https://doi.org/10.31516/2410-5333.065.14[1]
УДК 004.912

В. Г. Курило
аспірант спеціальності «Інформаційна, бібліотечна та архівна справа», Національний університет «Львівська політехніка», м. Львів, Україна
vasyl.h.kurylo@lpnu.ua
https://orcid.org/0000-0003-1819-2269

М. В. Комова
доктор наук із соціальних комунікацій, професор, доцент, кафедра соціальних комунікацій та інформаційної діяльності, Національний університет «Львівська політехніка», м. Львів, Україна
mariia.v.komova@lpnu.ua
https://orcid.org/0000-0002-4115-3690

ІНТЕГРАЦІЯ ПРОГРАМНИХ ЗАСОБІВ ОЦИФРУВАННЯ ДОКУМЕНТІВ В ЕЛЕКТРОННУ СИСТЕМУ АРХІВУ

Стаття присвячена дослідженню актуальних засобів, за допомогою яких можна здійснювати процес оцифрування документів в архівах та установах, де відбувається зберігання документів. Вибір оптимального засобу оцифрування документів в архівах за допомогою сучасних рішень дозволить швидко та якісно зберегти великі масиви інформації. Сучасні технології оцифрування розглядаються в контексті встановлення їхніх переваг та недоліків, а також класифікації документів у цифровому середовищі. Проаналізовано виклики, які виникають під час переходу від традиційного аналогового до цифрового зберігання документів, зокрема це проблеми з класифікацією оцифрованих документів, безпекою та збереженням інтегрітету інформації. Розроблено методичні рекомендації щодо вибору оптимальних засобів оцифрування з метою забезпечення ефективного зберігання й структурування документації в архівній сфері.

Ключові слова: *архів, оцифрування документів, класифікація документів, автоматичне розпізнавання, OCR.*

V. Kurylo
postgraduate student at the Department of Social Communications and Information Activity, Lviv Polytechnic National University, Lviv, Ukraine

M. Komova
Doctor of Sciences in Social Communications, professor, Assistant Professor, Department of Social Communications and Information Activity, Lviv Polytechnic National University, Lviv, Ukraine

INTEGRATION OF DOCUMENT DIGITALIZATION SOFTWARE IN THE ELECTRONIC SYSTEM OF THE ARCHIVE

The relevance of the research. The modern world is rapidly moving to a digital format for storing and processing information. Archives, like any other organization, must adapt to the trend by introducing digital technologies to preserve their documents. Choosing the appropriate means of digitizing documents in archives using modern

solutions will allow quickly and efficiently preserve large amounts of information. Researching, selecting, and using appropriate document classification algorithms will help speed up the process of sorting documents into the right categories.

The purpose of the article is to analyze the functionality of document digitization software for their integration into electronic archive systems.

The methodology. To accomplish this objective, various theoretical and scientific research approaches were utilized, including analysis, synthesis, induction, and deduction.

The results. The article highlights modern technologies and software designed for document digitization, explores existing problems that arise during document digitization and classification and developed recommendations for the integration of digitization systems into the archive.

The scientific novelty. The article highlights and summarizes recent scientific publications related to the development and enhancement of document digitization tools. It reviews modern research conducted by scientists in the field of handwritten text and image recognition. Additionally, it presents general methodological recommendations for integrating digitization systems into archives.

The practical significance. The article will be useful to scientists, software developers and archive workers. The presented research results are relevant and structured. This scientific article can become the basis for further research into document digitization systems.

The conclusion. A significant breakthrough occurred with the development of optical character recognition (OCR) technology. Building upon this, more sophisticated algorithms have been developed for recognizing handwritten text in various languages. Presently, there is ongoing research focused on achieving high-quality digitization of images, enabling efficient search capabilities within vast collections of digitized materials. Artificial intelligence has greatly simplified and accelerated work in this direction, but challenges persist with recognition accuracy. Currently, research efforts continue, and a reference database is being developed to facilitate greater automation of the classification process. The formulated methodology for integrating digitization tools into the archive enhances comprehension of the process and simplifies the preparatory tasks associated with integration.

