Proceedings from the Document Academy

Volume 9
Issue 2 *Proceedings from the 2022 Annual Meeting of the Document Academy*

Article 9

2022

Artificial Intelligence and the Preservation of Historic Documents

Gaute Barlindhaug
University of Tromsø, gaute.barlindhaug@uit.no

Recommended Citation
Barlindhaug, Gaute (2022) "Artificial Intelligence and the Preservation of Historic Documents," *Proceedings from the Document Academy*: Vol. 9 : Iss. 2 , Article 9.
DOI: https://doi.org/10.35492/docam/9/2/9
Available at: https://ideaexchange.uakron.edu/docam/vol9/iss2/9

In recent decades, different digitalization strategies have been presented as effective methods for both preserving and creating better accessibility to historical materials. This has also been the case in Norway, where the National Library has embarked on a large project, digitizing not only all printed matter but also as much photography, analog tapes, video cassettes and movies it can get its hands on. As part of such endeavors, several memory institutions have examined the possibilities for using Artificial Intelligence (AI) and Machine Learning (ML) to enhance our access to and understanding of the digitized material. Digital tools present a range of technical possibilities for interpreting and organizing large collections of digital information.

The aim of this text is not to go into the technical details of such technology, but rather look at how it is proposed to be used in the context of the Norwegian National Library in ways that enforce political and cultural ideas about cultural heritage preservation. From the archival community, criticism has been raised towards many attempts to digitize historic material, both concerning the technical quality of the result, but also the cultural impact of large-scale digitalization projects. One aspect is that of "digitization"—converting analog material into digital formats—but there are also questions concerning "digitalization" of the process of safekeeping and presenting of cultural heritage. Digitalization of cultural heritage preservation can impact which material we are able to experience, how we experience it, and subsequently establish to what extent we might understand the material.

Faced with this insight, the Norwegian state has embarked on an ambiguous project, ordering the National Library to digitize *all* cultural heritage. Skipping the appraisal process, this project aims to include all existing historic material in Norway into one unified Digital Assets Management System, giving equal attention to all objects. This may, however, create problems for organizations and search functions, since the collection will consist of different media and material, from different sources and based on differing organizational principles. At the risk of ending up with a large digital collection without the ability to contextualize and organize the different materials, the National Library has looked to AI and ML to solve the problem.

This text starts out by describing the archival challenges associated with digitization of historic material, then goes on to look at the Norwegian cultural heritage policies in light of this. Finally, I will compare how the National Library uses artificial intelligence, especially in its digitalization of historic photography, in comparison with other similar projects. The aim is to see how it chooses to use artificial intelligence to fulfill the practical and cultural goals of its work.

Challenges Concerning Digitization of Historic Material

Digitization projects executed by memory institutions in recent decades can roughly be divided into two broad categories: mass digitization and slow digitization. Both categories aim at preserving often fragile material and giving audiences and researchers better access to this, and both come with advantages and disadvantages (Dahlström & Hanson, 2019, p. 2; Hughes, 2004, pp. 8–11).

Critical digitization, also called slow digitization, implies an in-depth consideration of how the material in question might be transformed into the best possible digital format. The aim is to find the optimal technological solution to digitalize a single historic artifact or a defined historic collection. The term slow digitization was coined in reference to the digitization of older manuscripts, using technology as a tool for a slow forensic investigation rather than merely to produce color photocopies. Considering the original material, its content and context, different approaches can be chosen for different documents or collections (Prescot & Hughes, 2018). The most famous exponent for this praxis is probably the Google Arts and Culture Project, where a selection of famous art pieces are given in-depth scanning to reveal details not even visible to the naked eye. For painted artworks, Google has devised the gigapixel project, which scans the artifacts in million-pixel detail (Proctor, 2011, p. 215).

