Library Science. Record Studies. Informology.

Library Science. Record Studies. Informology. 20(4), 2024, 56-68
Journal homepage: https://bdi.com.ua/en/journals/journal
Received 01.08.2024 Revised 24.10.2024 Accepted 25.11.2024

UDC 004.01:[930.25:929.5] DOI: 10.63009/lsrsi/4.2024.56

Digitisation of genealogical documents based on automatic text recognition technology

Artur Spektor
Postgraduate Student
Lviv Polytechnic National University
79013, 12 Stepan Bandera Str., Lviv, Ukraine
https://orcid.org/0000-0003-4176-9177

Abstract. This study aimed to examine existing approaches and technologies for the digitisation of genealogical documents, drawing on international experience. This enabled more efficient organisation of digitisation processes and mechanisms for archive groups, their centralised storage, accelerated genealogical research, and improved user accessibility. The digitisation of archives had became a critically important aspect of preserving cultural heritage, particularly in the context of Russia's military aggression against Ukraine. The introduction of automatic text recognition technology had contributed to the optimisation of this process, facilitating access to information and enhancing the efficiency of research, particularly in the field of genealogy. The study analysed the operating principles of optical character recognition, its advantages, the features of ready-made solutions, and the functionality of software based on this technology. The strategy for digitisation in Ukraine was assessed, along with the challenges facing the archival sector in terms of digitisation and access to archive groups. The research also examined the outcomes of implementing automatic text recognition in leading archives worldwide, as well as the capabilities of online archives that offered contextual search functions. Particular attention was given to the opportunities afforded to researchers through the integration of such systems into archival operations, notably the ease of locating required information, the increased speed of data processing, and the provision of round-the-clock access to archival resources regardless of users' geographical location. The study also reviewed the research of scholars involved in the development and implementation of optical character recognition in archival institutions. Drawing on international experience, the potential of modern Optical Character Recognition technologies to modernise the archival sector in Ukraine was identified, with positive implications for genealogical research and the preservation of cultural heritage. The practical value of the study lies in demonstrating the effectiveness of information technologies in improving the digitisation process of archival documents and enhancing access to them. The proposed recommendations aim to optimise the organisation of digital archives, improve document storage and retrieval processes, and accelerate genealogical research. These developments will contribute to the preservation of cultural heritage and improve access to archival information for users

Keywords: archive group; scanning; genealogical research; Optical Character Recognition; information technologies; automation

Introduction

Since the late 1990s, there is been a big increase in interest in family history research worldwide, thanks to online services like Ancestry.com and FamilySearch. This trend started in Ukraine around 2010. Because of this, looking after, organising, and making archival documents accessible has become a really important part of how archives were developing (Logvynenko *et al.*, 2024). Ukraine had a huge number of unique and valuable archive groups that held information about historical figures. With Russia's aggression, these

Suggested Citation:
Spektor, A. (2024). Digitisation of genealogical documents based on automatic text recognition technology. *Library Science. Record Studies. Informology,* 20(4), 56-68. doi: 10.63009/lsrsi/4.2024.56.

*Corresponding author

groups were at risk, making it really important to efficiently turn them into digital copies and allowed people to access them remotely.

The issue of using text recognition technology and how well it works with Ukraine's archive groups has been looked at by K. Lipianina-Honcharenko *et al.* (2024). They pointed out how particularly effective it was and the need for more research. N. Korzhyk *et al.* (2023) noted that the work of archival institutions in Ukraine has faced challenges due to Russia's military aggression, including the destruction of buildings and the loss of some archive groups. Because of this, it was necessary to took several steps to preserve documents, which included speeding up the process of making archive groups digital. I. Khoma *et al.* (2023) highlighted that making documents digital was really important for protecting cultural heritage, education, and engaging with history. Digitising archives helped to keep and promote valuable cultural resources, made research easier, and allowed more people to access historical documents and other materials. O. Artemenkova (2022) looked at the theoretical ideas and methods developed by leading archival experts in Ukraine regarding the information tools used to improve, how effectively family history research was done. The academic O. Artemenkova (2023) also considered the role of information technology as a key tool for making family history research more popular in Ukrainian archives and argued for the need to create a single platform for accessing digitised family history data. The author identified, how archive websites work, categorised family history resources, and clarified the terminology used. The findings were important in practical terms as a foundation for making archival work digital, improving communication, and preserving national memory.

The author L. Kovalska (2019) analysed, how important looking back at the past through access to archival documents was for the development of Ukrainian society, stressing the significance of new methods in making archival information more widely available. The academic identified the need to introduce automated and online information systems to improve the quality of archival services and the development of archival practices, which opened up new possibilities for social communication. O. Rybachok (2018) identified the role of UNESCO and international organisations in creating strategies and methodological approaches for developing joined-up digital resources and also researched the historical background and stages of development of documentary cultural heritage in an international context. M. Friedewald *et al.* (2024) analysed the impact of digitisation on the accessibility of archival documents in the digital age, finding positive results such as improved ease of access and reduced burden on users. However, only a small part of archive groups had been digitised, highlighting the need for further efforts in this area. One of the main problems was data protection and copyright, which created legal restrictions and uncertainties, as well as challenges with the compatibility and organisation of digital records. For the effective preservation and access to digital collections, collaboration between archival institutions and new groups of users was important, as was the use of artificial intelligence technologies to improve metadata processing.

J. Nockels *et al.* (2024) developed guidelines to help researchers, data providers, platforms, and institutions understand, how the results of handwriting recognition technology interact with the wider information environment. The researchers found that the technology made it easier to access more materials, including languages that were at risk of disappearing. This allowed for a new focus on personal and private materials (diaries, letters), broadens access to historical voices not usually included in historical records, and increases the amount and variety of available material. A.L. Silva & A.L. Terra (2023), using Europeana as an example, noted that using the principles of linked data had a positive effect on the speed of digitisation and the preservation of cultural heritage for libraries, archives, and museums.

This research aimed to highlight the advantages of using automatic text recognition technologies, drawing on global experience in digitising genealogical documents. The scientific novelty of the study lies in determining the impact of automatic text recognition technologies on the speed of digitising archive groups and the effectiveness of conducting genealogical research based on the experience of leading archival institutions.