Keywords: *archive, document digitization, document classification, automatic recognition, OCR.*

Актуальність теми дослідження. Сучасний світ стрімко переходить до цифрового формату зберігання та опрацювання інформації. Архіви, як і будь-які інші організації, повинні адаптуватися до цього тренду, впроваджуючи цифрові технології для зберігання документів. Відбувається процес генерування великих обсягів документації, інформація з яких може бути цінною і важливою для майбутніх поколінь або для вирішення поточних питань. Вибір оптимального засобу оцифрування документів в архівах за допомогою сучасних рішень дозволить швидко та якісно зберегти великі масиви інформації.

Крім того, архіви повинні ефективно керувати збереженою документацією, забезпечуючи швидкий і легкий доступ до неї. Дослідження, вибір та застосування відповідних алгоритмів класифікації документів допоможе

пришвидшити процес розподілу документів на потрібні категорії й, як наслідок, збільшити пропускну здатність електронних баз в архівах.

Зважаючи на ситуацію в Україні, виникає потреба в гарантуванні безпеки зберігання документів в архівах. Застосування засобів оцифрування та зберігання цифрових копій паперових оригіналів дозволить значно зменшити ризики втратити їх під час нещасних випадків, а якісно підібрані алгоритми класифікації дозволять визначати цінність документів (юридичну, культурну) та виконати пріоритезацію їхнього зберігання в хмарних середовищах чи на віддалених серверах.

Постановка проблеми. У сучасному інформаційному суспільстві обсяги опрацьованих документів зростають експоненційно, що зумовлює появу певних викликів щодо зберігання документів, управління та доступу до них. Із такими складнощами стикаються не лише великі архіви, а й окремі архівні відділи на підприємствах чи навіть персональні архіви. Відтак, кожен окремий випадок потребує гнучкого підходу для задоволення індивідуальних потреб. В Україні відсутня єдина нормативно-правова база, яка регулювала б вимоги до виготовлення цифрових копій документів як кінцевого продукту. Це актуалізує необхідність дослідити наявні програмні засоби оцифрування та класифікації документів, здійснити аналіз рівня розвитку сучасних рішень й описати можливий процес інтеграції відповідних інструментів в архіви.

Аналіз останніх досліджень і публікацій. В актуальних наукових дослідженнях проблем оцифрування документів та класифікації їх в архівах розглядається термін оцифрування, який визначено як діяльність, у якій цифрова інформація про об'єкти та їхній контекст можуть бути об'єднані та згруповані в єдину систему (Navarrete & Owen, 2011).

Світовий досвід демонструє, що процес оцифрування паперових документів стає стандартною практикою в бібліотеках, архівах та музеях різних країн (Lischer-Katz, 2022). Глобальний тренд до диджиталізації стимулює науковців та розробників у різних куточках світу створювати нові й поліпшувати вже існуючі алгоритми оцифрування. Відтак, було розроблено програмне забезпечення на основі генетичних алгоритмів для розпізнавання китайського рукописного тексту (Liang et al., 2020) та арабських символів (Balaha et al., 2021). Продемонстровано можливість графічного розпізнання складних текстів, у зв'язку із чим можна передбачити вдосконалення технології розпізнавання інших нестандартних шрифтів.

Такі нові технології, як глибоке навчання та растрова сегментація зображень, розширюють функціонал процесів оцифрування (Shen et al., 2021). Наразі існує складність із пошуком зображень у великих масивах оцифрованих документів, проте було розроблено алгоритм створення постійних

виразів для тегання, за допомогою яких здійснюється пошук (Yurtsever et al., 2021). Використання такого методу може значно спростити роботу із цифровим графічним матеріалом.