There are however challenges concerning this critical approach. The cost and resources often associated with this approach means that only a small selection of historic materials is being subject to this level of detailed examination. Institutions collaborating with Google in their digitization project, usually only gets to select a few artifacts to be scanned with their gigapixel technology (Hylland, 2017, p. 72). This means that we are faced with what in archival terms is called a process of appraisal (Delsalle & Procter, 2018, p. 172). Some materials are elevated to the level of cultural and historic importance, justifying the costs, while other material are valued as less importance and excluded from such form of presentation and preservation. Since the technology needed for any digitization historically has been expansive, the process of appraisal through choosing the most valuable and most popular item to be digitized has become an established norm (Hughes, 2004, pp. 32, 40). From the archival communities, such selection processes have however been subject to a great deal of criticism in the past twenty years. The concern is that any such selections lack neutrality. When giving any memory institution the power to record and preserve some events and not others, one is also giving them the power to decide whose voices will be used to construct our history (Schwartz & Cook, 2002, p. 5; Harris, 2002, pp. 64–65). The fundamental problem with any selection process done in slow digitization is that it results in a biased view of history.

The cost historically associated with digitization has also in many cases forced state institutions to form partnerships with international and commercial ventures. Again, such partnerships have been criticized for pushing the selection process in an even more biased direction as well as resulting in restricted public access to the digitized material. Both the overrepresentation of materials from North America and large European museums, as well as commercial restrictions on the material has been pointed out as a problem (Breckenridge, 2014, p. 500; Kizhner et al., 2021, p. 626).

Turning to mass digitization, which attempts to digitize as much material as possible, the problems concerning biased selection processes can of course be avoided. However, this approach presents other challenges. First of all, the financial and human recourses associated with digitization—a reality that on an early stage led many institutions to choose only to digitize a small amount of material—has forced many to reduce technical quality when digitizing large volumes of material. Due to the often automated process, we can then be left with somewhat random results rather than those informed by careful and deliberate consideration (Dahlström & Hanson, 2019, p. 2).

But besides the concerned about technical quality of the actual digitized material, concerns can also be raised about the resulting organization of large digital collections. The challenge presents itself when you are joining material from different collections and sources into one major digitization process. In most cases, different materials and different collections are organized around different organizational principles. In a process of mass digitization, one will often need to establish a uniform system of metadata to search and organize the finale material. From the perspective of archival theory, an important principle has been that of provenance (Ridener, 2009, pp. 32–34). Material from one creator should never be split up or joined together with material from other sources. The reason is that the structure of the archive—its organizational principles—represent an important context for understanding the specific documents. In large digitization projects there is there for a danger that the construction of a new and unified principle for organization, can deprive the single documents of their original context. Some of the criticism against digitization from archival communities has earlier followed these lines, arguing that such a process might include removing the object from its original context (Oliver, 2012, p. 55). In the worst cases the result is a brilliant digital version of the documents themselves, but the knowledge of where these documents came from, why they were created, and what they meant could be lost.

The National Library of Norway and Their Digitalization of National Heritage Preservation

Since 2006 the National Library of Norway have been responsible for digitizing all the printed material ever published in Norway (Takle, 2009a, p. 2). All books, magazines and newspapers published in Norway have had copies deposited at the National Library. The process of scanning this material was undertaken with the aim of creating a complete online library consisting of all printed matter ever published in the country. New publications are now uploaded in digital format directly to the library, meaning that the collection is always up to date. For copyright reasons, temporary restrictions are imposed on access to the latest material, but all more historic publications can be directly accessed from Norwegian IP addresses.

In comparison to the concerns voiced earlier against such mass digitization projects, the National Library's digital collection is of great technical quality. Organizing and attaching metadata to such a sizable collection is also rather unproblematic. Since it only consists of printed and published matter, it is easy simply to connect it to existing bibliographical databases.

However, the Norwegian state wanted to digitize more of the cultural heritage. It wanted to move beyond the published material that already formed part of the traditional library and sought out more unique materials in museum collections, archives and in private ownership. The new aim was to digitize the entire cultural heritage of Norway, implying digitalizing access and preservation of all historic material. So, in 2020 the Ministry of Culture allocated money to establish a Center for Cultural Heritage Digitization at the National Library's branch in the small city of Mo i Rana (Ministry of Culture, 2021). The motivation for the establishment of this new center was twofold. As with the earlier digitization of published books, the aim was in part to increase access to the material through a public Digital Assets Management System for all types of cultural heritage. But the digitization was also motivated by a need to preserve visual and sonic material stored on analog media, such as analog tapes and photographic paper that was in danger of deterioration.

Compared to the earlier digitization performed by the National Library that only encompassed published material already in the library, the intention behind this new project was to preserve any material archived in recording studios, photo studios, television stations and newspapers and make it available for the future public. The Ministry has estimated that the process will take thirty years.