Materials and Methods

The research methodology was based on general scientific methods of analysis and synthesis, methods of comparative and content analysis of academic literature, and the use of modelling, grouping, and generalisation methods. The analysis of scientific publications and legal documents helped to identify current trends in the digitisation of archival documents, particularly in the field of genealogical research. A content analysis was carried out on academic articles, reports from archival institutions, and technical documentation regarding the implementation of automatic text recognition technologies (OCR – Optical Character Recognition). The stages of the research were:

- analysis of literature on the digitisation of genealogical documents using automatic text recognition technologies, which allowed for the formulation of the main approaches and problems in this area;
- study of the current state of digitisation processes of archive groups in Ukraine, including an examination of existing archival systems and infrastructure for document digitisation;
- familiarisation with the theoretical aspects of OCR technologies, explaining their operating principles, algorithms, and application possibilities;

- analysis of existing OCR-based solutions, studying their functional capabilities as well as limitations that affect the effectiveness of their use for processing archival data;
- examination of the experience of leading archival institutions and their practices in implementing and using OCR, as well as an evaluation of the effectiveness of online archives in improving access to archival documents and optimising genealogical research.

The main method for gathering materials was bibliographic and internet research, which allowed for an understanding of existing archive digitisation strategies in Ukraine and abroad. For the comparative analysis, several international archival institutions were selected, including the Arolsen Archives, the National Archives of the Netherlands, the National Archives of Zurich, and the National Library of Finland, all of which have implemented OCR technologies and capabilities for processing genealogical documents. Documentation from companies developing OCR technologies was also studied, and their technical features, particularly the possibilities for automating data processing, were analysed.

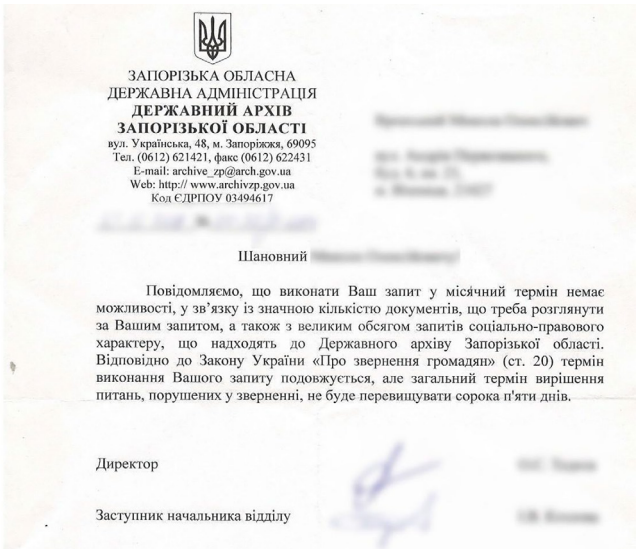
For the research, the Strategy for Digital Transformation of Ukraine was analysed. This Strategy outlined the directions for the development of digital services in the public sector, including the creation of a unified electronic archive and improved access to archival data. The Strategy also involved the implementation of digital technologies to simplify access to information and increase management efficiency (Order of the Cabinet of Ministers of Ukraine No. 1353-r, 2020). Furthermore, the work plan of the Ministry of Digital Transformation for 2023 was examined, which included measures for the development of open data and the improvement of legislation for the digitisation of archives (Report on the implementation…, 2023). These documents helped to outline the key directions for the development of archival practice in Ukraine.

Methods of grouping and generalisation were used to systematise and classify data related to the digitisation of archives and the implementation of OCR technology. Grouping allowed for the organisation of information according to specific criteria, such as types of archival documents, digitisation technologies, and the countries and institutions implementing these tools. Generalisation helped to formulate overall conclusions regarding the effectiveness of using OCR in genealogical research and to identify the main trends and challenges in the processes of digitising archive groups. The results of the research highlighted the advantages of a centralised implementation of OCR in institutions.

Results and Discussion

In line with the Strategy for Digital Transformation of Ukraine, digital services provided by government bodies and institutions were actively developing. One aspect of this strategy was information accessibility, specifically ensuring the retro-conversion of existing paper-based primary documents, including those in archives, to create a single, centralised electronic archive (Order of the Cabinet of Ministers of Ukraine No. 1353-r, 2020). For example, according to this policy, the Ministry of Digital Transformation set the following goals for 2024: 100% of public services should be accessible to citizens and businesses online, 95% of transport infrastructure, populated areas, and their social facilities should have access to high-speed internet, 6 million Ukrainians should be involved in digital skills development programmes, and the share of IT products in the country's GDP should be at least 10% (Digital transformations in Ukraine…, 2020). According to the work plan of the Ministry of Digital Transformation of Ukraine for 2023, measures were planned in the area of open data development, their compliance with European legislation, and the improvement of legislation in the field of public electronic registers and their implementation (Report on the implementation…, 2023).

The digitisation of archival documents and fonds was a pressing issue. It will allow for the preservation of documents from negative external factors (physical damage, loss, destruction) and provide the possibility of quick, free, and transparent access to documents (Onuchak, 2024). These factors will positively contribute to the development and simplification of genealogical research through the possibility of remote access to the archive groups of institutions and the ability to search quickly using information retrieval systems. This will allow a researcher, based on a search query, to receive all records held in archival institutions that mention the required surname within a specified period. This was unlike the current procedure, which involved writing a request to each individual archival institution, where documents might potentially be located and waiting 30 days for the request to be processed. Additionally, it should be noted that if a certain number of requests were exceeded, for example, 10 archival references simultaneously, the archive may extend the processing time for that request from 30 to 60 days (Fig. 1).However, as of 2024, there was not a unified strategy for the methods of processing digitised documents, their storage, and the analysis of data obtained from archive groups. Each archival institution had its own mechanisms for digitising archival documents, which could include scanned copies with added information about the document, photographs, or manually typed documents. To solve the problem of access to archive groups, optimise genealogical research, speed up the digitisation of archival documents, and simplify their search, Optical Character Recognition technology can be used (What is optical character…, 2024).



ЗАПОРІЗЬКА ОБЛАСНА
ДЕРЖАВНА АДМІНІСТРАЦІЯ
ДЕРЖАВНИЙ АРХІВ
ЗАПОРІЗЬКОЇ ОБЛАСТІ
вул. Українська, 48, м. Запоріжжя, 69095
Тел. (0612) 621421, факс (0612) 622431
E-mail: archive_zp@arch.gov.ua
Web: http:// www.archivzp.gov.ua
Код ЄДРПОУ 03494617

Шановний

Повідомляємо, що виконати Ваш запит у місячний термін немає можливості, у зв'язку із значною кількістю документів, що треба розглянути за Вашим запитом, а також з великим обсягом запитів соціально-правового характеру, що надходять до Державного архіву Запорізької області. Відповідно до Закону України «Про звернення громадян» (ст. 20) термін виконання Вашого запиту подовжується, але загальний термін вирішення питань, порушених у зверненні, не буде перевищувати сорока п'яти днів.