Стає очевидною потреба в порівнянні нових алгоритмів та методів, які мають перспективи замінити застарілі. Прикладом такого дослідження є зіставлення застосування методів гістограм і тензорного числення для сегментації рукописного тексту (Babczyński & Ptak, 2024). Гістограми являють собою графіки розподілу щільності тексту. Під час тензорного числення відбувається процес накопичення інформації про аналогічний текст. У результаті було виявлено, що обидва методи потребують вдосконалення, хоча й демонструють задовільну якість розпізнавання (≈ 70%).

Дослідження публікаційної діяльності науковців у сфері оцифрування документів свідчить, що тематика активно розвивається. Відбувається процес систематизації методів оцифрування, а це підкреслює важливість дослідження актуальних рішень та наявних викликів.

Мета статті — проаналізувати функціональні можливості програмних засобів оцифрування документів задля їх інтеграції в електронні системи архівів. Головними завданнями дослідження є:

– аналіз актуальних технологій та програмних засобів, призначених для оцифрування документів;
– дослідження викликів, які виникають під час оцифрування та класифікації документів;
– розроблення методичних рекомендацій щодо інтеграції систем оцифрування в електронні системи архівів.

Виклад основного матеріалу дослідження. Процес оцифрування документів являє собою перетворення паперових документів у відповідні цифрові формати. Прикладами таких форматів можуть бути .pdf, .doc або спеціалізовані бази даних, залежно від фінальних потреб та цифрового середовища. В Україні немає чітко сформованої нормативно-правової бази щодо оцифрування архівів (Ковтанюк, 2023), де зазначався б єдиний перелік технологій, а отже, й програмного забезпечення, яке слід використовувати в архівах. Український науково-дослідний інститут архівної справи та документознавства (УНДІАСД) виокремлює три основні методи оцифрування архівних документів:

– фотографічне копіювання;
– електрографічне копіювання;
– сканування (Гаранін та ін., 2012).

Для оцифрування можна застосовувати такі методи, як сканування, цифрові технології збору даних, оптичне розпізнавання символів (OCR). Найпоширенішим із цих методів є сканування, під час якого відбувається

перетворення інформації з фізичних носіїв на електронні зображення з використанням сканерів. Залежно від призначення сканери можуть бути спеціалізовані, промислові або для масового використання.

Проте для якіснішої роботи з документами потрібно зробити не тільки статичну цифрову копію, а й уможливити опрацювання (перегляд, редагування) інформації, яка міститься в ньому. Це стає можливим завдяки сучасним методам розпізнавання символів. Одним із таких є технологія оптичного розпізнавання символів (OCR). Вона здатна трансформувати скановані зображення з текстом, яке стає доступним для пошуку та редагування. Програмні засоби OCR здійснюють розпізнавання символів у зображенні і перетворюють їх на текст для машинного читання.

Наразі існує чимало інструментів оптичного розпізнавання символів — програмне забезпечення комерційного штибу чи з відкритим вихідним кодом. Виокремлюють також офлайн- та онлайн-інструменти.

Серед доступних, універсальних і популярних програмних засобів для автоматичного розпізнавання текстового друкованого вмісту можна означити такі, як Tesseract та OCRopus, які розповсюджуються безкоштовно. Має свій сегмент користувачів і програмний засіб ABBYY Finereader. До систем оптичного розпізнавання, які мають технічну підтримку, також належать Infty Reader, OCR.space, AFR. У попередні роки було створено й інші програми для розпізнавання тексту, проте багато з них вже не отримують підтримку або розвиток.

Технології OCR — актуальні і такі, що часто використовуються. Функціонування цих технологій базується на алгоритмах розпізнавання за допомогою шаблонів символів і нейронних мереж. Робота в шаблонах розпізнавання здійснюється на основі бібліотеки символів, за якими відбуваються пошук та порівняння на зображенні документа. Нейронні мережі ж є прогресивнішими та складнішими системами. Завдяки їхньому розвитку стали можливими складніші методи розпізнавання тексту, які вже не обмежуються впізнаванням окремих символів, а здатні розпізнавати слова, речення або навіть фрази в оцифрованому документі.

