They consider it unsustainable to maintain the analogue archival practices for the digital, for example the idea that for each record a detailed description has to be manually created.

As this thesis is situated in the Swedish archival context, attention also has to be given to the transforming Swedish archive in the digital age. In 2019, the report *From here to eternity* was published, presenting the findings of a committee appointed by the Ministry of Culture to investigate the archival sector in Sweden. The investigation focused on how society’s access to public records could be ensured now and in the future. The report gives insight in how the Swedish archival sector has changed in the digital age. It describes how technical developments have impacted the Swedish archives in terms of diversity and size of archival collections, with incoming media types requiring alternatives archival approaches and practices. Furthermore, the report highlights how the increase of the number of electronic records is changing the division between archival creator and archiving institution. A more proactive and continued management of digital records is needed, as archiving concerns for long-term preservation need to be taken into account from the moment the digital record or information is created. (SOU:2019)

In a response to the ongoing investigation, the Swedish National Archives (2018) published a document with their take on the state and future of the Swedish archival sector. In regard to the changing archive in the digital age, the increasing expectations and demands of different archive users for online access observed by Meehan (2014) are also experienced by the Swedish National Archives. The archive notices a growing interest in the use of digital data among companies, governmental agencies, private persons, and researchers. These actors expect to be able to access this data through the archive. This means that the demands of digitalization increases, and archival records are expected to be enriched with metadata and made machine readable through HTR and OCR. Another observation of the Swedish National Archives (2018) relates to the representativity of archival collections. As a larger part of human life is taking place online, the ability of archives to document and preserve all facets of Swedish society is threatened. Nowadays, people tend to organize themselves in online communities instead of formal organizations such as labor unions or local political parties. Making these movements accessible for future generations requires the adaption of new methodologies and technologies.

Hansen & Sundqvist (2016) have studied the consequences of the digital age in relation to Swedish private archives, with a specific focus on popular movement archives. The authors describe how private archives have been important agents in the development of the Swedish archival sector as they contribute to a more pluralistic archival memory. Their collections document spheres of society and parts of everyday life that are not represented in the public records of public archives. The private archives are facing similar challenges due to the digital transformation of society. In their article, Hansen & Sundqvist (2016) highlight the technological challenges private archives face. Many private archival institutions, for instance, lack the economic resources, technological infrastructure, and required technical expertise & know-how to receive and preserve digital records. Additionally, private archives do not have the same ability as public archives to engage in closer collaboration with the archive creator, as they "lack a legal framework to reinforce a standardized way of creating records using recommended formats for long-term preservation". (Hansen & Sundqvist 2016, p. 133)

In order to meet the challenges of the digital age described above, scholars, such as Esteva et al. (2013), have called for the adaptation of new practices, methodologies and computational tools within archival institutions. Colavizza et al. (2021, p. 1) state that "human archivists need the support of machine agents" to ensure that the evaluation of digital records meets the requirements of quality and trust. Both the committee (SOU:2019) and the Swedish National Archives (2018) subscribe to this potential of AI for handling the challenges presented to archives in the digital age. The technological innovations in the field of AI can assist archives in metadata enrichment, improving the quality of transcriptions, and automating certain archival tasks, such as the appraisal (*värderingsarbete*) and disposal (*gallring*) of records and information. In order to grasp the potential of AI for (Swedish) archival institutions and study implementations of AI, a starting definition of the term and related key concepts have to be presented.

Defining artificial intelligence

Defining the term artificial intelligence is a rather complex task. The term has countless meanings and interpretations and a clear universally accepted definition is missing. Its definitions also shift over time, with a newly introduced technology starting out as being considered as AI until it becomes commonplace and the next exciting technological innovation emerges.

The first experiments with artificial intelligence are often traced back to the *Dartmouth Summer Research Project on Artificial Intelligence* in 1956. At the start, the aim was to design machines that were able to reason and think at the same level as humans. Presently, the developments described as AI do not strive after this ideal of developing extremely intelligent machines that behave just like humans. Instead, the efforts are mainly focused on the analysis of large volumes of data, recognizing patterns, making predictions, and taking decisions. (Jordan 2019) The level of intelligence of contemporary machines or technologies can be measured and compared on a multidimensional spectrum. Scale, speed, autonomy, and generality are factors that are often used when differentiating instances of artificial intelligence. (Stanford University 2016)

Another way of defining AI is as a field of research and innovation, consisting of distinctive but interrelated disciplines, methods, and application areas. The areas include machine learning, deep learning, reinforcement learning, robotics, computer vision, and natural language processing. The interrelations and dependencies between these areas can be illustrated by the example of robotics. Robotics is closely related to natural language processing, as robots need to be able communicate and interact with people using human language. In the archival world, a combination of NLP techniques and machine learning is used to perform analysis of textual data and extract (meta)data, entities, or themes from archival records. (Hutchinson 2020) AI is an interdisciplinary field in which different disciplines come together, from computer science, statistics, data science, and engineering to the social sciences and humanities. (Jordan 2019)

As with many other parts of society and public life, the term AI has also made its way into the archival and recordkeeping sector. Rolan et al. (2019, p. 181) take a broad perspective and understand AI as "digital systems that automate or assist in activities that we associate with human thinking", for example decision-making, problem solving, and learning. From the recordkeeping and archival perspective, they highlight three AI systems that are used in archives and have varying levels of automation. These are rule-based systems, statistical models, and deep learning models. Colavizza et al. (2021) use artificial intelligence as a proxy for machine learning, as this field and application of AI is most closely related to recordkeeping and archives. The term is used instead of machine learning to capture the broad applications of AI to recordkeeping and not necessarily only the technical systems. The authors are primarily interested in the professional, cultural and societal consequences of AI for archival processes and archivists.

Most of the current implementations of AI within the archival sector relate to machine learning, a subfield of AI that designs algorithms to process large amount of data and make reliable predictions and decisions based on the output. Jaillant

(2022) writes that AI implementations in the archival sector are designed to perform low-level tasks. Identifying sensitive information, transcribing handwritten texts, or classifying images does not require the most advanced AI-technology. It can still lead to breakthroughs within the archival sector nonetheless:

> The value of AI is not its ability to perform complex high-level tasks that require contextualization, theorization or creativity. Instead, the value of AI comes from its capacity to process huge amounts of data very rapidly – something that no human can do single-handedly. (Jaillant 2022, p. 14)

Below, the key concepts and terms related to field of AI used in this thesis are introduced. As this thesis is interested in the understanding of the concept of AI within the different archival institutions, the explanations of the key concepts and terms presented below have to be seen as starting points for discussion rather than rigid and explicit definitions. The understanding and uses of AI and interrelated concepts may differ between archival institutions and archivists.

Artificial intelligence

In this thesis, artificial intelligence or the acronym AI is used in a similar fashion as Colavizza et al. (2021); to describe the impact of various AI-based techniques and implementations on the archival institution, its concepts, workflows, and the role of archivists. This means that AI is used as an overarching term for its many subfields relevant to the archival sector, such as machine learning, deep learning, natural language processing, and computer vision. This broad definition provides a space for the various understandings of AI concepts and uses of AI techniques of each archival institution. When discussing specific implementations of techniques from different areas within the field of AI, this thesis will try to use technical terms to the extent possible. Having a general understanding and familiarity with the common and potential AI techniques used within the archival sector is crucial if archivists are to be involved as stakeholders in the development and implementation of AI solutions. For this purpose, terms like machine learning, deep learning, natural language processing, and computer vision are introduced below.

Machine learning

Machine learning is a subfield of AI, and closely connected to the fields of statistics, computer science, and data science. The field of machine learning consists of a variety of methods through which computers “can learn from data without being explicitly programmed to generate a particular output”. (Cordell 2020, p. 4). Machine learning algorithms are based on statistical models with which patterns in data can be identified. Simply put, instead of telling a computer how to recognize faces, the computer trains on a large dataset after which it should be able to determine the features that make up a face. Machine learning methods underly the development of models for several other AI applications for computer vision, natural language processing, and speech recognition to name a few. (Jordan & Mitchell 2015)

Generally speaking, machine learning algorithms can be divided into two broad categories: supervised and unsupervised. Supervised machine learning algorithms are trained on labeled data, where the input data is already coupled with the correct output. In other words, there is an instructor guiding the algorithm on how to interpret the data. The algorithm trains on this data, finding out what features determine the relationship between the input and output. After this process, the algorithm should be able to correctly predict the right output for data it has not seen before. Unsupervised algorithms, on the other hand, do not need any labeled training data.

Deep Learning

Deep learning is a subfield of machine learning. Deep learning methods can extract high-level, abstract features out of data through several layers. Advancements in deep learning have been made possible by the availability of more powerful computers and large datasets. A deep learning model trained for object recognition for example can understand an image by looking at its pixels, the visible layer of an image, and extract features from subsequent hidden layers, such as the edges, corners & contours, and object parts of an image. By identifying these various layered features, the model is able to determine whether the image has a person, car, or animal on it. (Goodfellow et al. 2016)

Natural Language Processing

Natural Language Processing (NLP) is a field that deals with the question how computers can be used to understand and manipulate natural language to perform certain tasks. (Chowdhury 2003) Various NLP tasks and techniques are relevant for the archiving sector. NLP can be used to clean up and preprocess texts in order to remove any noise from the texts and the clear up ambiguities within human language. With the help of named entity recognition (NER), entities such as persons, companies, events, expressions of time, and geographical locations can be identified. Another useful NLP technique is regular expression with which email addresses, personal numbers, or credit card information can be identified in archival records. Lastly, NLP can be used to classify records based on its type. (Hutchinson 2020) Recent development in machine learning and deep learning have increased the capacity of computers to understand human language and perform more complicated tasks (Young et al. 2018)

Computer Vision

The field of Computer Vision deals with the question how computers can extract information from digital images and video. Examples of techniques developed within the field of computer vision relevant to the archival sector are character recognition (HTR & OCR), object recognition, and facial recognition.

AI & the cultural heritage sector

The interest in AI and machine learning applications is not limited to archives but stretches across the entire cultural heritage sector. In 2021, the Europeana Network Association published a report on the role of AI and machine learning in the cultural heritage sector. The report, consisting of a survey and interviews with cultural heritage professionals, focused on how cultural heritage and research institutions have adopted AI in their practices and the challenges and issues they faced during this process. The survey data showed that more than 90% percent of the 56 institutions that answered, were interested in at least one AI topic and over half of them (54%) had already some practical expertise in one or more AI topics. AI topics of interested included metadata quality, collections management, discovery & search, knowledge extraction, and visualizing collections. (Europeana 2021)

Furthermore, the report shows that AI and machine learning tools and technologies are applied by cultural heritage and research institutions for various reasons and on a wide range content types. The tools are primarily utilized for facilitating the accessibility, usability, and discoverability of digitized collections, through for example metadata extraction & enrichment and automatic indexation. AI tools are used on textual documents (OCR & HTR), images (classification, retrieval, object recognition, 3D reconstructions based on 2D images), and video and audio (audio processing, video segmentation, music recognition). Challenges and issues identified in the report relate to access to high-quality and general training data, skills and expertise of internal staff, scaling up from project to production level, legal and ethical considerations such as copyright, personal data protection, and the nature of objects connected to contentious past, for example colonialism. (Europeana 2021)

Commissioned by the US Library of Congress, Cordell (2020) published the report *Machine Learning + Libraries* in 2020 on the state of machine learning within libraries, and the opportunities and risks of machine learning applications. The report provides an overview of the vital questions, opportunities, and cautions related to machine learning, as well as recommendations for practitioners from all kinds of cultural heritage institutions interested in setting up a machine learning project. The opportunities of machine learning for libraries include crowdsourcing (building training datasets), discoverability of collections (through for example clustering & classification based on genre, theme, structure, shared topics, and linguistics structures), and library administration and outreach. The challenges include the construction and contextualization of datasets, expertise and literacy of staff in machine learning, the development of computational infrastructure, and finally understanding entire machine learning workflow and steps.

The same year, the Museums + AI Network, a collaborative network of British and American museum professionals and academics, published the *AI: A Museum Planning Toolkit* authored by Murphy & Villaespesa (2020). The aim of the toolkit is to help museum professionals understand the possibilities of technologies from the AI field and assist in the development of “strategically, ethically, and operationally robust project plans”. (Murphy & Villaespesa 2020, p. 1) The toolkit includes three case studies, from the American Museum of Natural History, the British National Gallery, and the Metropolitan Museum of Art. These institutions have experimented with NLP, machine learning, and/or computer vision for the analysis of visitor feedback, improvement of marketing and visitor experience, and automated tagging of artworks for increased accessibility. In addition to the case studies, the toolkit includes a 1) framework for discussing required AI capabilities (data, tools, resources, skills, organization, stakeholders), 2) workflow for the consideration of ethical implications and algorithmic biases during each cycle of the project, from data input (collection & clean up), to data training, testing/model development, application, data output, and evaluation, and a 3) glossary of AI terms and concepts.

AI & the archival sector

The capability of archives to implement AI and machine learning-based solutions has grown exponentially the last decade. In the archiving world, a slow transition is visible from incidental research and development projects and experiments towards feasible and operational integrations of AI in archival processes and workflows. (Colavizza et al. 2021) On the other hand, Rolan et al. (2019) mention that the lack of compelling case studies within the academic and professional settings may fail to inspire archives to start experimenting with AI. In this a selection of AI implementations carried out within the archival sector is introduced. While certainly not complete, the use cases presented here show the potential of AI in relation to a wide range of archival tasks and materials.