Директор

Заступник начальника відділу

Figure 1. Example of an archive extending the processing time for a request
Source: State Archive of Zaporizhzhia Region (2024)

Optical Character Recognition was a set of technologies that converted images of text into a machine-readable text format. For example, it can turn a scanned archival document into text data that can be analysed by other software. This allowed access to all the data in the document, the ability to edit it, search for text fragments, and automate search processes based on the document's metadata. The principle of how OCR technology works involved the following stages: image acquisition, pre-processing, text recognition, conversion to text format, and saving the results obtained (Fig. 2).

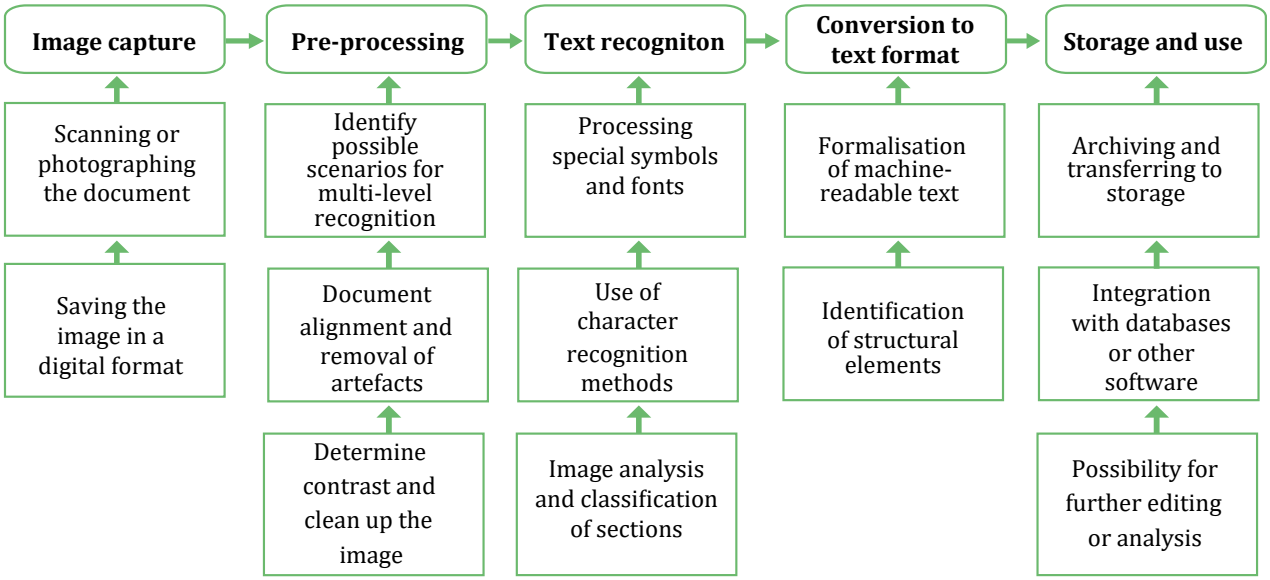


Figure 2. The process of digitising a document using OCR
Source: based on What is OCR (Optical Character Recognition)? (2024), What is optical character recognition (OCR)? (2024)

To obtain the image, all necessary documents were scanned. Next, the OCR system converted the digital document into a two-colour or black-and-white version. The resulting image was analysed for light and dark areas, where the programme subsequently identified the dark areas as characters to be recognised and the light areas as the background (What is optical character..., 2024). In the pre-processing stage, the document was cleaned of unnecessary pixels. This involved correcting skew to fix any misalignment of the image during scanning and removing graphic lines and frames that were part of the printed image (What is OCR

(Optical Character Recognition)..., 2024). It was worth noting that this process was very important, when processing archival documents, as the paper usually loose its properties over time and became thinner, leading to more digital noise. Parts of symbols from the reverse side of the document may also be present.

Text recognition involved the identification and processing of letters, numbers, or symbols. This stage typically targeted one character, word, or block of text at a time. The characters were then identified using one of two algorithms: pattern recognition or feature recognition. The elements were identified through one of the following algorithms:

- pattern recognition (or pattern matching): the OCR was pre-trained on examples of text in various fonts and formats to recognise characters by comparing them to a template in the digital document or image file. Each unique combination of shape, size, and font was called a glyph. For this to work, the characters had to be in a font that the OCR programme had already been trained on;
- feature recognition (detection or extraction): this was used, when the OCR programme analysed a font it had not been trained on. The OCR applied rules about the specific characteristics of a particular letter or number to recognise characters in the digital document. Features included the number of lines at an angle, line intersections, loops, or curves in a symbol. For example, the capital letter "A" was stored as two diagonal lines intersecting with a horizontal line in the middle. Once a symbol was identified, it was converted into an American Standard Code for Information Interchange (ASCII) code, which computer systems used for further manipulation (What is optical character..., 2024).

Subsequently, OCR analysed the structure of the obtained image. During this process, the page was divided into elements such as blocks of text, tables, or images. The lines were then split into words and subsequently into characters. After the characters were extracted, the programme compared them to a set of image templates. Once all possible matches were processed, the programme returned the recognised text (What is optical character..., 2024).

After all the image processing steps were completed, depending on the capabilities and features of the software, it was possible to preview the results and make certain corrections to the resulting text or to run the analysis again with a different set of parameters. This stage was important for ensuring the high quality of the final text, as it allowed the user to identify and correct any errors that may have occurred during the automatic recognition, especially in the case of handwritten documents, where certain elements could be mistakenly identified as digital noise. The resulting text was saved to a digital file, which was then archived and stored.