Якщо порівнювати прогресивний функціонал програмного забезпечення, то OCRopus має можливість виконувати автоматичну корекцію переносів рядка тексту за допомогою алгоритмів глибокого навчання, аналізувати структуру сторінки та бінарну морфологію. Завдяки інструментам програми можна моделювати рядки тексту й виконувати їх пошук. Виконання таких функцій особливо важливе в разі необхідності якісного оцифрування документів з нетиповим розміщенням тексту на сторінці, рукописів, маргіналій.

Комерційне програмне забезпечення ABBYY Finereader здатне якісно сканувати та розпізнавати текст завдяки широкому функціоналу. У програмі можна виконувати елементарне редагування зображення, розбиття

сторінки на частини. Це важливо, коли оцифрований документ має складну структуру або складається з різних типів інформаційного наповнення. Для детальної корекції тексту програма пропонує функції виправлення слів відповідно до інтегрованого словника, визначення переносів слів. Стає очевидною зацікавленість розробників ABBYY Finereader у комфорті користування програмою, адже її інтерфейс інтуїтивно зрозумілий. Це зумовлює швидке освоєння програми та відкриває перспективи використання різними установами.

Вагомим чинником, який вливає на доступність програмних засобів оцифрування, є їхня здатність до інтеграції у вже наявні електронні системи архівів. Відтак, Tesseract є програмним засобом, керувати яким можна за допомогою командного рядка. Він сумісний з такими популярними мовами програмування, як JavaScript, Python. Останні версії Tesseract не лише розпізнають текст, а й присвоюють словам індивідуальні координати для більш надійного позиціювання та редагування.

Раніше згадані програмні засоби мають здатність інтеграції в різні установи. Проте слід відзначити спеціалізовані рішення для бібліотек та архівів, наприклад Transkribus. Це проєкт, який позиціює себе як технологія розпізнавання тексту за допомогою нейронних мереж та машинного навчання. Уважається, що це одна з перших платформ, яка зробила цю технологію доступною (Muehlberger et al., 2019). Це самостійний програмний засіб, орієнтований на такі категорії користувачів: гуманітарії, архівісти, фахівці з комп'ютерних наук. Алгоритм роботи програми базується на методі сегментації відсканованого зображення на складові елементи, застосування створених нейронною мережею моделей та пошуку ключових слів для кращого підбору моделей. Проєкт продовжує вдосконалюватися і розвиватися для поліпшення якості розпізнавання та оцифрування текстових документів.

Засоби, які використовуються для оцифрування архівних документів, повинні мати можливість якісно оцифровувати та розпізнавати не лише текст, а й зображення: ілюстрації, герби, печатки, інші документи та їхні компоненти, які мають юридичну й історичну цінність. У такому випадку доводиться використовувати різне програмне забезпечення, залежно від ситуації. Для якісного розпізнавання виключно нетекстових елементів потрібно вдосконалити наявні рішення, розробити методики та створити референтні тестові колекції (Colesnicov et al., 2020).

Для сучасної архівної справи важливим є не лише якісне оцифрування вмісту документів, а й їхня подальша класифікація за змістом, контекстом та призначенням. Цифрові документи, які містять додаткову інформацію (метадані), дозволяють класифікувати та організовувати їх у базах даних. Такими метаданими можуть бути ключові слова, дати створення та типи

документів. Алгоритми машинного навчання та штучного інтелекту (ШІ) постійно покращують й удосконалюють механізми класифікації документів, автоматизуючи рутинні завдання. ШІ за допомогою технологій опрацювання природної мови (NLP) може аналізувати вміст оцифрованих документів, призначати їм категорії відповідно до попередньо визначених правил.

Одним з викликів, який постав перед розробниками програмного забезпечення, є поліпшення алгоритмів розпізнавання схожих паралельних елементів тексту в різних документах для вдосконалення автоматичної класифікації (Harris et al., 2019). Складність полягає в тому, що семантично схожі уривки тексту можуть мати лексичні та синтаксичні варіації. Неточність класифікації оцифрованих документів може виникати у зв'язку з використанням діалектів чи перефразуванням тексту.