Colavizza et al. (2021) review academic literature on the impact of AI on archival institutions and archival processes. Their review gives a broad overview of the different sections of the archival workflow where AI can be integrated. The sections where AI can be of assistance include 1) the appraisal of large volumes of unstructured and non-categorized documents and data, 2) the identification of sensitive or personal information hidden within digital archival material, 3) organizing and describing archival collections through a combination of digitization (OCR, HTR), indexation, and metadata extraction, 4) access and retrieval of archival documents and data, and 5) capturing new kinds of archival documents and data through web scraping and crawling.

Hutchinson (2020) covers similar applications of AI as listed above, but from the perspective of the field of Natural Language Processing (NLP). In combination with machine learning, NLP has the potential to contribute to the automatization of archival workflows through techniques such as named entity recognition and topic modeling. The author focuses on how NLP and machine learning techniques can be applied to several parts of the archival workflow, including appraisal and selection, description and access, and the sensitivity review of archival records. Hutchinson also includes a discussion of (open source) software tools specifically designed for the archival sector, such as ePADD, BitCurator NLP, and ArchExtract. These software tools allow archivists to use a combination of NLP techniques through a user-friendly interface. While the tools discussed show potential, Hutchinson (2020) gives suggestions for future development, highlighting the requirement of usable, interoperable, flexible, iterative, and configurable tools.

Rolan et al. (2019) describe the arrival and impact of AI technologies in the context of the Australian archival sector. The authors discuss several AI initiatives, focused on creating AI systems for the automatic appraisal of email, the detection of sensitive and protected information, and the disposal and retention of archival records. Based on these study cases, they conclude that there is an emerging capability of AI technologies to impact the working environments of archives in the future. The technologies implemented at the Australian archives and governmental agencies, however, are not production ready. Successful implementation requires a lot of data preparation and access to economical and technical resources. A broader understanding of AI and its impact on the archiving world is necessary in order for archival institutions to take fully take advantage of the capabilities of AI.

Alongside textual archival material, AI-based technologies have a wide range of use cases for audiovisual material. Kern et al. (2020) explore the possibilities AI technology has to offer for image processing and recognition. They have experimented with different image processing technologies to detect and recognize faces, persons, and objects in images from the Robben Island Mayibuye Archives. These images form an important part of the history of South Africa but remain largely inaccessible due to the fact that most images are undigitized and unlabeled. Although the size, quality, and condition of historical images posed several difficulties, the project achieved promising results.

Bocyte & Oomen (2020) discuss the possibilities that AI-based technology offers for another type of audiovisual material, namely video. They describe the potential for video analysis through video segmentation, video summarization, and speech analysis. These techniques can be used to create specialized interfaces for audiences to explore the audiovisual heritage preserved within archives. Use cases include AI-technology being implemented to adapt and publish content for broadcasters and media archives, deliver personalized videos to end-users, and retrieve video segments for creators.

Impact of AI on the archive & the archival profession

The digitization of archives & society and the increasing use of AI and machine learning in archival processes transforms the conceptualization of archival institutions. The understanding of archives is transitioning from collections of text to be read towards collections of data to be mined or made sense of. In this thesis, the archive-as-data paradigm is utilized as a theoretical lens to capture and analyze the transformations in the Swedish archival sector and the archival profession as a result of the introduction of AI and machine learning concepts and technologies.

In the context of the archiving world, Cook (2013, p. 97) defines paradigms as "*frameworks* for thinking about archives, or archival *mindsets,* ways of imaging archives and archiving" [original italics]. These frameworks or mindsets are not rigid or mutually exclusive according to Cook, as strands of previous paradigms tend to carry over into the next. Paradigms can be used to capture and reflect on the emerging transformations and challenges facing archives today:

> [W]e can see them as liberating, authorizing us to develop new directions in light of the astonishing changes to archiving today from theory, technology, and society, and the expectations and demands each occasions. (Cook 2013, p. 117)

Cook (2013) describes the shifts in archival identity and practice from the nineteenth century up until the present with the four paradigms of evidence, memory, identity, and community. Together, these paradigms capture the developments of the archive and the archival profession, with archivists transforming from impartial guardians of evidence of public/governmental records, active appraisers and selectors, mediators of societal identity, to facilitators of community archiving.

The archive-as-data paradigm introduced by Mordell (2019, p. 146) is a fifth addition to Cook's paradigms and "encapsulates the reframing of digital archives as data and the attendant implications for archival identity". It offers a framework for understanding and reflecting on the implications and impact of the use of AI within the Swedish archival sector, on archival concepts, the archival processes, and the roles and responsibilities of the archivist. The following sections introduces reflections on the consequences of viewing archival records as data, the changing archival profession, the need for collaboration, and the ethical considerations.

Archival records as data

Digitalized and born-digital archival collections and records are not automatically data, but become data through datafication and opaque human-decision making. (Mordell 2019) The process of turning archival collections into data takes time and resources. Rolan et al. (2019) highlight that experimenting with AI and machine learning in archival context requires preparatory groundwork and sizeable investments. It requires the identification of suitable training and test data, the cleaning and preparation of this data, and access to sufficient computational power to process the data. The resources and time needed for the development of large, high-quality training and test datasets and the configuration of the algorithm may be too high, especially for smaller archival institutions. Rolan et al. (2019) point out the fact that AI-solutions are difficult to translate and apply to other datasets, making sizeable investments of time and resources hard to justify. The heterogeneity of archival records and metadata across institutions, collections, and file and media types require specialized datasets for each application. Lastly, having access to data is not sufficient for successful adaptation of AI-based technologies according to Rolan et al. (2019). Archival institutions may lack the domain expertise and knowledge needed to frame archival questions into AI-solutions and interpretate the outcome of the algorithms. For this a direct connection with the data science sector is required.

Mordell (2019) challenges the perceived neutrality and objectivity surrounding the concept of data and the use of computational methods, by investigating the hidden assumptions and biases that are brought along with the conceptualization of archives-as-data. The same neutrality and objectivity are often perceived of the use of computational processes, algorithms, and machine learning. The view that machine learning methods are more objective or neutral compared to human decision-making, as they ‘simply’ discover existing patterns in data, is also questioned by Cordell (2020). Even if the underlying algorithms of computational methods act autonomously, Cordell argues that the assumptions and biases of those who have programmed the instructions or rules of the algorithms and/or the biases of those who have created the (archival) data in the first place, will inevitably seep through in the final output of AI and machine learning methods.

Roles and responsibilities of archivists

The archive-as-data paradigm provides a framework for reflecting on the archival profession, and the roles and responsibilities of archivists. The conceptualization of archives-as-data brings the risk of erasing the agency of archivists in archival work.

> Reinforcing the conceptual relationship between data and objectivity, an archive-as-data-as-raw-material frame may likewise serve to minimize the archivist’s agency in arrangement and description: rather than constructing meaning, the archivist merely discovers what is innate in the records. (Mordell 2019, p. 150-151)

This view is challenged by Mordell and other scholars, who argue instead, that the arrival of computational tools based on AI and machine learning prompts archivists to assume new roles and responsibilities. In the changing digital environment of the archival sector, archivists are expected to acquire new skills and knowledge. Theimer (2018, p. 9) suggests that archivists need to become “masters of data”. This means understanding how AI and machine learning tools work, how they are applied, and what uses and possibilities they offer. Additionally, Ranade (2016) points out that archivists need to be able to assess the accuracy and reliability of the output of machine learning algorithms and explain and present the results in understandable ways to different archival audiences. Connected to this, Bunn (2020) introduces the concept of explainable artificial intelligence (XAI) to the debates on AI and recordkeeping within archival institutions. The opacity or black box of AI technologies are posing challenges to archival traditions of transparency and access to trustworthy records. According to Bunn, archivists should be able to explain how the outcome of the algorithm came into being.

Mordell (2019) sees an opportunity for archivists to be critically involved in the transformation of archives as collections of data and the implementation of computational methods. Archivists are able to keep a critical eye on the interpretative decisions made during the datafication process of archival collections. The perceived objectivity and/or neutrality of data for example is something archivists are able to disprove. Mordell (2019) points out that the application of computational methods opens up new venues for archival processing but may also close other venues, by privileging certain types of records or performing analyses based on incomplete or biased training data. As experts on the collections of the institution and the biases in the data, archivists can ensure that newly arrived technologies contribute to the archival workflow and activities in a positive way. Jaillant (2022), however points out that archivists are currently not often considered as stakeholders in the development and implementation of AI technology. This is problematic considering the roles and responsibilities archivists could have in promoting algorithmic transparency and accountability, highlighted by both Ranade (2016) and Bunn (2020).

Collaboration

As digital materials are making their entry into the archival collections, archival practices and tools are becoming increasingly dependent on technology and computational thinking. (Underwood et al. 2018) The archival science, therefore, needs to have an understanding of how digital records are formed, processed, and stored in order to continue to preserve and make accessible authentic records. (Marciano et al. 2018) The archive-as-data paradigm addresses the implications of the integration of computational methods and the collaboration between the disciplines of archival science, computer & data science, linguistics, and digital humanities. According to Mordell (2019), the use of computational methods based on AI and machine learning especially calls for stronger collaboration between archivists and computer scientists.

> As the use of computational tools begins to shape a greater extent of archival work, it becomes increasingly important for the two intellectual traditions to come together and recognize areas of mutual concern. (Mordell 2019, p. 157-158)

Due the subjectivity of machine learning, Cordell (2020) sees the need for experts from other disciplines and fields, such as the cultural heritage sector, to contribute to AI implementations. Interdisciplinary collaboration between AI experts, computer & data scientists, and scholars & cultural heritage professionals is necessary to be able to successfully and responsibly operationalize domain-specific questions within the cultural heritage sector into machine learning tasks.

To facilitate the collaboration and knowledge exchange between disciplines, Marciano et al. (2018, p. 179) suggest the formation of the computational archival science (CAS), which they define as:

> A transdisciplinary field concerned with the application of computational methods and resources to large-scale records/archives processing, analysis, storage, long-term preservation, and access, with the aim of improving efficiency, productivity, and precision in support of appraisal, arrangement and description, preservation, and access decisions.

Marciano et al. (2018) envision CAS to be a space for mutual exchange between the archival science and the computer & data science. In this space, knowledge and expertise can exchanged on shared issues and problems faced in both disciplines, for example related to the documentation of the provenance of data. The formation of CAS, however, is not intended to replace the ‘traditional’ archival science. With CAS, the archival sciences and sector can capitalize on the potential of applying computational methods within archival processing and develop AI in harmony with archival practices, principles, and concepts. At the same, the collaboration can strengthen the computer and data sciences and sector through expertise within the archival institutions, on data provenance and biases.

From the other side, the AI/machine learning field and community, the potential of archival concepts and traditions is also recognized. Jo & Gebru (2020) call for an interdisciplinary subfield within the AI and machine learning field, similar to CAS, where (ethical) issues regarding how to engage with human information, construct datasets, and preserve data can be addressed. According to them, archival approaches to inclusivity, consent, power, transparency, and ethics & privacy can be used to strengthen the field of AI and machine learning.

> By showing the rigor applied to various aspects of the data collection and annotation process in archives (…) we hope to convince the ML community that an interdisciplinary should be formed focused on data gathering, sharing, annotation, ethics monitoring, and record-keeping processes (Jo & Gebru 2020, p. 307)

Ethics and AI

The implementation of AI into archives poses several ethical concerns, related to bias in data and the prioritization of certain records. These concerns should not be underestimated or ignored by looking only at the great potential AI is supposed to have in regard to archives. Recently a European regulatory framework regarding the use of AI within the European Union has been developed. The framework is intended for EU countries to profit from the economic and societal benefits of AI and to increase human wellbeing, while mitigating the risks and harm AI poses for individuals and society, end goal to. The framework includes a topology of AI systems, from prohibited, high-risk, and low-risk systems. As of yet, it is unclear how the AI systems used in archival institutions are to be classified within the framework of the European Union. The regulations will lead to new demand on openness, explainability, and transparency of AI systems used within the public sector and public archival institutions. (European Commission 2021)

AI in Sweden

The national government of Sweden aims to be the leading country in the world to utilize the opportunities of digital transformation. The investment and development of AI-technology are part of this leading role. In 2018 the Ministry of Enterprise and Innovation published the National Approach to Artificial Intelligence. The document outlines the direction in which AI-efforts in Sweden should be heading and identifies the conditions necessary to use AI in ethical, secure, and sustainable ways. To reap the benefits of AI in Sweden, all parts of society need to be involved in its development. Collaboration is needed between the public sector, private sector, and research infrastructures within Sweden and abroad:

> For Sweden to reap the benefits of AI, all sectors of society must be involved; this is not an issue that the state, municipalities, county councils, academia, or private companies can deal with on their own. (Ministry of Enterprise and Innovation 2018, p. 5)

In regard to the public sector, the national government has high expectations. AI-applications are seen to have the potential to meet societal challenges in the future and create a more effective and relevant public sector, with services developed in the interest of citizens. The public sector can contribute to meeting the prerequisites of AI applications, by making available large volumes of high-quality data and developing national digital infrastructures. The unique position of the public sector is highlighted in the national approach, which states that governmental agencies have unique access to large volumes of data.