Using OCR offered the following advantages: reduced costs for searching, providing, and analysing documents, when requested; faster processes for conducting genealogical research, specifically reducing the time to obtain and search documents with the possibility of context-based searching across all available digitised archival groups; centralisation and standardisation of data format and the ability to export it in various formats (JSON, CSV, SQL tables, machine-readable text) for further use; and the ability to store data in cloud storage or remote servers, protecting data from fire, loss, or damage, as well as quick copying to electronic media if needed (What is optical character..., 2024). There are several popular software solutions based on OCR:

- Tesseract – a neural network-based software tool focused on online recognition as well as character pattern recognition. The programme can recognise over 100 languages, including Ukrainian, and supports various input data formats. Advantages include the fact that it is free, open-source software with the ability to modify it and add new languages. Disadvantages include the lack of a graphical interface (third-party solutions are needed for this) and the complexity of installation and configuration (Tesseract OCR, 2024).
- OCR4all – free software designed for working with handwritten documents, but it can also handle printed text. The workflow is structured so that all operations and tools are in one consistent interface and are as user-friendly as possible. A drawback is the need to create a language model for Ukrainian (User guide – introduction, 2024).
- Google Cloud Vision and Document AI – solutions from Google that allow for text recognition in documents and also use AI to create workflows for analysing, describing, and structuring documents. They have very well-written documentation and a user-friendly interface, with Ukrainian language support, but are fully paid software products (Vision AI: Extract insights from..., 2024).
- Transkribus – a platform that enables automatic text recognition, editing, collaboration, and, if necessary, the training of specialised artificial intelligence to digitise and interpret historical documents of any kind. Transkribus already has a pre-trained model for Ukrainian handwriting. The platform offers 100 free scans each month, after which scans become paid, but they have individual organisational plans for scientific and cultural institutions that include more features (Unlock the past with Transkribus, 2024).
- Amazon texttract and Amazon rekognition – these are machine learning services that use OCR to automatically extract handwritten text and data from scanned documents. The service can also analyse distorted text and attempt to normalise it (Amazon rekognition, 2024).
- IBM cloud pak for business automation – this is a modular suite of integrated software components designed for work automation. It includes a Document Processing module that allows for obtaining data from

documents regardless of format, classifying them, and finding and interacting with the necessary fields within the document. The main disadvantage is the price and the need to purchase the entire software package (IBM cloud pak for..., 2022).

The Arolsen Archives was an international centre on Nazi persecution, whose mission was to protect documents and preserve them for future generations. This involved digitisation, preservation, the addition of keywords, and detailed archival descriptions to make them more suitable for a wide range of purposes, including historical research, genealogical research, and searching in the online archive. Since 1998, staff have been digitising documents in Bad Arolsen. As of 2024, between 85% and 90% of the groups have already been scanned – a rate that only a small number of other archives can match. Digital documents and processes not only help to speed up the process of responding to enquiries, but they also provided much better access to documents, whether in the reading rooms in Bad Arolsen, on the premises of selected partner institutions, or in the Bad Arolsen online archive. The archive actively indexed documents due to the significant number of enquiries from journalists, academics, and educators, who were interested in key topics, specific locations, nationalities, or victim groups. Because of this, they have actively started using OCR to record the entire content of documents. Private companies such as the genealogical portal Ancestry have also been involved, which facilitates the quick and easy searching of as many documents as possible. In 2019, Ancestry processed lists of displaced persons as well as a large collection of Allied documents about formerly persecuted individuals, making them easier to find in the online archive (Documentation and archiving, 2024).

The Arolsen Archives have organised a user-friendly online archive, where individuals can search for information by topic, full name, or specific words within documents, thanks to completed descriptions or the prior processing of documents using OCR. There was a convenient filter that allowed users to refine their search query (Fig. 3). It was possible to download a search report with information on the filtered records. As an additional feature, users can leave comments on records and share them via social media. When clicking on a relevant record, a detailed card was displayed with the full title, reference link and code, document creation date, number of documents, volume and content of the collection, the direct source of acquisition or transfer, the language of the documents, subject indexes, and an annotation with usage rules, as some documents may have been transferred to the archive by other institutions.

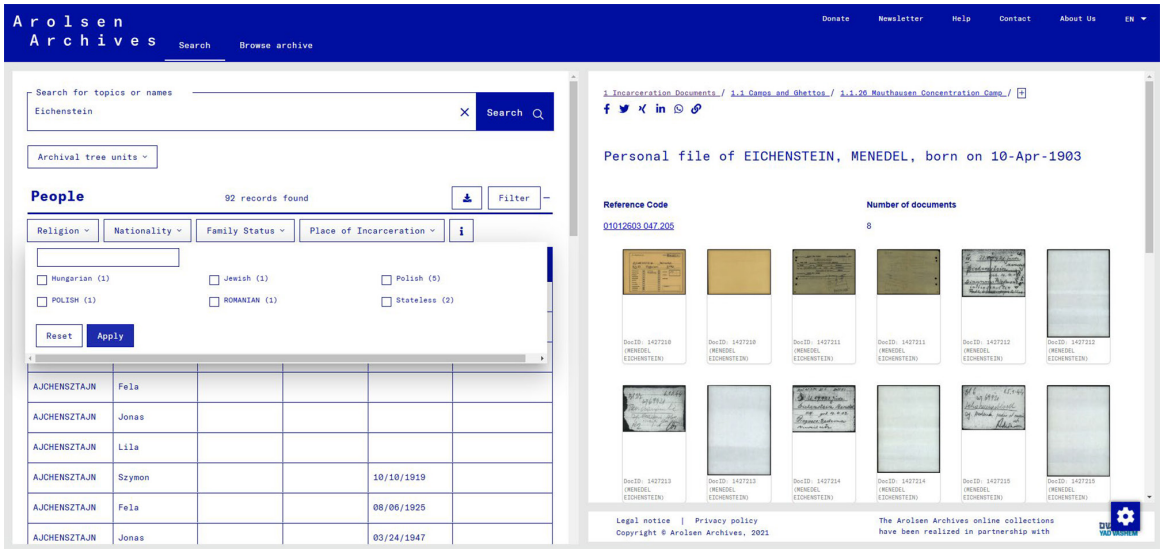


Figure 3. Example of a surname search query in the Arolsen Archives
Source: based on Documentation and archiving (2024)

An example of the successful use of technologies based on automatic text recognition in state archives and specific archive groups was the experience of the State Archives of Zurich, which digitised more than 50000 pages of Zurich council meeting minutes from the 18th century using the Transkribus programme (Unlock the past with Transkribus, 2024). To train the model, the following strategy was developed: for each volume of material, 1-2 pages were transcribed manually as training material, and the remaining pages were recognised based on the information obtained. Due to the inconsistent fonts, the automatic recognition of handwritten text in the initially processed 18th-century volumes showed an error rate of 5 to 8%, which provided both good searchability and good readability. In the edited volumes up to 1700, the error rate was 3%-5%. Digital documents can be viewed and accessed through the online catalogue of the Zurich Archives,

where it was possible to search by text or by category (Zurich council manuals 1642-1798, 2024).

On the archive's webpage, a search function was implemented with various parameters, including keywords, contextual search within the documents themselves, tag-based searching, and a wide range of filters for more precise identification of the required information. It was important to note that selecting one filter automatically adjusts the number of available options for other filters. This was done to ensure that the system only processes correct data. When accessing a selected document, an interactive preview of the digital copy of the document and its full text, obtained using OCR Transkribus, was available (Fig. 4).