На нашу думку, перспективним напрямом досліджень є напрацювання методологічних підходів до визначення відмінностей між текстами та оцінювання їхньої контекстної важливості. На основі цих даних розробники програмного забезпечення створюють референтні бази й вдосконалюють алгоритми.

Дослідження засобів оцифрування та класифікації виявили, що сучасний рівень розвитку інструментарію недосконалий. Здебільшого програми спеціалізуються на статичному оцифруванні й розпізнаванні тексту. Алгоритми для розпізнавання рукописного тексту розроблено, проте все ще виникають певні труднощі з якісним розпізнаванням та маркуванням зображень. Тому є практична потреба в розробленні узагальненої методології щодо інтеграції систем оцифрування в електронні системи архівів. Ці рекомендації сприятимуть оптимальному та ефективному впровадженню цифрових технологій у роботу архівів.

Першочергово слід виконати аналіз потреб архіву, проаналізувати поточні процеси зберігання й оцінити перспективи інтеграції системи оцифрування. Необхідно дослідити структуру архівного фонду та типів документів, оцінити кількість друкованого тексту, рукописного тексту й зображень. Обґрунтованість і повнота зібраних фактичних відомостей про кількісні та якісні характеристики архівного фонду, який підлягає оцифруванню, зумовлюють ефективність вибору інструментів. Наступним кроком є вибір системи оцифрування, зважаючи не лише на потреби архіву, а й на фінансові можливості, доцільність використання спеціалізованих комплексних рішень. Ринок програмних засобів містить широкий спектр різної за можливостями та доступністю продукції, яка забезпечить реалізацію завдань цифровізації архівів. Враховуючи функціональні характеристики обраної системи, варто розробити детальний поетапний план інтеграції засобів оцифрування, який передбачає підготування документів, опрацювання та

архівацію. Підготування інфраструктури архіву до інтеграції системи оцифрування передбачає наявність і готовність обладнання, мережі та програмного забезпечення. На ринку є програмне забезпечення, яке встановлюється окремо та працює незалежно від інших програм, або програми-плагіни, які є надбудовами до іншого архівного програмного забезпечення. Підвищення кваліфікації персоналу із цифрової грамотності на навчальних курсах, розроблення навчальних та інструктивних матеріалів забезпечить ефективність упровадження засобами оцифрування, освоєння нового програмного забезпечення.

Тестування системи доцільно виконати на пілотній базі перед повним впровадженням й перевірити коректність роботи інструментів, оцінити якість оцифрування та розпізнавання рукописного тексту й зображень. Інтегрувавши засоби оцифрування в електронну систему архіву, необхідно виконувати моніторинг та підтримання програмних засобів, здійснювати регулярне оновлення системи з метою забезпечення її надійності й ефективності.

Такий алгоритм цифровізації є актуальним для архівів різного типу, оскільки сучасні системи оцифрування документів володіють доволі широким функціоналом та можливостями до інтеграції.

Висновки. Оцифрування як явище в українській та світовій архівній практиці базується на стрімкому розвитку технологій, які стосуються засобів та методів зберігання інформації. Метод оптичного розпізнавання символів уможливив опрацювання з оцифрованим текстом. На його основі створюються складніші алгоритми для розпізнавання рукописного тексту різними мовами. Ведуться активні дослідження щодо якісного оцифрування зображень з подальшою можливістю їх пошуку у великому масиві оцифрованого матеріалу. Штучний інтелект значно оптимізував роботу в цьому напрямі, проте досі існують труднощі з точністю розпізнання.

Не менш важливою для архівів є можливість не лише оцифрувати та розпізнати інформацію, а й класифікувати її для подальшого розподілу на категорії. Досі не вдалось автоматизувати цей процес, оскільки програмні засоби повинні не тільки розпізнати інформацію, а й зрозуміти контекст. Нині тривають активні дослідження та напрацьовується база референтних даних, за допомогою яких стане можливим вищий рівень автоматизації процесу класифікації.