Purpose of the study

The previous section shows that there is a growing potential for the integration of AI-based tools and methods within archives. The maturity and advancement of these tools means that archives can start integrating AI into parts of the archival workflows and processes more permanently. At the same time, however, economic and technical obstacles as well as ethical challenges still have to be overcome before AI can be implemented in a reliable and responsible way. Strengthening the collaboration between archivists, digital humanities researchers, and experts from the computer/data sciences is one of the requirements for overcoming these economic, technical, and ethical challenges. Archivists can play an important role in addressing the ethical issues surrounding AI, but for this to happen they first need to be recognized as stakeholders at the table of AI innovations.

The purpose of this study is to survey and analyze the AI landscape within the Swedish archival sector. Utilizing the archive-as-data paradigm, the study explores the archival sector from three main angles: 1) the current uses of AI and machine learning, 2) the opportunities and obstacles of (future) AI implementations, and 3) the impact of the AI on the archival profession in Sweden. In order to fulfil the purpose of the study, the following research questions have been formulated:

- How are Swedish archival institutions currently using computational tools and methods based on AI and machine learning?
- What opportunities and obstacles for the implementation of AI/machine learning into archival workflow and processes do practitioners from Swedish archival institutions identify?
- How will the introduction of AI and machine learning concepts and methods impact the archival profession and the role of archivists?

The study contributes to the research field on archival institutions in the digital age, as it highlights the transformations caused by the introduction of AI and machine learning. By focusing on the perspectives and experiences of practitioners of different types of archival institutions, this study highlights how AI is impacting workflows, practices and responsibilities of practitioners at large and small, and public and private Swedish archival institutions. It gives the opportunity to contrast and compare the perspectives on and experiences with AI and machine learning of archival institutions within the local Swedish archival context, the international archival context, and the broader cultural heritage sector.

By presenting and analyzing AI case studies from different archival institutions and highlighting the current state-of-the-art of AI tools for archival institutions, this study also strives to be a useful resource for practitioners in the Swedish archival sector. The study aims to further the practitioners' understanding of both the opportunities and challenges presented by AI. Ideally, the study will support archivists engage in closer collaborations with computer and data scientists and become a stakeholder in the development and implementation of AI technology within and, potentially even, outside of the archival institution.

Background: the archival sector in Sweden

To be able to put the experiences of the interviewees and archival institutions into perspective, an introduction is given on the Swedish archival sector.

The archival sector in Sweden is regulated by the Archives Act (*Arkivlag* 1990:782) and the Archives Enactment (*Arkivförordning* 1991:446) in which the sections of the Archives Act are further clarified and specified. Additionally, *Riksarkivet* or the Swedish National Archives publishes regulations and recommendations on matters that are not covered by the act and the enactment, for example regarding certain media and formats. The legal regulation of archives applies to archival institutions, governmental agencies, or private bodies such as churches that manage public archival documents and records called *allmän handling*. It does not concern private archival institutions that manage private documents and records. Private archive creators have no legal obligation to preserve their documents on a long term.[1] The Archives Act and Archives Enactment are strongly connected to the principle of openness (*Offentlighetsprincipen*), engrained in the Swedish constitution, and the Freedom of the Press Ordinance (*Tryckfrihetsförordningen*) which together ensure that citizens have the democratic right to access and read (almost) all records of every governmental agency. (SOU 2019:58)

Public versus private archival institutions

The archival sector of Sweden can be divided up in two broad categories: public archival institutions (*arkivmyndigheter*) and private archival institutions (*enskilda arkivinstitutioner*), although the categories are not mutually exclusive.

The public archival sector consists of all governmental agencies and archival institutions that preserve public records from the national government, municipalities, and counties. Some of these archival institutions also preserve private archival collections which, as previously mentioned, do not fall under the legal documents and regulations. Private archival material is preserved to ensure a broader representation of developments and movements in society, which are not documented in public records. Examples of public archival institutions are the Swedish National Archives, municipality archives, and region archives. An example of a governmental agency with an archival function is the Institute for Language and Folklore (*Institutet för Språk och Folkminnen*). In some instances, public archival institutions, such as the Swedish National Archives or *Stadsarkivet*, the Stockholm City Archive also regulate and control the creation of archival collections at governmental agencies through regulations and supervision. In Sweden the legal division between an archive creator (*arkivbildare*) and archiving institutions is not as clearcut. There is not a strong separation between records management and archives, as can be the case in other countries.

[1] For businesses and organizations there may be requirements to preserve financial and budgetary documents related to the bookkeeping for a certain amount of years.

The private archival sector consists of archival institutions managing and preserving the archival collections of private creators and institutions. These archival institutions have the same mission as their public counterparts, to collect, preserve, and make available archival collections. Often in the case of private archival institutions, this is done for the members, owners or clients and not primarily for the general public. The members or clients often pay a membership fee as well as a shelf meter fee, in return for the organization and preservation of their archival collections. Ownership of the records does not transfer to the archival institution when a client or member deposits material. The member remains in control of the material and decides what happens to it and who is allowed to access it. In some cases, the archival collections are gifted. Then the archival institution decides how it is made accessible. There are a wide variety of public archival institutions, collecting different kinds of private archival material from associations, people's movements, and businesses on a local, regional, and/or national scale. Examples of private archival institutions are *Arbetarrörelsearkivet och bibliotek*, the Swedish Labour Movement's Archives and Library and *Föreningsarkivet Värmland*, the Association's Archive in Värmland. (SOU 2019:58)

Throughout the study 'archival institutions' and 'archives' are used as umbrella terms for all public archival agencies, private archival institutions, and governmental agencies performing archival tasks within the Swedish archival sector.

National archival database (NAD)

The *Nationella arkivdatabasen* or the National archival database (NAD) is an online database and information system for the entire Swedish archival sector, developed by Riksarkivet. In the database, information is collected about institutions and archival collections from both public and private actors within the archival sector and the wider cultural heritage sector. NAD functions as a one-way entry point into the available archival material in Sweden. The aim is to have metadata and descriptions of all organizations that preserve archival material in Sweden. Currently archival and cultural heritage institutions are free to decide whether they want to register their archival collections and material in the database. They also decide the level of detail for the descriptions and information added. (Swedish National Archives 2019)

Material and method

The empirical data for this study was collected through semi-structured interviews with archivists and other information/data specialists working at Swedish archival institutions. A general interview guide (see appendix) was created with questions organized around the core themes of the study:

- Archival practices for digitized and born-digital material
- Previous and current experimentations and implementations of AI
- Opportunities and obstacles for AI integration in archival practices
- Impact of AI on the archive and archival profession

Depending on the institution and expertise of the interviewee, questions were added and/or modified. For example, additional questions were asked specifically relating to a certain AI project or modified to the specific type of archival institution. The interview questions and themes were not always discussed in the exact same order. The flexible semi-structured nature of the interview guide provided a way to stay in control during the interview, ensuring that each core theme was addressed in the end, while at the same time leaving the opportunity open to delve into the interesting replies, observations, and perspectives of the interviewee. (Galletta 2013)

Interviewees were invited to participate in the study by e-mail. For each institution, people were approached that were deemed to have relevant experiences and expertise in relation to the thesis topic. This could be due to their involvement in a certain project or based on their function title. Sometimes, the initial person I reached out to recommended interviewing another one of their colleagues instead. Based on the recommendations of the first interviewee from the Swedish National Archives, an additional interview was conducted in order to get a broader perspective on the topic of the study within this specific organization.

In total, five interviews were conducted. The interviews lasted between 45 minutes and 1 hour and 45 minutes, with an average duration of an hour. Three of the interviews were conducted in-person, at the archival institution. The other two interviews were conducted online through Zoom Meetings, which is developed by Zoom Video Corporations.

Although online interviews bring along certain obstacles and drawbacks, recent advancements in communication technology and the availability of stable internet connections have made online interviews a viable option alongside in-person interviews. (Irani 2019) Online interviews bring the risk of technical nuisances that can interrupt the flow of the interview, distracting both the interviewee and interviewer. Additionally, not being in the same physical space might result in the interviewer missing certain changes in body languages or emotional cues that the interviewee could be showing in relation to the topic or a certain question being brought up. (Gray et al. 2020) In order to minimize the chance of any interruptions or technical issues arising, several measures were taken. A week before the scheduled interview, a Zoom invitation was sent out to the interviewee with a direct link to the meeting. This link was once again shared on the day of the interview. The audio recording functionality of Zoom was tested beforehand and a quiet place with a stable internet connection was chosen to conduct the interview.

The empirical data from the interviews was complemented with information related to the archival workflows/practices of the institution and, if applicable, AI project reports and descriptions, either retrieved from the institution's website or from documents provided by the interviewee after the interview. This information was helpful for the formulation of relevant interview questions beforehand and the contextualization of the interview data afterwards.

Selection of archival institutions

Four archival institutions have been selected to capture a variety of experiences with digitized and born-digital collections and the development and implementation of AI-based tools. Together, these four give an adequate representation of large and small, public and private, and local, regional, and national archival institutions within the Swedish archival sector. In total, five interviews were conducted with archivists, the head of one archival institution, and a data scientist. In the case of the Swedish National Archives two interviews were conducted.

The Swedish National Archives

The Swedish National Archives or *Riksarkivet* operates nationwide with multiple offices across the country. The public archival agency predominantly manages governmental records from national and regional agencies, but also private archival collections. Additionally, the Swedish National Archives issues regulations and general advice (RA-FS and RA-MS) for archival institutions managing public or state records.

The Stockholm City Archive

The Stockholm City Archive or *Stadsarkivet* is the public archival agency for the Stockholm municipality (*Stockholms Stad*) and Stockholm County (*Stockholms län*). The archive predominantly manages public records from regional and local agencies, but also several private archival collections.

Popular Movement's Archive

The Popular Movement's Archive or *Folkrörelsearkivet för Uppsala län* is a private archival institution located in Uppsala. It was founded in 1987 to preserve archival collections and records of local associations (*föreningar*) within Uppsala County. The archive preserves the archival collections on behalf of their members, the local associations.

Centre for Business History

The Centre for Business History or *Centrum för Näringslivshistoria* is a private archival institution with offices in both Stockholm and Uppsala. It was founded in 1974 to preserve archival collections and records of Swedish companies and business. Similar to the Popular Movement's Archive, the Centre for Business History manages the archival records on behalf of their members, the Swedish companies and businesses.

Methodology

The qualitative analysis of the interview data was based on the data analysis spiral described in Creswell & Poth (2017). The analysis proceeded in the following subsequent steps. All interviews were audio recorded and transcribed in its entirety for the purpose of data analysis and ensuring accuracy of quotes. The transcripts of the interviews formed the basis of the data analysis. These were read through repeatedly to get familiar with the content. In this process any emerging ideas and impressions were written down. Next, the transcripts were described using codes or labels and categories of codes. Codes could capture reoccurring patterns, surprising comments, and connections to the literature and concepts. Important codes were then merged together into themes. The themes that emerged during the coding process were then related to each other, the theoretical concepts, and the findings of the literature review. As the thesis includes data from multiple case studies it was important to be aware of codes and themes within the specific case study's context. Attention was also given to the cross-case analysis focusing on similarities and differences in the type of archival institution, archival practices, and experience with AI. The discussion section of the thesis will interpretate and generalize the themes identified in the interviews and contrast and compare the results of each case study. Lastly, the surfacing themes were brought under the three sections of the thesis, the current implementations of AI, opportunities and obstacles of AI, and the impact of AI on the archival profession. The nature of archival data, for example, was mentioned during several interviews and placed as a common theme under obstacles of AI implementations.

Ethical considerations

The research topic and research design both raise ethical considerations that need to be addressed throughout the thesis writing. In regards to the research design, the data collection method in the form of semi-structured interviews brings along ethical concerns that need to be sufficiently addressed. The confidentiality of interviewees has been protected to the fullest extent possible. The purpose of the study did not require the collection of detailed personal information of the interviewees. Therefore, the personal information stored only included the name, contact details, archival institution, and position of the interviewee. This information was deemed necessary for correspondence and contextualization of the interview. During the transcribing of the interviews, all personal information was moved to another document or deleted from the transcript. In the final thesis, the interviewees have been anonymized. Only a general description has been included of the participating archival institution and the positions of the interviewees.

Oral consent was sought at the start of each interview, for the voluntary participation, the collection of personal information and data, and the recording and transcribing of the interview. An electronic information letter was sent out in advance, informing the interviewees about the purpose of the study and explaining how the data and personal information collected during the interview would be handled and stored. This information was repeated at the start of the interview. Interviewees had the right to withdraw from the study at any given moment. Additionally, all quotes used in the thesis are reviewed for accuracy and cleared by the interviewees. Once the thesis is approved and uploaded to DiVA, a summary of the main results and conclusions of the study as well as a copy of the entire study will be shared with the participating interviewees and archival institutions. I hope the summary and final thesis will be beneficial for them.

Ethical considerations also arise in regards to research topic. As became apparent in the theoretical framework section, the use of AI and machine learning and its implementation have certain societal implications and ethical consequences. While the ethical consequences of AI implementations in the archival sector are not as grave as for example the use AI in governmental decision-making, this study still advocates for mindful implementations of AI in the archival sector with awareness of the benefits and associated risks. The potential of AI is best fulfilled in human-machine framework, where archivists or other specialists remain in control over the final decisions.