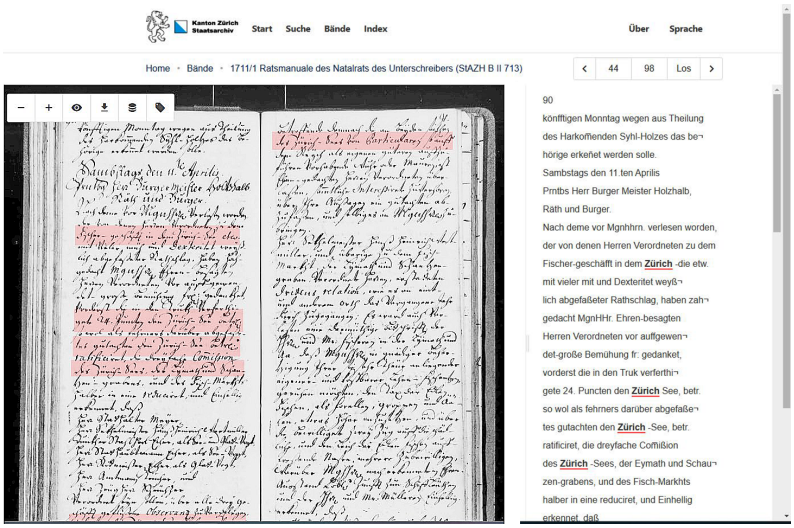


Figure 4. Example of a document digitised using Transkribus at the National Archives of Zurich
Source: based on Zurich council manuals 1642-1798 (2024)

As part of a large-scale project to digitise fonds, the National Archives of the Netherlands aimed to scan approximately 10% of its collection annually, amounting to over 10 million scans. The digitised archival documents, including notarial deeds and records of the Dutch East India Company, contained important information for genealogical research, such as records of marriages, inheritances, family relationships, and professional activities. Thanks to the recognition of contextual entities within the documents, it was easy to construct search queries that aid in tracing ancestry and the socio-economic context of families. To make access to historical documents easier, the archive also transcribed some of the materials. In particular, using Transkribus software, 3 million documents, including handwritten ones, were digitised and were now freely available on their online resource. At the beginning of their work with Transkribus, they experienced a character error rate of 20%. However, after the model processed 6000 pages of training data, the rate improved to 7%, which was even better than they had anticipated (Unlock the past with Transkribus, 2024). As part of the work in digitising these fonds, an artificial intelligence model called "Dutch Handwriting 17th-19th century" was developed. This model contained 1.5 million words and could be used by any Transkribus user working with similar documents, with an error rate of 4%-10%. This model was trained to work with manuscripts written in Dutch from the 17th to the 19th centuries (Dutch handwriting 17th-19th century..., 2023). On the archive's online platform, there was a user-friendly filter with a search function for document context and information retrieval queries (help was available on how to construct such queries with different parameters). The filter also allowed users to specify a time period for the search and the archive department, where the search will take place (Fig. 5). When results were returned, the user sees brief information about the document and its inventory number, allowing for a quick assessment of the search results. When a user selects a relevant document, they were taken to a page with its digital copy and the text, which can be interacted with. Three viewing modes were available: original, transcription, and advanced transcription (Fig. 6). The document was accompanied by a full set of metadata, the option to download the original and a JSON structure file, as well as navigation through the collection. An example of the digitisation of specific collections was the experience of Jessica Sherrill (Cook), a PhD candidate in the English department at the University of California, Los Angeles. She worked on digitising Ada Lovelace's archive, which comprised approximately 14000 pages. She developed her own Lovelace AI model, which actively developed for Transkribus to work more effectively with the documents (Creating a digital scholarly edition..., 2021).

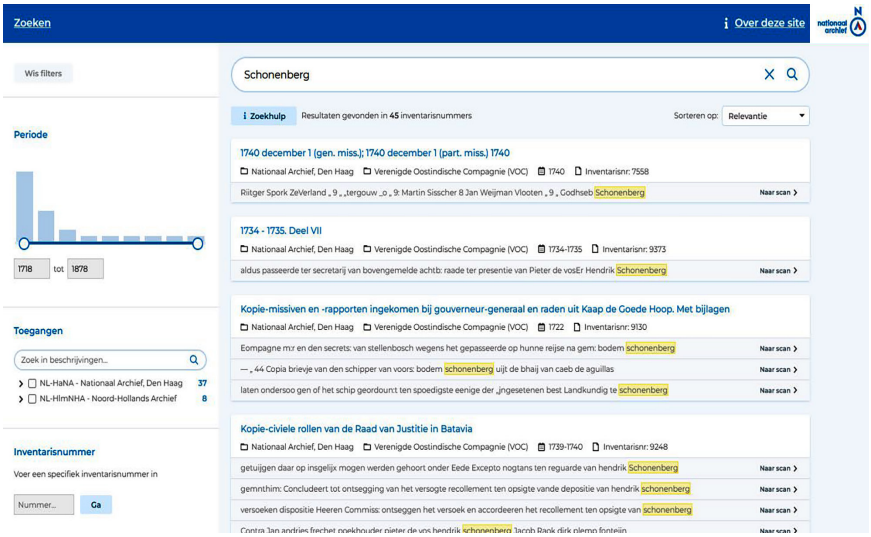


Figure 5. Search mechanism on the website of the National Archives of the Netherlands
Source: based on National Archives (2024)

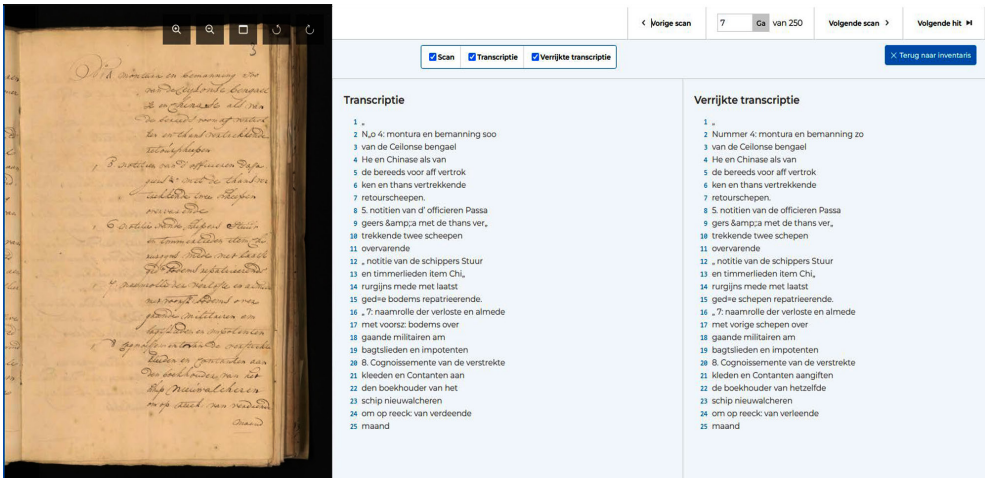


Figure 6. Example of Transkribus working with a document from the National Archives of the Netherlands
Source: based on National Archives (2024)