Хоча в Україні в останні роки проводиться активна цифровізація, велика кількість архівів та установ досі не використовують нові функціональні системні засоби для оцифрування, розпізнавання й класифікації документів. Загальна методологія інтеграції засобів оцифрування в роботу архівів сприяє кращому розумінню процесу та спрощує підготовчі роботи при інтеграції.

Перспективи подальших досліджень. Актуальним напрямом подальших досліджень є розроблення систем розпізнавання вмісту, систем, які будуть здатні розуміти контекст документів. Це надасть змогу ефективніше здійснювати пошук необхідної інформації, а також допоможе створювати засоби для автоматичної класифікації й типології архівних документів.

Водночас є потреба в дослідженнях вітчизняного досвіду щодо оцифрування архівних фондів. Варто виконати аналіз якості та точності оцифрування текстів і зображень в архівах України. Це дозволить оцінювати й порівнювати різні засоби оцифрування та їхні можливості під час масового використання.

Список посилань

Гаранін, О., Христова, Н., & Срібняк, І. (2012). *Вплив копіювально-розмножувальної техніки на збереженість архівних документів*. УНДІАСД, Київ. Отримано з https://undiasd.archives.gov.ua/doc/mr-vpluvtex.pdf

Ковтанюк, Ю. (2023). Нормативно-правове регулювання оцифрування фондів закладів культури як вимога розвитку державної інтеграції електронних інформаційних ресурсів національної історико-культурної спадщини. *Рукописна та книжкова спадщина України, (31),* 379–406. doi:10.15407/rksu.31.379

Babczyński, T., & Ptak, R. (2024). Direct Tensor Voting in line segmentation of handwritten documents. *International Journal of Electronics and Telecommunications*, 95–102. doi:10.24425/ijet.2024.149519

Balaha, H. M., Ali, H. A., Youssef, E. K., Elsayed, A. E., Samak, R. A., Abdelhaleem, M. S., ... & Mohammed, M. M. (2021). Recognizing arabic handwritten characters using deep learning and genetic algorithms. *Multimedia Tools and Applications, 80,* 32473-32509. doi:10.1007/s11042-021-11185-4

Colesnicov, A., Malahov, L., Cojocaru, S., & Burtseva, L. (2020). Semi-automated workflow for recognition of printed documents with heterogeneous content. *Computer Science Journal of Moldova, 84 (3),* 223–240. Retrieved from https://ibn.idsi.md/sites/default/files/imag_file/v28-n3-%28pp223-240%29.pdf

Harris, M., Levene, M., Zhang, D., & Levene, D. (2019). Comparing “parallel passages” in digital archives. *Journal of Documentation, 76 (1),* 271–289. doi:10.1108/jd-10-2018-0175

Liang, J., Wang, H., & Li, X. (2020). Task design and assignment of full-text generation on mass chinese historical archives in digital humanities: a crowdsourcing approach. *Aslib Journal of Information Management*, *72* (2), 262–286. doi:10.1108/AJIM-09-2019-0245

Lischer-Katz, Z. (2022). The emergence of digital reformatting in the history of preservation knowledge: 1823–2015. *Journal of Documentation*, *78 (6),* 1249–1277. doi:10.1108/jd-04-2021-0080

Muehlberger, G., Seaward, L., Terras, M., Oliveira, S. A., Bosch, V., Bryan, M., ... & Zagoris, K. (2019). Transforming scholarship in the archives through

handwritten text recognition: Transkribus as a case study. *Journal of documentation, 75 (5),* 954–976. doi:10.1108/jd-07-2018-0114

Navarrete, T., & Mackenzie Owen, J. (2011). Museum libraries: how digitization can enhance the value of the museum. *Palabra clave, 1 (1).* Retrieved from https://www.researchgate.net/publication/265069291

