

1 СТВОРЕННЯ ЕЛЕКТРОННОГО ПРЕДСТАВЛЕННЯ ФРАГМЕНТА ДРУКОВАНОГО ВИДАННЯ

1.1 Мета роботи

Вивчення особливостей здійснення процесу створення електронної версії друкованого видання.

Після виконання лабораторної роботи студент повинен знати:

1. сутність процесу та варіанти програмних засобів створення електронного представлення друкованого видання.
2. сутність процесу оптичного розпізнавання тексту, особливості здійснення сканування та підготовки оригіналів до процесу перетворення видання у електронне представлення;
3. специфіку форматів файлів, які використовуються для збереження цифрових версій друкованих видань;

уміти:

1. обґрунтований вибір програмних засобів створення електронного представлення друкованого видання;
2. створювати електронне представлення друкованих видань з вбудованою тестовою частиною їх вмісту, придатною для пошуку або копіюванню;
3. створювати електронні версії друкованих видань та документів з можливістю навігації та редагування їх вмісту;
4. створювати електронні версії друкованих видань та документів з озвучуванням фрагментів тексту;
5. конвертувати електронні видання у різні формати.

1.2 Методичні вказівки з організації самостійної роботи

Вибрати друковане видання та принести оригінал. Друковане видання має містити текст на декількох мовах; зображення з обтіканням текстом; схеми, діаграми або графіки, що містять написи; складні таблиці; нумерацію сторінок або колонтитули; використовувати декілька шрифтів або декоративні шрифти.

Узгодити з викладачем, які саме частини повинні увійти до електронного представлення фрагмента друкованого видання.

Вибрати та обґрунтувати програмні засобами, які будуть використовуватись для створення електронне представлення видання.

Робота з текстом. Оптичне розпізнавання символів (англ. Optical Character Recognition, OCR) – це технологія, яка дозволяє перетворювати різні типи документів, такі як відскановані документи, PDF-файли або фото з цифрової камери, в формати, що редагуються, та з можливістю пошуку.

Одним з сучасних засобів для такого перетворення є програмне забезпечення ABBYY FineReader. Процес розпізнавання в цілому складається з трьох етапів: відкриття (або сканування) документа, розпізнавання і збереження в найбільш підходящому форматі (DOC, RTF, XLS, PDF, DjVu, HTML, TXT та інші) або перенесення даних безпосередньо в офісні програми, такі як Microsoft Word, Excel або додатки для перегляду PDF.

Після запуску програми FineReader з'являється діалог створення нової задачі (рис. 1.1), де можна вибрати один з автоматичних режиму перетворення у заданий формат або перейти до OCR-редактора для просунутої конвертації з перевіркою результатів. Після натискання на тип задачі з'явиться діалогове вікно вибору файлів для розпізнавання.