As part of the NewsEye project, the National Library of Finland, in collaboration with READCOOP, successfully improved the text recognition of nearly two million pages of historical Finnish newspapers using Transkribus technology, with funding from the European Union. This resulted in higher text recognition accuracy, making historical sources more accessible and userfriendly. The digitised Finnish newspapers contained birth, marriage, and death announcements, obituaries, court reports, and lists of residents, which helped to establish family connections and find ancestors. Thanks to the improved text recognition, this data became easily accessible, significantly simplifying genealogical research and information retrieval. Updated versions of the newspapers gradually replaced older ones in the digital library, starting in the summer of 2021, and the library plans further processing of newspapers published after 1914 (Unlock the past with Transkribus, 2024). An example of a digitised newspaper with contextual search capabilities and an interactive original document was shown in Figure 7. The digitisation and indexing of archival documents have become an important step in ensuring their accessibility to the wider public and have improved the research process, including genealogical research. The Arolsen Archives, the State Archives of Zurich, and the National Archives of the Netherlands have demonstrated successful examples of using OCR technologies to digitise archive groups. Their online catalogues provide convenient searching and access to digital copies of documents and the context of the documents, which has allowed for faster research and access to previously inaccessible information. This has also positively impacted the interconnectedness of data between documents within the same archive group and, overall, the preservation of historical and cultural heritage for future generations.

Figure 7. Example of a digitised document from the website of the National Library of Finland
Source: based on Savon Työmies (1920)

I. Tiurmenko *et al.* (2022) researched the digitisation processes of documents in regional state archives of Ukraine, which intensified from 2016 and peaked in 2019-2020. They noted that, although all archives already have electronic resources, the digitisation process remained unsystematic due to the lack of a unified state policy. The authors pointed out the chaotic selection of documents, problems with the structure of archive websites, and the need to expand digital collections to increase the accessibility of cultural heritage. M. Sokil *et al.* (2024) highlighted the significant losses of Ukraine's cultural heritage as a result of the Russian invasion and analysed measures for its preservation, including the creation of digital models of architectural sites.

The authors L. Salamanca *et al.* (2024) proposed a developed automation mechanism for processing and structuring the content of archival records using automatic text recognition technology, using the example of records from the Swiss Parliament from 1891 to 1995. The result of their work was the processing of over 200000 pages of documents, which exceeded the budget of most projects that used manual processing. The authors also noted that the developed mechanism not only made it possible to link documents that Swiss Members of Parliament discussed over the years, but also connected these draft laws with parliamentary speeches, legislative proposals, or votes.

M. Paliienko (2023) emphasised that despite the war and limited funding, the digital transformation of Ukrainian archives was actively continuing, contributing to the preservation of documentary heritage and the expansion of international cooperation. The integration of Ukrainian archival science into the global space, the development of educational programmes, and the attraction of financial and technological support have become important. A. Tikhonov & A. Rabus (2024) presented a universal AI model for recognising handwritten Ukrainian text on the HTR Transkribus platform, which achieved a CER of 4.2% and can be used for the mass digitisation of cultural heritage.

The research by Yu. Kovtaniuk (2023) highlighted the need to create a unified legal framework in Ukraine for the digitisation of cultural institution collections, with an emphasis on the integration of international standards and best practices. The absence of such a framework has led to technological incompatibility of electronic resources, causing inefficiency and potential risks to the preservation of fonds. S. Ferro *et al.* (2023) investigated the process of digitising historical documents using automatic handwriting recognition, applying a specific neural network (CRNN). They showed that with data augmentation techniques and fine-tuning on modern handwriting, transcription accuracy with an error rate of less than 10% can be achieved, even with a limited amount of labelled data. However, they emphasised that this method was only a supporting tool for experts, not a complete solution for the digitisation of historical documents.

S. Martínez-Cardama & A.R. Pacios (2022) analysed the strategic plans and vision statements of 159 national archives, including archives belonging to regional branches of the International Council on Archives (ICA). They found that most archives focus on preserving and digitising their collections, as well as providing access to them. However, many websites lack strategic documents, which limits citizens' access to information about future plans. One of the main problems was the digital divide, which complicated the process of digitising archives, especially in developing countries. The research highlighted the need to create digital policies and regularly update strategic plans to ensure the transparency of archives and accessibility for the public.

The National Library of Israel has actively invested in the digitisation of Hebrew manuscripts through the NLI Ktiv platform, using automatic text recognition technologies, particularly Transkribus, to transcribe

Hebrew manuscripts such as 15th-century Sephardic semi-cursive script. This process has significantly improved access to the textual content of manuscripts that were previously inaccessible and has facilitated the mass digitisation of cultural heritage. The results showed that even with small investments, great results can be achieved by using around 15000 transcribed words to train the model. This has made the mass digitisation of unpublished manuscripts easier, opening up wide possibilities for future research (Prebor, 2024).

S. Spina (2023) considered the impact of artificial intelligence on the digitisation processes of archival heritage, particularly regarding the automatic recognition of manuscripts, their correction, and normalisation. The influence of digitisation on the re-evaluation by scholars of the role of the archive and history for processing large amounts of data was emphasised. The article provided an analysis of two artificial intelligence-based systems for text digitisation, namely Transkribus and ChatGPT.

The analysis of academic studies had shown an intensification of the digitisation processes of archive groups both in Ukraine and abroad. Researchers have emphasised the importance of implementing modern technologies, particularly OCR and artificial intelligence-based solutions, to automate the processing of archival documents and processes, preserve cultural heritage, and ensure open access to them. The need to create a unified legal framework, strategic planning, and the development of digital archival infrastructure has been highlighted. International experience has demonstrated the effectiveness of using such tools, which have significantly accelerated the pace of digitisation and ensured the preservation of archive groups.

Conclusions

In modern world, where information played a crucial role, converting paper documents into digital formats had become a necessary step for preserving historical and cultural heritage. The digitisation and indexing of archival documents have become important for ensuring their accessibility to a wide audience, including researchers, historians, and genealogists. Many unique and valuable archive groups have been put at risk of destruction or loss. The effective digitisation of these documents and the provision of remote access to them have become important tasks for the state.

