УДК 681.3.06.686.1.03

ТЕХНОЛОГІЯ ВІДНОВЛЕННЯ РАРИТЕТНИХ НАУКОВО-ТЕХНІЧНИХ ВИДАНЬ

© О. П. Коханівський, к.ф.-м.н., доцент, Д. С. Суглобов, НТУУ «КПІ», Київ, Україна

Предлагается автоматизированная технология восстановления раритетных научно-технических изданий.

It is offered the automated technology of restoring rare scientific and technical editions.

Постановка проблеми

Людству у спадок дістався величезний об'єм культурного надбання у вигляді книг, картин і інших твердих носіїв інформації, які, на жаль, відзначаються недовговічністю. Хоча до наших часів дійшли папіруси Стародавнього Єгипту, праці вчених Європи епохи Відродження, а через них досягнення Стародавньої Греції та Риму, але сталося це за збігом обставин або завдяки дбайливому до них ставленню. Папір — матеріал, який хоч і може на протязі десятків, а то й сотні років, зберігати свою структуру та фарбу, проте дуже подразливий до вогню, теплового випромінювання, вологості, постійного використання. Часті пожежі, війни, а то й просто нещасні випадки, на жаль, інколи призводять до втрати безцінних знань, які потім дуже важко відновити.

Задачею дослідження є розробка ефективної технології створення електронних копій раритетних науково-технічних видань. Нами не ставиться завдання створення електронної версії через розпізнавання текстової частини видань, оскільки раритетні науково-технічні видання містять велику кількість математичних, хімічних та ін. формул, які найчастіше неадекватно відтворюються при такому відновленні. Копії повинні бути якісними, по можливості не містити дефектів оригіналу. Також треба вибрати зручний формат для їх подальшого зберігання та використання. Такі копії повинні бути доступними для масової аудиторії через створення спеціальних електронних бібліотек, які доступні користувачам у мережі Internet.

Аналіз попередніх досліджень

До появи цифрового копіювання документів в бібліотеках найчастіше застосовували для відновлення та зберігання раритетних видань метод фотокопіювання. Видання встановлювали на спеціальний стенд і за допомогою фотокамери проводили фотографування розворотів видань. Цей процес є досить довгим, матеріало- та трудомістким, особливо за рахунок процесу проявлення та отримання зображень копій. Копії видань є достатньо якісними, але

вони передають й дефекти оригіналу. Вони недосить зручні у використанні. З часом їхні властивості погіршуються. Вони не стійкі до впливу температури, вологості, тощо. Треба зазначити, що якість та точність відображення інформаційної частини видання дуже залежить від умов зйомки, типу плівки, параметрів камери та ін. Якщо неправильно підібрати освітлення, витримку, чутливість плівки, фокусування та інші показники, отримані копії можуть бути неякісними: затемненими, нечіткими, неконтрастними, тощо. Це вимагає професійного підходу до фотокопіювання. Також на якість відображення впливає процес проявлення фотоплівки та отримання зображень на фотопапері.

Відомі й інші способи копіювання видань, наприклад, ксерографія. Цей спосіб відомий більш ніж 50 років. Він дозволяє за короткий час отримати багато копій видання. Але, на жаль, такі копії дуже поступаються за якістю іншим способам копіювання. Текст на ксероксі може відтворюватись досить прийнятно, але цього не можна сказати про півтонові малюнки. Вони спотворюються під час копіювання. Не найкращим чином передаються переходи між тонами. Також ксерокс недостатньо передає затемнені ділянки оригінала, наприклад, такі, які виникають під час нещільного прилягання паперу до робочого скла ксерокса. Якщо текст потрапляє в таку ділянку, він може спотворитись, або навіть зникнути. Ксерокс має визначений формат для сканування оригіналів, тому копіювання видань з великими розмірами може бути незручним. До того ж і фарба тримається на паперовому носії нестійко. Тому такі копії мають дуже короткий термін використання [1].

Сьогодні для копіювання, зберігання, відновлення та друку різних паперових документів використовують сполучення сканера, комп'ютера та принтера. Сканери дозволяють якісно оцифрувати оригінали. Користувач може провести різні маніпуляції зі сканованими зображеннями: повернути, обрізати, коригувати, прибрати дефекти, провести розпізнання текстової частини та багато іншого. Найголовніше те, що цифрові копії видань можна безліч разів копіювати та досить довго зберігати на різних носіях інформації. Такі копії за бажанням неважко роздрукувати на різних вивідних пристроях: струминному, лазерному чи навіть на цифровій друкарській машині [2].