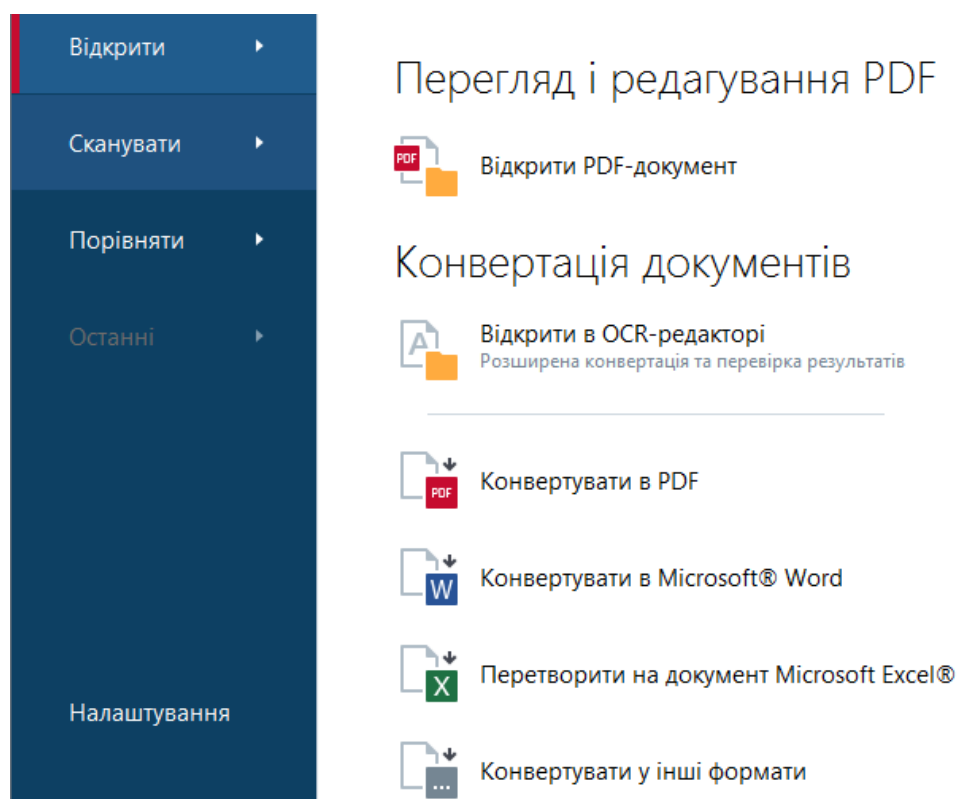


Рисунок 1.1 – Створення нової задачі

Автоматичний режим розпізнавання підходить для документів з простою одноманітною структурою. В такому випадку можна швидко отримати якісний результат з мінімальною кількістю налаштувань. Режим «Конвертувати в інші формати» (рис. 1.2) надає гнучкі налаштування, де можна обрати будь-який

підтримуваний вихідний формат. У полі «Мова розпізнавання» обов’язково треба вибрати мову або групу мов, які використовуються у оригіналі видання. Об’єднувати всі файли необхідно, щоб отримати єдиний документ, коли оригінал міститься у декількох файлах.