Limitations of study design

The design of the study naturally comes with certain limitations. First of all, the interviewees' views and perspectives may not be representative for the breadth of opinions and views within the entire archival institution. Their colleagues can have different takes for example on how AI could impact the archival workflows and the role of the archivist in the future. Likewise, it can be argued that the sampling of archival institutions is not fully representative for the entire Swedish archival sector. The decision was made to select four cases to be able to show a diverse picture of AI implementations at different kinds of archival institutions, but also allow room for in-depth analyses of each individual institution and interviews with multiple archivists or information/data specialists if necessary. This means that the selection of types of archives and the geographical spread of institutions could have been expanded. Minority or community archives, such as the Archives and Library of the Queer Movement (*Queerrörelsens Arkiv och Bibliotek*) in Gothenburg or special collections archives with audiovisual material, such as the former Swedish National Archive of Recorded Sound and Moving Images (*Statens ljud- och bild arkiv)*, which nowadays is a part of the Swedish National Library, would have been valuable additions to this study.

Secondly, the selected archival institutions from the public and private archival sector do not necessarily have to reflect the experiences of other institutions belonging to the same type of archival institution. In other words, not every popular movement archive or business archive is the same. These institutions do not form monolithic groups, as there may be similarities and differences in terms of the nature of the collections (analogue versus digital), access to resources, number of members, size and composition of the staff, etcetera. The results of this study therefore have to be regarded as indicative of the experiences of the different types of archival institutions within the Swedish archival sector, rather than an accurate representation of the entire type.

Lastly, it can be argued that the division of interviews between the four archival institutions is unbalanced. Two of the five interviews were conducted with representatives from the Swedish National Archives which led to the overrepresentation of this organization in the study and analysis section compared to the other three organizations. A higher number of interviewees for the Swedish National Archives can be legitimized due to its size and the advancement of AI implementations. Additional interviews with representatives from the other three archival institutions, however, could have enriched and broadened the understanding of the experiences with and outlook on the use of AI and machine learning within these organizations.

Study and analysis

In this section, the data gathered through the semi-structured interviews with archive practitioners and the analysis of organizational documents, project plans, and webpages are presented. The section is divided up in three parts, corresponding with the research questions of the thesis. Together the parts contribute to achieving the overall purpose of the study, which is to survey and analyze the AI landscape within the Swedish archival sector. The three parts are: 1) the current implementations of AI within the archival institutions, 2) the opportunities and obstacles for AI within the archival sector, and 3) the impact of AI on the archival profession.

I. Current implementations of AI

The implementations of AI technologies vary between the archival institutions. The Swedish National Archives can be considered as the frontrunner within the Swedish archival sector when it comes to the implementation of AI within different parts of the archival workflow. On the opposite side of the spectrum, the Centre for Business History can be found where AI is currently not implemented or experimented with. In general, the archival institutions actively engaging with AI are still in the stage of experimentation and project-based AI implementations, with the Swedish National Archives being the closest to moving towards more permanent integration of AI in their archival workflows. Below, the activities of each archival institution in the field of AI are introduced.

The Swedish National Archives

The Swedish National Archives can be considered as the frontrunner within the Swedish archival sector when it comes to the exploration of the possibilities to implement AI within the workflows and practices. The organisation has a designated unit for AI innovations called *Artificiell intelligens inom Riksarkivet* (AIRA) within the Research & Development department.

The AI journey of the Swedish National Archives started in 2018 with a KAM investigation. The investigation focused on how AI techniques could be integrated into the modernization and digitalization plans of the organization. It explored the practical and theoretical implications of integrating AI into the archival workflow and practices. The final report lists AI techniques that have reached a certain level of maturity so that they could be implemented in archival workflows and activities in a positive way directly. The areas in which AI could be applied included digitization, open data & big data, accessibility and searches, enquiry handling, deliveries and inventory, customer service, and digital preservation. Examples of the techniques explored in the investigation are text recognition through OCR and HTR technology, image recognition, visualization techniques, and NLP applications.

Through the KAM investigation, the Swedish National Archives was able to identify how different AI areas and technologies interrelate and build upon each other. This way a prioritization of different AI applications and technologies could be made. In the two projects that followed, AIRA I and AIRA II, the focus was directed towards building the foundations of AI implementations. The development of language models for historical archival material and HTR/OCR techniques & pipelines form the basis for the extraction of metadata through indexation. This in turn opens up the possibilities for other AI implementations that focus on creating new ways in which archivists and users can access and search the collections.

The Swedish National Archives is currently developing a generalizable methodology for automated indexation pipelines of archival material, using a combination of OCR and HTR technologies. A pipeline is constructed using different AI techniques in a certain order, starting with layout analysis and segmentation, followed by image preprocessing, transcription, and validation. This pipeline has been tested on two types of archival materials, popular records and property books. The organization receives over 30.000 enquiries from the public regarding real estate. The aim of the indexation pipeline is to improve the information retrieval from the property books and handle enquiries from the public more effectively. In the near future, the Swedish National Archives hopes to automatize the indexation of the property books, a process that has been ongoing manually for ten years.

The pipeline has a modular design structure, giving the ability to switch out individual modules. This structure has several benefits. As many AI applications consist of the same steps, individual modules from the pipeline can be more easily integrated into other pipelines of future AI projects. Furthermore, if a better HTR or segmentation module is developed/trained, this module can replace the original without having to ‘reprogram’ the entire pipeline.

The Swedish National Archives uses open-source models from research and GitHub repositories and adapts these AI architectures to the Swedish historical data. This way the organization is able to stay up-to-date and profit from the most recent developments within the fast-developing field of AI. Different architectures have certain strengths and benefits. Several architectures can be combined to give the best possible results. Newly developed architectures perform better, for example the attention-based model called MMOCR is not as sensitive to the order in which the first name and surname appear. This is important when working with the census or popular records in which names are sometimes written in different orders.

The language models and datasets that the Swedish National Archives uses, are developed within Swe-Clarin. Swe-Clarin is a research infrastructure that aims to make “extensive language-based materials available as primary research data” and offer “state-of-the-art language technology as an e-research tool” to researchers from the humanities and social sciences. (Bodin et al. 2016, p. 2) The projects of Swe-Clarin generally strive to be a collaboration between language technologists, representatives from the Swedish academic community as well as the data owner, typically the memory institution where the data is kept. (Bodin et al. 2016) In this framework, the Swedish National Archives is creating and developing language models and training and evaluation datasets for textual material from the 1730s to the 1910s. These models and datasets form the base for future NLP and AI implementations. The Swedish National Library (*Kungliga Biblioteket*) and the linguistic departments of several universities are important partners in this. The language models developed by the Swedish National Library for modern Swedish are specialized for old Swedish and historical textual material. With the help of Gothenburg University and their network of citizen scientists or volunteers, an evaluation dataset is compiled with annotated texts that can be used for the training of several NLP language models and other AI tools in the future.

Lastly, the Swedish National Archives is part of eSAM, a collaborative network of national governmental agencies focused on digitalization in the public sector. Within eSAM, there is a group that works specifically with AI implementations and questions. The group discusses several topics and issues that all governmental agencies have to address when working with AI. The legal ramifications of the use of AI in governmental agencies are explored, for example when purchasing AI services from an external company, how to handle GDPR and data security concerns, and how to share AI resources, such as training datasets. Many AI innovations are applicable to the public sector and can be used in archival settings with some modifications. The Swedish National Courts Administration, *Domstolsverket*, for example, created an NLP-based system for redacting personal names and information, using named entity recognition. This AI application could be very useful for the archival sector and the Swedish National Archives.

The Stockholm City Archive

The Stockholm City Archive has experimented, tested and worked with AI technologies in different kinds of settings, in collaborations with other regional governmental agencies, cultural heritage institutions, and universities.

The first experiment with AI at the Stockholm City Archive aimed to optimize the retrieval of school records. Each year, the institution receives thousands request from the public wanting to get access to their school grades. Due to GDPR regulations, these records cannot be published online, so in collaboration with another governmental agency within Stockholm County, an AI system was developed. With the help of the AI system, this administrative task was to be automated and reduce the economic resources needed. The expectations and optimism regarding the performance of the AI system, however, were not met during this experiment, as it took a long time for the system to deliver the right document.

Currently the Stockholm City Archive is also using the HTR tool Transkribus developed by READ-COOP to transcribe 18th century documents regarding the history of Stockholm. The documents come from the collections of the chambre of commerce, court, and police. (Transkribus n.d) In 2022, the institution also provides free courses in reading 18th century handwriting. (Stockholm City Archive n.d.) The use of Transkribus is seen as a way to involve the general public, as so-called citizen scientists. This way the City Archive aligns itself with the political goals of the board of the city. At the same time, allowing citizen scientists to participate in the archival activities presents the City Archive the opportunity to transcribe large collections of handwritten texts, which they without the help of the citizen scientists could not afford.

In collaboration with the Stockholm City Museum and researchers from Uppsala University and Stockholm University, the archive is working on a project called City Faces. The project, funded by the Swedish Research Council, focuses on training an AI to recognize faces on 19th and 20th century portrait photographs. Trainings data is created by connecting portrait photographs with the city's popular records. The idea is that the algorithm will be able to recognize people on other photographs from the collection of the archive and the museum. Alongside portrait photographs, the project is also looking into the possibility of identifying buildings using architectural drawings. The project hopes to open up new research possibilities and lead to new insights on urban life in the 19th and 20th century. (Stockholmia 2022)

The Popular Movements' Archive in Uppsala

The Popular Movements' Archive is mainly engaged with AI-technology in the context of the project called Labour's Memory in collaboration with the Swedish Labour Movement's Archive and Library in Stockholm and two international archives from Germany and the Netherlands. The aim of Labour's Memory is to digitize, index, and make searchable annual and financial reports from trade union organisations operating on a local, regional, national, and international level between 1880 and 2020. The project runs from 2021 to 2023 and is funded by the Riksbankens Jubilieumsfond, a Swedish foundation supporting research in the humanities and social sciences. The project makes use of several AI technologies, applications and tools. These are developed in collaboration with four researchers from Uppsala University, affiliated with the departments of Information Technology, Linguistics and Philology, and Archives, Libraries and Museums. (Swedish Labour Movement's Archives and Library 2020)

The project offers the Popular Movement's Archive a unique opportunity to learn and develop skills and technologies that they otherwise would not have the resources and time for. After the project has ended, the aim is to make the developed programs publicly available as open-source repositories.

Part of the Labour's Memory project includes the digitization of documents from different local labour unions. As the time span of the project includes documents from 1880 until 2020, the digitization concerns both handwritten and typed text. In collaboration with a data scientist from Uppsala University, the archival institution is developing deep learning models for HTR purposes. The aim is to create a model that does not require a lot of training data in the form of transcribed texts and can easily handle the different handwriting styles in the documents. Currently this part of the project is at a standstill as the data scientist accepted another job and the vacancy within in the project is expected to be filled next year.

The Popular Movement's Archive had previously experimented with the HTR tool Transkribus, but the demands of Transkribus could not be met due to the diversity in handwriting styles across the documents and required amount of manual transcribed pages. It would take too much time and resources to transcribe texts as training data for every unique handwriting style in the archival material would be required. There was not enough consistency in the handwriting for Transkribus to be functional, every meeting another person of the trade union may have taken notes. At the end of the project, the aim is to have integrated HTR and OCR functionalities into one user-friendly program that would not require any knowledge about programming and the command line. This would include the cleaning and preparation of the document as well as the transcribing of the document.

Another part of the Labour's Memory project where AI comes into play is the searchability of the newly digitized material. In collaboration with linguists from Uppsala University, the Popular Movements' Archive is working on improving searchability. Search queries of archive users are to take into account the changes of words in terms of semantics/meaning and spelling. If a modern term is used for a search, the hits are to include the historical equivalents and spellings tied to the modern word. Another area is the development of methods to meet the GDPR requirements when publishing the contemporary annual and financial reports. This includes the blurring of the appearances of personal information, names or faces in images. The exact method for how to best deal with GDPR and making digitized accessible online has not been determined. The linguists will develop a search protocol which will be executed through some sort of script. This part of the project relates to the realm of NLP.

Centre for Business History

The Centre for Business History is the only archival institution out of the four not actively engaged with any implementations of AI. The institution does make use of technologies based on AI and computer vision, namely OCR to make documents machine readable using book scanners from a commercial party with pre-installed software. The decision not to make use of AI technology is partly steered by the members of the Centre, owning their archival material. Currently, members do not see any financial value or return from applying AI into the archival workflows. This aspect will be further explored in the next section about the opportunities and obstacles for AI within the Swedish archival sector.

Implementing AI for born-digital material

Interestingly, the four archival institutions are currently not actively exploring AI for born-digital archival material. Instead, the digitization and accessibility of older analogue material is mainly prioritized. The Labour's Memory project forms the exception, as the time span of 1880 to 2020 includes both digitized analogue and born-digital archival materials The influx of large volumes of born-digital material naturally poses challenges to the Swedish archival institutions, but these are tackled in other ways, by engaging in closer collaborations with the archive creator.