The use of automated text recognition technology has accelerated the pace of digitising archive groups and provided opportunities to improve the working processes of archival institutions. Depending on the chosen platform, it had simplified access to archive groups, provided a convenient tool for searching information within archive groups and documents, automatic classification of documents, and the ability to analyse large amounts of data for various research purposes.

The use of OCR had been particularly valuable from the perspective of genealogical research. Using information-retrieval queries, researchers had been able to automatically search for the necessary data and conduct in-depth analysis based on it. The use of OCR has made it possible to process large volumes of data in a short time, which has significantly sped up the information retrieval process. It also provided the ability to access digitised documents remotely, and the digitised data can be used to conduct statistical analysis and identify patterns in family relationships. This has greatly facilitated and accelerated the process of finding information about ancestors and building a family tree.

Considering the positive experiences of the Arolsen Archives, the State Archives of Zurich, and the National Archives of the Netherlands, adopting similar approaches in Ukrainian archival institutions would allow for the significant potential of modern technologies to be used to update archival practices. Applying their experience in Ukraine could substantially improve the processes of digitisation, searching, and access to archival documents, opening up new possibilities for research.

Further research into the digitisation of genealogical documents will involve studying approaches to building a centralised repository of archival documents, taking into account the experience of global institutions, and developing methodological recommendations for the implementation of such a system in Ukraine.

Acknowledgements

None.

Conflict of Interest

None.

References

[1] Amazon rekognition. (2024). *Amazon Web Services.* Retrieved from https://aws.amazon.com/rekognition/?nc1=h_ls.

[2] Artemenkova, O. (2022). Information tools of genealogical research in the archives of Ukraine. *Visnyk of Kharkiv State Academy of Culture*, 61, 81-93. doi: 10.31516/2410-5333.061.08.

[3] Artemenkova, O. (2023). *Information technologies as a tool for popularizing genealogical research in the archives of Ukraine*. (Doctoral dissertation, Kyiv National University of Culture and Arts, Kyiv, Ukraine).

[4] Creating a digital scholarly edition of the Lovelace papers with Jessica Cook. (2021). *Transkribus*. Retrieved from https://www.transkribus.org/success-story/lovelace-cook.

[5] *Digital transformations in Ukraine: Do domestic institutional conditions meet external challenges and the European agenda?* (2020). Chernihiv: Polissya Foundation for International and Regional Studies.

[6] Documentation and archiving. (2024). *Arolsen Archives.* Retrieved from https://arolsen-archives.org/en/about-us/what-we-do/documentation-and-archiving/.
[7] *Dutch handwriting 17th-19th century. Free public AI model for handwritten text recognition with Transkribus*. (2023). The Hague: National Archives Netherlands.
[8] Ferro, S., Pelillo, M., & Traviglia, A. (2023). AI-assisted digitalisation of historical documents. In *The international archives of the photogrammetry, remote sensing and spatial information sciences. 29th CIPA symposium "Documenting, understanding, preserving cultural heritage: Humanities and digital technologies for shaping the future"* (Vol. XLVIII-M-2-2023, pp. 557-562). Florence, Italy. doi: 10.5194/isprs-archives-XLVIII-M-2-2023-557-2023.
[9] Friedewald, M., Székely, I., & Karaboga, M. (2024). Preserving the past, enabling the future: Assessing the European policy on access to archives in the digital age. *Preservation, Digital Technology & Culture*, 53(2), 61-71. doi: 10.1515/pdtc-2024-0003.
[10] IBM cloud pak for business automation. (2022). *IBM.* Retrieved from https://www.ibm.com/products/cloud-pak-for-business-automation.
[11] Khoma, I., Vovk, N., Holoshchuk, R., & Muravska, S. (2023). Promoting the Ukrainian education and culture centre "Oseredok" through the digitization of Ukrainian studies archival collections in Canada. In *SCIA-2023: 2nd international workshop on social communication and information activity in digital humanities*. Lviv, Ukraine.
[12] Korzhyk, N., Solianyk, A., Borysova, A., & Aleksander, M. (2023). State archives in Ukraine during the russian aggression: Challenges and achievements. In *SCIA-2023: 2nd international workshop on social communication and information activity in digital humanities*. Lviv, Ukraine.
[13] Kovalska, L. (2019). Document communication of archive information users. *Intercultural Communication*, 6, 231-248. doi: 10.13166/inco/103415.
[14] Kovtaniuk, Yu. (2023). Normative and legal regulation of digitization of fonds of cultural institutions as requirement for development of state integration electronic information resources of national historical and cultural heritage. *Manuscript and Book Heritage of Ukraine*, 31, 379-406. doi: 10.15407/rksu.31.379.
[15] Lipianina-Honcharenko, K., Yarych, V., Ivasechko, A., Filinyuk, A., Yurkiv, K., & Lebid, T. (2024). Evaluating the effectiveness of attention-gated-CNNBGRU models for historical manuscript recognition in Ukraine. In *The first international workshop of young scientists on artificial intelligence for sustainable development*. Ternopil, Ukraine.
[16] Logvynenko, B., *et al.* (2024). Anatolii Khromov: On digitisation, preservation and openness of archives. *Ukrainer.* Retrieved from https://www.ukrainer.net/anatoliy-khromov/.
[17] Martínez-Cardama, S., & Pacios, A.R. (2022). National archives' priorities: An international overview. *Archival Science*, 22, 1-42. doi: 10.1007/s10502-021-09367-y.
[18] National archives. (2024). Retrieved from https://www.nationaalarchief.nl/.
[19] Nockels, J., Gooding, P., & Terras, M. (2024). The implications of handwritten text recognition for accessing the past at scale. *Journal of Documentation*, 80(7), 148-167. doi: 10.1108/JD-09-2023-0183.
[20] Onuchak, V. (2024). Electronic archives in Ukraine: How the state plans to store electronic documents. *Vchasno. EDO*. Retrieved from https://vchasno.ua/elektronni-arhivy-ukrainy/.
[21] Order of the Cabinet of the Ministers of Ukraine No. 1353-r "On Approval of the Strategy for Digital Transformation of the Social Sphere". (2020, October). Retrieved from https://zakon.rada.gov.ua/laws/show/1353-2020-%D1%80#n10.
[22] Paliienko, M. (2023). Rethinking approaches to archival theory and practice in Ukraine in the context of digital transformation of the society. *Atlantic +*, 33(2), 83-99.
[23] Prebor, G. (2024). From digitization and images to text and content: Transkribus as a case study. *Manuscript Studies: A Journal of the Schoenberg Institute for Manuscript Studies*, 9(1), 72-89. doi: 10.1353/mns.2024.a930877.
[24] Report on the implementation of the Ministry of Digital Transformation of Ukraine's work plan for 2023. (2023). *Official website of the Ministry of Digital Transformation of Ukraine*. Retrieved from https://thedigital.gov.ua/community/reports.
[25] Rybachok, O. (2018). *International integrated digital resources of documentary heritage of archives, libraries, museums: Stages of creation, development strategies (1980s-2010s)*. (PhD dissertation, Vernadsky National Library of Ukraine, Kyiv, Ukraine).
[26] Salamanca, L., Brandenberger, L., Gasser, L., Schlosser, S., Balode, M., Jung, V., Perez-Cruz, F., & Schweitzer, F. (2024). Processing large-scale archival records: The case of the Swiss parliamentary records. *Swiss Political Science Review*, 30(2), 140-153. doi: 10.1111/spsr.12590.