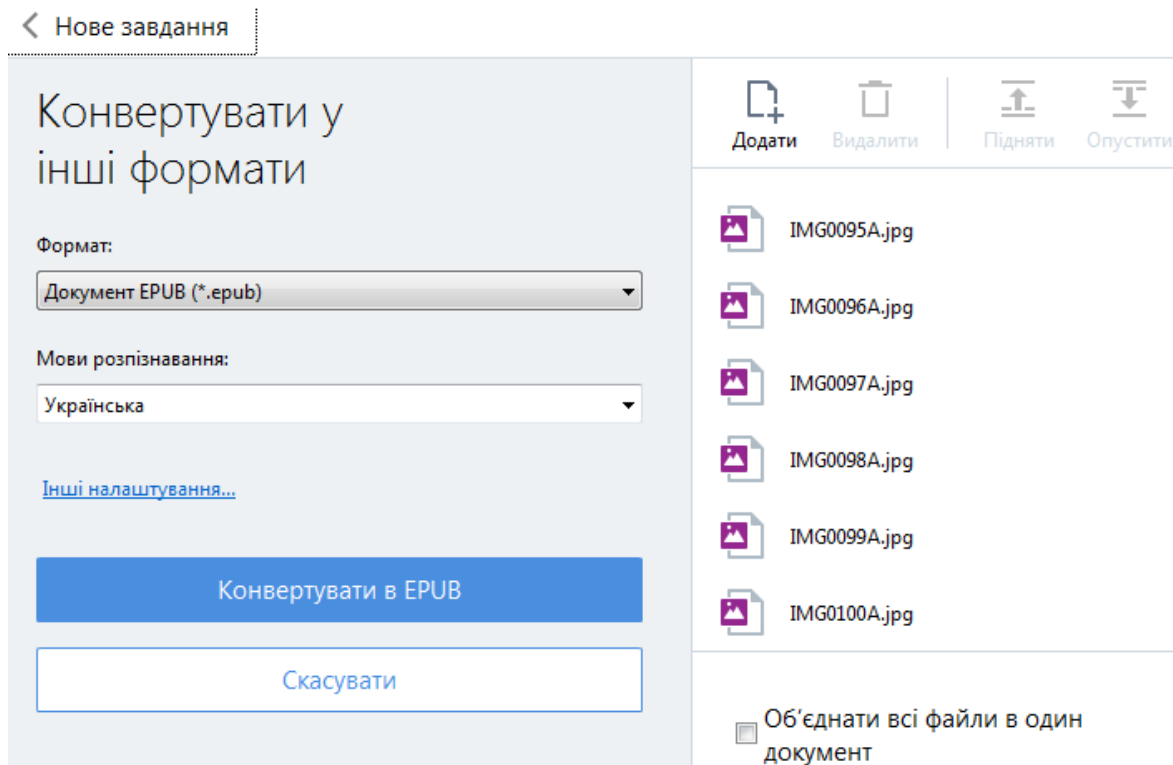


Рисунок 1.2 – Налаштування автоматичного конвертування

Діалогове вікно «Інші налаштування» (рис 1.3, 1.4) дає змогу редагувати параметри кожного вихідного формату. Слід перевірити чи буде зберігатися текстова інформація, у якому вигляді та кодуванні.

Тип документа	Налаштування PDF
PDF	Якість зображень: Збалансована
DOC(X) / RTF / ODT	☐ Створити PDF/A документи Версія: PDF/A-2b
XLS(X)	☐ Створити PDF/UA документи (потрібне розпізнавання)
PPTX	☐ Захищати документи паролем Налаштування...
CSV	☐ Видаляти об'єкти і дані Вибрати елементи...
TXT	☐ Приводити зображення сторінок до одного розміру: A4
HTML	☒ Стискати зображення за допомогою MRC (потрібне розпізнавання)
EPUB / FB2	☐ Застосувати технологію ABBYY PreciseScan для згладжування символів
DjVu	Налаштування PDF з можливістю пошуку
	Текст під зображенням сторінки
	☐ Створювати закладки за знайденими заголовками
	☒ Створити теги PDF
	Використовувати шрифти Windows ☐ Вбудовувати шрифти

Рисунок 1.3 – Параметри PDF документа

Наприклад у форматі PDF налаштування можливості пошуку визначає необхідність розпізнавати текст на рисунках у документі та положення текстового шару відносно до сканованого зображення, що впливає на вигляд електронного представлення друкованого видання. Налаштування «Створювати закладки за знайденим заголовкам» дозволить спробувати автоматично додати до електронного видання такі інтерактивні елементи як зміст та закладки, за якими можна швидко переходити до необхідної частини документа.

Тип документа	Налаштування DjVu
PDF	Режим збереження: Текст під зображенням сторінки
DOC(X) / RTF / ODT	Якість зображення: Висока якість
XLS(X)	Багатошаровість:
PPTX	☒ Автоматично
CSV	☐ Завжди використовувати
TXT	☐ Ніколи не використовувати
HTML	
EPUB / FB2	
DjVu	

Рисунок 1.4 – Параметри DjVu документа

У форматі DjVu є схожі налаштування. «Режим збереження» визначає необхідність додавати текстовий шар у документ. Багатошаровість дає змогу зменшити розмір документа, розділивши оригінальне зображення сторінки на шари, які стискаються за допомогою різних алгоритмів відповідно до

особливостей формату. В автоматичному режимі програма сама визначає доцільність такого поділу.

В робочій області OCR-редакторі (рис. 1.5) зверху знаходиться головна панель інструментів та панель швидкого доступу, зліва розташована панель сторінок для управління послідовністю та кількістю сторінок у документі.