Останнім часом широкого розвитку набули цифрові фотоапарати. Вони дозволяють, на відміну від сканерів, швидко провести оцифрування раритетних видань різних форматів. Це дещо нагадує процес звичайного фотокопіювання, але з деякими відмінностями: користувач легко контролює процес зйомки; відсутній довгий та трудомісткий процес проявлення. Користувачу достатньо підключити камеру до комп'ютера та перенести туди оцифровані зображення розворотів раритетних видань [3].

Мета роботи

При створенні електронних версій раритетних видань нами було поставлено завдання адекватного відображення оригінала при його переведені в електронну форму і забезпечення зручної навігації по електронній копії. Бажано забезпечити зменшення розміру файлів, які будуть зберігати такі видання, без погіршення якості.

Технологія повинна забезпечити відновлення паперових раритетних видань різних форматів та розмірів, з різними ступенями пошкодження.

Технологія повинна реалізовуватись за допомогою нескладного в освоєнні програмного забезпечення. Бажано, щоб технологія по можливості працювала на технічних засобах будь-яких виробників з різними технічними характеристиками, оскільки в залежності від місця реалізації не завжди можливо досягти точної конфігурації технічної системи для виконання технології. До того ж треба забезпечити можливість використання різних способів введення-виведення інформації у технологію.

При розробці технології треба вибрати необхідне програмне забезпечення для введення, обробки та відновлення сторінок раритетного видання. Визначити методи та способи усунення різних дефектів, спотворень та інше. Також треба вибрати необхідні формати файлів, які будуть використовуватись при записі, зберіганні та перетворення інформації.

Для створення технології треба визначити оптимальні режими для відновлення раритетних видань, витрати часу, необхідні умови для якісного оцифрування сторінок (освітлення, стан носія, механічні та технічні умови введення інформації та інше).

У роботі ставиться мета запропонувати технологію автоматизованого відновлення раритетних видань, причому за допомогою поширених технічних та програмних засобів.

Результати дослідження

Протестовано та вибрано необхідне технічне та програмне забезпечення, яке необхідне для відновлення раритетного видання. Технічне забезпечення вибирається із врахуванням таких технічних показників: вага, розміри, стан паперового носія, тощо. Саме від них залежить, як ми будемо проводити оцифрування розворотів видання: за допомогою сканера чи цифрового фотоапарата. Цифровий фотоапарат доцільно використовувати в таких випадках, якщо:

— формат видання більший за робоче скло сканера;

— видання своєю вагою може пошкодити робоче скло сканера;

— розвороти нещільно прилягають до робочого скла і це призводить до викривлення інформаційної частини видання;

— паперовий носій дуже нестійкий.

В інших випадках видання можна оцифровувати за допомогою сканера.

Для зберігання та обробки сканованих зображень використовується персональний комп'ютер (ПК) з можливостями гра-

фічної станції. Він повинен мати потужний процесор, великі масиви оперативної пам'яті, великий об'єм жорсткого диску та ін [3].

В якості об'єкта для дослідження ми взяли раритетне науково-технічне видання 30-річної давності випуску, книгу авторів Л. Еліота, У. Уілкокса «Фізика». Відновлення ми проводили по такій технологічній схемі (див. рис. 1), яка узагальнює процес відновлення раритетного видання. Кожний етап включає в себе декілька операцій з використанням необхідних програмних продуктів. В залежності від складності відновлення видання кількість операцій та етапів можуть змінюватись.

Користувач також може змінювати перелік програм, що використовуються. Для цієї технології найкраще мати сучасні засоби: потужні системи введення зображень, новітні ПК та більш досконале програмне забезпечення.

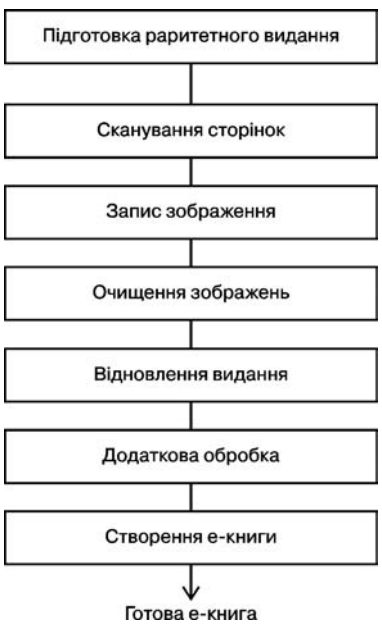

Рис. 1. Спрощена технологічна схема відновлення раритетного видання

Програмна реалізація пропонованої технології включає такі програмні продукти:

— TWAIN-драйвер сканера;

— програма прийому та запису сканованих сторінок зі сканеру (IrfanView 4.10);

— програма обробки графічних зображень (Adobe Photoshop CS3);

— програма для відновлення сканованих сторінок (Scan-Kromsator 5.6);

— програми для створення е-книг, їх перегляду та корекції (Adobe Acrobat 7.0 Professional, DjVu Small, WinDjView-0.4.3);

— програма швидкого перегляду зображень (Adobe Bridge CS3).

