УДК 004.01:[930.25:929.5]

Спектор А. Ю., аспірант,
Національний університет «Львівська Політехніка»

АНАЛІЗ ПІДХОДІВ ЦИФРОВОГО АРХІВУВАННЯ ГЕНЕАЛОГІЧНИХ ДОКУМЕНТІВ

Ключові слова: архівування, інформаційні технології, автоматизація.

Вступ. Сучасне інформаційне суспільство потребує методів та засобів, які можуть вдосконалити, пришвидшити, покращити та автоматизувати процеси зберігання та обробки інформації. Одним із таких напрямів є оцифрування генеалогічних документів, яке допоможе вирішити такі проблеми: систематизація великих обсягів інформації та їх аналіз, проведення дослідження на основі отриманих даних, швидкий доступ до інформації, пришвидшення проведення генеалогічних досліджень та покращення їх якості.

Актуальність. Аналіз наявних інформаційних технологій оцифрування генеалогічних документів для подальшої їх уніфікації та обробки.

В Україні зараз іде активна фаза диджиталізації архівних фондів, що дає змогу отримати ідентичну оригіналу електронну копію, забезпечити збереження інформації навіть у випадку втрати оригіналу, накопичення документів у Національному архівному фонді.

Згідно з рекомендаціями Державної архівної служби України існують такі засоби для оцифрування архівних документів:

- фотографічне копіювання (мікрофільмування) – це процес, під час якого використовують контрастний фотопапір та плівки, апарати для рефлексного фотографування та друку. Мікрофільмування застосовують під час створення копій страхового фонду на документи НАФ, що забезпечує довготривале збереження інформації. Має такі переваги: забезпечення цілісності документа, тривалий термін зберігання, стандартизація та можливість переведення у цифровий формат. З недоліків можна виділити такі: низька швидкість оброблення інформації, механічне навантаження на палітурку документа, залежність від якості апаратури;
- електрографічне копіювання є найпростішим способом, який широко застосовують архівні установи переважно для виготовлення копій на вимогу користувачів документами НАФ. До переваг можна віднести: високу оперативність, порівняно невисоку вартість копіювання. З недоліків можна виділити часткову втрату інформації під час копіювання зображень та значне механічне навантаження на оригінал документа;
- електронне копіювання (оцифрування) – проводиться за допомогою різних видів сканерів або цифрової фотокамери. До переваг належать: зручність та швидкість копіювання документа або його частини без втрати якості, зменшення зносу оригіналів, можливість доступу користувачів до документів та просто організації, можливість необмеженого тиражування копій, просторова і спектральна чуттєвість окремих сенсорів, що використовуються під час

оцифрування. До недоліків оцифрування віднесено часту зміну технічної та програмної бази програмної й апаратної частини [1].

Найбільш ефективним та комплексним рішенням для архівування генеалогічних документів є електронне копіювання через свою універсальність та можливість зберігати дані з документів у цифровому вигляді з подальшою їх систематизацією та уніфікацією.

Для отримання даних з оцифрованих копій можна використовувати два такі підходи: використання ручного вводу даних із цифрових копій до інформаційної системи або використання автоматичних рішень на основі наявних технологій оптичного розпізнавання символів.

Оптичне розпізнавання символів (ОРС) – це технологія, яка дає змогу виділяти символи, слова, речення та поєднувати їх у готові дані. Певні рішення використовують для своєї роботи штучний інтелект (ШІ), що дає змогу більш точно та якісно опрацьовувати дані, а за допомогою машинного навчання (МН) постійно вдосконалювати технологію [2].

Document AI та Vision AI від компанії Google є дуже потужними рішеннями для диджиталізації документів на основі ШІ та МН. Document AI дає змогу структурувати дані документів, аналізувати їх та шукати патерни для автоматизації процесів чи класифікувати їх на основі попередньо підготовлених моделей. Vision AI дає змогу налаштувати власні програми для аналізу зображення за допомогою великого набору попередньо навчених API, AutoML чи спеціальних моделей. Перевагою використання рішення від Google є: розгортання інфраструктури і всіх процесів у хмарі, великий набір інструментів, якісно написана документація та наявність підтримки, але недоліком є його велика ціна [3, 4].

OCR4all – це повністю безкоштовне програмне рішення з відкритим кодом, яке дає змогу повністю автоматизувати робочий процес з розпізнавання тексту – і друкованого, і рукописного. З переваг можна виділити легкість встановлення та використання для нетехнічних користувачів, кросплатформність та можливість коригувати результати на кожному етапі роботи. З недоліків можна виділити необхідність у потужному апаратному забезпеченні, невелику кількість документації та час на налаштування правил для обробки документа [5].