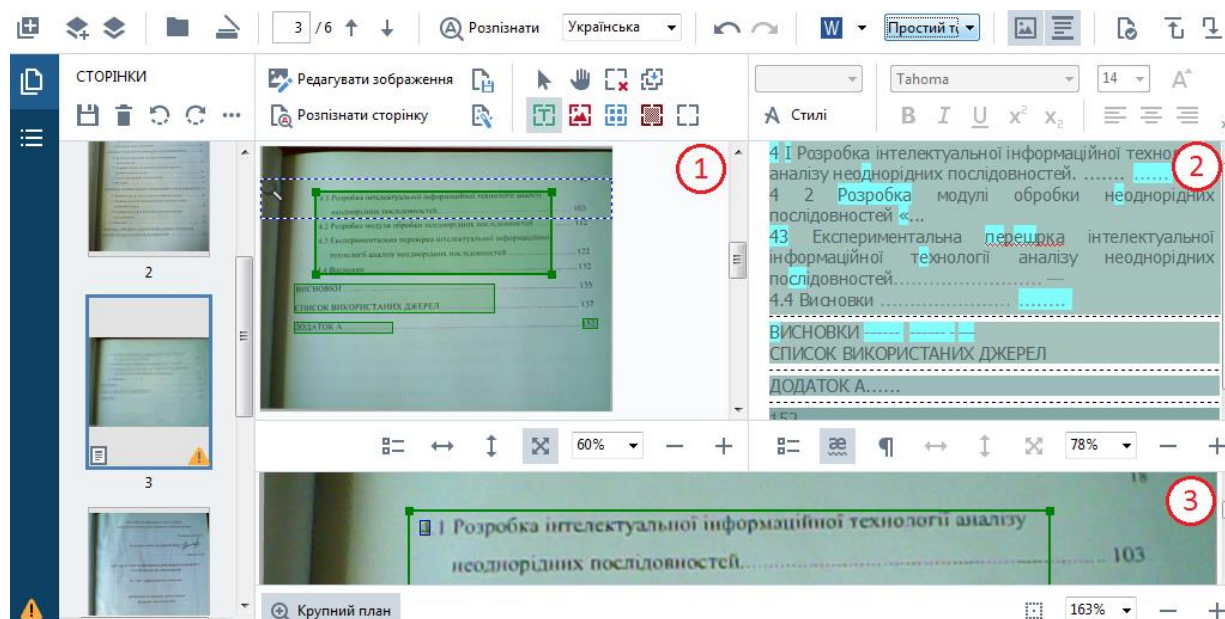


Рисунок 1.5 – FineReader OCR-редактор

У вікні зображень (1) показується зображення поточної сторінки. У цьому вікні можна редагувати області розпізнавання та їх властивості. Вікно тексту (2) відображає розпізнаний текст, який можна редагувати, перевіряти орфографію, змінювати форматування. У вікні Крупний план (3) знаходиться збільшене зображення поточного рядка або ділянки зображення, що обробляється. У цьому вікні зручно корегувати положення зон розпізнавання, порівняти невпевнено визначений символ з оригіналом.

На панелі інструментів розташована кнопка запуску розпізнавання всіх сторінок документа, вибір мови, формату для збереження та типу збереження. Існують наступні типи (режими) збереження.

- Точна копія. Дозволяє отримати документ, оформлення якого буде повністю відповідати оригіналу. Рекомендується використовувати для документів складного оформлення, наприклад рекламних брошур. Однак цей режим не передбачає внесення значних виправлень у текст і оформлення.
- Копія, що редагується. Дозволяє отримати документ, оформлення якого може дещо відрізнятись від оригіналу. Документ, отриманий за допомогою цього режиму, легко редагується.

- Форматований текст. В отриманому документі зберігається написання та розмір шрифту, розбиття на абзаци, але не зберігається розміщення об'єктів на сторінці та міжрядкові інтервали. Таким чином, буде отриманий суцільний текст із вирівнюванням за лівим краєм. Для тексту, у якому порядок читання справа наліво, вирівнювання буде за правим краєм.
- Простий текст. У цьому режимі форматування тексту не збережеться.
- Гнучка копія. Дозволяє отримати HTML-документ, у якому зберігається приблизне розташування об'єктів на сторінці.

Перетворення оригіналу складається з аналізу структури сторінок та розпізнавання. За замовчування всі етапи виконуються автоматично. Під час аналізу вміст сторінки поділяється на типові зони: текст, зображення, таблиця, фонове зображення, штрих-код. На одній сторінці може бути кілька зон різних типів. Зони мають порядкові номери, які відповідають послідовності вставки їх вмісту у загальний документ. Для документів складної структури автоматичний аналіз часто виконується некоректно. У такому випадку необхідно скорегувати або самостійно задати та налаштувати зони для розпізнавання. На панелі швидкого доступу є спеціальні інструменти для роботи з кожним типом зон.