Shen, Z., Zhang, R., Dell, M., Lee, B. C. G., Carlson, J., & Li, W. (2021). Layoutparser: A unified toolkit for deep learning based document image analysis. In *Document Analysis and Recognition–ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part I 16* (pp. 131–146). Springer International Publishing. doi:10.1007/978-3-030-86549-8_9

Yurtsever, M. M. E., Özcan, M., Taruz, Z., Eken, S., & Sayar, A. (2022). Figure search by text in large scale digital document collections. *Concurrency and Computation: Practice and Experience, 34 (1),* e6529. doi:10.1002/cpe.6529

References

Haranin, O., Khristova, N., & Sribniak, I. (2012). *The influence of copying and reproduction technology on the preservation of archival documents*. URIAARM, Kyiv. Retrieved from https://undiasd.archives.gov.ua/doc/mr-vpluvtex.pdf. [In Ukrainian].

Kovtaniuk, Y. (2023). Normative and legal regulation of digitization of funds of cultural institutions as requirement for development of state integration electronic information resources of national historical and cultural heritage. *Rukopisna ta knižkova spadśina Ukraïni, (31),* 379–406. doi:10.15407/rksu.31.379. [In Ukrainian].

Babczyński, T., & Ptak, R. (2024). Direct Tensor Voting in line segmentation of handwritten documents. *International Journal of Electronics and Telecommunications*, 95–102. doi:10.24425/ijet.2024.149519. [In English].

Balaha, H. M., Ali, H. A., Youssef, E. K., Elsayed, A. E., Samak, R. A., Abdelhaleem, M. S., ... & Mohammed, M. M. (2021). Recognizing arabic handwritten characters using deep learning and genetic algorithms. *Multimedia Tools and Applications*, *80*, 32473-32509. doi:10.1007/s11042-021-11185-4. [In English].

Colesnicov, A., Malahov, L., Cojocaru, S., & Burtseva, L. (2020). Semi-automated workflow for recognition of printed documents with heterogeneous content. *Computer Science Journal of Moldova*, *84* (3), 223–240. Retrieved from https://ibn.idsi.md/sites/default/files/imag_file/v28-n3-%28pp223-240%29.pdf. [In English].

Harris, M., Levene, M., Zhang, D., & Levene, D. (2019). Comparing "parallel passages" in digital archives. *Journal of Documentation, 76 (1),* 271–289. doi:10.1108/jd-10-2018-0175. [In English].

Liang, J., Wang, H., & Li, X. (2020). Task design and assignment of full-text generation on mass chinese historical archives in digital humanities: a crowdsourcing

approach. *Aslib Journal of Information Management*, *72 (2),* 262–286. doi:10.1108/AJIM-09-2019-0245. [In English].

Lischer-Katz, Z. (2022). The emergence of digital reformatting in the history of preservation knowledge: 1823–2015. *Journal of Documentation*, *78 (6),* 1249–1277. doi:10.1108/jd-04-2021-0080. [In English].

Muehlberger, G., Seaward, L., Terras, M., Oliveira, S. A., Bosch, V., Bryan, M., ... & Zagoris, K. (2019). Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study. *Journal of documentation, 75 (5),* 954–976. doi:10.1108/jd-07-2018-0114. [In English].

Navarrete, T., & Mackenzie Owen, J. (2011). Museum libraries: how digitization can enhance the value of the museum. *Palabra clave*, *1 (1).* Retrieved from https://www.researchgate.net/publication/265069291. [In English].

Shen, Z., Zhang, R., Dell, M., Lee, B. C. G., Carlson, J., & Li, W. (2021). Layoutparser: A unified toolkit for deep learning based document image analysis. In *Document Analysis and Recognition–ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part I 16* (pp. 131–146). Springer International Publishing. doi:10.1007/978-3-030-86549-8_9. [In English].

Yurtsever, M. M. E., Özcan, M., Taruz, Z., Eken, S., & Sayar, A. (2022). Figure search by text in large scale digital document collections. *Concurrency and Computation: Practice and Experience*, *34 (1),* e6529. doi:10.1002/cpe.6529. [In English].

Надійшла до редколегії 04.03.2024