In its online presentation of the digitization project, it becomes clear that the National Library faces many of the same challenges that have been voiced earlier concerning mass digitization—that of dealing with metadata for a large amount of heterogenous material. How will it manage to organize material from collections

with very different organizational principles and metadata standards? However, it argues that this is a trade-off it is willing to make in order to achieve other goals. As stated on its web page, collecting and digitizing cultural heritage before its current material manifestation deteriorates is a greater concern than the recording of metadata (National Library, 2022). In the first phase of the digitization, only metadata for basic retrieval will be added. The plan is to work to enrich the metadata after the material has been digitized. Firstly, after the digitization the original creators or the owner of the material are invited to add more data through an online platform (Ministry of Culture, 2021). The second phase is however what is most interesting; here it is suggested that AI and ML will be applied in order to further enrich the metadata and organize the material.

What I find interesting is that AI and ML are presented as strategies to achieve specific cultural policy goals. The challenges facing this the National Library's digitization project come from a political decision to bypass any biased appraisal process, instead underscoring all cultural heritage as having equal importance. Norwegian cultural politics has been based around an expanded concept of the democratization of culture. This has not only aimed at giving the public access to historic material, but also at placing a significant emphasis on including everybody's history into the nation's memory institutions (Hylland, 2017, p. 67; Mangset & Hylland, 2017, pp. 150–154; Røssaak, 2022, pp. 166, 168). Following this ideal, the digitization process aims both at giving the public greater access to material hidden away in museums and archives around the country as well as seeing that absolutely nothing and nobody is left out of the digitalized cultural heritage. The enormous amount of diverse material therefore being digitized presents vast challenges when it comes to organizing it in such a way to enable the public to search and retrieve material from the collection. The concepts of AI and ML have been put forward to resolve this challenge, and attempts are being made to use such technology to provide the public with better access to the material in the collection even when metadata is scarce.

The National Library's approach to digitize cultural heritage without a strong focus metadata stands in sharp contrast to much of what has been argued by archivists. When digitizing documents like public records, the importance of digitizing the indexes and descriptions of the collection together with the documents has been proposed as an important strategy to establish a basis for their digital organization and retrieval (Colavizza et al., 2019, p. 3). This is not something the National Library has simply neglected but rather a decision made in the face of the vast differences in organizational standards surrounding much of the material they are seeking to collect. Government records, which tend to be the focal point for many archivists, have always had strict organization and sets of metadata. But outside of government institutions, in newspapers and small museums, amongst photographers and recordings studios—the areas the National Library is now

aiming to digitize—the situation is often quite different. Parts of the material may be stored in numbered boxes but without any description indicating the contents. Sound recordings might be dated but only name who the recording engineer was, not describing who is playing. In effect there may be little structure amongst the material around which to build an organizational principle for the digitized content.

When turning to AI and ML as means to organize and interpreted the material in the collection, The National Library is also at odds with parts of the archival community. The possibility, for instance, to run full text searches across different bodies of digitized material has given research a powerful tool, but it has also been argued that this also will result in a loss of context. Where the documents originate from—their "provenance"—is not necessarily taken into consideration when documents are digitally reproduced and distributed (Breckenridge, 2014, p. 501). The new search tools that can be applied to digital material, such as full text search within each document, have been met with skepticism. These tools represent a departure from earlier archival research methods, which relied on the classification and organization done by the creator of the documents. It has been said that search based on machine ranking of relevance, as is the common strategy in digital information retrieval, does not necessarily "understand" the meaning of such distinct data elements and their relevance to specific topics (Andresen, 2019, p. 66; Cole & Hackett, 2012, p. 113). This stands in opposition to many of the concepts of semantic technology that have been pushed by information retrieval and tech communities for years. Most search engines and other digital systems that index information are based on some degree of semantic analysis of the text in the documents, using Latent Semantic Analysis (LSA), a vector-based mathematical system for establishing likeness between digital documents (Croft et al. 2011, p. 401; Deerwester et al., 1990; Landauer et al., 1998). In contrast to much of the classification and search done within traditional archival work, digital search engines do not necessarily depend on the metadata creators have attached to their documents, but rather an interpretation of the content of the documents performed by algorithms.