Наступний етап – розпізнавання. Він відбувається в межах визначених зон згідно їх типу та зробленим налаштуванням. Результат відображається у вікні тексту. Фрагменти тексту з помилками перетворення помічаються червоним, невпевнено розпізнані виділяються блакитним. Перед збереження результату необхідно перевірити текст та виправити помилки розпізнавання та форматування.

Якщо багато помилок в тексті, то треба змінити параметри перетворення, при необхідності відредагувати зображення оригіналу (яскравість, контраст, кількість кольорів), наостанок пересканувати друковане видання якісніше.

Робота з графікою. Потужні інструменти растрових редакторів (наприклад Photoshop) можуть покращити, відновити та виправити колір та тональність (яскравість та контраст) зображень. Для корегування зображень електронного видання виконайте послідовно декілька кроків:

- за допомогою гістограми перевірте якість та градаційну шкалу зображення;
- застосуйте налаштування кольору і відтінків, перевірте. Створіть коригуючий шар, який забезпечує більшу гнучкість та не призводить до втрати будь-якої інформації про зображення;

- налаштуйте баланс кольору, щоб видалити небажані кольори або виправте занадто насичені або недостатньо насичені кольори;

- налаштуйте градаційну шкалу. Почніть налаштування тонів з регулювання значень дуже світлих тонів та дуже темних тіней зображення, встановивши загальну градаційну шкалу для зображення. Цей процес відомий як встановлення світлих тонів та тіней або встановлення білих та чорних точок. Встановлення світлих тонів та тіней зазвичай відповідно перерозподіляє пікселі напівтонів, однак може знадобитися ручне налаштування;

- внесіть інші колірні зміни за необхідністю. Після виправлення загального балансу кольорів зображення можна внести додаткові зміни для покращення кольорів або створення спеціальних ефектів;

- зробіть границі зображення різкішими. Для цього можна скористатися фільтрами. Ступінь збільшення різкості залежить від якості зображення, створеного цифровою камерою або сканером.

Створення навігації. Посилання дають змогу швидко перейти до інших місць у тому самому документі, до інших електронних документів, у тому числі до вкладених файлів, або до веб-сайтів. Посилання можна використовувати для запуску дій або для забезпечення читачеві миттєвого доступу до інформації за темою. Можна також додавати дії, щоб відтворити звуковий файл або фільм.

Для створення посилань у Adobe Acrobat виберіть команди меню Інструменти > Редагувати PDF > Посилання > Додати або редагувати. Вказівник миші приймає форму перехрестя, і всі наявні в документі посилання, зокрема невидимі, стають тимчасово видимими. Обведіть прямокутником ділянку, що на ній бажаєте створити посилання. Це область, в якій посилання є активним. В діалоговому вікні «Створити посилання» виберіть потрібні параметри вигляду посилання. Виберіть одну з таких дій посилання.

- Перейти до посторінкового перегляду. Натисніть кнопку «Далі», щоб задати номер сторінки та масштаб перегляду в поточному документі або в іншому документі, наприклад у вкладеному файлі, а потім натисніть кнопку «Установити посилання».

- Відкрити файл. Виберіть цільовий файл та клацніть на команді «Вилучити». Якщо файл є PDF-файлом, вкажіть спосіб відкриття документа (наприклад, в новому вікні або в наявному вікні), потім натисніть «ОК».

- Відкрити веб-сторінку. Укажіть URL-адресу веб-сторінки відсилання.

- Власне посилання. Натисніть кнопку «Далі», щоб відкрити діалогове вікно «Властивості посилання». У цьому діалоговому вікні можна вказати будь-

яку дію, таку як читання статті або виконання команди меню, яку треба пов'язати з посиланням.

Додавання вкладення. Можна додавати PDF-файли та файли інших типів як вкладення до PDF-файлів. Якщо перемістити PDF-файл в інше місце, вкладені файли будуть переміщені разом із ним. Вкладені файли можуть містити посилання на документ або з документа вищого рівня, або посилатися на інші вкладені файли.