Підготовка оригіналу до сканування складається з наступних операцій:

— розміщення книги на робочому склі сканера;

— задання необхідних параметрів для запису файлів;

— задання необхідних параметрів для сканування;

— безпосереднє сканування та запис;

— контроль записаних сканованих файлів.

Перед початком сканування видання необхідно очистити поверхню книги від пилу та бруду при наявності. Їхня наявність може спотворити результати сканування, а також пошкодити поверхню робочого скла сканера.

Необхідно також забезпечити максимальне прилягання розворотів видання до робочого

скла сканера. Для цього потрібно несильно розігнути книгу, так щоб її не пошкодити.

Перед скануванням розкриту книгу необов'язково максимально рівно розміщувати на робочому склі сканера, паралельно його краям. При скануванні треба уникати надмірного притискання книги до робочого скла сканера. За допомогою програми IrfanView ми встановили необхідні параметри для пакетного сканування розворотів, а саме, місце для запису сканованих файлів на жорсткому диску, параметри записуваного файлу, його назву та індекси, параметри для процесу сканування (модель — GrayScale, роздільна здатність — 300 dpi, область сканування — все робоче поле сканера). Зазначимо, що для якісного збереження отриманих розворотів ми використовували формат TIFF з LZW-компресією. Він дозволяє зберігати максимальний об'єм графічної інформації, а метод LZW-ущільнення дозволяє ущільнювати файли у 2-3 рази без втрати інформації.

Далі проводимо етап сканування для кожного розвороту видання. При цьому автоматично проводиться етап запису отриманого зображення на жорсткий диск. Після виконання цих етапів для кожного розвороту ми отримали сукупність сканованих зображень, записаних у форматі TIFF з LZW-компресією (див. рис. 2). Якщо розворот видання сканований незадовільно (наприклад, має завеликий перекос), тоді необхідно повторити сканування цього розвороту.

Тепер можна провести очищення зображень та підготовку їх до відновлення. В нашому досліді скановані розвороти мали недостатню контрастність, а також наявність шуму по всій площі розвороту, що може незадовільно вплинути на подальший етап відновлення. Для того, щоб виправити ці недоліки, ми використали можливості програми Adobe Photoshop CS3 для підвищення контрастності та автоматизації цього етапу для всіх розворотів. В результаті ми от-

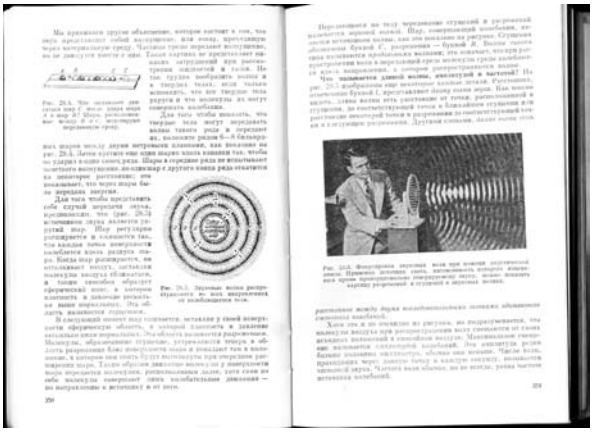

Рис. 2. Приклад сканованого зображення розвороту раритетного видання

римали розвороти з більшою чіткістю та кращою читабельністю.

Після проведення етапу очищення зображень переходимо до етапу відновлення видання. Наші скановані розвороти можуть мати різну орієнтацію та перекос. Основним завданням є отримання окремих сторінок видання. Тому потрібно розділити розвороти. Відновлення можна було б провести і вручну за допомогою різних програм обробки графічних зображень. Але це дуже трудомісткий етап, що включає велику кількість операцій і вимагає занадто багато часу.