Artificial Intelligence and the Enrichment of Metadata

There have been several examples and trials with AI and ML being used in the organization, analysis, and tagging of digitized archival material in recent years. There are, however, special circumstances regarding the cultural and political goal of the National Library's work that impact the size and the diversity of the collection, eventually setting a specific scope for what they can and are willing to achieve with the use of AI and ML. In this last part of the text, I want to examine the National Library's use of AI and ML in comparison with other similar examples, comparing above all what the different institutions use the technology

for and what challenges they seek to resolve with the technology. As mentioned, the digitization work performed at National Library is estimated to take thirty years, so the project is only in its initial phase both in regard to material being digitized and the use of AI and ML. Besides printed material, digitized photography is for now the most prominent part of the library's digital collection, so I will focus on AI and ML in relation to this category of material.

As mentioned earlier, there is a long history within information retrieval of the use of sematic technology like LSA in the analysis of digital text. Attempts have even been made to use this to categorize for instance e-mails to determine what should be archived as important government records (Rolan et al., 2019, p. 188). In this specific instance, the technology has been put to work on a classical archival challenge of appraisal, deciding what to preserve and what to discard. How this technical works, is that you train an algorithm using a set of documents already categorized and let the computer compare remaining uncategorized documents to these, establishing a probability of what category they should belong to.

When turning to visual material, there is a growing field of research within Computer Vision, using algorithms to identify and subsequently tag material in relation to different categories. On example is using AI to help categorize digitized pictures from FARL's photo archive of art (Han et al., 2022, p. 32). In this case a specific organizational principle was already established through a specific taxonomical description of the material. The AI is merely doing the work of placing the pictures in the different categories.

Another example is the CAMPI project at Carnegie Mellon University, tasked with using AI to add metadata to the database of photography documenting campus life. The goal was to design an Asset Manager that could fulfill the needs of the users wanting access to these historic pictures. The material was frequently used as illustrations for publications and student work about the university, and there was often a need to easily retrieve pictures from established categories like "Athletics," "Buildings," or "Football players." Some pictures were already categorized, so the algorithm was used to identify other pictures that were similar and could need the same tag (Lincol et al., 2022, p. 18).

What unifies the above examples is that AI operates to help categorize the pictures in accordance with an already established organizational principle. The purpose of the collection is already defined at a previous stage and the digitization and use of AI is in line with goals predefined in an analog world. This approach takes a similar path to the digitization of public records: to digitize the existing indexing and description as the basis for the organization of the digitized material (Colavizza et al., 2019, p. 3). The result is an Asset Manager that to a certain extent functions around making established forms of organization and retrieval easier and less time-consuming.

The problem facing the digitization project in Mo i Rana lies in the enormous variety of materials, sound, moving images, photography, and text. All of these materials also originate from different archives and collections; if organized, the indexing will follow very different categories that reflect the initial purposes from which the material originated. Consequently, there are no single predefined set of categories that can guide the overall process of digital organization. Compared to the two other examples, there is no defined dataset that can be used to teach the AI what to do with the other material. Any pre-established categories originally made for one part of the material would probably not make any sense in relation to other parts of the digitized material considering all the different medias included in the project. From an archival perspective, forcing a new principle onto the totality of the collection, would put the actual material in danger by breaking down the actual context surrounding each material.

This lack of any unifying organizational principle and categories for metadata are one of the challenges that AI and ML hopefully can solve. From my visits to the National Library, I know that among the material now being digitized are several archives of negatives from different newspapers and photographers. It has turned out that much of this material has little metadata attached to it; some roles of negatives are kept in boxes with perhaps just a number on the lid and no explanation of when or where the pictures were taken. Making it possible to search and identify such material is of great important for the cultural heritage.

Turning to the use of AI and ML, at this point the sheer size and heterogeneity of the digitized material can be of advantage. In addition to the 622,000 photography that by now has been digitized into the collection, there National Library also have 4,279,000 digitized newspapers in the collection, also including pictures. The crucial point with using AI and ML to search in collection is that the metadata is not there to search; as mentioned, one has to dig into the actual content of the documents. Even if some archivists have doubts about how good such technology is in understanding and interpreting the content of documents, when there is little metadata to start with or there is little correspondence between the different collections, this possibility perhaps becomes the most promising approach.