The Swedish National Archives has not prioritized developing AI solutions for born-digital material, due to a number of reasons. The archival institution receives the most enquiries from the public in regard to the older analogue archival material compared to the born-digital. Designing AI solutions for handling these enquiries was regarded as more pressing and would result in freeing up valuable time and resources. Due to limited resources and AI expertise, increasing the accessibility and searchability of older material was prioritized. Additionally, the Swedish National Archives is in the process of developing a new strategy for the preservation of born-digital material. Once this strategy is created, AI solutions for born-digital material can be determined.

The Stockholm City Archive works closely with the regional governmental agencies it receives digital archival material from, providing detailed guidelines and instructions for how information should be stored and organized. The majority of the archival tasks are actually done by the regional governmental agencies before the delivery to the Stockholm City Archive. These tasks include the appraisal, arranging, and description of records. After the delivery the archival institution is responsible for the preservation and accessibility. The workflow is organized and divided this way because the agency is considered to be the most knowledgeable on what the records contain and can therefore process them more effectively.

Even though the Popular Movements' Archives currently does not have the possibility to receive born-digital archival material, members are already recommended to start thinking about the archiving and preservation of digital material directly after its creation. This is done in order to make future deliveries and deposits of (born-)digital material easier. The Popular Movements' Archive has created a policy document in which members are given instructions on how digital records should be stored, arranged, and named. The document also includes guidelines on sustainable formats for long-term preservation of different types of media, such as images, text documents, video, audio, presentations, and databases. (Popular Movements' Archive 2021)

The Centre for Business History prefers to work with the company before the delivery of (born-digital) material is made, so that the documents for example arrive with informative titles including the type of document, title, and year of creation. The company is asked to collect and organize the material, it deems as interesting or worthwhile to preserve. This way of working is not always possible, however, especially when the company in question is discontinued and the remaining employees of the company are not familiar with the older archival material. In these situations, the archivists of the Centre for Business History need to go through the archival material and see what it contains and what is worth to preserve, if the delivering member is willing to invest the required resources for this.

Interpretations and definitions of AI

The motivations behind the use of AI within the archival institutions are diverse. AI solutions are primarily utilized for improving the accessibility and searchability of digitalized archival material. This is done for a multitude of reasons and with different audiences in mind. Some AI solutions target researchers, by opening up new venues of exploration and research. Other applications have a strong economic dimension. This is visible in the case of the Stockholm City Archive and the Swedish National Archives where more effective handling of enquiries were prioritized in the first experimentations and implementations of AI. For these institutions, the attractiveness of AI partly stems from its ability to reduce the economic resources required for performing certain archival tasks, such as handling public enquiries.

In some cases, activities and projects archival institutions were not brought up in relation to the topic of AI by the interviewee. For example, when a project with the corresponding AI technique were mentioned in relation to making digital archival material accessible rather than the specific technique underlying it.

In line with this, the presence of shifting definitions and classifications of AI, as discussed in the theoretical framework section, are also present in the archival sector. For the Centre for Business History, it could be argued that the use of commercial book scanners counts as the organization making use of AI technology. Many archival institutions today make use of OCR technologies to digitalize documents and books, or in other words making these texts machine readable. The field of OCR has developed in such a way that archivists can do this in an automated way using commercial book scanners with pre-installed proprietary software. The question whether these digitalization efforts ‘belong’ to the field of AI depends on which definition of AI is handled. This example of OCR through book scanners means that applications that are regarded as AI today may have developed and become common place in the archival sector in the future, to the extent that they are no longer considered as belonging to the field of AI.

II. Opportunities and obstacles for AI implementation

This section addresses the opportunities and obstacles of AI implementations. Even though not all archival institutions are equally involved in AI, each interviewee identified opportunities and obstacles for their organisation. The opportunities relate to a wide range of archival tasks and processes, from the ingest of incoming archival material to the accessibility of archival material for the user. The obstacles relate foremost to the diversity and nature of historical archival material, the state of AI techniques and tools, and ownership of archival material.

Opportunities

Besides the projects and experiments with AI described in the section *Current Implementations of AI*, the interviewees at the four archival institutions identify opportunities for future AI implementations and applications. These address different parts of the archival workflows, ranging from the ingest of digital archival collections to providing access to the end-users. Compared to the current (and previous) implementations of AI, it is interesting to note that the opportunities relate to a wider variety of archival content, including both digitized and born-digital material as well as textual and audio-visual material.

An interviewee of the Swedish National Archives sees potential for AI in the ingest and migrations of public records from government agencies. An AI-based system could support archivists by checking whether the delivery of a government agencies is complete. Any missing records or collections can be identified and the system could provide answers or explanations to where these missing parts may be.

Another area where the interviewee of the Swedish National Archives identifies a potential is in the description of archival collections and records. Currently when archivists describe archival collections, they are dependent on their own level of expertise on the time period, type of archival material, and archive creator. An AI system coupled to a large database with a large number of contextual sources can give suggestions to archivists on what should go into the archival descriptions, based on the content of the collection. According to the interviewee, this can make archival institutions less dependent on the knowledge and expertise that they have within the organization. This is especially relevant for when archivists encounter rare archival materials that they have no previous experience with.

Regarding the description of images, the interviewee of the Centre for Business History identifies potential of image recognition and analysis. Their audio-visual collections include between three to four million images, which the institution will never be able to describe manually. Similar to the project of the Stockholm City Archive, the Centre would welcome the help of AI to describe the images in their collections and enrich the metadata connected to the image. The AI could be used to identify people, locations, or objects on the image. This information is then added as metadata, making the images more searchable for clients and researchers.

The digital preservation of archival material is another area in which potential of AI is identified. Compared to analogue paper material, the deterioration of digital archival material is not visible. The interviewee of the Stockholm City Archive believes AI technology can be used to design flexible preservation formats and ensure continuous conversions. Some file formats migrations and conversions may exclude unique usages of the original or previous format. Instead of converting all digital files to another format at once, for example PDF/A, the interviewee sees the opportunity of AI to provide the archivist or user archival records in different or flexible formats, depending on the intended use.

The accessibility and searchability of digital archival collections are considered as another area in which AI can be applied in the future. The interviewee of the Stockholm City Archive experiences that the current structure adopted by archival institutions based on inventories is not suitable or intuitive for users to access the archival collections. Users access archives with the idea of finding information and records on a certain theme, event, or phenomenon, and do not tend to immediately think in terms of archival or organizational structures. In addition to this, the interviewee points out the inconsistencies within the descriptions of archival institutions and their inventories within NAD, the Swedish national archival database, making navigation of archival collections more difficult. AI can be used to add metadata to the collections in a systematic way and improve the searchability.

> Sometimes archives are like this. If you want to buy a litre of milk at the store, you have to draw the code to the milk before you can get it. Because otherwise no one knows what you are talking about. They don't know milk; they just know the barcode. Sometimes archives are that way. You know, I am interested in witch hunts in the 1680s. Well tell me what archives you want to see and what series. No one thinks that way. (Stockholm City Archive 2022)

An interviewee of the Swedish National Archives thinks along similar lines, and hopes that in the future archival collections can be searched in different ways. For example, accessing different kinds of materials and formats with one search query.

> I think it would be really cool to create a system like you can search our collections and you write in cat and you get all the images of cats, you get all the archival documents where cats are mentioned and you also get all the tv material where you either see a cat on like a video clip or you hear a [meow]. (Swedish National Archives 2022)

For this to be realized, the entire archival collection would have to be indexed using AI, and connections would have to made between words so that the search word 'cat' is associated with cat-related words such as 'tabby' or 'calico', which are two different breeds of cats. This would require a combination of different AI technologies for textual and audio-visual materials.

> If you get a lot of the texts in a digitalized format, in a text format rather than an image, you can use NLP technologies to improve the searchability. You can use semantic search, you can use entity driven search, you can extract metadata from the texts and search on that metadata so basically make a google for archival material, it won't be that good but it will be more accessible. (Swedish National Archive 2022)

These opportunities brought up by the interviewees of the Stockholm City Archive and the Swedish National Archives connect to the digital transformation of society, described in the introduction and previous research. Users want to explore archival collections for records and information in similar ways as they use search engines such as Google Search, and the archival institutions are trying to meet these needs.

Lastly, the increasing availability of open-source AI models and architectures that are published in research papers and repositories is seen as an opportunity for archival institutions. Access to these resources provides archival institutions with the flexibility to modify and train models for their own AI implementations.

> [Open-source solutions are] more adaptable to a specific project because you can design it yourself. And that is usually how you work with open-source repos, you fork it to your local GPU station and then you work with it, you change somethings and keep somethings and then there are abundance of open-source code for AI solutions. (Swedish National Archive 2022)

Sharing open-source models and programs among archival institutions is crucial for the archival sector, as most institutions do not have a lot of resources available to develop models on their own. The Popular Movements' Archive plans to share the fruits of the Labour's Memory project in the hopes of being able to profit when other archival institutions are carrying out similar kinds of project:

> I think this is a really important to be generous and share knowledge. The problem is too great so we won't solve it one archive at a time. (Popular Movements' Archive 2022)

> I am personally very much a fan of open-source cause it feels especially now all archives have so little money we really need to share what we develop (…). For now we are in luck, we have this money, but it will run out sooner or later so if we share maybe someone else has luck with money in the future and they can share to us instead. (…) Especially with the HTR and OCR programs it would be great if they could be downloaded from GitHub so whoever can make use of it. (Popular Movements' Archive 2022)

Obstacles

Despite the identified opportunities for AI in different parts of the archival workflow and activities, the interviewees of the four archival institutions mentioned obstacles that have to be overcome before AI technology can realize its full potential. The obstacles relate to the character of the historical material, the ownership, security and legal concerns, state of tools, and the performance of AI technology.

Character or nature of historical archival material

The nature and quality of historical archival material is one of the pressing obstacles for current and future applications of AI technologies. Language models and training and evaluation datasets are needed for historical documents written in older Swedish. These resources are necessary for HTR and OCR applications, which in their turn form the foundation for further NLP implementations such as named entity recognition. Creating language models and datasets is not an easy task. Training and evaluation datasets have to be created based on manually annotated transcribed archival material, which requires both time and resources that most of the archival institutions do not have access to. Compared to contemporary archival material, the nature and quality of historical archival material complicates the training of language models and the creation of training and evaluation datasets.

> You need to appreciate that it is a different NLP problem, for instance creating a language model for modern texts and creating one for historical texts are different tasks. There are different difficulties with it. You can't just take the modern version and apply it to historical texts; you need to adapt to the historical aspect of it. (Swedish National Archives 2022)

In order for HTR and OCR technology to provide the best results, language models and datasets are to be trained on and specialized for the specific kind of material. AI and machine learning models do not perform well on material that it has never seen or trained on. Ideally, models are specialized for the genre or type of document. The layout may differ between types of archival material. There is for example a difference in how newspaper articles and popular records registers are constructed, which impacts the way in which the AI should be trained to look for text.

The model and datasets also need to be specialized for the language within the material. The language in historical documents changed more often than present-day Swedish does, both in terms of spelling and in terms of semantics, the meaning of the words. Having many spelling variants in the same dataset creates problems when the algorithm is executing a statistical analysis of the dataset. Spelling variants of the same word are not recognized as being the same by the algorithm. The same counts for words that have changed meaning over time. The language in historical documents also differs across genres or types of documents, which impacts the statistical probability of words appearing in the documents and probability of words appearing next to each.

> When we worked with the historical material, we see that on the one hand we need to specialize the model for the time period, but we also probably need to specialize the model for the genre of documents. If it's novels, the statistical probability of which words are most likely to be used are going to be different from when it is a legal document or if it is a document from the government agency that specializes in mining. (Swedish National Archives 2022)

The Centre for Business History receives unique archival material each time another company delivers material, with for example a different layout or structure. This may pose problems for the AI technologies to understand them. This could be a different for public archival institutions that receive similar or the same archival material each year from government agencies.

> I think there is a lot of potential where you have certain kind of documents that should be added to the archive every year, then I think machine learning can easily handle that material because they know what to look for, they know what to do with each kind of material they see. But in our case, everything we receive is so different from each other that it would be, I think (…), very hard to train an AI to know what to do with it, unless it is something it has seen before. (Centre for Business History 2022)

Lastly, the quality of historical material is lower compared to modern Swedish texts. Archival documents may contain inconsistencies or errors in spelling of names and places. The documents can have both handwritten and typed texts together so a combination of HTR and OCR would be needed to digitalize them. Then, there also may be irregularities in how information is noted down in forms or tables. An example of this last difficulty is the order in which the full name of a person is written down, in for example property books. If a certain kind of archival document has the variants Gijs Aangenendt and Aangenendt, Gijs in them, the AI model needs to be able to differentiate between the two and understand that Gijs is not a surname in the second variant.

Tools

The character of historical material and the need for specialized language models and architectures creates problems for implementing AI through tools. HTR is an area of AI which has great relevance for the archival sector. For archival institutions without much expertise on programming cloud-based tools and services exist that perform most of the programming tasks. For HTR, the Transkribus platform is the most prominent one. Although, Transkribus allows any archival institution to transcribe handwritten archival material, the design and set-up of the program poses obstacles to widespread and large-scale application in the archival sector.