Transkribus – це платформа ШІ, яка підтримує роботу з історичними документами. Transkribus дає змогу автоматично розпізнавати текст, макет і структуру документів за допомогою потужностей ШІ. Також є можливість навчити власні моделі ШІ, які будуть відповідати конкретним документам. З переваг: потужна підтримка різними архівними установами Європейського союзу, наявність двох публічних моделей для української мови (Друкована українська 20 століття та Український родовий почерк), можливість об’єднуватись з іншими організаціями та спільно працювати над колекціями. З недоліків можна виділити високу ціну на послуги, але є певні пільгові умови для організацій [6]. Приклад використання Transkribus зображено на рис. 1.

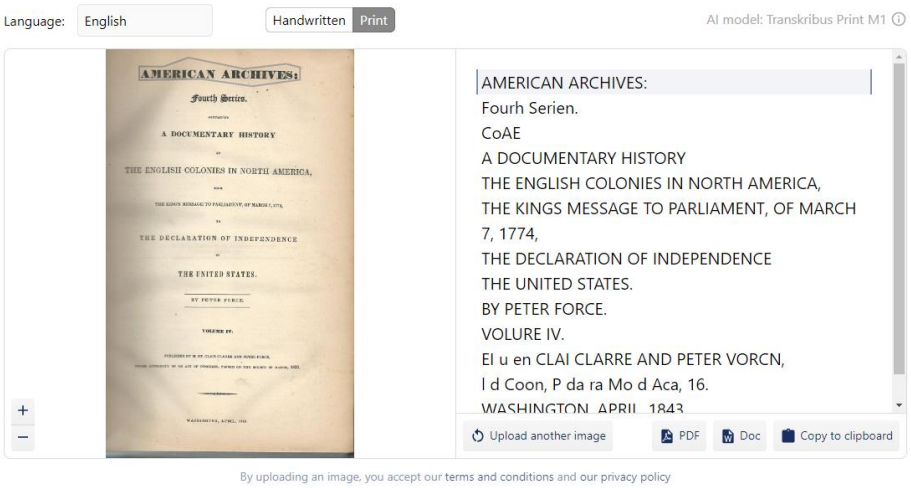

Рисунок 1 – Розпізнавання тексту за допомогою Transkribus

Також існують більш прості програмні рішення, які дають змогу вирішити певне завдання або допоможуть автоматизувати частину роботи. До них можна віднести: Tesseract – бібліотеку оптичного розпізнавання символів на основі нейронної мережі, OpenCV – фреймворк для аналізу, класифікації та обробки зображень, ABBYY FineReader – програма для оптичного розпізнавання символів тощо.

Висновки

Отже, опираючись на аналіз доступних методів та засобів архівування документів, можна зазначити, що оцифрування документів є найефективнишим способом цифрового архівування документів, воно дає змогу комплексно вирішити проблеми автоматизації отримання даних з документів, більш ефективного зберігання, розповсюдження та доступності для користувачів, їх уніфікації та систематизації. Тому впровадження в установах технологій як-от ОРС дасть змогу пришвидшити диджиталізацію архівних фондів. З погляду генеалогічних досліджень наявність уніфікованих та систематизованих цифрових фондів дасть змогу прискорити пошук, обробку відомостей та покращити якість досліджень. Але під час впровадження даних технологій потрібно врахувати низку факторів, як-от: доцільність, ціна та наявність необхідного апаратного забезпечення.

Список використаних джерел

1. Копіювання документів у архівних установах України: методичні рекомендації. URL: https://undiasd.archives.gov.ua/doc/mr_copy_docs.pdf (дата звернення: 10.11.2023).
2. What Is Optical Character Recognition (OCR)? URL: https://www.ibm.com/blog/optical-character-recognition/ (дата звернення: 10.11.2023).
3. Document AI. URL: https://cloud.google.com/document-ai?hl=en (дата звернення: 10.11.2023).
4. Vision AI. URL: https://cloud.google.com/vision?hl=en (дата звернення: 10.11.2023).
5. What is OCR4all. URL: https://www.ocr4all.org/about/ocr4all (дата звернення: 10.11.2023).
Transkribus. URL: https://readcoop.eu/transkribus (дата звернення: 10.11.2023).