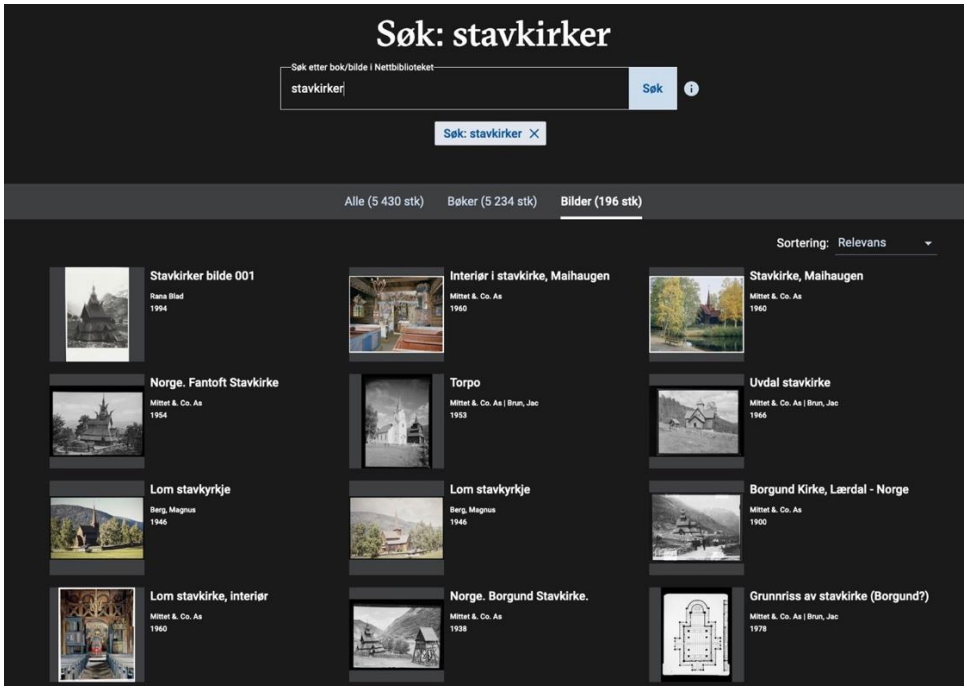
The National Library is working on looking for similarities between different pictures. The same process that can be used to match pictures to established categories can be used to find similarities between any other pictures. One can choose a picture and the algorithm will retrieve pictures with some degree of similarity to the chosen picture. In the CAMPI project, this function was also implemented into the search possibility, letting the user browse through similar pictures in the collection (Lincoln et al., 2022, p. 5). The technology has also been used on digitized collections of paintings, letting the user explore similarities and differences between different artists and art collections (Lincoln et al., 2019).

However, the National Library is using this technology on material that is so much bigger and diverse then the other examples mentioned here. CAMPI was only working on 20,000 images, and all its pictures were collected for the same purpose. Finding similarities between digitized material in National Library collections enables connections to be made between pictures that may originate from very different sources, establishing likeness across material generated for different purposes and therefore possible labeled with different metadata.