According to several interviewees, Transkribus has more use cases for smaller projects and research initiatives where a limited collection of material that needs to be transcribed. For (larger) archival institutions, Transkribus can be too expensive to use if they want to transcribe millions of pages, as they are required to buy credits for each project or on a monthly basis through subscription. In terms of the design of the platform, an interviewee of the Swedish National Archives notices that the architecture behind Transkribus tool is not as advanced and does not capitalize on the most recent developments in the field of HTR and AI. Another interviewee mentions that the segmentation model of Transkribus is not suitable for the transcription of documents with tabular data.

> A problem with Transkribus for instance, they only have one segmentation method and that is done for text lines. You can't train your own segmentation model for instance if you want to analyze tabular handwritten data. (Swedish National Archives 2022)

As the source code is not fully accessible for paying members, the archival institutions cannot finetune the tool for their own archival material as much as they could when using open-source AI solutions.

Ownership

For private archival institutions, the ownership of the archival material can be an obstacle for the implementation of AI technologies. The Centre for Business History preserves archival material on behalf of their members, the companies and businesses. The collections, however, remain in the hands of the company that deposited it at the institution.

The company determines what archival tasks are performed on the collections; the level of organization, description, and accessibility of archival material depends on the wishes and willingness of the company. Their interest for their own historical material does not stretch to the application of AI and machine learning. They do not consider it worth the investment and do not perceive the financial value of using these techniques. In order for the companies to invest in AI and machine learning there needs to be some sort of benefit or economic result. The companies and their employees perceive a nostalgic value of their collections, rather than a financial value. Somehow the implementations of AI need to be financed, companies do not think it is worth it to invest in HTR techniques.

> It is hard for the company to perceive the value in that today: "Why do we need to be able to read old economical or financial reports? We can just ask you to do it if we need to see anything", which they most often don't. (---) It is more a fun thing to see than something that can

> be used in the company today. So, it has perceived nostalgic value but not a financial value. (Centre for Business History 2022)

There might be specific companies or members that are willing to have an AI train on their archival collections in the future, for example companies that are no longer functioning or active. *Brandförsäkringsverket*, Sweden's oldest insurance company, is given as an example by the interviewee. The board of directors of this company already makes their archival collections accessible online through the digital archive platform of the Centre for Business History.

> I think you would have to find some specific company or members that are willing to share some material to train an AI on. And maybe there are some, we have for example the oldest insurance company in Sweden which isn't functioning anymore. They are just a board of directors nowadays and they have already shared all their old insurances online open to anyone. (Centre for Business History 2022)

Due to the ownership of the archival material, documents cannot leave the security or premise of the Centre for Business History. Uploading documents to a cloud or webservice for converting and analyzing collections is not possible as data cannot leave the safety of the organization. It is unclear what happens with the documents and the information contained in them when they are uploaded to cloud-based services of an external party.

> But we have to be a bit careful, since the companies own their material and everything… There could be company secrets inside of them, (…) the documents can never leave our safety or our protection. I mean there are a lot of webservices (…), we can't use those because we don't know what happens with the original document. (Centre for Business History 2022)

Furthermore, archival material from different companies cannot be brought together into the same datasets for the training and evaluation of AI models.

> But again, then you would have to access the archival material of several different companies at once. And that would be even scarier from a security and accessibility point. I know some companies that would say absolutely no to something like that. (Centre for Business History 2022)

GDPR and legal obligations

Security concerns also play a role for public archival institutions. The General Data Protection Regulation (GDPR) prevents archival institutions from uploading archival material to cloud based or web based computational tools and services. This counts especially for born-digital material that contains personal information of citizens that are still alive. For these archival documents, AI technologies have to be implemented on the premise of archival institution. Public archival institutions need to have a legal basis, called *rättslig grund* in Swedish, for implementations of AI and machine learning, where there is a difference between developing a prototype and actual implementation.

Performance of AI technology

Expectations of AI may not always align with the performance of the technologies in reality. It is costly to develop an AI system, and it needs to work in order for it to be economic valid. Awareness is needed of what AI can and cannot do as well as realistic expectations and knowledge about the possibilities.

> I think really to be honest when we started with trying this AI thing which was about three years ago, we were more optimistic how fast it could work. Because it is not that fast really. I thought you know, when you program that you get it back immediately. No, it takes several hours to process and as I am not very technical, I have not really dwelled into why it works so slowly. (Stockholm City Archive 2022)

III. AI and the archival profession

The impact of the digital transformation of society and the archival sector is visible in the activities and workflows of all four archival institutions. Using the archive-as-data paradigm as theoretical lens, this section will address the question how AI will impact the archival profession. It discusses the roles, skills, and responsibilities archivists need to assume and acquire in the digital day and age of archiving.

Computer and programming skills of archivists

The digitization of the archival sector means that archivists have to develop new technical and programming skills related to the management of electronic archives and digital preservation. Data science concepts and methods have become more prominent within the studied Swedish archival institutions, even in the institutions not actively working with AI. The interviewees highlight the need for the development of new digital skills related to handling the long-term preservation of digital archival materials, such as finding suitable formats and performing file conversions.

> You have to learn a lot about formats when working with digital material. What works in the long term, what is an acceptable format for digital material. (…) a digital document has to be converted to a proper format and then saved in a proper way. That is definitely a skill you have to learn. (Centre for Business History 2022)

> I have studied some data science while at the Centre for Business History to complement my archive education. (…) It was mostly about file formats and ways of handling data (…) programming languages and stuff like that. (Centre for Business History 2022)

Acquiring the necessary skills to make use of available software for the preservation of digital archival material can be challenging. Further education is a possibility for archivists to acquire certain required skills, but the learning curve can be steep and includes experimentation and self-learning.

> So suddenly I need to write a PHP script, I need to know how to handle python environments. I need to be able to do Linux and bash, and I need to install these things on the servers. (Popular Movements' Archive 2022)

> We are in this moment in time it feels like a lot software (…) feels like it is made by technicians for technicians and it is really hard if you are not an IT person to use it. (Popular Movements' Archive 2022)

There are programs and software for the archival sector available that take part of the work out of the hands of the archivists, although there can be risks associated with this. The interviewee at the Popular Movements' Archive warns for relying solely on software and advocates for acquiring knowledge within the organization on the inner workings of the programs upon which the software is build.

> When you have the program doing all that [identifying file types, migrating files, G.A.], you don't really need to know what it is actually doing. I understand the attraction of "one-click archive solutions" since it demands way less from the archivists, but I think not understanding what is going on "under the hood" so to say is dangerous from a preservation perspective. Digital archivists can't just delegate all responsibility to a machine and call it a day. Instead we need to monitor the changing landscape of digital preservation, collaborate with other professionals in our field and create and update our preservation policies that in turn govern how our programs for digital archiving process and preserve incoming data. (Popular Movement's Archive 2022)

In terms of AI implementations, the question is how much an archivist should know about AI. Mordell (2019), Theimer (2018), and Ranade (2016) all emphasize the need for archivists to have an understanding of AI, both to explain and presents its outputs to archival users and be aware of the ethical consequences of the use of AI, such as potential biases in datasets.

The interviewees of the Swedish National Archives considered it beneficial if archivists have conceptual knowledge of the current state of the field of AI and the implementation possibilities of the common AI technologies within the archival sector. Furthermore, the importance of regular refreshing knowledge is considered important as the AI field develops fast. Archivists with a solid understanding of AI are important to understand how certain technologies train on datasets and foresee and handle the appearance of biases in the data.

> What is really important for them to understand how the most common AI techniques work cause they need to understand how they use data otherwise they wouldn't understand what potential biases and what risks they could get so you need to understand the biases and how the model is trained so you can correctly interpret the output, the suggestions, the predictions that you get. (Swedish National Archives 2022)

Acquiring knowledge on AI implementations and technologies, however, can be difficult according to the interviewees of the Swedish National Archives. Finding resources on AI that are non-technical can be hard to find and recommending technical works may backfire and spark resistance against the arrival of AI. An interviewee from the Swedish National Archives, therefore, advocates for the distribution of knowledge through demonstrations and interactive workshops.

Having knowledge on AI would facilitate the collaboration between archivists and data scientists. Dialogues between archivists and data scientists at the Swedish National Archives focus on how certain tasks in the archival workflow and processes are currently performed and how AI or machine learning could be used to support the archivist during this. This dialogue is important for the data scientist to understand how the archivist works, how they access the archival material, how the data or archival material is constructed, and how the archivist would like to work with the archival material in an ideal world.

> I think the most important parts are how do you work now, how do you access the data, what is the interface and how would you want it to work in an ideal world because then you can start thinking about the problem and approximate a solution: 'what could AI and machine learning do to facilitate this?' (Swedish National Archives 2022)

Another aspect of discussion is in what parts of the implementation process the archivists should be involved. As became apparent in the discussing of the opportunities and obstacles for AI, not every archival institution will be able to afford to hire its own data or computer scientist. At the same time, recent developments in programming tools and AI and machine learning software libraries allow archivists to implement AI technologies on their own, through self-study and learning, using their personal computer. (Lee 2019) This raises the question what the differences are between a programming archivist and an archiving data scientist.

One of the interviewees of the Swedish National Archives acknowledges the availability of useful software for archivists to experiment with. The NLP framework spaCy is mentioned as relatively user friendly with many use cases within archives. On the other hand, having someone with a background in data science in the archival institution opens up more opportunities than only relying on a programming archivist according to one of the interviewees of the Swedish National Archives. The data scientist would have an understanding of the theoretical foundation and a strong familiarity with the key concepts of the field. This knowledge can be used to identify the possibilities of newly developed AI architectures published in recent research papers or open-source repositories, and then having the practical know-how necessary to apply these architectures in a new research project and adapt it to the archival context.

> I think the difference perhaps is that you often aim towards practical applications straight away instead of grounding the theory when you self-study which means you have a little less understanding of the concepts underlying the practical applications which means you don't have the same ability to see possibilities like if you learn how to use a framework to train a model without learning the architecture of the model for instance. (Swedish National Archives 2022)

There is a risk that the archivists and thus the archival institution remains working with tools, software and AI architectures that they are familiar with and take less advantage of the latest advancements in the field compared to a data scientist.

> You don’t have the same ability to choose the frameworks you use and to judge which is ones are the most promising for this particular project, for this particular data. (Swedish National Archives 2022)

At the same time, an interviewee of the Swedish National Archives, argues that archivists and other information specialists bring experiences to the table valuable for AI innovations. The potential of these people may even be under appreciated in projects where AI is implemented. They know what kind of data or information the organization is sitting on and improve the reusability of existing data. This counts for archival institutions, but also governmental agencies trying to implement AI in areas unrelated to archiving.

> [Archivists] are the ones who understand what information you already have in the organization, they are the ones who understand what biases there are, what kind of flaws there are in the data and differences in like how the data was created and what it contains during different time periods. That is really important information to have if you are training AI models based on previous data. (Swedish National Archives 2022)

> And also, it is really important to have archivists in the projects because the potential for data reuse is much higher and creating new training data is always what is most costly in these projects so it is like if you have an archivist in the project, you could start using metadata from the systems that you have in place already. (Swedish National Archives 2022)

One of the interviewees of the Swedish National Archives highlights the importance of ‘two-way street’ interactions between archivists and AI specialists. Both parties need to understand the potential of each other. Without proper dialogues between archivists and AI experts/data scientists, the success of AI implementations projects can be negatively impacted.

> So, I think it is like on the one hand the archivists need to understand some about the AI, but also the AI people need to understand the potential of archivists and information professionals cause right now there are not really talking to each other in most agencies and that is holding these projects back really. (Swedish National Archives 2022)

Human versus machine

The arrival of AI also raises discussions on impact of automation on the archival profession and the difference between human archivists and machines. The core roles and responsibilities of archivists and information professionals are not directly in danger according to an interviewee of the Swedish National Archives, as AI will not be able to execute tasks hundred percent correctly every time, mainly due to its performance capabilities and the inconsistent quality of historical data. Archivist and AI, or human and machine, therefore need to work in a tandem: with archivists performing the final check and the AI technology providing options in case of uncertainty. AI systems have the ability, for example, to present multiple alternatives, arranged on the basis of so-called confidence scores, a percentage indicating the likelihood of that particular answer being the right one.

> We use AI for indexation and for smarter searches but we still have our staff look through the material and say okay yes this is correct, send, cause I don't think it would be secure for us to automate it completely. I think AI is too stupid to work that out. There are too many mistakes in the archive, the information quality in some places is too low so I don't think it would be safe to do that. (Swedish National Archives 2022)

> We are creating a program to even if the AI is not certain, it is going to give suggestions to the people who are indexing. So instead of writing every letter, from scratch, they just need to pick from a list. (Swedish National Archives 2022)

The automation of certain tasks does not mean that archivists are no longer needed or that the archival profession is disappearing in the long term. The arrival of AI changes the nature of the profession, by introducing collaboration with machines. An interviewee of the Swedish National Archives emphasize that AI implementations can free up resources, that can instead be invested in more traditional archival tasks and duties, that archival institutions are not able to otherwise.

