УДК 004.912

ТЕХНОЛОГІЯ ВІДНОВЛЕННЯ В ЕЛЕКТРОННОМУ ВИДІ РАРИТЕТНИХ ВИДАНЬ ДОВІДКОВОГО ХАРАКТЕРУ

© О. П. Кохановський, к.ф.-м.н., доцент,
О. Л. Бриндзя, магістрантка,
НТУУ «КПІ», Київ, Україна

Были рассмотрены и сопоставлены различные технологии и программное обеспечение для восстановления в электронных форматах печатного издания справочного типа.

Various technologies and software for recovering printed reference books to electronic formats were examined and compared.

Постановка проблеми

Нині є дуже актуальним перетворення довідкових видань в електронні. Оскільки багато зразків корисної літератури стали недоступними, зокрема, через малий тираж, тим більше, дуже актуально мати електронну версію корисного словника, в якому впроваджено пошукову систему. Це дуже зручно і значно скорочує час на пошук необхідної інформації [1, 2].

Масштаби оцифрування у світі. У 2010 році компанія Google та міністерство культури Італії підписали рекордну угоду, що передбачає сканування та розміщення на сайті Google Books одного мільйона старовинних книг, серед яких тексти Данте та Мандзоні. Оцифрування книг XVIII–XIX століть, не покритих авторськими правами, коштуватиме 100 млн євро, усі витрати візьме на себе компанія Google.

Передбачалося, що робота займе приблизно два роки. Серед видань будуть як наукові роботи з ілюстраціями та літографіями, так і художня література. У відсканованих текстах буде опція «шукати за словом», що значно полегшить роботу філологів-дослідників. Після сканування книги стануть безкоштовно доступними для всіх і назавжди.

На сьогодні сайт Google Books має у своєму розпорядженні близько 12 млн. видань сотнями мов світу [3].

Стан електронних бібліотек в Україні. Президент України Віктор Янукович виступає з ініціативою створення повномасштабної електронної бібліотеки, що буде містити найрізноманітнішу інформацію про Україну, її історію, культуру, географію та сьогодення.

Віктор Янукович наголосив, що вже сьогодні на різних електронних ресурсах існує чимало найрізноманітнішої інформації про Україну, її історію, культуру, географію та сьо-

годення. Проте проблема в тому, що ця інформація розпорошена, несистемна й дуже часто її надзвичайно важко знайти. «Нам треба скоординувати зусилля і відкрити для цієї роботи бібліотеки та архіви. Ми повинні зробити ті перші кроки, що допоможуть не лише створити культурний продукт, а головне — усвідомити, що він є, і він конкурентоздатний. Знаю і вірю, що саме Інтернет, саме нові технології в перспективі не лише дадуть поштовх розвитку книговиданню, кіно, музиці, театру, музейній справі, а й сприятимуть розвитку громадянського суспільства загалом», — наголосив Віктор Янукович [4].

Отже, мотивація до створення електронних та паперових версій одного документа дуже висока і актуальна. Це питання охоплює все більші масштаби і потребує постійного дослідження та пошуку карколомних рішень.

Аналіз попередніх досліджень

Для визначення шляхів оптимального, зручного та доступного способу розробки технології відновлення раритетного видання з метою забезпечення потреб споживача було проведено пошук та дослідження новітніх технологій, які вже добре зарекомендували себе і мають застосування.

Існує багато способів представлення електронних документів — HTML, PDF, DjVu та ін. Тому питання про формат представлення попередньо оцифрованої друкованої інформації було детально досліджено, щоб навести достовірні аргументи вибору того чи іншого формату.

У результаті проведеного інтернет-дослідження виявилось, що 7 осіб із 10 надають перевагу формату PDF (створений в 1991 році корпорацією Adobe) перед DjVu (створений у 1996 компанією AT&T), аргументуючи свій вибір популярністю та поширеністю формату.

Проте, врахувавши всі особливості форматів, виявилось, що їх не можна вважати суперниками. Вони призначенні для вирішення різних задач, а тому не змагаються, а доповнюють один одного. Обидва формати, і DjVu, і PDF, широко використовуються при створенні електронних документів [5, 6].

Мета роботи

На даному етапі оцифровування друкованих видань з метою створення електронних бібліотек існує чимало технологій з різноманітними програмними та апаратними засобами.