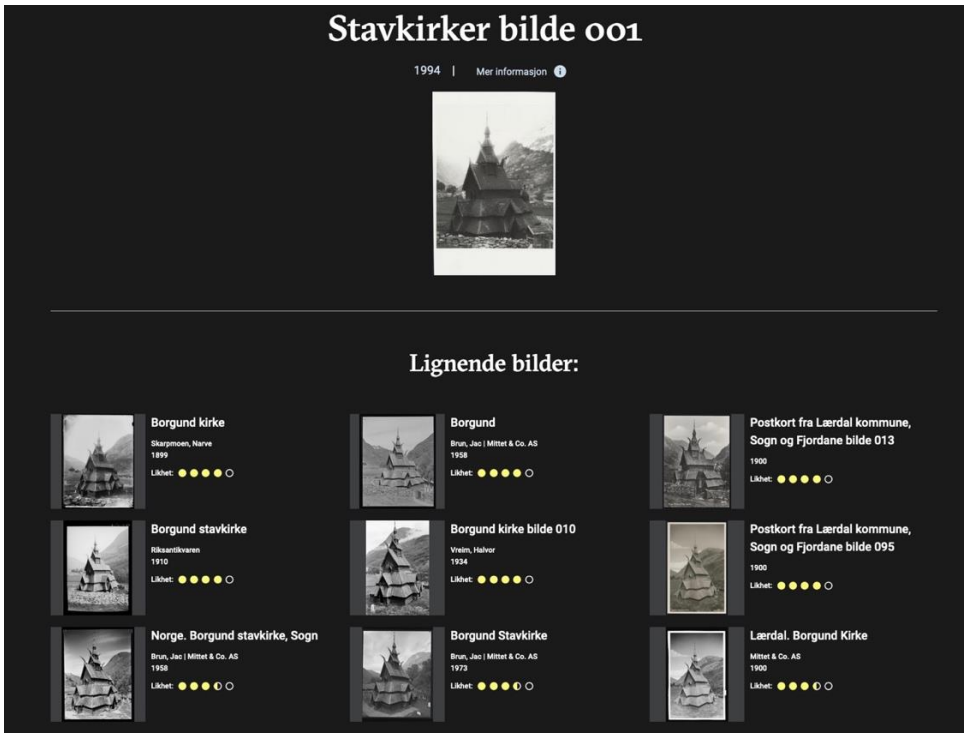
In order to explore such possibilities for making new connections between existing digitized material, the National Library has now launched an experimental search function in its digital collection called "Maken" ("Similar") (https://www.nb.no/maken/).[1] This enables users to select a document, either text or image, click "Maken," and bring up documents that the computer has identified as similar. This is a connection not made based on metadata but on an interpretation of the document itself. At the present time, not all documents have been analyzed by the computer, so the function does not cover all newly digitized negatives, but this experiment exemplifies a method that can be used to acquire a better context to the material we don't yet know much about. Material for which little metadata has been registered or perhaps never existed in the first place, can be linked to other material, sometimes enabling us to establish more information about the content of that first material.

To illustrate how this method works, I chose a picture of a stave church (probably the of the most famous examples of Norwegian cultural heritage). The picture I selected was only labeled "stave church 001," but when I press the "Maken" button I get other pictures of the same church but with a different metadata attached to it. Looking through these pictures, I can identify the selected picture as Borgun Stave Church in Lærdahl, built around 1200 (see Pictures 1 and 2).

[1] Due to copyright, the online services from the National Library are only available from Norwegian IP addresses.



Picture 1



Picture 2

What this small trial exemplifies is that through the likeness of pictures, one can connect information from different sources to describe a motif. Across time, different pictures have been taken of the same object, and due to the initial circumstances for using and taking these pictures, they have been labeled with different information and descriptions. A future possibility now being worked on is connecting the digitized photography to the picture material in the digitized newspaper. For this to work, one first has to extract pictures and headlines and corresponding articles from the newspaper. A similar process has been done within the Chronical America initiative, extracting visual content from 16 million pages of digitized newspapers (Lee et al., 2020). Since the Norwegian collection is over four million full newspapers, this process is still not complete. When finished, it will give the user the possibility to match digitized photography with pictures in the newspapers and use the conjoining textual information to gain more insight into the photography. What makes this especially useful is that many of the unmarked negatives undergoing digitization originate from old newspaper picture archives, so there will be a great deal of overlap between the picture materials.

From my perspective I also believe this search for likeness between published pictures in papers and unmarked negatives can be used to give some context to other unknown photography in the collection. From working in dark rooms and developing negatives I know that analog negatives have that quality that links pictures in time. The roll of negatives always shows pictures taken in succession. From my own experience I know a professional photographer usually shoot an entire roll in a day. After taking pictures, one would often develop the negatives the same evening. Therefore, identifying one image, might tell us something about who, when and where the other pictures on the same roll were taken. Even if we don't find any direct visual likeness between pictures, we can use known images from the same roll of negatives to perhaps establish a context for the other images too.

In the future there are several possibilities but also challenges in the use of AI and ML at the National Library. One challenge is to establish links between visual and textual material. The National Library's digitization project has the unique quality that it spans many different materials categories—picture, text, sound, and moving images. Would there be a possibility for AI and ML to find likenesses between pictures and text documents? There are of course experiments using ML to extract information from pictures on more general terms, identifying if the picture is showing buildings, airplanes, or red light. In recent years, Google has released its Cloud Vision API, a ready-trained algorithm enabling users to generate a textual description of their images (Chen et al. 2017; Mulfari et al., 2016). One imagines that it could be possible for the computer to recognize what Borgun Stave Church looks like and bring up pictures of it as it simultaneously retrieves results from a full text search. But this might take a lot of training. For

now one can use the metadata attached to some pictures and through that establish a link to relevant textual documents manually by simply typing in a new search.