> The National Archives is pretty understaffed (…) and we get tens of thousands of enquiries from the public every year. Just when it comes to the real estate, we get like 33.000 enquiries every year. If we could automate (…) the information retrieval for those, it would mean that we don't need to put staff towards working with answering questions from the public. (Swedish National Archives 2022)

In terms of the differences between human and machine, the interviewee of the Stockholm City Archive states that AI systems have an advantage over human archivists in performing repetitive archival tasks that take a long time to complete. For these tasks, the AI can be more accurate, for example in finding school records.

> If I have a bad day and I am going to find your grades and I don't find them I tell you well they are not here. But (…) if you get that answer from the AI, it is probably more right than if you get it from me. (Stockholm City Archive 2022)

Discussion

In this section, the findings from the analysis of the four archival institutions are summarized and discussed in relation to the research questions, the current research field, and the theoretical framework, namely the definitions of AI, the archive-as-data paradigm, and the ethical considerations. In addition, important themes and observations are lifted on the design of AI tools for the archival sector and the nature of the Swedish archival sector versus the international archiving community and cultural heritage institutions.

The AI landscape in the Swedish archival sector

This study has analysed the experiences with and outlook on AI of archivists and other professionals from four different archival institutions. The level and depth of engagement with AI techniques and tools varies between the archival institutions studied, with the Swedish National Archives and the Centre for Business History placed on opposite sides of the spectrum.

Current and future implementations of AI

A range of AI applications are explored by the Swedish archival institutions, that combine the power of machine/deep learning algorithms with AI techniques from the field of computer vision and NLP. The identified applicability of AI addresses many parts of the archival workflow and practices of the four archival institutions. The AI techniques currently used and tested relate to the 1) digitalization of handwritten and typed documents, 2) accessibility and searchability of archival material and images for the general public, researchers and archivists, and the 3) sensitivity review of the content of documents. Opportunities for future AI implementations relate to 1) the ingest or migration of archival records from government agencies to the archival institute, 2) the description of images and (rare) documents, 3) the preservation of digital records and information, and 4) the improvement of the accessibility and searchability of archival collections in the electronic archives.

The identified AI techniques and application areas align with the interest of other cultural heritage institutions, as studied in the Europeana (2021) report and Cordell (2020) analysis of the potential for (primarily) libraries. Compared to the potential AI implementations within archives discussed in research field, however, the current usage of AI within the Swedish archival sector most often relates to analogue archival material that has been digitalized instead of born-digital material, such as e-mails. Examples are nineteenth and twentieth century photographs, digitized high school grades, popular records, real estate books, and annual and financial reports of local labour unions.[2] This can partly be explained by the interest of the general public and researchers in digitized archival collections and the issues of personal data protection when it comes to born-digital material. In regard to born-digital collections, a potential of AI was recognized by the interviewees. The current solution for effective processing of large volumes of born-digital records, on the other hand, seems to be the strengthening of the connection with archive creator. All four archival institutions are, or strive to be, more involved from the moment born-digital records and information are created, to ensure sustainable preservation in the long run. As argued by Hansen & Sundqvist (2016), this is more easily done for public archival institutions in the case of archive creators and public records due to the existence of a legal framework.

In short, the findings on practical AI implementation show that the four archival institutions identify a potential for AI techniques to be implemented in the archival workflows and practices now and in the future. The transition from experiments to permanent use of AI described by Colavizza et al. (2021), however, has yet to materialize in the Swedish archival sector. In order to use AI on an operational scale, archival institutions need to have access to a developed electronic archive, large collections of digitalized documents and/or images, suitable AI architectures and language models for the material, and the technical knowledge and expertise to implement AI in a sustainable and ethical way. The Swedish National Archives is the closest to realizing this, as they are in the process of creating and developing the building blocks needed for structural AI implementations.

Obstacles and challenges for AI implementation

The study has identified several key obstacles that prevent the archival institutions from meeting the requirements for implementing AI on a larger scale. Many of these obstacles and challenges align with obstacles other archival institutions experience abroad and institutions of the European cultural heritage sector (Rolan et al. 2019; Europeana 2021). Obstacles relevant for the entire cultural heritage sector relate to the access to and quality of (historical) data, the construction of general (training and evaluation) datasets, the right AI/machine learning expertise, and the need for personal data protection. Having said that, there are obstacles surfaced from the analysis that specifically relate to the Swedish archival context, mainly relating to the division of public and private archival institutions. Private archival institutions, such as the Centre for Business History and the Popular Movements' Archive, do not possess ownership over all their archival material. A large part of the material is deposited at the institutions instead of transferred. In the case of the Centre for Business History companies exercise strict control over what archival tasks are performed on the collections. They would determine whether the Centre would be allowed to apply AI tools and techniques on their collection, and this is

[2] The Labour's Memory project includes both digitized and born-digital archival material.

currently not the case, due to a combination of financial and security reasons. Lastly, it should not be forgotten that (most) Swedish archives have limited financial resources, especially the private ones. Prioritizations have to be made and considering the steep initial investments of AI archival institutions need to be sure it would a worthwhile in the long run. Exploring AI and machine learning possibilities is not immediately the first item upon the list, when archives are occupied with setting up sustainable electronic archives. Even the Swedish National Archives needs to prioritize AI applications that would result in the most return of investment.

AI and the changing archival profession

The third part of the analysis and study section focused on AI and the changing archival profession. The archive-as-data paradigm provided a lens through which the changing Swedish archival profession could be studied. The findings show that Swedish archival profession is becoming increasingly dependent on technology as observed by Underwood et al. (2018). The interviewees experience a need for using computational tools and acquiring computational and programming skills related to the long-term preservation of digital archival material. This speaks for the increasing importance of data science and computer science concepts and methods and concepts in the archival sector. In regard to the implementation of AI, archivists are not designing their own solutions. Instead, they interact with AI either through an interface, for example Transkribus, or with the output of an AI system that has been created by AI experts inhouse or externally via universities.

Understanding of AI

Archivists are expected to have a conceptual understanding of AI and a familiarity with common techniques applicable for the archival sector. This way archivists can contribute to the implementation of AI in a sustainable and ethical way. Strengths of archivists include the expert knowledge they possess on the hidden gems of historical data within archival collections and the potential limitations and biases of this data. Furthermore, the archivist can facilitate to the reuse of data, potentially reducing the overall costs of AI projects.

Collaboration in the AI landscape

Research indicates that AI implementations and projects within archives and the broader cultural heritage sector require interdisciplinary collaboration between various stakeholders and actors (Marciano et al. 2018; Jo & Gebru 2020) The AI projects and experiments within the Swedish archival sector are not just conducted by archivists within the confines of the individual archival institutions. A wide range of actors are involved in the various stages of AI experiments and projects.

The archival institutions collaborate with cultural heritage institutions, from other archives to libraries and museums. The Stockholm City Archive collaborates with the Stockholm City Museum in the City Faces project, as both cultural heritage institutions have large collections of nineteenth and twentieth century photographs. Within Labour's Memory, the Popular Movements' Archive is partnering with the Swedish Labour Movement's Archives and Library, but also two international archival institutions from Germany and the Netherlands. Together these four institutions plan to present archival material of the Swedish, German, and Dutch labour unions and movements in a shared digital database. The Swedish National Archives has an important partner within Swe-Clarin in the form of the Swedish National Library in the development of language models and resources.

Universities are important partners in the implementation of AI within the archival sector. Within Swe-Clarin, the Swedish National Archives collaborates with various linguistic departments of several Swedish universities. For the Stockholm City Archive and the Popular Movements' Archive, the universities provide the knowledge and expertise needed for AI implementations. The departments of computer science, data science, and linguistics bring the knowledge and know-how of computational methods and AI/machine learning fields such as computer vision and NLP. These institutions are dependent on the researchers in order to accomplish the goals of the project. This dependency on external knowledge can pose problems. In the case of Labour's Memory, the data scientist from Uppsala University took on another job elsewhere. The vacancy could not be filled immediately, as there are not many with the right expertise. A replacement is expected to arrive in 2023.

Citizens or archival users contribute to the development of AI within archival institutions through crowdsourcing. Citizen scientists help the Stockholm City Archive to manually transcribe handwritten documents through the interface of Transkribus. These transcribed texts form dataset for the training and evaluation of the machine learning algorithm. Afterwards, the citizen scientists can also help with identifying and correcting errors in the automatic transcriptions. Similarly, the Swedish National Archives and Swe-Clarin call upon the network of volunteers connected to the Gothenburg University for the compiling of evaluation datasets for historical texts useful for implementations of various NLP techniques.

Lastly, public archival institutions collaborate with other governmental agencies. The City Archive developed an AI solution for the automated retrieval of school records together with another regional agency within Stockholm County. The Swedish National Archives is included in the eSAM network, where various large governmental agencies discuss and exchange experiences on issues facing the public sector related to digitization and AI.

AI solutions for the archival sector

The studied archival institutions make use of AI tools that require different levels of skills and expertise. As became apparent from the interviews, AI tools are used for the transcription of handwritten textual documents. The different application of HTR in multiple archival institutions gives the possibility to reflect on the design of AI solutions and tools. Hutchinson (2020) outlines five design requirements of NLP tools for archival purposes. These can be used as a starting point to review the HTR tools used by archival institutions.

- Usable: a tool that aligns with the technical expertise of the user, the archivist, responsible for carrying out the archival task.
- Interoperable: a tool that incorporates functionalities fitting for different parts of the workflow and offers the ability to share dictionaries, training models, and resources across collections and institutions.
- Flexible: a tool that allows the user to view the results on various levels (i.e. document and corpus level) and export the results for further independent processing and analysis.
- Iterative: a tool that has the ability to refine models, identify and correct false positives and missed labels.
- Configurable: a tool that allows customization in terms of how the model is applied, control over the level of pre-processing, and is applicable for various NLP tasks.

Three archival institutions have been using HTR in some capacity. At the Swedish National Archives, HTR is a part of the general pipeline for automated indexation of archival material such as property books and popular records. The Stockholm City Archive has been working with Transkribus. The Popular Movements' Archive experimented with Transkribus and are now developing specialized HTR models in collaboration with researchers from Uppsala University within the Labour's Memory project.

While the scope of this study does not allow a thorough analysis of the Transkribus platform, the experiences of the archival institutions with HTR do give some insight in how the platform meets the design requirements for use in archival settings. In terms of the configurability of Transkribus, it has been noted by the interviewees that the tool only has one segmentation model that does not function as accurately on handwritten documents with a tabular structure. In terms of interoperability, Transkribus users have no access to the source code which therefore cannot be used in other contexts. For the requirement of usability, Transkribus is user-friendly and aligns with the technical expertise of the archivist. Another issue with Transkribus was the number of transcribed texts needed to train the model, which the Popular Movements' Archive could not realize.

The Swedish National Archives' pipeline for indexation through OCR and HTR has the benefit of having a modular structure. If technological innovations in the field have resulted in improvements of certain elements of the pipeline, the new and improved elements can be fitted in. Similarly, if the organization wants to use a certain part of the pipeline, for example the segmentation section, this can easily be transferred to another AI application. High level of expertise and knowledge is necessary to build pipelines and maintain them, not possible for every archival institution to do this.

The aim of the Labour's Memory project is to create an interface through which archivists can perform HTR and OCR preparation and tasks. Middle way between the Swedish National Archives and the Stockholm City Archive. Take advantage of the newer improvements of the field while also having a user-friendly interface that archivists can use. However, once the project is over, the question remains if and how the interface and underlying codes are kept up-to-date. The organization most likely will not have the technical knowledge necessary to make improvements to the architecture on their own and exchange sections of code that have stopped working or are no longer supported.

Public versus private archival sector

With both public archival institutions and private archival institutions represented included in this study, it is possible to reflect on how the character of the Swedish archival sector impacts the ability of archival institutions to implement AI. Public and private institutions differ in multiple regards.

Firstly, the two differ in terms of the character of the archival collections. Private archival institutions were originally created to document the parts of society that were not clearly represented in the public archival institutions, such as the labour unions, businesses, and various associations. While public archival institutions have started to actively collect archival material from private actors, the majority of records is received from public agencies and companies.

Second, there might be a difference in terms of access to (economic) resources. Compared to their public counterparts, private archival institutions generally have less resources available. Income is generated through membership fees, shelve fees, and funding from public agencies.

Thirdly, the private and public archives differ in terms of target audiences and ownership of the archival material. The collections at private archival institutions, often, remain in the hands of its original owner, the member: the association, labour union, or the company. Therefore, archival tasks are performed for the owner instead of the general public and researchers. The ownership of the material can be an obstacle for private archival institutions to make use of the potential of AI and collaborate with other actors in the AI landscape.

There are serious risks connected to the implementation of AI if private archival institutions are not resourceful enough, in terms of economic and human resources, to train their own models on archival collections. Models only trained on archival material from public archival institutions, such as the Swedish National Archives, can lead to a bias. The different parts of the Swedish archival sector make it so that the nature of the material differs between the institutions. The structure and nature of archival material can differ across the sections of the archival sector, so if the Swedish National Archives want to take a leading role in the development of open-source AI solutions for the archival sector, considerations have to be made if the material upon which the models are trained are suited for the archival material of private archival institutions. If possible, models could be trained on material from private institutions in the future. The question of ownership, however, complicates this potential collaboration between public and private archival institutions.

Conclusion

The aim of the thesis has been to survey and analyze the AI landscape within the Swedish archival sector. Through the analyses of four archival institutions and the semi-structured interviews with archive professionals, this study has illuminated the current state of AI, opportunities and obstacles of AI implementations, and the impact of AI on the Swedish archival profession.