Як приклад наведемо пілотний проект для Національної Історичної Бібліотеки України. Йдеться про стародруки: «Историческая пъснь о походь на половцовъ» 1800 р., «Слово о плъку Игореве, Игоря сына Святъславля, внука Ольгова» 1934 р., «Новый Завет с Псалтырью» 1580 р.

Їхнє сканування здійснювалося за допомогою найсучаснішого спеціалізованого сканера PLANSCAN-C для безпечного сканування старовинних книжок без додаткового освітлення, яке містить шкідливе ультрафіолетове випромінювання. Обробка сканованих зображень сторінок здійснювалася з використанням спеціалізованого програмного забезпечення BookRestorer, призначеного для професійного створення електронних книг.

Проте зазначене програмне забезпечення досить дороге і вимагає деякої підготовки до його використання, тому метою роботи є дослідження існуючих програмних та апаратних засобів з подальшим обґрунтуванням та створенням найбільш доцільного та доступного технологічного процесу відновлення друкованого видання.

Результати проведених досліджень

Процес створення електронної версії друкованої книги досить складний і багатоступінчастий, і зрозуміло, що одним з найважливіших і обов'язкових компонентів цього процесу є сканування.

Літературний огляд дозволив проаналізувати переваги та недоліки сканерів з різними оптичними системами CCD та CIS [7]. Було приділено увагу розгляду спеціалізованого книжкового сканера [8], а також визначено режими сканування для друкованого видання без великої кількості ілюстрацій:

— режим: Grayscale;
— роздільна здатність: 300 dpi;
— різкість: Low або Medium, спеціальні параметри не використовувати.

Формат вихідного файлу: Uncompressed (нестиснутий) TIFF [9].

Було проаналізовано багато способів опрацювання зображень, проте оскільки видання за характером інформації – текстове та чорно-біле, недоцільно використовувати складне, дороге програмне забезпечення, адже бажаного результату можна досягти і за допомогою простих доступних програм. Вибір було зупинено на використанні безкоштовної спеціалізованої програми Scan Tailor, яка дає можливість значно автоматизувати процес оцифровування з достатньою якістю.

Етапи роботи зі Scan Tailor

1. Сканування розворотів книги. Для часткової автоматизації цього процесу було використано програму IrfanView. Режим сканування – «відтінки сірого», роздільність — 300 dpi, формат зображень — TIFF.

2. Створення нового проекту.

3. Виправлення орієнтації сторінок.

4 Розрізання сторінок. Scan Tailor автоматично намагається визначити межі між сторінками. При наявності не-

точностей їх можна виправити вручну.

5. Вирівнювання сторінок – відбувається автоматично. Для зручності на відскановані сторінки при цьому накладається «аркуш в клітинку».

6. Визначення корисної області, щоб відсікти зайві порожні місця. Це досить довгий процес, який може тривати від одиниць до десятків хвилин, залежно від обсягу видання.

7. Створення макету сторінки (розмір полів у відсканованого тексту і вирівнювання на сторінці). Ця операція займає кілька секунд.

8. Виведення результуючих файлів. Тут можна задати деяке коригування на зразок видалення плям. У результаті програма створює папку out, куди зберігає отримані файли теж у форматі TIFF [10].

Таким чином організовується процес обробки зображень за допомогою програми Scan Tailor.

Для задоволення більшого кола споживачів оброблені зображення було конвертовано в єдиний файл у форматі DjVu (за допомогою DjVu Small) та у форматі PDF (за допомогою Adobe Acrobat). Для прискорення роботи з отриманим електронним словником було здійснено огляд ряду програм, які допомагають оснастити його навігацією: DjvuDic, Bookmark-Tool, PdfFactory, Pdf&Djvu Bookmarker тощо [11, 12].