На даний час створено нові програмні продукти, що працюють на алгоритмах розпізнавання образів (OCR). Для деяких етапів автоматизації процесу відновлення ми використали програму ScanKromsator 5.6. Робота проходить в діалоговому режимі з образами файлів. Спочатку програма в попередньому (підготовчому) режимі перевертає розвороти, визначає змістовну частину, її межі та положення, проставляє автоматично необхідні направляючі та ін. Користувач може вносити певні корективи у процес підготовки відновлення. Далі програма у пакетному режимі виконує необхідну послідовність операцій:

— Правильне позиціювання розворотів;

— Розділення розворотів на окремі сторінки;

— Вирівнювання сторінок;

— Виділення змістовної частини;

— Додавання необхідних полів;

— Задання параметрів для запису відновлених сторінок.

Для запису відновлених сторінок ми обрали формат TIFF з LZW-компресією. Сторінки можна переводити в модель B/W (Bitmap), якщо вони містять тільки текст та формули. Це дозволяє зменшити розмір файлів в декілька разів. Тестоване видання містить багато фотографій та рисунків. Для них більш прийнятною є модель Grayscale [4].

Результатом виконання цього етапу є відновлені сторінки раритетного видання (див. рис. 3).

Наступним етапом в нашій технологічній схемі може бути

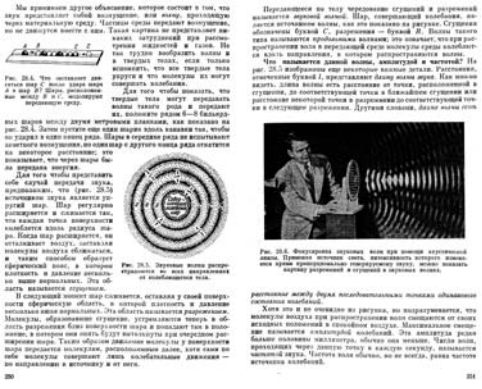

Рис. 3. Приклад відновлених сторінок раритетного видання

додаткова обробка. В додаткову обробку можуть входити: підчищення зображень та переведення у інші графічні формати.

Далі йде етап об'єднання оброблених сканованих зображень в єдиний файл з метою створення ефективної навігації для спрощення пошуку необхідної інформації. В цьому випадку користувач може вибрати різні формати, в залежності від складності видання та його призначення.

В тестованому виданні використано два формати: DjVu та PDF. Перший формат дозволяє добре стискувати електронний документ з малою втратою читабельності. При цьому текст і контрастні малюнки зберігаються з роздільною здатністю 300 dpi, все інше вважається фоном і зберігається із зниженою роздільною здатністю. У форматі DjVu розмір файлу відновленого видання становить декілька мегабайт, що є прийнятним для використання його в Internet.

Другий формат дозволяє зберігати е-книгу з різними ступенями якості, від прийнятного для Internet до е-книг, призначених для високоякісного друку. Також у форматі PDF створювати зручну навігацію по файлу, створювати гіперпосилання на інші файли і т. ін. [2].

Висновки

Результатом нашого дослідження є розробка технології відновлення науково-технічних раритетних видань. Ця технологія потребує використання доступного апаратного забезпечення та поширених програм обробки графічних файлів. Вона має модульну прозору структуру, що дозволяє заміняти програми та устаткування більш кращими.

Технологія в значній мірі автоматизована, що значно спрощує процес відновлення видань. В результаті її виконання можна отримати документи, придатні для архівування та подальшого використання. Е-книги можна розміщувати в Internet, формувати електронні бібліотеки. Для їх читання та друкування існує ряд ефективних програм. Для формату PDF — це Adobe Acrobat Reader, більшість веб-броузерів. Для формату DjVu — WinDjView.

Е-книги можна доповнювати не тільки засобами навігації, а й навіть засобами захисту від несанкціонованого втручання. В е-книги за бажанням можна впровадити текстовий шар з метою наближення їх до текстових документів [4].

1. Г. Киппхан. Энциклопедия по печатным средствам информации. Технологии и способы производства / Г. Киппхан. — М. : МГУП, 2003. — 1280 с. 2. Айриг С. Сканирование — профессиональный подход / С. Айриг, Э.Айриг ; пер. с англ. — Мн. : ООО «Попурри», 1997. — 176 с. : ил. 3. Айриг С. Подготовка цифровых изображений для печати / С. Айриг, Э. Айриг ; пер. с англ. — Мн.: ООО «Попурри», 1997. — 192 с. : ил. 4. Материалы по сканированию и оцифровке бумажных книг // Djvu-soft. — Режим доступу : http://www.djvu-soft.narod.ru/scan/, 12.03.2009.

Рецензент — В. Т. Мартинюк, к.т.н.,
доцент, НТУУ «КПІ»

Надійшла до редакції 28.05.09