Based on the case studies and discussion, it can be concluded that there is an interest in and high potential of the use of AI within the Swedish archival sector. The AI efforts identified are currently in an experimental and project phase, with the Swedish National Archives being the closest to operational use of AI. As this study has shown, multi-disciplinary collaboration and knowledge exchange with stakeholders are two important aspects in overcoming obstacles of AI implementation in the Swedish archival sector. The increasing use of computational methods and tools based on AI will require archive professionals to acquire new knowledge, skills, and responsibilities. The traditional strengths of archivists remain relevant in the AI collaboration, for ensuring sustainable and ethical implementations.

Based on the findings of this thesis, some suggestions can be made to accelerate the use and implementation of AI in the Swedish archival sector. To move from project and experiments to more structured implementations of AI, it is firstly recommended that AI resources and tools are adapted to the diversity of Swedish archival material and made available open-source. This is important to ensure that all archival institutions, regardless of type and access to resources, can learn and build upon each other's expertise. The Swedish National Archives could take a leading role in the dissemination of AI resources. Secondly, archival institutions should develop strategies and identify the key areas for AI implementations. This way, the fruits of project-based collaborations with AI experts and researchers can be engrained in the organization on the long term. Lastly, archival institutions are to systematically support their archive professionals to become proficient in AI for archives to ensure sustainable and ethical services to stakeholders and society.

Due to the design and scope of the study, a limit of the number of interviews could be carried out with archivists and other information/data specialists at each institution. Future research can broaden the understanding of the status of AI in the Swedish archival sector further in various ways. Partners from the cultural heritage sector, higher education institutions, and the public sector played a role in all of the projects described in this study to various extents. The perspective of these actors has not been included in the scope of the study and therefore also not in the data collection. A future study could illuminate the network of archival institutions for AI collaborations more clearly. A study of this kind could include the members of the Swe-Clarin research infrastructure such as the Swedish National Archives and

Library and the (computational) linguistics departments of universities. Another direction would be to focus on the collaboration and knowledge exchange between governmental agencies and public archival institutions. As the archival duties of the agency and the archival institution become more intertwined there might be AI initiatives connected to archival tasks performed within governmental agencies in the future.

In order to get a deeper understanding of the status of AI in the Swedish archival sector, an in-depth study targeting a specific AI project within an archival institution can be useful to study the interaction between archivists, IT professionals, researchers and computer and data scientists more closely. Such a study could utilize a mixed-method approach, with reoccurring interviews and observations during the course of the project. As the study focuses on the knowledge exchange and communication between stakeholders in AI projects, suggestions can be made to enhance the way AI implementations are designed, organized, and executed.

Bibliography

In author's possession

Transcription of interview with the City Archive Stockholm, 2022-03-17.
Transcription of interview with the Swedish National Archives, 2022-03-22.
Transcription of interview with the Popular Movements' Archives, 2022-03-28
Transcription of interview with the Centre for Business History 2022-03-30.
Transcription of interview with the Swedish National Archives, 2022-04-01.

Literature

Bocyte, Rasa & Oomen, Johan (2020), "Content Adaptation, Personalisation and Fine-grained Retrieval: Applying AI to Support Engagement with and Reuse of Archival Content at Scale", *Proceedings of the 12th International Conference on Agents and Artificial Intelligence,* pp. 506-511.

Borin, Lars, Tahmasebi, Nina, Volodina, Elena, Ekman, Stefan, Jordan, Caspar, Viklund, Jon, Megyesi, Beàta, Näsman, Jesper, Palmér, Anne, Wirén, Mats, Björkenstam, Kristina, Grigonytė, Gintarė, Capková, Sofia & Kosiński, Thomasz (2016), "Swe-Clarin: Language Resources and Technology for Digital Humanities", *Extended Papers of the International Symposium on Digital Humanities,* pp. 29-51.

Bunn, Jenny (2020), "Working in contexts for which transparency is important: A recordkeeping view of explainable artificial intelligence (XAI)", *Records Management Journal* 30 (2), pp. 143-153.

Chowdhury, Gobinda (2003), "Natural language processing", *Annual Review of Information Science and Technology* 37, pp. 51-89.

Creswell, John & Poth, Cheryl (2017), *Qualitative Inquiry and Research Design. Choosing among five approaches*, Thousand Oaks: SAGE Publications, Fourth Edition.

Cook, Terry (2013), "Evidence, memory, identity, and community: four shifting archival paradigms", *Archival Science* 13, pp. 95-120.

Elragal, Ahmed & Päivärinta, Tero, (2017), "Opening Digital Archives and Collections with Emerging Data Analytics Technology: A Research Agenda", *Tidsskriftet Arkiv* 8 (1), pp. 1-15.

Esteva, Maria, Xu, Weijia, Tang, Jeffrey Felix & Padmanabhan, Karhik Anantha (2014) Data mining for "big archives" analysis: A case study, *Proceedings of the American Society for Information Science and Technology* 50 (1), pp. 1-10.

Colavizza, Giovanni, Blanke, Tobias, Jeurgens, Charles & Noordegraaf, Julia (2021), "Archives and AI: An Overview of Current Debates and Future Perspectives", *Journal on Computing and Cultural Heritage* 15 (1), Article 4, pp. 1-15.

Cordell, Ryan (2020), *Machine Learning + Libraries. A Report on the State of the Field*. Commissioned by LC Labs, Library of Congress. 14 July 2020.

Galletta, Anne (2013) *Mastering the Semi-Structured Interview and Beyond. From Research Design to Analysis and Publication*, New York: New York University Press.

Goodfellow, Ian, Bengio, Yoshua & Courville, Aaron (2016), *Deep Learning,* Cambridge, Massachusetts: MIT Press.

Goudarouli, Eirini, Sexton, Anna & Sheridan, John (2019), "The Challenge of the Digital and the future Archive: Through the Lens of The National Archives UK", *Philosophy & Technology* 32, pp. 173-183.

Gray, Lisa, Wong-Wylie, Gina, Rempel, Gwen & Cook, Karen (2020), "Expanding Qualitative Research Interviewing Strategies: Zoom Video Communications", *The Qualitative Report* 25(5), pp. 1292-1301.

Hutchinson, Tim (2020), "Natural language processing and machine learning as practical toolsets for archival processing", *Archival Processing* 30 (2), pp. 155-174.

Irani, Elliane (2019), "The Use of Videoreferencing for Qualitative Interviewing: Opportunities, Challenges, and Considerations", *Clinical Nursing Research* 28(1), pp. 3-8.

Jaillant, Lise (2022), “Introduction”, in *Archives, Access, and Artificial Intelligence, Working with born-digital and digitized archival collections,* ed.: Lise Jaillant, Bielefeld: Bielefeld University Press, pp. 7-28.

Jordan, Michael & Mitchell, Tom (2015), “Machine learning: Trends, perspectives, and prospects”, *Science* 349(6245), pp. 255-260.

Jordan, Michael (2019), “Artificial Intelligence – The Revolution Hasn’t Happened Yet”, *Harvard Data Science Review* 1 (1), pp. 1-9.

Kern, Daria, Zweng, Manuel, Sello, Stanley, Bagula, Antoine, & Klauck, Ulrich (2020), “Archiving 4.0: Application of Image Processing and Machine Learning for the Robben Island Mayibuye Archives”, 2020 International SAUPEC/RobMech/PRASA Conference, pp. 1-6.

Lee, Benjamin (2019), “Machine learning, template matching, and the International Tracing Service digital archive: Automating the retrieval of death certificate reference cards from 40 million document scans”, *Digital Scholarship in the Humanities* 43(3), pp. 513-535.

Marciano, Richard, Lemieux, Victoria, Hedges, Mark, Esteva, Maria, Underwood, William, Kurtz, Michael & Conrad, Mark (2018), “Archival Records and Training in the Age of Big Data”, in *Re-Envisioning the MLS: Perspectives on the Future of Library and Information Science Education; Volume 44B*, eds.: Johanna Percell, Lindsay Sarin, Paul Jeager & John Bertot Percell, Bingley: Emerald Publishing Limited, pp. 179-199.

Meehan, Jennifer, (2014) “Arrangement and description: between theory and practice”, in: *Archives and Recordkeeping. Theory into practice*, ed. Caroline Brown, London: Facet Publishing, pp. 63-100.

Millar, Laura (2017), *Archives. Principles and practices*, London: Facet Publishing, Second Edition.

Mordell, Devon (2019), “Critical Questions for Archives as (Big) Data”, *Archivaria* 87, pp. 140-161.

Moss, Michael, Thomas, David & Gollins, Tim (2018), “The Reconfiguration of the Archive as Data to be Mined”. *Archivaria* 86, pp. 118-151.

Ranade, Sonia (2016), “Traces through Time. A probabilistic approach to connected archival data”, *2016 IEEE International Conference on Big Data (Big Data),* pp. 3260-3265.

Rolan, Gregory, Humphries, Glen, Jeffrey, Lisa, Samaras, Evanthia, Antsoupova, Tatiana & Stuart, Katharine (2019), ”More human than human? Artificial intelligence in the archive”, *Archives and Manuscripts* 47 (2), pp. 179-203.

Stanford University (2016), *Artificial Intelligence and life in 2030; One hundred year study on artificial intelligence*, Report of the 2015 Study Panel, September 2016.

Thibodeau, Kenneth (2016), “Breaking Down the Invisible Wall to Enrich Archival Science and Practice”, *2016 IEEE International Conference on Big Data (Big Data)*, pp. 3277-3282.

Underwood, William, Weintrop, David, Kurtz, Michael & Marciano, Richard (2018), “Introducing Computational Thinking into Archival Science Education”*, 2018 IEEE International Conference on Big Data,* pp. 2761-2765.

Young, Tom, Hazarika, Devamanyu, Poria, Soujanya & Cambria, Erik (2018), “Recent Trends in Deep Learning Based Natural Language Processing”, *IEEE Computational Intelligence Magazine* 13(3), pp. 55-75.

Reports

Cordell, Ryan (2020), *Machine Learning + Libraries. A Report on the State of the Field*. Commissioned by LC Labs, Library of Congress. 14 July 2020.

European Commission (2021) *Proposal for a Regulation of the European Parliament and of the Council. Laying down harmonized rules on artificial intelligence (Artificial Intelligence Act) and amending certain union legislative acts* 2021/0106, 2021-04-21.

Europeana Network Association (2021) *AI in relation to GLAMs Task Force. Report and recommendations*.

Ministry of Enterprise and Innovation (2018), *National approach to artificial intelligence,* article number N2018.14.

Swedish National Archives (2018) *Arkivutredningens frågor till arkivsektorn*, 2018-11-23.

Swedish National Archives (2019) *Nationell arkivdatabas*, *NAD*, 2019-06-04.

Statens Offentliga Utredningar (2019) *Härifrån till evigheten. En långsiktig arkivpolitik för förvaltning och kulturarv. Betänkande av Arkivutredningen*. SOU 2019:58.

Popular Movements' Archive (2021), Formatpolicy för långsiktigt digitalt bevarande [Formatpolicy for long-term digital preservation] 2021-11-16.

Webpages

Stockholm City Archive n.d. Stockholm City Archive website > Start > Visningar & event > Gå på event > Renskriv 1700-talet [2022-04-25].

Stockholmia 2022, *Stadens ansikten*, Stockholmia - forskning och förlag, viewed 24 April 2022, < https://stockholmia.stockholm.se/forskning/projekt/stadens-ansikten/>.

Swedish Labour Movement's Archives and Library (2020), *SEK 12 million to digitize trade union's annual reports*, Swedish Labour Movement's Archives and Library, viewed 15 May 2022, <https://www.arbark.se/en/2020/11/sek-12-million-to-digitize-trade-unions-annual-reports/>.

Transkribus n.d. *Transkriberade handskrifter. Stadsarkivet*, Transkribus, viewed 25 April 2022, < https://transkribus.eu/r/stockholm-city-archives/#/>.

Appendix: Interview Guide

Interview with archivist:

- Could you briefly introduce yourself?
    - Educational background and position at the organization

- How has the arrival of digitized and born-digital archival records impacted your organization and your role as archivist?

- How would you define AI and machine learning in the context of archives?

- How is your organization currently working with AI and machine learning?

- How do you look at the potential for integrating AI and machine learning into the archival workflow and processes?
    - Are there any parts of the organization, i.e. workflow and archival tasks, where you see the most potential?

- Do you identify any obstacles for integrating AI and machine learning into the archival workflow and processes?
    - What kind of obstacles, and in what stages of the workflow?

- How will the use of AI and machine learning impact your organization?
    - Workflow, archival concepts, archival profession

- Do you have anything to add?

Interview with data scientist:

- Could you briefly introduce yourself?
    - Educational background and position at the organization.

- How did you get interested in working with archival data as a data scientist?

- What (additional) skills and knowledge should a data scientist possess when working with historical data within an archival context?
    - Did you need to adjust to working with historical texts and data?

    - Does the nature of historical data require another approach?

- What skills and knowledge should an archivist develop or possess when working with historical data and AI/machine learning?

- How can the mutual understanding between archivists and data scientists be improved?

- How can the archival sciences and data sciences benefit from each other?

- Do you have anything to add?