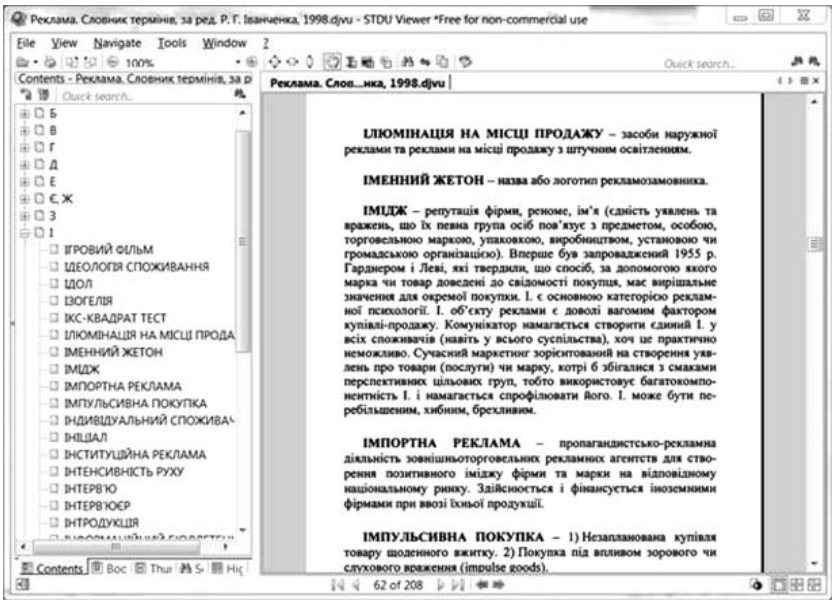
Вибір було зупинено на останній. Адже Pdf&Djvu Bookmarker — програма, призначена для автоматизованого створення дерева змісту відразу для DjVu і PDF-файлів, тому алгоритм процесу накладання пошукової системи на електронне видання значно спрощується.

Висновки

Проаналізовано та систематизовано інформацію про існуючі програмні та апаратні засоби для створення факсимільної копії друкованого видання в електронному вигляді. Розроблено оптимальну технологію отримання електронної версії книги з урахуванням доступності, часу, зручності та простоти.

Розглянута технологія тестувалася на прикладі друкованого видання «Реклама. Словник термінів» Р. Г. Іванченка. Характерною особливістю є те, що у процесі створення електронної версії видання не було використано традиційної технології розпізнавання. Графічний формат DjVu дав змогу у сотні разів зменшити розмір словника порівняно зі сканованими зображеннями.

При допомозі програми Pdf&Djvu Bookmarker у словник було впроваджено ефективну систему пошуку та навігації (рис.).



Вигляд електронної версії словника з впровадженим деревом навігації

1. Вуль В. А. Электронные издания: учебник / В. А. Вуль. – М.— СПб. : Петербургский институт печати, 2001 — 308 с. 2. Композиция изданий: Особенности проектирования различных типов изданий: учебное пособие/ С. М. Болховитинова [и др.]; ред. С. М. Болховитиновой. — М. : МГУП, 2000. — 166 с. 3. Прес-центр. Уряд Італії та Google підписали угоду про сканування 1 млн книг [Електронний ресурс] — Режим доступу : http://presscenter.ukrinform.ua/news-40904.html?p=8. 4. Президент України Віктор Янукович. Офіційне інтернет-представництво. Президент виступає з ініціативою створення повномасштабної електронної бібліотеки [Електронний ресурс] — Режим доступу : http://www.president.gov.ua/news/19564.html. 5. Википедия. Portable Document Format [Електронний ресурс] — Режим доступу : http://ru.wikipedia.org/wiki/Portable_Document_Format. 6. Википедия. DjVu [Електронний ресурс] — Режим доступу : http://ru.wikipedia.org/wiki/DjVu. 7. Леонтьев Б. К. Секреты сканирования на ПК / Б. К. Леонтьев. — ООО «Бизнессофт», Литературное агентство «Бук-Пресс», 2006. — 44 с. 8. Пирит. Отдел сканеров. Книжные сканеры [Електронний ресурс] — Режим доступу : http://www.docscan.ru. 9. Википедия. TIFF [Електронний ресурс] — Режим доступу : http://ru.wikipedia.org/wiki/TIFF. 10. Scan Tailor. О проекте [Електронний ресурс] — Режим доступу : http://scantailor.sourceforge.net/?q=ru/about. 11. Кутовенко А. О. Все для DjVu/ Алексей Кутовенко // UPGRADE — 2010. — № 49.— С. 36–39. 12. Создаем PDF-файлы. Сайт Ивана Чередниченко [Електронний ресурс] — Режим доступу : http://chpas.narod.ru/article_3.html.

Рецензент — О. М. Величко, д.т.н.,
професор, НТУУ «КПІ»

Надійшла до редакції 07.02.12 р.