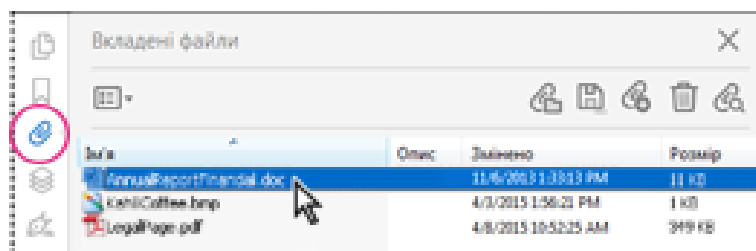


Рисунок 1.6 – Панель «Вкладені файли». Для додавання, видалення та перегляду вкладених файлів.

Для додавання вкладених файлів у Adobe Acrobat виберіть команди меню Інструменти > Редагувати PDF > Більше > Вкласти файл. У діалоговому вікні «Додати файли» виберіть файл, який потрібно вкласти, і натисніть кнопку «Відкрити».

Щоб вкладені файли можна було переглядати у ранніх версіях, виконайте одну з таких дій: Виберіть «Перегляд» > «Показати/сховати» > «Панелі навігації» > «Вкладення», а потім виберіть пункт «Показувати вкладення за замовчуванням» у меню параметрів (вибрано за замовчуванням). Виберіть пункт меню «Файл» > «Властивості», перейдіть на вкладку «Вигляд при відкриванні», виберіть пункт «Панель вкладень та сторінка» у меню «Вкладки навігації» і натисніть «ОК». Збережіть PDF-документ. Щоб додати до вкладеного файлу опис, що допомагає розрізняти схожі файли на панелі «Вкладені файли», виділіть вкладений файл і в меню параметрів виберіть команду «Редагувати опис». Відредагуйте текст опису, а потім збережіть файл.

1.3 Порядок виконання роботи і методичні вказівки до її виконання

1. Проаналізувати друковане видання, визначити які саме частини повинні увійти до електронного представлення фрагмента друкованого видання.

2. Вибрати та обґрунтувати програмні засоби які будуть використовуватись для створення електронного представлення фрагмента друкованого видання. Визначити параметри сканування та обґрунтувати їх.
3. За допомогою OCR-редактора виконати аналіз сторінок оригіналу. Самостійно скорегувати зони для розпізнавання та налаштувати їх властивості. Виконати розпізнавання та перевірити результат.
4. Відкоригувати зображення, що містяться в електронному виданні, обкладинку та титульні сторінки.
5. Створити електронні представлення обраного друкованого видання у форматах PDF, DjVu, DOC, EPUB, HTML, FB2. В усіх варіантах представлення має працювати пошук по тексту та його копіювання. Порівняти результати.
6. В версію у форматі PDF додайте інтерактивний зміст (переходи до інших місць у тому самому документі) та посилання до веб-сайтів.
7. Озвучте фрагмент тексту друкованого видання та збережіть його. Додайте створений аудіо файл як вкладення до PDF.

1.4 Зміст звіту

Стандартний титульний аркуш з темою та номером лабораторної роботи.

Мета лабораторної роботи і формулювання завдання.

Порядок виконання сформульованого завдання з коротким описом виконуваної послідовності дій, зазначення використаних вікон і показ кінцевих результатів роботи.

Висновки по роботі.

Контрольні запитання і завдання

1. Назвіть основні етапи створення електронного представлення друкованого видання.
2. Які параметри впливають на якість розпізнавання друкованого видання?
3. Опишіть принципи вибору параметрів сканування друкованого видання для якісного перетворення в електронне представлення.
4. Порівняйте можливості форматів PDF, DjVu та EPUB.
5. Чим відрізняються формати DOC, RTF і TXT?
6. На які типи зон розбивається сторінка перед розпізнаванням?
7. Назвіть програмні засоби створення електронного представлення друкованого видання.
8. Яке призначення текстового шару у PDF та DjVu документах?
8. Як створити інтерактивний зміст у електронному виданні.
9. Як додати до PDF-файлів файли інших типів, щоб переміщувати вкладені файли разом із ним?