[27] Savon Työmies. (1920). *Sanomlehdet*. Retrieved from https://digi.kansalliskirjasto.fi/sanomalehti/binding/3112293?page=1.
[28] Silva, A.L., & Terra, A.L. (2023). Cultural heritage on the semantic web: The Europeana data model. *IFLA Journal*, 50(1), 93-107. doi: 10.1177/03400352231202506.
[29] Sokil, M., Syerov, Y., & Boiko, V. (2024). From destruction to digitization: Safeguarding Ukraine's cultural and archival heritage in wartime. In P. Štarchoň, S. Fedushko & K. Gubíniová (Eds.), *Data-centric business and applications. Lecture notes on data engineering and communications technologies* (Vol. 208, pp. 253-280). Cham: Springer. doi: 10.1007/978-3-031-59131-0_12.
[30] Spina, S. (2023). Artificial intelligence in archival and historical scholarship workflow: HTS and ChatGPT. *Digital Humanities*, 16, 125-140. doi: 10.6092/issn.2532-8816/17205.
[31] State Archive of Zaporizhzhia Region. (2024). Retrieved from https://archivzp.gov.ua/uk/.
[32] Tesseract OCR. (2024). *GitHub*. Retrieved from https://github.com/tesseract-ocr/tesseract.
[33] Tikhonov, A., & Rabus, A. (2024). Handwritten text recognition of Ukrainian manuscripts in the 21st century: Possibilities, challenges, and the future of the first generic AI-based model. *Kyiv-Mohyla Humanities Journal*, 11, 226-247. doi: 10.18523/2313-4895.11.2024.226-247.
[34] Tiurmenko, I., Bozhuk, L., Struk, I., & Syerov, Y. (2022). Digital documentary collections of national cultural heritage on the Ukrainian regional state archives websites. In N. Kryvinska & M. Greguš (Eds.), *Developments in information & knowledge management for business applications. Studies in systems, decision and control* (Vol. 421, pp. 449-470). Cham: Springer. doi: 10.1007/978-3-030-97008-6_20.
[35] Unlock the past with Transkribus. (2024). *Transkribus*. Retrieved from https://www.transkribus.org/.
[36] User guide – introduction. (2024). *OCR4all.org*. Retrieved from https://www.ocr4all.org/guide/user-guide/introduction.
[37] Vision AI: Extract insights from images, documents, and videos. (2024). *Cloud Vision API*. Retrieved from https://cloud.google.com/vision?hl=en.
[38] What is OCR (Optical Character Recognition)? (2024). *Amazon Web Services*. Retrieved from https://aws.amazon.com/what-is/ocr/.
[39] What is optical character recognition (OCR)? (2024). *IBM*. Retrieved from https://www.ibm.com/think/topics/optical-character-recognition.
[40] Zurich council manuals 1642-1798. (2024). *Canton of Zurich State Archives*. Retrieved from https://ratsmanuale-zuerich.transkribus.eu/.

Цифровізація генеалогічних документів на основі технології автоматичного розпізнавання тексту

Артур Спектор
Аспірант
Національний університет «Львівська політехніка»
79013, вул. Степана Бандери, 12, м. Львів, Україна
https://orcid.org/0000-0003-4176-9177

Анотація. Метою дослідження було висвітлення наявних підходів та технологій щодо цифровізації генеалогічних документів, спираючись на світовий досвід. Це надало змогу більш ефективніше організувати процеси та механізми оцифрування архівних фондів, їх централізоване зберігання, пришвидшення проведення генеалогічних досліджень та доступність для користувачів. Оцифрування архівів стало критично важливим аспектом для збереження культурної спадщини, особливо в умовах воєнної агресії росії проти України. Впровадження технологій автоматичного розпізнавання тексту сприяло оптимізації цього процесу, полегшуючи доступ до інформації та підвищуючи ефективність досліджень, зокрема генеалогічних. У роботі було проаналізовано принципи роботи Optical Character Recognition, його переваги, особливості готових рішень і функціональні можливості програмного забезпечення на його основі. Було оцінено стратегію цифровізації в Україні, проблеми архівної галузі в сфері оцифрування та доступу до архівних фондів. Досліджено результати впровадження автоматичного розпізнавання тексту в провідних архівах світу, а також можливості онлайн-архівів з функціоналом контекстного пошуку. Приділено увагу тим можливостям, які відкриваються перед дослідниками завдяки впровадженню подібних систем у роботу архівів, зокрема зручності здійснення пошуку необхідної інформації, підвищенню швидкості обробки даних, а також забезпеченню цілодобового доступу до архівних ресурсів незалежно від географічного розташування користувачів. Також, було проаналізовано роботи вчених, які займалися розробкою та впровадженням Optical Character Recognition в архівних установах. На основі міжнародного досвіду визначено потенціал сучасних технологій Optical Character Recognition для модернізації архівної справи в Україні, що може позитивно вплинути на генеалогічні дослідження та збереження культурної спадщини. Практична цінність дослідження полягає в підтвердженні ефективності застосування інформаційних технологій для покращення процесу оцифрування архівних документів та забезпечення кращого доступу до них. Запропоновані рекомендації допоможуть оптимізувати організацію цифрових архівів, вдосконалити процеси зберігання та пошуку документів, а також прискорити генеалогічні дослідження. Це сприятиме збереженню культурної спадщини та покращенню доступу до архівної інформації для користувачів

Ключові слова: архівні фонди; оцифрування; генеалогічні дослідження; Optical Character Recognition; інформаційні технології; автоматизація