The heterogeneity and volume of digitized material in the National Library also represents a specific challenge to establishing a functional asset manager. From a user perspective the material in the National Library has so many angles of approach that strict organization and categorization might create problems. What the user's information needs are and what aspect of a picture actually interests them are questions that are difficult to answer. Do they want to know the name of the church, or is it the landscape behind it that is the object of their attention? It might also be that some wish to use the material to research different photographic techniques, and really do not care so much about the motif at all. Tagging and organization of archival material shapes and restricts our engagement with that material. The positive thing about the "Maken" search engine is that it has a rather fluent approach to search, enabling the user to shape and explore the content according to needs. One does not have to engage with a broad system of classification that has been imposed onto the material in retrospect. AI and ML does not necessarily need to establish metadata categories in a traditional archival sense. New, more fluid approaches might also be developed. By making their own connections, users can build new information about the context and the meaning of the documents in question, resulting in new interpretations of historic material.

Conclusion

The National Library's digitization of cultural heritage preservation is born out of a political desire to create an inclusive digital repository that encompasses the entirety of Norwegian cultural heritage. Through this, the hope is that it will leave the future public with an unbiased collection of historic material that is open for further research. This process does, however, create challenges concerning digitalization of search, organization, and retrieval within on large Asset Manager. A broad selection of different material, from different creators and collectors, cannot be organized around a unified principle for metadata without risking losing some of the original contexts surrounding the different materials. The use of AI and ML has been suggested as one possible solution to resolve such issues. In recent years several other digital collections have experimented with using such technology to analyze documents for categorization, search, and retrieval purposes. However, the majority of these trials have focused on already defined collections on previously established organizational principles. The purpose and use of these collections were to a large extent defined before they were digitized. The digitization of cultural heritage performed by the National Library in Norway is establishing a new, much broader collection which is gathering material from a range of creators and institutions. The purpose is to give the public broad access to

all cultural heritage, implying numerous angles of approach when it comes to search and organization.

The work of completing the digitization of all Norwegian cultural heritage is estimated to take thirty years, so the project is still in a very early phase, especially with regard to using AI and ML to analyze and organize the material. One experiment they have performed is a search engine that enables the user to look for similarities between documents, including digitized photography. By using this, the user can explore different pictures of the same motifs and compare different metadata tags describing these images. Through this, one is able to gather information about material that has little or no metadata attached to it, letting the user identify unknown pictures in the collection. This project exemplifies is the possibility to tailor new search functions, not around predefined organizational principles, but rather the individual users' information needs.

References

Andresen, H. (2019). Fremfinning og bruk av digitalt skapt materiale i arkivdepotene. *Norsk Arkivforum, 25,* 45-69.

Breckenridge, K. (2014). The politics of the parallel archive: digital imperialism and the future of record-keeping in the age of digital reproduction. *Journal of Southern African Studies, 40*(3), 499–519.

Chen, S. H., & Chen, Y. H. (2017, April). A content-based image retrieval method based on the google cloud vision API and wordnet. In *Asian conference on intelligent information and database systems* (pp. 651-662). Springer, Cham.

Colavizza, G., Ehrmann, M., & Bortoluzzi, F. (2019). Index-driven digitization and indexation of historical archives. *Frontiers in Digital Humanities, 6,* 4.

Cole, R., & Hackett, C. (2010). Search vs. Research: Full-Text Repositories, Granularity, and the Concept of 'Source' in the Digital Environment. In Avery & Holmlund (Eds.) *Better off forgetting?: essays on archives, public policy, and collective memory*. University of Toronto Press.

Croft, W. B., Metzler, D., & Strohman, T. (2011). *Search engines: Information retrieval in practice*. Pearson Education.

Dahlström, M., & Hansson, J. (2019). Documentary Provenance and Digitized Collections: Concepts and Problems. *Proceedings from the Document Academy*, *6*(1), 8.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, *41*(6), 391–407.

Delsalle, P., & Procter, M. (2018). *A history of archival practice*. Routledge.

Han, X. Y., Papyan, V., Prokop, E., Donoho, D. L. & Johnson, C. R. (2022). Artificial Intelligence and Discovering the Digitized Photoarchive In Jaillant, L. (Ed.), *Archives, access and artificial intelligence: working with born-digital and digitized archival collections* (pp. 29-61). Bielefeld University Press.

Harris, V. (2002). The archival sliver: power, memory, and archives in South Africa. *Archival science*, *2*(1), 63–86.

Hughes, L. M. (2004). *Digitizing collections: strategic issues for the information manager* (Vol. 2). Facet Publishing.

Hylland, O. M. (2017). Even Better than the real Thing? Digital Copies and Digital Museums in a Digital Cultural Policy. *Culture Unbound*, Volume 9, Issue 1, 2017: 62–84. Published by Linköping University Electronic Press: http://www.cultureunbound.ep.liu.se

Kizhner, I., Terras, M., Rumyantsev, M., Khokhlova, V., Demeshkova, E., Rudov, I., & Afanasieva, J. (2021). Digital cultural colonialism: measuring bias in aggregated digitized content held in Google Arts and Culture. *Digital Scholarship in the Humanities*, *36*(3), 607–640.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse processes*, *25*(2–3), 259–284.

Lee, B. C. G., Mears, J., Jakeway, E., Ferriter, M., Adams, C., Yarasavage, N., ... & Weld, D. S. (2020). The newspaper navigator dataset: extracting and analyzing visual content from 16 million historic newspaper pages in chronicling America. *arXiv preprint arXiv:2005.01583*. Cornell University.

Lincoln, M., Levin, G., Conell, S. R., Huang, L. (2019) "National Neighbors: Distant Viewing the National Gallery of Art's Collection of Collections", (November 2019) https://nga-neighbors.library.cmu.edu

Lincoln, M, Corrin, J., Davis, E., Weingart, S. B. (2020): CAMPI: Computer-Aided Metadata Generation for Photo archives Initiative. Carnegie Mellon University. Preprint. https://doi.org/10.1184/R1/12791807.v2

Mangset, P., & Hylland, O. M. (2017). *Kulturpolitikk. Organisering, legitimering og praksis*. Universitetsforlaget.

Ministry of Culture (2021). *Kulturdepartementets oppdrag til nasjonal biblioteket om digitalisering av kulturarvsmaterialer i arkiver og museer*. https://abmdig.no/content/uploads/sites/14/2021/11/Om-Kulturdepartementets-oppdrag-til-Nasjonalbiblioteket-om-digitalisering-av-kulturarvsmater.pdf

Mulfari, D., Celesti, A., Fazio, M., Villari, M., & Puliafito, A. (2016, June). Using Google Cloud Vision in assistive technology scenarios. In *2016 IEEE symposium on computers and communication (ISCC)*, 214-219 IEEE.

National Library (2022) *Metadata—data om data.* https://abmdig.no/senter-for-kulturarvdigitalisering/praktisk-gjennomforing/metadata-eller-data-om-data/

Oliver, G. (2012). The digital archive. In Hughes, L. M. (Ed.) *Evaluating and measuring the value, use and impact of digital collections*. Facet Publishing.

Prescott, A., & Hughes, L. M. (2018). Why do we digitize? The case for slow digitization. *Archive Journal*.

Proctor, N. (2011). The Google Art Project: A new generation of museums on the web?. *Curator: The Museum Journal*, *54*(2), 215–221.

Ridener, J. (2009). From polders to postmodernism: A concise history of archival theory.

Rolan, G., Humphries, G., Jeffrey, L., Samaras, E., Antsoupova, T., & Stuart, K. (2019). More human than human? Artificial intelligence in the archive. *Archives and Manuscripts*, *47*(2), 179-203.

Røssaak, E. (2022). Nasjonalbibliotekets massedigitalisering: En ny tilgang på fortid blir til. In Røssaak. E. & Gran, A. (Eds) *Mangfold i spill: Digitalisering av kultur og medier i Norge.* Universitetsforlaget. 152–172.

Schwartz, J. M., & Cook, T. (2002). Archives, records, and power: The making of modern memory. *Archival science*, *2*(1), 1–19.

Takle, M. (2009a). The Norwegian national digital library. *Ariadne*, *60*.

Takle, M. (2009b). *Det nasjonale i Nasjonalbiblioteket*. Novus.